Regression

& Time Series

Linear Regression

- Linear Regression is a very simple approach for Supervised Learning

- It is a useful tool for predicting quantitative response

- Least square is the most commonly used approach to fit the model

Simple Linear Regression (SLR)

- Simple Linear Regression predicts Y on the basis of single Predictor Variable X
- It assumes that there is approximately linear relationship between X and Y.

$$Y \approx \beta_0 + \beta_1 X$$

- The above equation is described as regressing Y on X (or Y onto X)
- The symbol ≈ can be read as “is approximately modeled as”
- We can regress sales onto TV by fitting the model

$$\text{sales} \approx \beta_0 + \beta_1 * \text{TV}$$

SLR coefficients and error term

- We can also write the linear relationship as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model
- Together, β_0 and β_1 are known as the model coefficients or parameters
- β_0 is the expected value of Y when X=0 and β_1 is the average increase in Y associated with a one unit increase in X
- The error term is the catch all for what we miss with the simple model
  - The true relationship is probably not linear
  - There may be other variables that cause variation in Y
  - There may be measurement error
- We typically assume that the error term is independent of X

Fitting Simple Linear Regression

- We fit the linear model using training data and estimate model coefficients as $\hat{\beta}_0$ and $\hat{\beta}_1$
- We can predict future sales on the basis of a particular value of TV advertising by computing

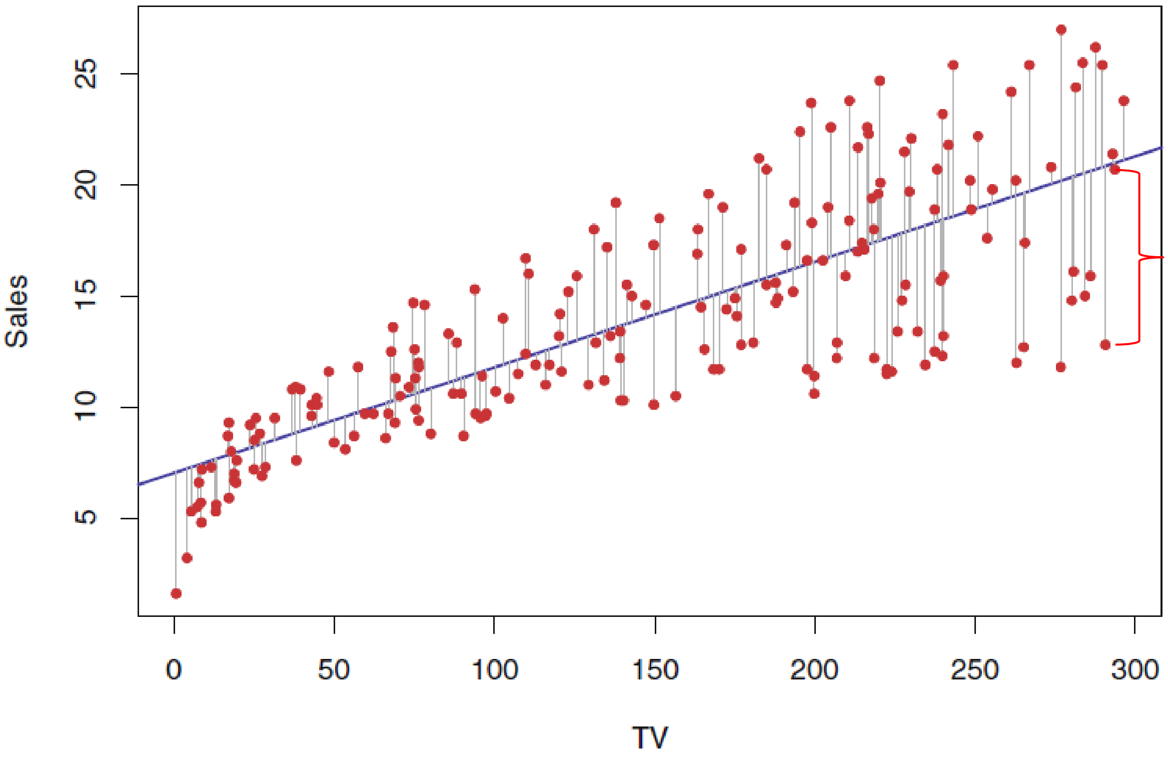
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x$$

- Where $\hat{y}$ indicates a prediction of Y on the basis of X=x
- Hat symbol **^** is used to denote the estimated value of
  - An unknown parameter or coefficient
  - Predicted value of the response

Estimating SLR Coefficients using Least Square

- Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well
- We want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to all the data points
- The most common approach of measuring closeness involves minimizing the least squares criterion
- *i*th residual $e_i = y_i - \hat{y}_i = y_i – (\hat{\beta}_0 + \hat{\beta}_1 *x)$ is the difference between *i*th observed and predicted response value
- The Residual Sum of Squares is defined as $RSS = e_1^2 + e_2^2 + \ldots. + e_n^2$
- The least squares approach choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

Estimating SLR Coefficients using Least Square

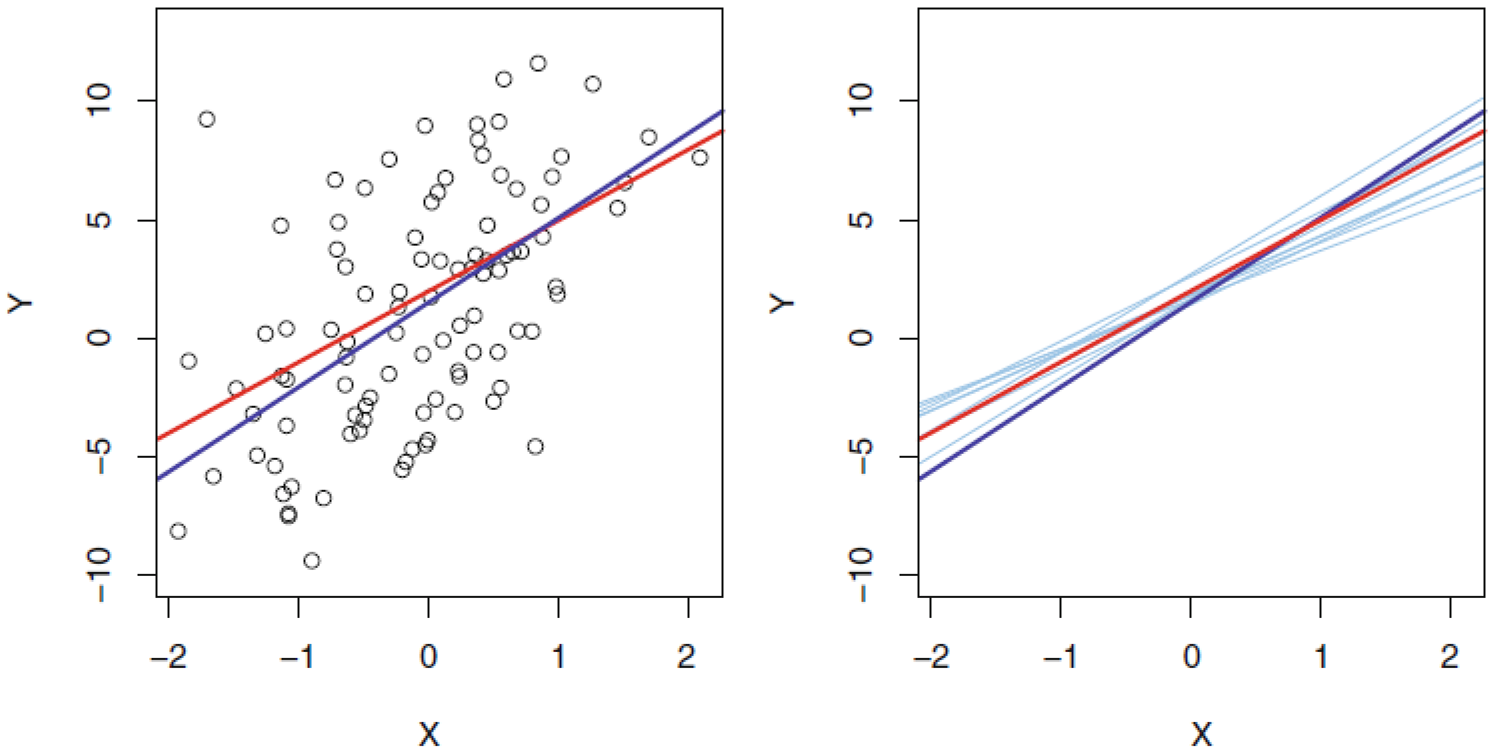


i th residual

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{\sum_{i=1}^{n}(xi - \bar{x})^2};$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimating SLR Coefficients using Least Square



- In left diagram red line shows the true relationship and blue line is the least regression line for this simulated data set
- Right diagram shows multiple least regression lines generated from separate data sets, average of which will be very close to the red line

Assessing accuracy of the coefficients: Confidence Intervals

- Standard Error (SE) tells us the average amount this estimate differs from average value
- SE is used to construct confidence intervals (CI)
- A 95% CI is defined as a range of values such that with 95% confidence, the range will contain the true value of the parameter
- The range is defined in terms of lower and upper limits computed from the data and takes following approximate form

$$\mathrm{CI}_{\hat{\beta}_1} = \hat{\beta}_1 \pm 2*\mathrm{SE}(\hat{\beta}_1)$$

$$\mathrm{CI}_{\hat{\beta}_0} = \hat{\beta}_0 \pm 2*\mathrm{SE}(\hat{\beta}_0)$$

- p-value is another measure using which we can assess significance of the coefficients. A small p-value (<0.05) means that the coefficients are significant

Assessing accuracy of the Model: Error Term

- We want to quantify the extent to which the model fits the data
- The quality of linear regression fit is typically assessed using two related quantities: residual standard error (RSE) and R^2 statistic

Residual Standard Error:

- Due to presence of error term ε in the model, even if we know true regression line, we will not be able to perfectly predict Y from X
- RSE is an estimate of standard deviation of ε and it is average amount that the response will deviate from the true regression line

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

- RSE is considered as lack of fit of the model. A large value indicating that the model doesn't fit the data well

Assessing accuracy of the Model: RSE and R^2

R^2 Statistic

- RSE provides an absolute measure of lack of fit of the model to the data
- Since it is measured in the unit of Y, it is not always clear what is good RSE
- R^2 takes the form of proportion, values between 0 and 1, and is independent of the scale of Y
- It is the proportion of the variance explained

$$R^2 = 1 - \frac{RSS}{TSS}$$

- TSS is the total sum of squares $TSS = \sum(y_i - \bar{y})^2$
- R^2 measures the proportion of variability in Y that can be explained by X. A high value of R^2 indicates good model fit
- R^2 statistic is a measure of the linear relationship between X and Y

Multiple Linear Regression (MLR)

- For p different predictors, SLR is extended to MLR and takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots. + \beta_p X_p + \varepsilon$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed. Out example thus becomes:

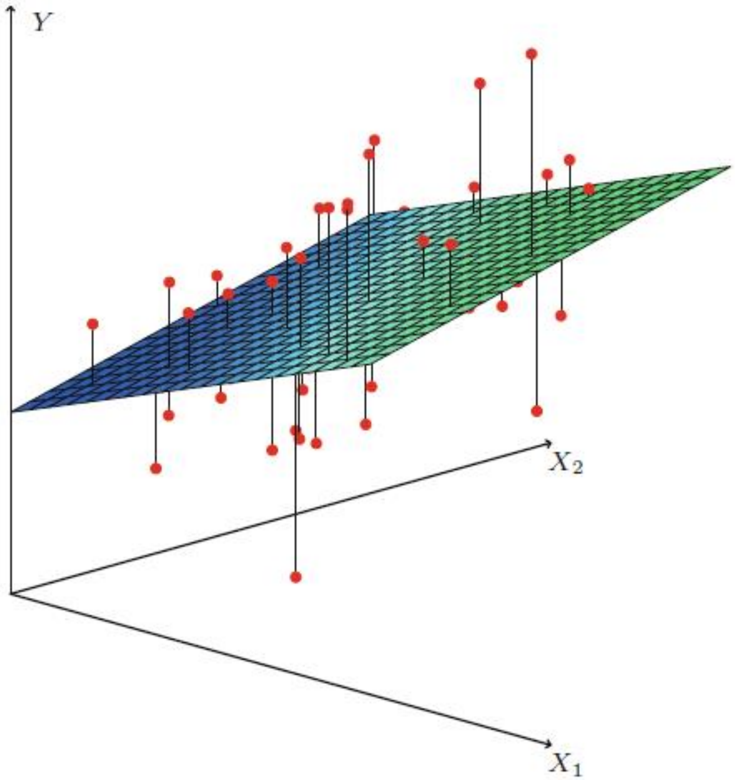
$$sales = \beta_0 + \beta_1{*}TV + \beta_2{*}Radio + \beta_3{*}Newspaper + \varepsilon$$

- Given estimates $\hat{\beta}_0 , \hat{\beta}_1, \ldots , \hat{\beta}_p$ we can make prediction using formula

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots\ldots. + \hat{\beta}_p X_p$$

- We choose parameters so as to minimize sum of squared residuals RSS

Multiple Linear Regression (MLR)



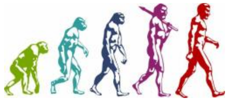
- MLR setting where p = 2. Least square line becomes a plane.
- The plane is chosen to minimize the sum of squared distance between each observation and the plane

MLR: Some important considerations

- We use F statistic and associated p-value to determine if there is relationship between Response and Predictors. F > 1 and low p-value is desired.
- We can do variable selection by : a) checking p-values of MLR and b) by trying out different models
- RSE and R^2 is used to check the quality of model fit. For MLR
- R^2 will always increase when more variables are added to the model
- For MLR, RSE is given by:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

- Models with more variables can have higher RSE if the decrease in RSS is small relative to increase in p
- Plot of data to check for synergy or interaction effect



MLR: Some important considerations

- We can compute a Confidence Interval in order to determine how close $\hat{Y}$ (= $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 +.. + \hat{\beta}_p X_p$) is to f(X) (= $\beta_0 + \beta_1 X_1 + \beta_2 X_2 +.. + \beta_p X_p$). The inaccuracy in the coefficient estimates is related to reducible error.
- We compute a prediction interval to compute how much will Y vary from $\hat{Y}$.
- Prediction Interval is always wider than Confidence Interval because they additionally incorporate uncertainty due to (irreducible) error ε
- Confidence Interval quantify the uncertainty surrounding average Y, whereas Prediction Interval quantify uncertainty surrounding Y itself
- MLR can be extended to qualitative predictors
- MLR can be extended to include interaction effect by adding an interaction term $\beta_j X_1 X_2$
- MLR can be extended for non-linear relationship by adding terms like $\beta_j X_1^2$, $\beta_j X_1^3$, etc

Qualitative Predictors – two levels

- We had assumed that all variables in our linear regression model are quantitative. But in practice they may be qualitative
- If a qualitative variable (also known as Factor) has only two levels, we create a dummy variable that takes two possible numerical values.
- Example: If we want to investigate differences in Credit card balance between Male and Female

$$x_i = \begin{cases} 1 & if\ ith\ person\ is\ female \\ 0 & if\ ith\ person\ is\ male \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & if\ ith\ person\ is\ female \\ \beta_0 + \varepsilon_i & if\ ith\ person\ is\ male \end{cases}$$

- β_0 = Average credit card balance amongst male and
- $\beta_0 + \beta_1$ = Average credit card balance amongst females

Qualitative Predictors – more than two levels

- If a Factor has more than two levels, we create multiple dummy variables each of which takes two possible numerical values.
- If we want to investigate differences in Credit card balance due to different ethnicity: Asian, Caucasian and Afro-American

$$x_{i1} = \begin{cases} 1 & \mathit{if\ ith\ person\ is\ Asian} \\ 0 & \mathit{if\ ith\ person\ is\ not\ Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \mathit{if\ ith\ person\ is\ Caucasian} \\ 0 & \mathit{if\ ith\ person\ is\ not\ Caucasian} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \mathit{if\ ith\ person\ is\ Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \mathit{if\ ith\ person\ is\ Caucasian} \\ \beta_0 + \varepsilon_i & \mathit{if\ ith\ person\ is\ Afro - American} \end{cases}$$

- The number of dummy variables will be one less than the number of levels. The level with no dummy variable is known as baseline

Extensions of the Linear Model

- The standard linear regression model provides interpretable results and works quite well on many real-world problems
- However, it makes highly restrictive assumptions that are often violated in practice
- Two of the most important assumptions state that the relationship between the predictors and response are *additive* and *linear*
- The additive assumption means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors
- The linear assumption states that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j

Removing the additive assumption: Interaction Effect

- Linear Model can be extended by allowing for interaction effects, by including additional predictor called an interaction term
- This term is constructed by computing the product of predictors
- The resulting model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

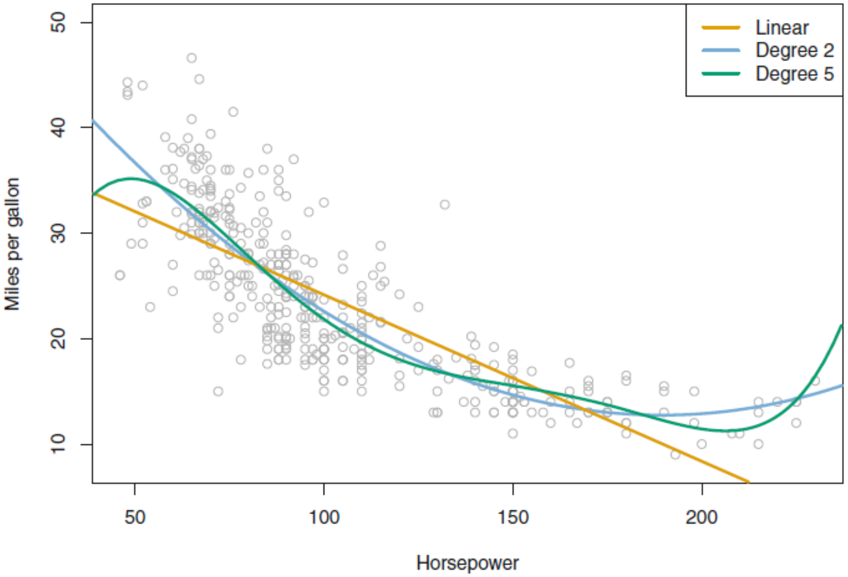
- The model can be re-written as:

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2)X_1 + \beta_2 X_2 + \varepsilon = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$
- Since $\tilde{\beta}_1$ changes with X_2 the effect of X_1 on Y is no longer constant. Adjusting X_2 will change the impact of X_1 on Y
- Example: units= 1.2 + (3.4 + 1.4 X workers) X lines + 0.22 X workers
- Here, the more workers we have, stronger will be the effect of lines. Each additional line will increase no of units produced by 3.4 + 1.4 X workers

Extending the linear assumption: Polynomial Regression

- In some cases, the true relationship between the response and the predictors may be non-linear
- A simple way to extend the linear model to accommodate non-linear relationships is by using *polynomial regression*
- Example: $\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$
- Above equation is simply a multiple linear regression model with X_1 = horsepower and X_2 = horsepower^2
- So we can use standard linear regression to estimate β_0, β_1 & β_2 in order to produce a non-linear fit



Linear Regression: Potential Problems

- Non-linearity of the response-predictor relationship
- Correlation of error term (basic assumption violated, confidence interval, prediction intervals and p-value will be narrower than they should be)
- Non constant variation of error term (Funnel shape in residual plot: transform to functions like logY, $\sqrt{Y}$)
- Outliers (y_i is far from predicted value: increase in RSE, CI and p-value)
- High-Leverage points (unusual value for x_i: least square line heavily impacted invalidating the entire fit)
- Collinearity (results in great deal of uncertainty in coefficient estimates and reduces their accuracy. As a result Standard Error and p-value increases)

Shrinkage Methods

- In Shrinkage Methods, we fit a model containing all p predictors using a technique that *constraints* or *regularizes* the coefficient estimates, or equivalently, that *shrinks* the coefficient estimates towards zero relative to the least square estimates
- This shrinkage (also known as regularization) has the effect of reducing variance
- Depending on type of shrinkage, some of the coefficient can be estimated to be exactly zero
- Hence Shrinking method can also perform variable selection
- Two best known techniques for shrinking the coefficient estimates towards zero are:
  - Ridge Regression
  - Lasso

Ridge Regression

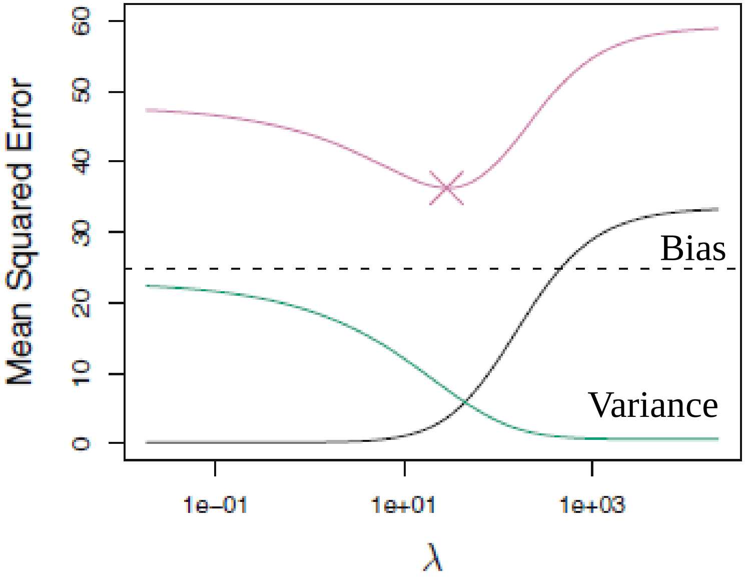
- Ridge Regression is very similar to Least Squares except that the coefficients are estimated by minimizing a slightly different quantity:

$$RSS+ \lambda\ (\beta_1^2 + \beta_2^2 + \beta_3^2 +.... + \beta_p^2\ ) = RSS + \lambda \sum_{i=1}^{p} \beta_j^2$$

- Where $\lambda >= 0$ is tuning parameter & RSS = $\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=1}^{p} \beta_j x_{ij})^2$
- First Term: Ridge regression seeks to fit the data well by making the RSS small
- Second Term: Also known as *Shrinkage Penalty* is small when $\beta_{1,}\ \beta_2$, β_j are close to zero and has the effect of shrinking the estimates of β_j towards zero. Shrinkage penalty is not applied to intercept β_0
- The tuning parameter λ controls the relative impact of the two terms
- When $\lambda = 0$, penalty term has no effect and it will produce LSE
- When λ -> ∞, the impact of shrinkage penalty grows and the ridge regression coefficient estimates will approach zero

Why Ridge Regression woks

- Ridge regression's advantage is rooted in Bias-Variance tradeoff when LSE has high variance (e.g. p is nearly same as n)
- At the LSE estimate, which correspond to ridge regression with λ=0, the variance is high but there is no bias.
- As λ increases, the shrinkage of Ridge regression coefficient estimates leads to a substantial reduction in the variance of the predictions, at the expense of slight increase in bias.
- Beyond a point, the decrease in variance due to increasing λ slows
- The shrinkage on the coefficient causes them to be significantly underestimated, resulting in large increase in bias



Lasso

- Ridge Regression has one disadvantage: It will include all p predictors in the final model.
- The penalty term will shrink all of the p coefficient towards zero, but it will not set any of them exactly equal to zero (unless $\lambda = \infty$).
- This creates a problem in model interpretability when p is large
- We might wish to build a model which includes only most important predictors
- *Lasso* is an alternative to Ridge regression which overcomes this challenge

Lasso

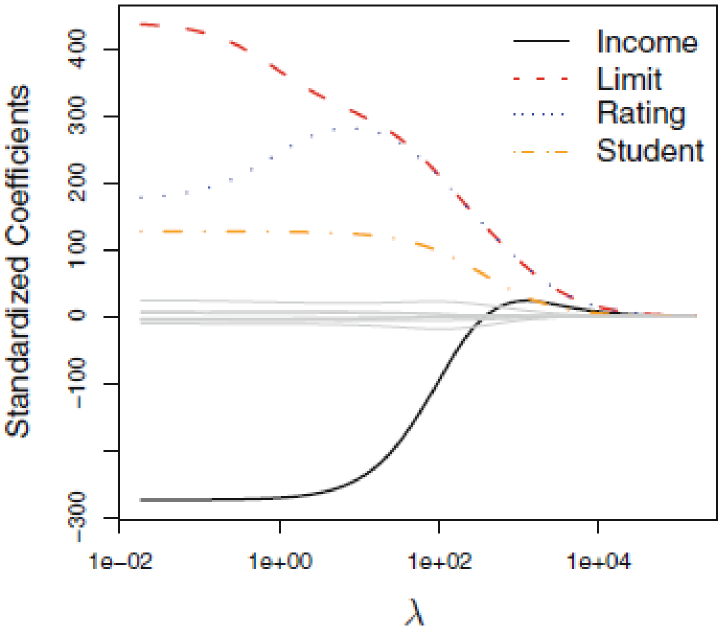
- Lasso is very similar to Ridge Regression except that the coefficients β_j^2 in Ridge regression penalty is replaced by $|\beta_j|$ in the Lasso penalty

$$RSS + \lambda\,(|\beta_1| + |\beta_2| + |\beta_3| + ... + |\beta_p|\,);$$

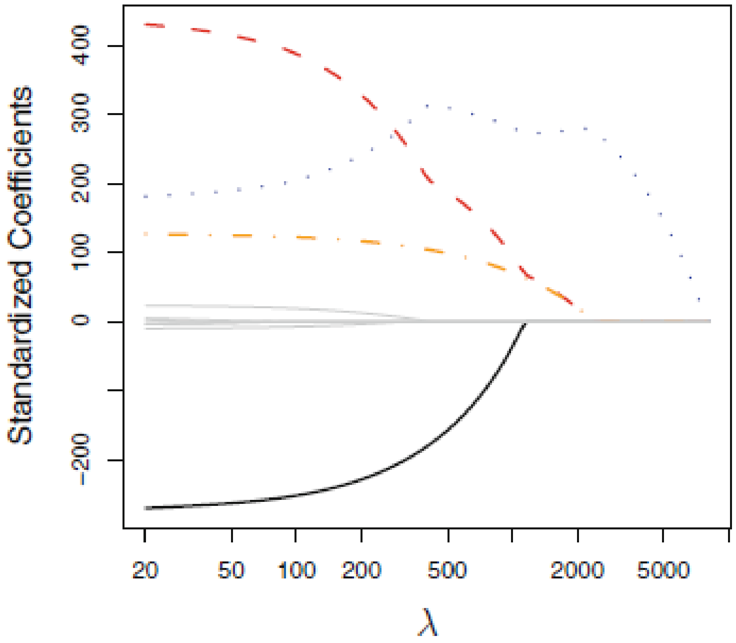
- As with Ridge regression, the lasso shrinks the coefficient estimates towards zero
- However, in the case of Lasso, the penalty has the effect of forcing some of the coefficients exactly equal to zero when λ is sufficiently large.
- Lasso performs variable selection hence model generated from the Lasso are much easier to interpret.
- Cross Validation is an effective mechanism to check the performance of both the models and select the better one.

Ridge Regression and Lasso

- When λ is sufficiently large, Ridge regression penalty forces the coefficients to become very small whereas in the case of Lasso, the penalty has the effect of forcing some of the coefficients exactly equal to zero



Ridge Regression

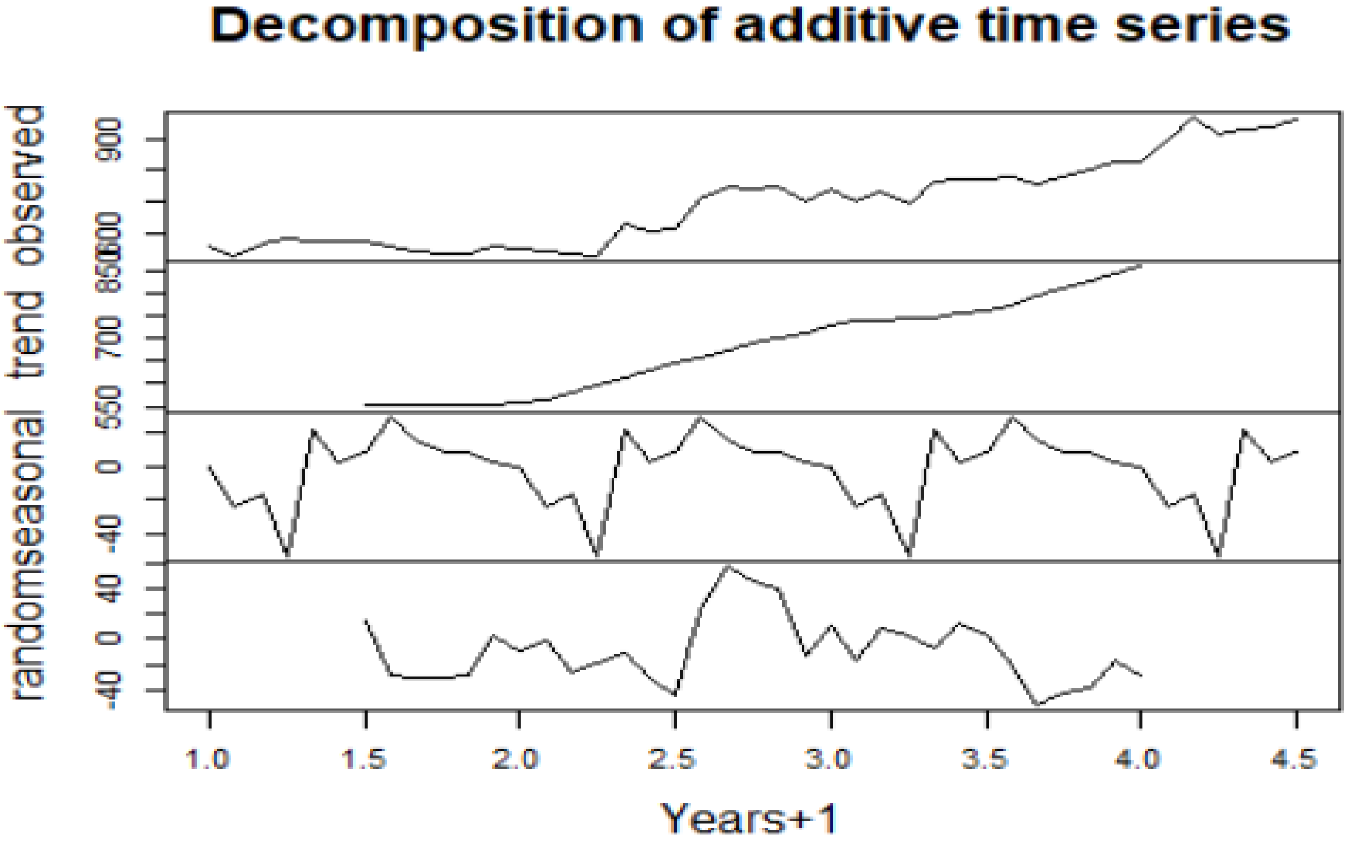


Lasso

Time Series

- Performance of y over time is time series (y_t at time t)
- Data is dependent over time
- Specific Pattern Type
  - Trends: Consistent long term increasing or decreasing pattern
  - Seasonal: Patterns related to time of week, month, quarter, year, festivals, agricultural (usually within a year – Short/Mid term)
  - Cycles: Patterns that are beyond a year – Long term
- Trends, Seasons and Cycles are components of time series
- Subsampling into training/test set is complicated as observations cannot be taken randomly
- Goal is to predict one or more observation in future
- Standard prediction functions can be used for forecasting

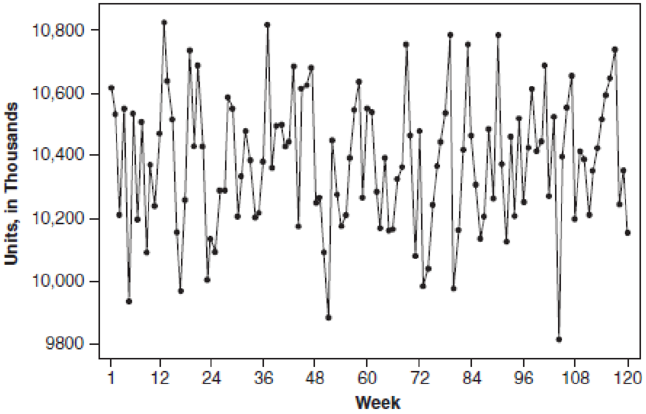
Time Series components



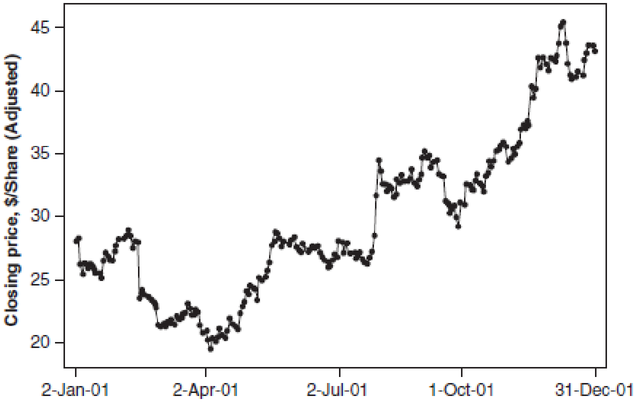
Decomposition of Additive Time Series

Static and Dynamic time series

- Static (or Stationary) Time Series have patterns that do not change over time.
  1. Distribution of y does not depend on t
  2. Roughly horizontal and constant variance
  3. No pattern predictable in long term



- Dynamic (or Non Stationary) Time Series have patterns that do change over time and estimates are updated using neighbouring values
  1. Time series has no natural mean
  2. Exhibits a pattern for both mean level and slope

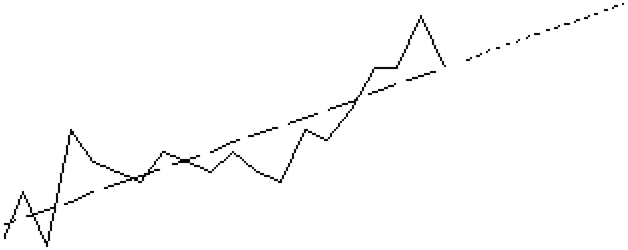


Time Series

- **Simple forecasting and smoothing methods** are based on the idea that reliable forecasts can be achieved by modelling patterns in the data that are usually visible in a time series plot, and then extrapolating those patterns to the future
- Your choice of method should be based upon whether the patterns are
  a. static (constant in time) or dynamic (changes in time),
  b. the nature of the trend and seasonal components, and
  c. how far ahead that you wish to forecast.
- These methods are generally easy and quick to apply.

Trend Analysis

- Fits a general trend model to time series data. Models may be linear, quadratic, exponential growth or decay. Use this procedure to fit trend when there is no seasonal component in your series. Used when:
- Data with constant trend, and
- Data with no seasonal pattern
- Long range forecasting



- Profile: extension of trend line, Continuation of trend line fit to data
- Trend analysis by default uses the **linear trend** model: $\mathbf{Y_t = b_0 + (b_1 * t) + e_t}$

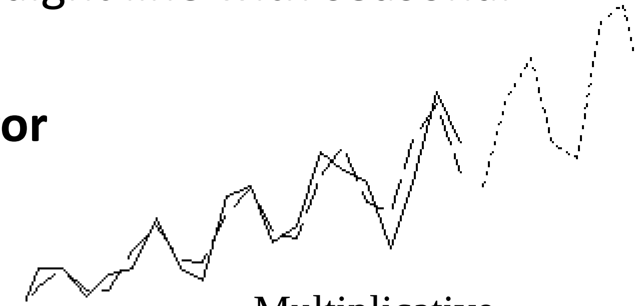
  In this model, b_1 represents the average change from one period to the next.
- The **quadratic trend** model which can account for simple curvature in the data, is: $\mathbf{Y_t = b_0 + b_1 * t + (b_2 * t^2) + e_t}$
- The **exponential growth** trend model accounts for exponential growth or decay. For example, a savings account might exhibit exponential growth. The model is: $\mathbf{Y_t = b_0 * b_1^t * e_t}$

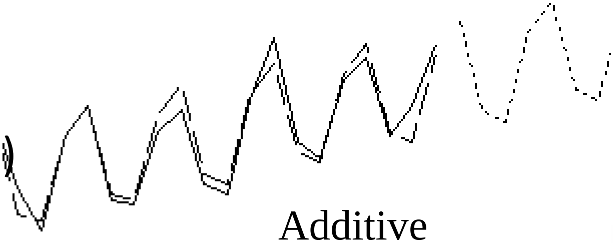
Decomposition

- Separates the times series into linear trend and seasonal components, as well as error and provide forecasts.
- Use this procedure to forecast when there is a seasonal component in your series or if you simply want to examine the nature of the component parts
- Choose whether the seasonal component is additive or multiplicative with the trend.

Decomposition

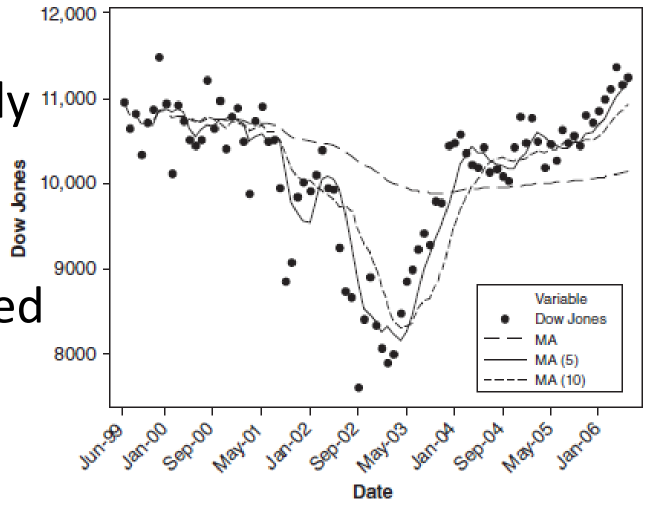
- **Multiplicative model** is used when the size of the seasonal pattern in the data depends on the level of the data. This model assumes that as the data increase, so does the seasonal pattern. Most time series exhibit such a pattern. Forecast Straight line multiplied by seasonal pattern
- The multiplicative model is : **Yt = Trend * Seasonal * Error**
- In **additive model** , the effects of individual factors are differentiated and added together to model the data. Forecasted by Straight line with seasonal pattern added
- The additive model is: **Yt = Trend + Seasonal + Error**
- where Yt is the observation at time t.
- Use when:
- Data with either no trend or constant trend, and
- Data with constant seasonal pattern
- Long range forecasting
- Size of seasonal pattern proportional to data (multiplicative)
- Size of seasonal pattern not proportional to data (additive)
- Profile: trend with seasonal pattern


Multiplicative


Additive

Moving Average

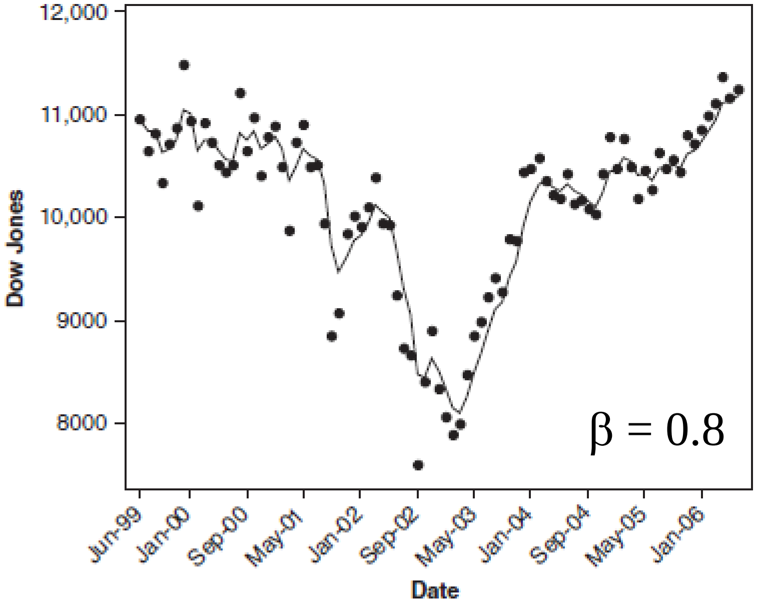
- Smoothens your data by averaging consecutive observations in a series. This procedure can be a likely choice when your data do not have a trend or seasonal component.
- **Length (or Span)**: A positive integer to indicate desired length for the moving average. With non-seasonal time series, it is common to use short moving averages to smooth the series. The length you select may depend on the amount of noise in the series.
- A longer moving average filters out more noise, but is also less sensitive to changes in the series. With seasonal series, it is common to use a moving average of length equal to the length of an annual cycle.
- The fitted value pattern lags behind the data pattern. This is because the fitted values are the moving averages from the previous time unit



- **Use for:**
- Data with no trend, and
- Data with no seasonal pattern
- Short term forecasting
- **Forecast profile:** Flat line
- **Length:** short

Single (First Order) Exponential Smoothing

- Single exponential smoothing smoothens your data by computing exponentially weighted averages and provides short-term forecasts. This procedure works best for data without a trend or seasonal component. The single dynamic component in a moving average model is the level. The fitted value at time t is the smoothed value at



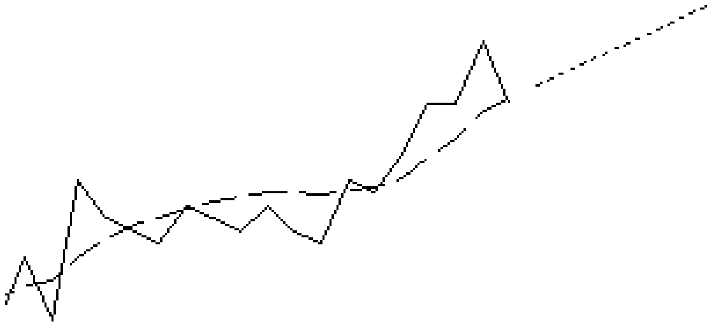
- **Use for:**
- Data with no trend, and
- Data with no seasonal pattern
- Short term forecasting
- **Length:** short
- **Profile:** flat line
- Model can be expressed as:

$$\tilde{y}_t = (1 - \beta) * y_t + \beta * \tilde{y}_{t-1}$$

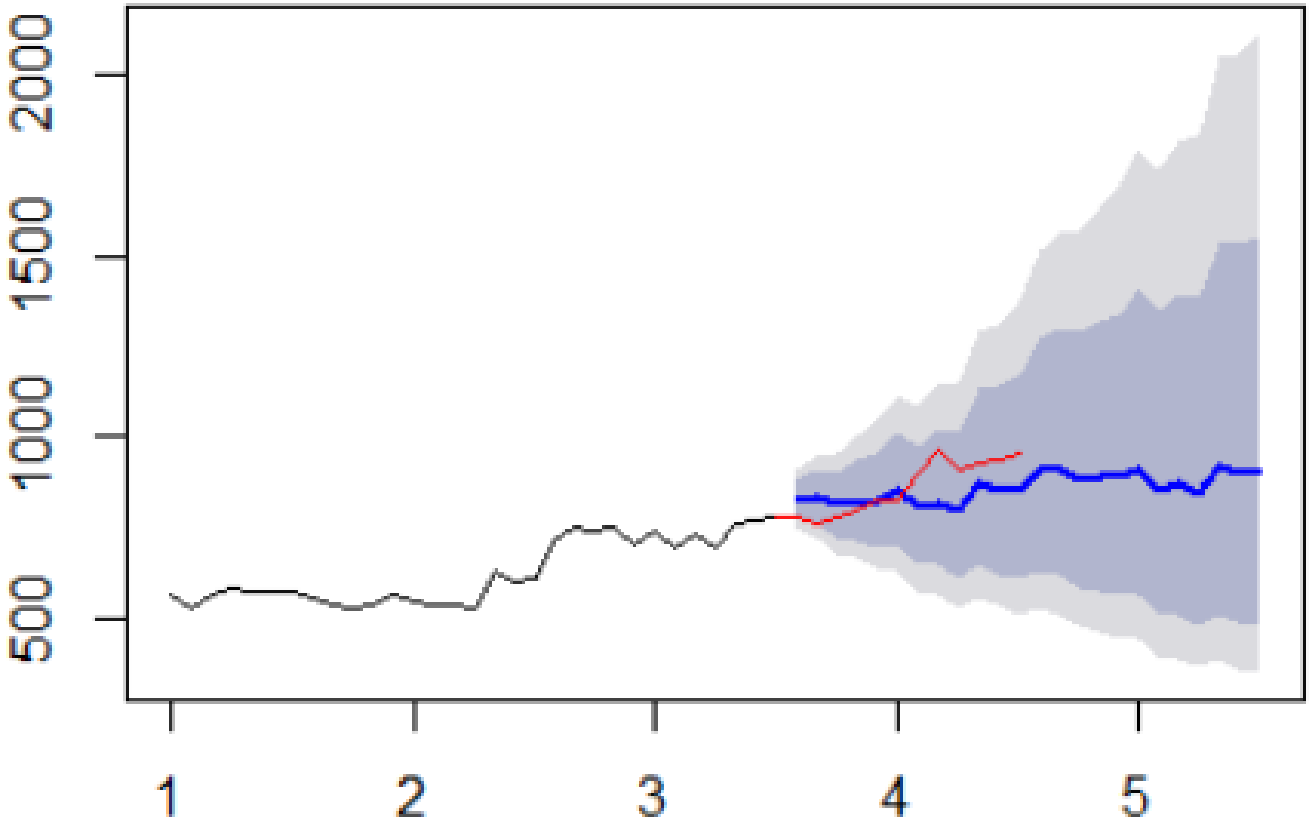
$$\tilde{y}_t = (1 - \beta) * y_t + \beta * (1 - \beta) * (y_{t-1} + \beta * y_{t-2} + \cdots + \beta^{t-2} * y_1)$$

Double (Second Order) Exponential Smoothing

- Double exponential smoothing smoothens your data by Holt double exponential smoothing and provides short-term forecasts. This procedure can work well when a trend is present but it can also serve as a general smoothing method.
- Dynamic estimates are calculated for two components: level and trend. Double exponential smoothing employs a level component and a trend component at each period. It uses two weights, or smoothing parameters, to update the components at each period.
- **Use for:**
- Data with constant or non-constant trend, and
- Data with no seasonal pattern
- Short term forecasting
- **Forecast profile:**
- Straight line with slope equal to last trend estimate
- Length: short



Exponential Smoothing with CI



Exponential Smoothing
(Prediction interval with 80%
and 95% Confidence Level)

Error Analysis

- There are three measures of accuracy of the fitted model: MAPE, MAD, and MSD for each of the simple forecasting and smoothing methods. For all three measures, the smaller the value, the better the fit of the model. Use these statistics to compare the fits of the different methods. In equations, y_t equals the actual value, $\hat{y}_t$ equals the forecast value, and n equals the number of forecasts.

- **MAPE**, or Mean Absolute Percentage Error, measures the accuracy of fitted time series values. It expresses accuracy as a percentage.

$$\text{MAPE} = \frac{\sum |(y_t - \hat{y}_t)/yt|}{n} * 100 \qquad (y_t \neq 0)$$

- **MAD**, which stands for Mean Absolute Deviation, measures the accuracy of fitted time series values. It expresses accuracy in the same units as the data, which helps conceptualize the amount of error.

$$\text{MAD} = \frac{\sum |y_t - \hat{y}_t|}{n}$$

- **MSD** stands for Mean Squared Deviation. MSD is always computed using the same denominator, n, regardless of the model, so you can compare MSD values across models. MSD is a more sensitive measure of an unusually large forecast error than MAD

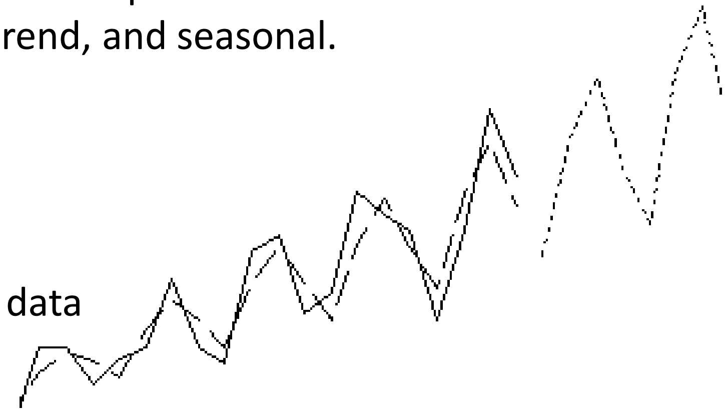
$$\text{MSD} = \frac{\sum |y_t - \hat{y}_t|^2}{n}$$

Other Methods

Winter’s Method:

- Winters' Method smoothens your data by Holt-Winters exponential smoothing and provides short to medium-range forecasting. You can use this procedure when both trend and seasonality are present, with these two components being either additive or multiplicative. Winters' Method calculates dynamic estimates for three components: level, trend, and seasonal.
- **Use for:**
- Data with or without trend
- Data with seasonal pattern
- Size of seasonal pattern

  Additive: not proportional, Multiplicative: proportional) to data
- Short to medium range forecasting



- **Forecast profile:**
- Straight line multiplied by seasonal pattern for multiplicative (added for additive)

ARIMA

- Use ARIMA to model time series behaviour and to generate forecasts. ARIMA fits a Box-Jenkins ARIMA model to a time series. ARIMA stands for Autoregressive Integrated Moving Average with each term representing steps taken in the model construction until only random noise remains. ARIMA modelling differs from the other time series methods in the fact that ARIMA modelling uses correlational techniques. ARIMA can be used to model patterns that may not be visible in plotted data.

Thank You

Abhinav Srivastava