# Regression Analysis

Leason Goals:

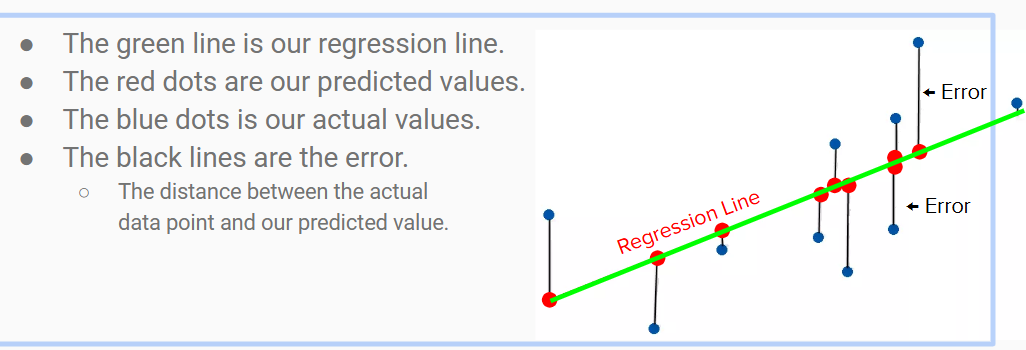
1. **Have a basic understanding of how to use and evaluate**
   1. **Linear Regression**
   2. **Random Forest Regression**
2. **Be able to build and save a model to predict new data.**

What do we predict using with regression?

* Discrete values
* Continuous values

What is linear Regression?

* Linear Reg is a basic predictive analytics technique that uses historical data to predict an output variable
* Why Linear Regression?
  + Explains to us how variables relate to each other.
* Type of Linear Regression
  + Simple Reg (2 dimensional at most)
  + Multiple Reg (Data in >2 dimensions)
* What is the Mean Absolute Error (MAE):



* How to calculate Mean Absolute Error (MAE)?
  + Take the absolute value of all of the errors
  + Sum them all up
  + Take the average
  + Note: A MAE score of 0 means there is NO error, which is good. The higher the error the wors the model is.
  + Ex: if you MAE is 0.5 we can say that your predicted value is on average 0.5 unites away from the true value.
* What is R-squared:
  + Measure the strength of relationship between your model and the dependent variables on a 0-100% scale.
* Why R-squared Scores?
  + The **larger** the R-squared, the **better** the regression model fits your observations.  
    \*\* *R-Squared values for human behavior related data, such as rating wine quality, are inherently lower due to the subjective nature, (ie; 30% is good).  In situations with less subjectivity, like a chemical reaction, higher R-Squared values are expected (ie; 90% is good where 30% would be not so good.)*
* Limitation of Linear Reg:
  + There must be linear relationship between independent and dependent variables
  + Linear reg is very sensitive to Outliers (affected the reg line and the forecasted value)
* Two common applications of regression:
  + Explaining variance: How much variation in one data series is caused by another?
  + Making predictions: How much does a move in one series impact another?

Data dimension:

1. Data in one Dimension

Oil Prices “Low” Average = $45/barrel “High”

Unidimensional data points can be represented using a line, such as a number line. However, it is analyzed using statistics such as mean, median (central of tendency indicates the average value of a distribution of figures) , standard deviation (How much the variation from there is from the average)

1. Data in Two Dimensions

Oil Prices

Government

Bond Yields

Bidimensional data can be represented in a plane:

For Linear regression: We are looking for the best fit line. We can make a line goes through each point, but this is not an optimal way to represent the relationship between the variables (features)

* A curve has a “good fit” if the distances of points from the curve are small. (We’ll talk about these type of distances)
* A linear line won’t all the time serve what you want.
* A challenge of overfitting occurs when finding a very complicated curve often hurts the predictive accuracy of linear regression model
* The goal is finding the “best” such straight line is called linear Regression
* Regression not only gives us the equation of this line, it also signals how reliable the line is.
* R-Squared (measure of goodness of fit): Helps us understand how strongly one variable explains another.

**High R2**

Oil Prices

Government

Bond Yields

* High R2 =High quality of fit

**Low R2**

Oil Prices

Government

Bond Yields

Let’s we have (x,y=g(x)):

Given a new value of x, use the line to predict the corresponding value of y (Using linear regression model).

**Predicted value of y**

**Out-of-sample**

**value of x**

Oil Prices

Government

Bond Yields

* Regression also allows to specify prediction intervals (Similar to confidence interval)







**Predicted value of y**

**Out-of-sample**

**value of x**

Oil Prices

Government

Bond Yields

* Predicting the value g(x) using linear regression line with 95% Prediction Interval (confidence)

What about other dimensions?

Data in N Dimensions

“Linear Reg can easily be extended to n-dimensional data”

Oil Prices

Government

Bond Yields

Share Index

**Why using regression?**

* Powerful: Perfectly suited to two common use-cases
* Versatile: Easily extended to nonlinear relationships we can also transform Nonlinear data
* Deep: The first “crossover hit” from ML

Mean? Average

Median?

Given the numbers: 44, 56,45,67,66,78,87,45,58

* Arrange the data: 44,45,45,56,58,66,67,78,87 Odd number of data and the median is 58
* If we have even number, we add them up and divide by 2

Range = Xmax – Xmin

Variance: is the second-most important number to summarize the data

V=

Outliers: You have to take care of them while you are wrangling data

Standard Deviation is the square root of variance

# Connecting the dots with Regression

Regression Equations: y=A + Bx

Y1 = A + Bx1 + e1 (errors)

Y2= A + Bx2 + e2

Y3= A + Bx3 + e3

…

Yn = A + Bxn  + en

errors refer to as residue of the regression

**(Xi,Yi)**

Y

X

A



**(Xi,)**

Regression Line:

Y = A + Bx

Residuals of a regression are the difference between actual and fitted values of the regression variable

Y=[y1,y2,y3,…yn]

Y’=[y’1,y’2,y’3…y’n]

e=[e1,e2,e3…en]

Improving regression models sometimes you need to minimize the errors or residuals

* To find the “best fit” line we need to make some assumptions about regression residuals

# Understanding simple Regression Models

* Use simple Reg models to:
  + Explain variance
  + Make forecasts

Cause & effect principal: X causes Y ( Regression idea)

Cause 🡪 Independent variable (X 🡪 Explanatory variable)

Effect 🡪 Dependent variable (Y)

# Minimizing Least Square Error

* The “best fit” line is the one where the sum of the squares of the lengths of these dotted lines is minimum (The goal of regression problem)

Ideally, residuals should:

* Have zero mean
* Common variance
* be independent of each other
* be independence of x
* be normally distributed

DEMO

Check Excel file:

Note: To determine the values of A and B (regression coefficients)

* Method of moments
* Method of least squares
* Maximum likelihood estimation

Best model developers look for something called BLUE solution: Best Linear Unbiased Estimator

Best🡪 Coefficients have minimum variance e.g. estimated with relatively high certainty

Unbiased🡪 Residuals have zero mean, are uncorrelated to each other and have equal variance

Correlation Vs Causation

Never use regression to establish a cause-effect relationship

Precedes: X happens before Y

Accompanies: X and Y happen together

Causes: X causes Y (Regression best scenario)

R2: The percentage of total variance explained by the regression. Usually, the higher the R2, the better quality of the regression (upper bound 100%)

R2 = ESS / TSS (TSS = Variance(y) ESS=Variance(y’) RSS=Variance(e)