Machine Learning & Predictive modelling

Agenda

Key Takeaways-

- Various types of Analytics
- What is Machine Learning?
- Types of Machine Learning
- Regression, its types and implementation
- Simple, Multiple linear regression model
- Polynomial, Exponential regression model
- Evaluation measures for regression

Analytics

With data comes analytics, in order to make decisions, get insights, discover hidden patterns/trends, gain competitive edge etc.

The big data revolution has given birth to different kinds of analytics. The most popular are -

1. Descriptive Analytics
2. Diagnostic Analytics
3. Predictive Analytics
4. Prescriptive Analytics

Descriptive Analytics

- Describing or summarising the existing/past data to better understand ***what is going on*** or ***what has happened***.

- Helps in crunching the massive data into understandable chunks.

"The simplest class of analytics, one that allows you to condense big data into smaller, more useful nuggets of information." - Dr. Michael Wu

- Techniques used are metrics reports, descriptive statistics, data aggregation, data mining etc.

Diagnostic Analytics

- To determine ***why something happened in the past***.
- Diagnostic analytics takes a deeper look at data to understand the root causes of the events.
- Helpful in determining what factors and events contributed to the outcome.
- Techniques used are attribute importance, correlation, principle components analysis, sensitivity analysis etc.

Predictive Analytics

- Predictive analytics tells ***what is likely to happen***.
- It is used to predict the future outcomes/trends.

“Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature." - Dr. Michael Wu

- Techniques used are quantitative analysis, predictive modelling, machine learning algorithms etc..
- The most popular tools/languages for predictive analytics are Python, R, Julia, etc.

Prescriptive Analytics

- To literally prescribe ***what action to take*** to eliminate a future problem or take full advantage of a promising trend.

- It takes the forecasts and likelihoods from predictive analytics one step further by creating advised solutions that will also align with the goals, limitations and influencing factors of the organization.

- Techniques used are recommendation systems,artificial intelligence, neural networks etc.

Machine Learning

In 1959, Arthur Samuel, a pioneer in the field of machine learning (ML) defined it as the "*field of study that gives computers the ability to learn without being explicitly programmed*".

ML algorithms are the specific algorithms which learn from the past/observational data, automatically detect the patterns/trends/regularities from the given data and make predictions based on them.

For example, in order to predict a car resale price, we first feed cars-sales data to the machine and let it learn various patterns in the data and predict resale price.

So Machine Learning is a concept which allows the machine to learn from the past data.

Types of Machine Learning

Machine Learning techniques are broadly classified as follows -

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

Supervised Learning

Supervised learning is often done under someone's supervision (target/output variable).

In supervised learning, the machine is presented with labelled data, where each input record has a corresponding labelled output. And, the machine learns/approximates a mapping from the input to the output.

For example, to predict the car resale price, consider the below dataset.

Manufacturer	Model	Registration Year	Kms driven	Ownership type	Price
Maruti	Wagon R	2012	42000	First	210000
Nissan	Sunny	2010	54000	Second	270000
Hyundai	Xcent	2015	28000	First	430000
Ford	Figo	2017	12000	Second	390000
Honda	City	2014	35000	Second	440000

Supervised Learning [Contd.]

Here, the machine learning algorithm learns the mapping from the input variables to the output variable. It predicts the resale price for any new car instance.

Manufacturer	Model	Registration Year	Kms driven	Ownership type	Price
Volkswagen	Polo	2013	32000	Second	345000

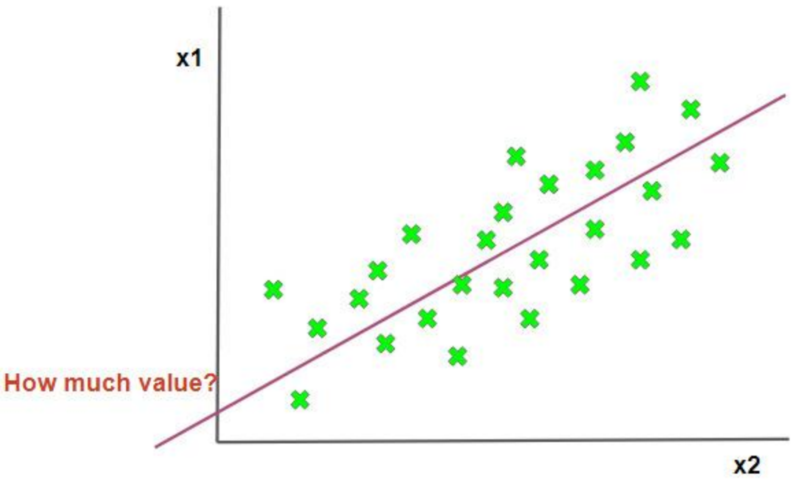
Some other examples of Supervised Learning -

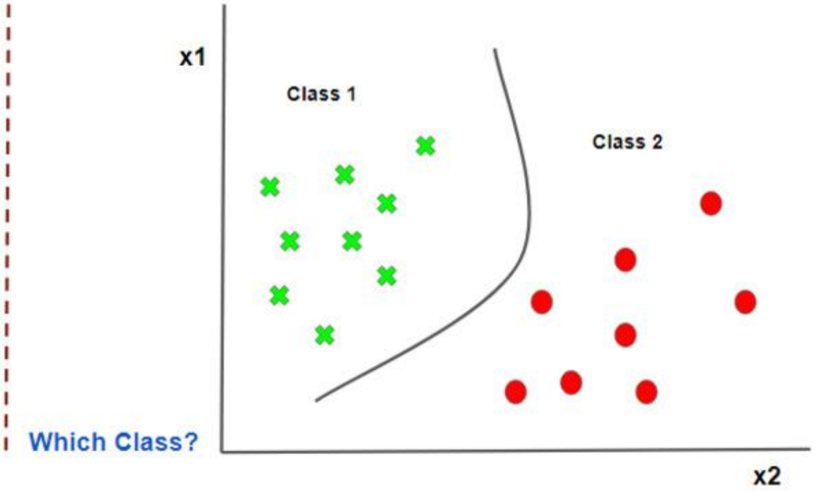
- Identifying an email as spam or ham.
- Predicting the delivery time of order placed.
- Determining whether the customer will be defaulter or not.
- Predicting the amount of water required in a city.

Types of Supervised Learning

Based on the type of target variable, supervised learning can be further categorised as -

- Regression - When the target variable is continuous/numeric/quantitative in nature.
  E.g. Price of mobile phones, employee salaries, loan amount etc.
- Classification - When the target variable is categorical/qualitative in nature.
  E.g. whether a transaction is fraudulent or not, email spam or ham.





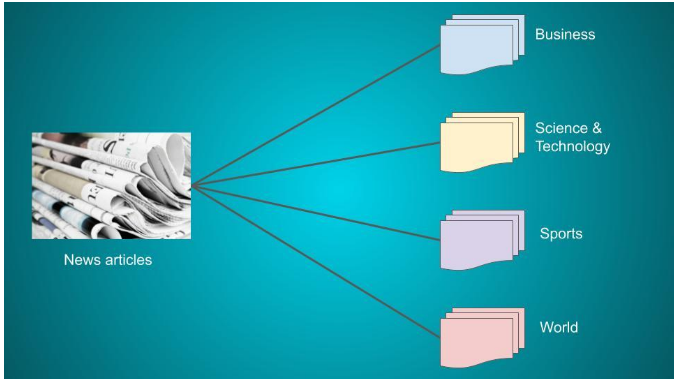
Unsupervised Learning

Unsupervised Learning has no explicit output/target variable, i.e. works without supervision.

So it discovers knowledge, hidden structures or relationship in the unlabelled data. For e.g. it can learn to group or organize data in such a way that similar objects are in the same group.

Similarly, news articles can be put together based on the topics like sports, business, technology, politics etc. This approach is known as Clustering.

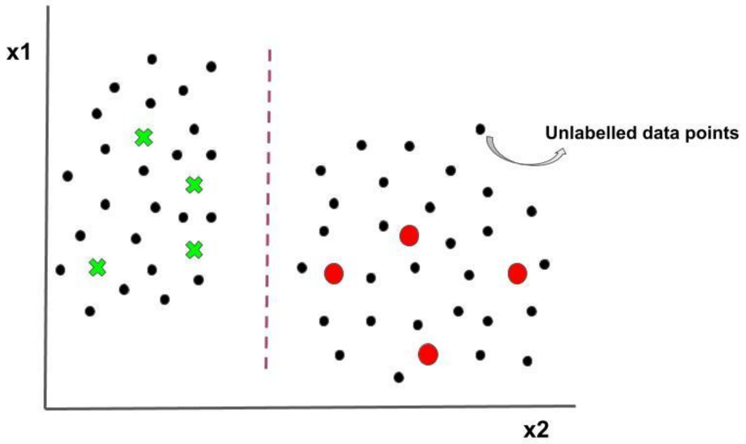
Clustering is a technique to group a set of objects in such a way that objects in the same group are much more similar to each other than to those in other groups.



Semi-Supervised Learning

Hybrid learning problems which fall between Supervised and Unsupervised learning.
In Semi-supervised learning, only a small amount of data is labelled whereas most of the data is unlabelled.

Here, **either** unsupervised learning can be used to discover and learn the structure in the input variables.
OR supervised learning can be used to make best guess(prediction) about the unlabelled data.

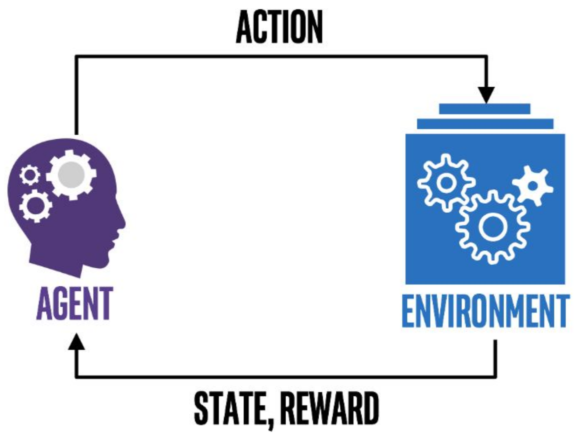


Reinforcement Learning

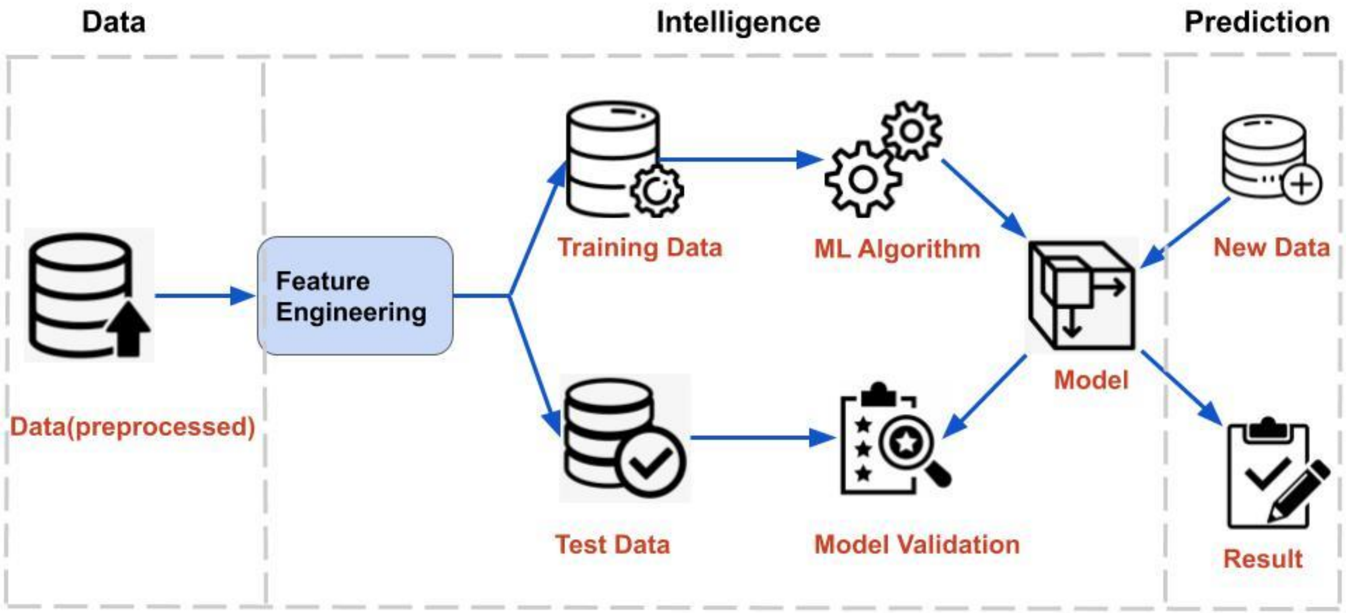
Reinforcement learning is a goal-oriented learning based on the interaction of an agent with the environment.

It describes a class of problems where an agent operates in an environment and must learn to operate using feedback in terms of punishment or rewards.
The agent attempts to maximize the accumulated rewards over time.

Applications - in Google's Alpha Go, In robotics for industrial automation etc.



Machine Learning Process



Quiz 1

Which of the following falls under Supervised Learning?

- Identifying whether a patient has brain tumour or not based on brain scan images.
- Grouping/Segmenting the customers based on past behaviour.
- Predicting the house price.
- Sentiment Analysis

Quiz 2

Which phase of Machine Learning process examines usefulness/fitness of the built model?

- Feature Engineering
- Model Validation
- Model Fitting
- Train-test splitting

Regression

- Regression is a statistical model to estimate the relationship between the independent variables (X) and dependent variable (y).

- The relationship can be either linear or non-linear.

- A regression model is represented as

  $$\mathbf{y = f(X)}$$

  where, y is the target/dependent/response variable and X is a set of predictors/independent variables (x1, x2, x3………………..xn).

Types of Regression

Regression can be further categorized as -

1. Simple Linear Regression
2. Multiple Linear Regression
3. Polynomial Regression
4. Exponential Regression
5. Lasso Regression
6. Ridge Regression

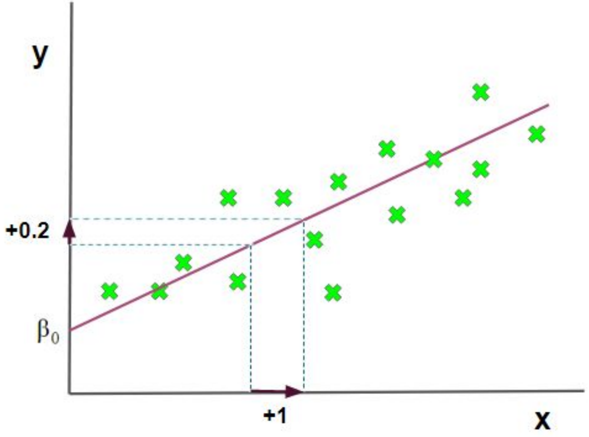
Simple Linear Regression

If the linear regression model involves only one predictor variable then it is known as Simple Linear Regression.

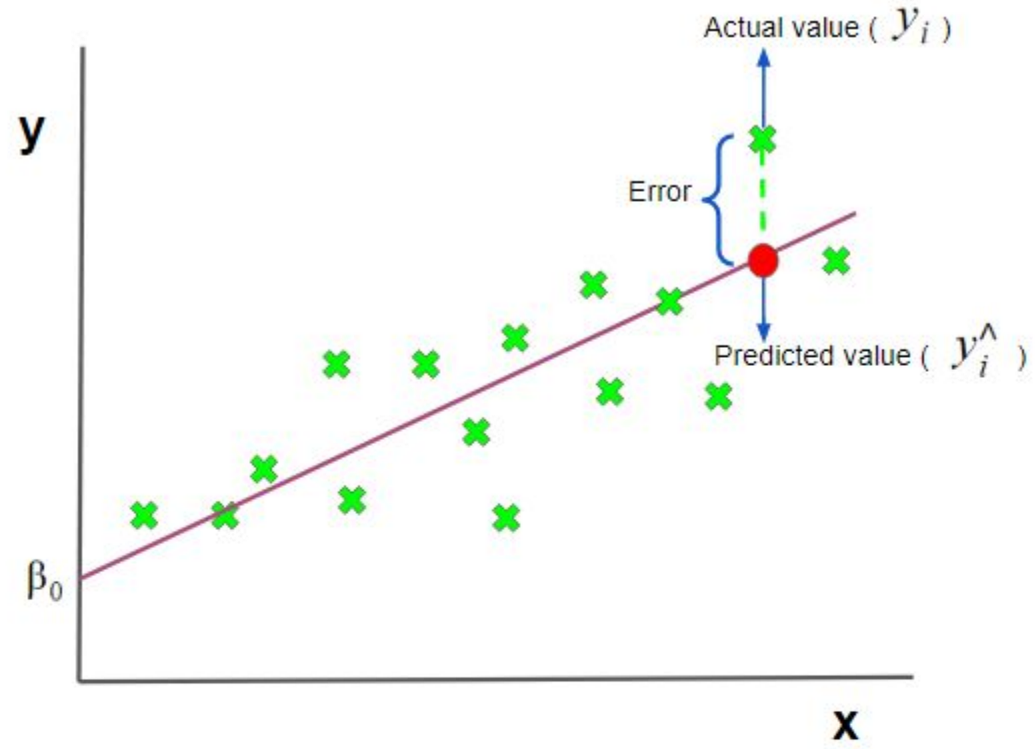
Mathematically,

$$y = f(X) = \beta_0 + \beta_1 x$$

Here, β_0 is the intercept and β_1 is the regression coefficient. This equation is analogous to line equation (y = mx + c). The slope indicates how the **y** varies with one unit change in **x**.



Simple Linear Regression



Error = $(y_i - y_i^\wedge)$

Finding optimal values of the coefficients

The error needs to be minimum for a good regression model. Optimal values of the coefficients are determined in such a way that the error(cost function) is minimum.

Cost (loss) function = $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

The square term in the cost function avoids the chances of positive and negative errors cancelling each other.

There are multiple ways to determine regression coefficients values such as

1. Differentiation (Differentiating the cost function w.r.t β_0 to get β_1 and vice-versa).
2. Closed form solution ($\beta = (X^T X)^{-1} X^T y$)
3. Gradient Descent

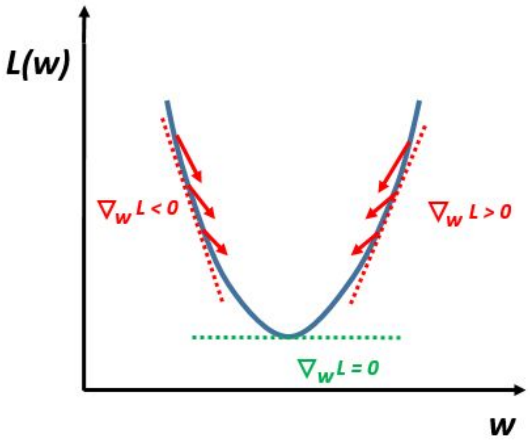
Gradient Descent (Generalized)

In Gradient descent, we calculate derivatives of loss w.r.t the parameters(coefficients) and update the parameters in the opposite direction of the gradient until the loss is minimized.

$$w = w - \eta\nabla w\,;\; b = b - \eta\nabla b$$

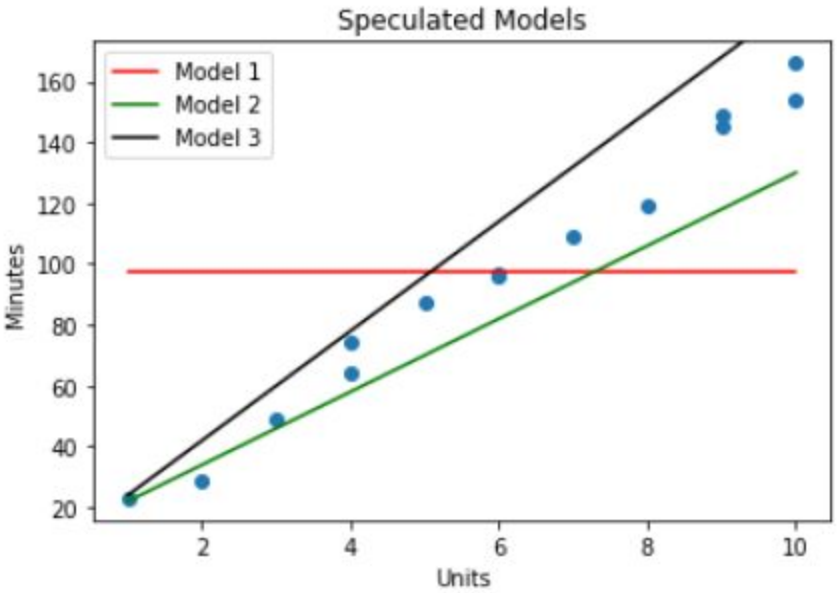
$$\text{where } \nabla w = \frac{\partial L(w)}{\partial w} \text{ and } \nabla b = \frac{\partial L(b)}{\partial b}$$

If the gradient is negative then **descent**(dive) towards the positive side and if the gradient is positive then descent towards the negative side until the minimal value of gradient is found.



Quiz 3

Choose the worst fit model out of below 3.



- Model 2
- Model 1
- Model 3
- All are good

Types of Error or Evaluation metrics in Regression

- Mean Squared Error (MSE) is the average squared difference between the actual values and predicted values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE) is the square root of average squared difference between the actual values and predicted values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Mean Absolute Error (MAE) is the absolute difference between the actual values and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

- Mean Absolute Percentage Error (MAPE) is the percentage equivalent of MAE.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Evaluation metrics in Regression [Contd.]

Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)
	RMSE & MSE share many properties with MSE because RMSE is simply the square root of MSE.		MAPE& MAE share many properties with MAE because MAPE is the percentage equivalent to MAE..
MSE is highly biased for higher values.	RMSE is better in terms of reflecting performance when dealing with large error values.	MAE is less biased for higher values. It may not adequately reflect the performance when dealing with large error values.	
	RMSE tends to be higher than MAE as the sample size goes up.	MAE is less than RMSE as the sample size goes up.	
MSE penalize large errors.	RMSE penalize large errors.	MAE doesn’t necessarily penalize large errors.	

Quiz 4

Choose the correct statement(s).

- Mean Squared Error (MSE) is less affected by the outliers.
- All four types of errors range from 0 to infinity.
- In Mean Absolute Error (MAE), positive and negative error values may cancel each other.
- Mean Squared Error (MSE) is a square of Root Mean Squared Error (RMSE).

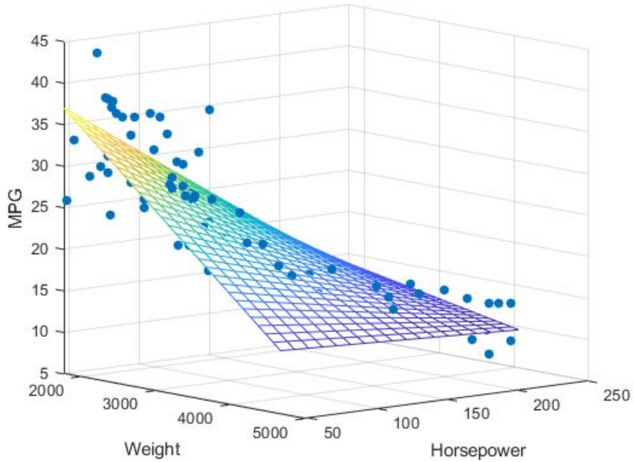
Multiple Linear Regression

If the linear regression model involves multiple predictor variables then it is known as Multiple Linear Regression.

Mathematically,

$$y = f(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots\dots\dots\dots\dots \beta_n x_n$$

Here, β_0 is the intercept and β_1, β_2, β_n are the regression coefficients.

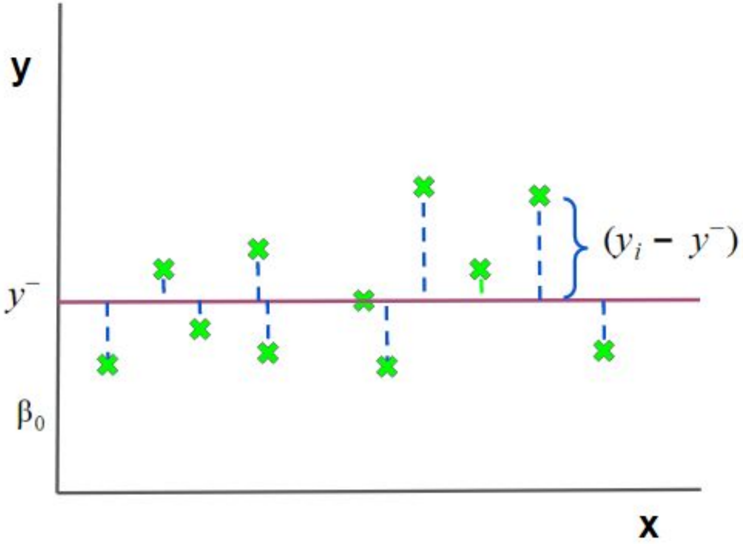


R-squared or Coefficient of Determination

The usefulness/fitness of a linear regression model can be determined using the coefficient of Determination (R^2).

Mathematically,

$$R^2 = 1 - \frac{SSE}{SST}$$



$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R-squared [Contd.]

- R-squared indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.
  For e.g. an R^2 value of 0.98 indicates that 98% variability in the dependent variable can be explained by the predictors variables collectively.

- R-squared generally lies between 0 to 1.

$$R^2 = 1 - \frac{SSE}{SST}$$

Case I - SSE = 0 then R^2 = **1** i.e. the ideal (best possible) model.

Case II - SSE = SST then R^2 = **0** i.e. the built model is equivalent to simple mean model.

Case III - SSE < SST then **0** < R^2 < **1**.

Case IV - SSE > SST 0 then R^2 = negative i.e. the built model is worse than the simple mean model.

Adjusted R-squared

Issues with R-squared

- With addition of every predictor variable, R^2 value keeps on increasing since there always exists a very small of correlation between the dependent variable and the predictor.

- The disadvantage with R^2 is that it assumes every predictor variable in the model explains variations in the dependent variable.

So R^2 doesn't tell exactly whether the model performance increases or decreases with addition of a new predictor variable.

Therefore, we use **Adjusted R-squared** as it takes number of predictor variables in account.

$$Adj.\, R^2 = 1 - \frac{(1 - R^2)(N - 1)}{(N - p - 1)}$$

Where, **p** = number of predictors in consideration

N = Number of data points

Adjusted R-squared Intuition

Case 1 - Profit = f(R&D Spend)

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.947
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	849.8
Date:	Tue, 19 Jan 2021	**Prob (F-statistic):**	3.50e-32

Case 2 - Profit = f(R&D Spend, Marketing Spend)

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.950
Model:	OLS	**Adj. R-squared:**	0.948
Method:	Least Squares	**F-statistic:**	450.8
Date:	Tue, 19 Jan 2021	**Prob (F-statistic):**	2.16e-31

Case 3 - Profit = f(R&D Spend, Marketing Spend, Administration)

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.948
Method:	Least Squares	**F-statistic:**	296.0
Date:	Tue, 19 Jan 2021	**Prob (F-statistic):**	4.53e-30

Case 4 - Profit = f(R&D Spend, Marketing Spend, Administration, State)

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	169.9
Date:	Tue, 19 Jan 2021	**Prob (F-statistic):**	1.34e-27

Quiz 5

Choose the correct statement(s).

- R-squared decreases with addition of new predictors.
- Adjusted R-squared never increases with addition of new predictors.
- Adjusted R-squared starts decreasing (or remain same) after a certain threshold.
- None of the above

Multicollinearity

In multiple linear regression, it could be possible that a single or group of predictors derives the another predictor i.e the predictors of the model are highly correlated. This phenomenon is called multicollinearity.

As the name suggests, multicollinearity is the collinearity between the variables (predictors).

Assessing Multicollinearity

Variance Inflation Factor (VIF) is used to determine if the predictors in the model are independent of each other or not.

$$VIF = \frac{1}{(1 - R_i^2)}$$

Where, R_i^2 is the coefficient of determination while predicting the candidate predictor using rest of the predictors.

If VIF > 5, then predictors are said to be correlated.

Quiz 6

Choose the correct statement(s).

- VIF ranges from 1 to infinity.
- VIF values of other features varies (although to a smaller extent). after dropping a variable with highest VIF.
- VIF is preferred over correlation as it can capture the relationship of a predictor with a group of other predictors.
- All of the above

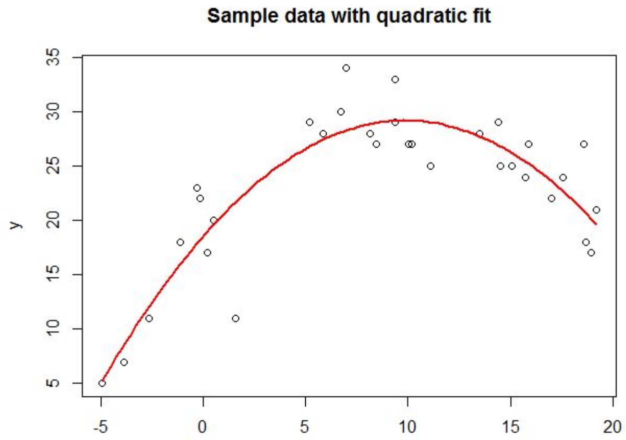
Polynomial Regression

*"**Polynomial regression** is a form of regression analysis in which the relationship between the independent variable **x** and the dependent variable **y** is modelled as an nth degree polynomial in **x**."* **-** Wikipedia

Mathematically,

$$y = f(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \ldots\ldots\ldots\ldots\ldots\ldots \beta_n x_n^n$$

Here, β_0 is the intercept and β_1, β_2, β_n are the regression coefficients.



Exponential Regression

An exponential regression model maps the equation of exponential function that best fits for a set of data.

Mathematically,

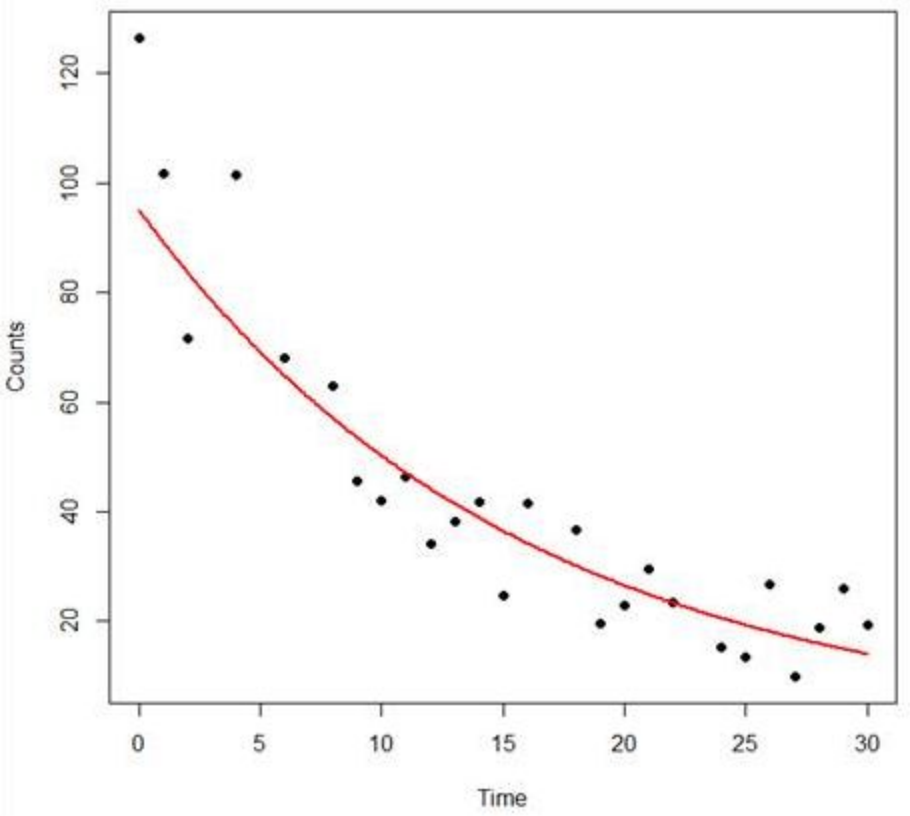
$$y = f(X) = \beta_0 e^{\beta_1 x}$$

Here, β_0 is the intercept and β_1 is the regression coefficient.

Taking natural logarithm on both the sides forms a linear regression equation.

$$log(y) = log(\beta_0) + \beta_1 x$$

Exponential Regression [Contd.]



Linear Regression Assumptions

There are five assumptions associated with a linear regression model:

1. **Linearity** : There should be a linear relationship between the dependent/target variable and independent/predictor variables.

2. **No or little multicollinearity** : The predictor variables are assumed to be independent of each other.

3. **Homoscedasticity** : The error (residual) terms should have constant variance.

4. **Normality** : The error terms should be normally distributed.

5. **Little or No autocorrelation** : Autocorrelation occurs when the residual errors are dependent on each other so the errors of the model should be statistically independent of each other.

Validating Linear Regression Assumptions

1.	**Linearity**	Scatter plot or Residual plot
2.	**No or little multicollinearity**	VIF
3.	**Homoscedasticity**	Scale-location plot
4.	**Normality**	Graphical test (Histogram, Q-Q plot), Numeric test (Shapiro-Wilk test, K-S test)

Quiz 7

The normality check of the errors can be examined using -.

- Q-Q plot
- Shapiro-Wilk test
- Chi-Square test
- Histogram

Quiz 8

Which of the following is(are) valid assumptions about linear regression?.

- The residuals of the model should be normally distributed.
- There should be little or less auto-correlation between the residuals.
- Predictors should be dependent of each other.
- All of the above

Analysing the coefficients

The regression coefficients indicate the relationship between each independent variable and target variable.

How to assess the significance of a feature?

Using hypothesis testing and based on p-values for each of the feature we assess whether the feature is significant or not.

p-value < 0.05	Reject the null hypothesis, means the feature has some significance and need to be retained.
p-value > 0.05	Accept the null hypothesis, means the feature is not significant and can be removed.

Feature Selection

- Feature Selection is a process of selecting the most significant and relevant features from a vast set of features in the given dataset.

- For a dataset with **d features**, if we apply the hit and trial method with all possible combinations of features then total (**2^d – 1**) models need to be evaluated for a significant set of features.

  So It is a time-consuming approach, therefore, we use feature selection techniques to find out the smallest set of features more efficiently.

There are three types of feature selection techniques :

1. Filter methods
2. Wrapper methods
3. Embedded methods

Feature Selection using Wrapper methods

Wrapper method follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.

The evaluation criterion is simply the performance measure which depends on the type of problem,

For e.g. For regression evaluation criterion can be p-values, R-squared, Adjusted R-squared etc.

Similarly for classification the evaluation criterion can be accuracy, precision, recall, f1-score, etc.

Most commonly used techniques under wrapper methods are:

1. Forward selection
2. Backward elimination
3. Bi-directional elimination(Stepwise Selection)

Forward Selection

In forward selection, we start with a null model and then start fitting the model with each individual feature one at a time and select the feature with the minimum p-value.

The steps for the forward selection technique are as follows :

1. Choose a significance level (e.g. SL = 0.05 with a 95% confidence).
2. Fit all possible simple regression models by considering one feature at a time. Total **’n’** models are possible. Select the feature with the lowest p-value.
3. Fit all possible models with one extra feature added to the previously selected feature(s).
4. Again, select the feature with a minimum p-value. if p_value < significance level then go to Step 3, otherwise terminate the process.

Backward Elimination

In backward elimination, we start with the full model (including all the independent variables) and then remove the insignificant feature with the highest p-value(> significance level). This process repeats again and again until we have the final set of significant features.

The steps involved in backward elimination are as follows:

1. Choose a significance level (e.g. SL = 0.05 with a 95% confidence).
2. Fit a full model including all the features.
3. Consider the feature with the highest p-value. If the p-value > significance level then go to Step 4, otherwise terminate the process.
4. Remove the feature which is under consideration.
5. Fit a model without this feature. Repeat the entire process from Step 3.

Bi-directional elimination (Stepwise Selection)

It is similar to forward selection but the difference is while adding a new feature it also checks the significance of already added features and if it finds any of the already selected features insignificant then it simply removes that particular feature through backward elimination.

Hence, It is a combination of forward selection and backward elimination.

The steps involved in bi-directional elimination are as follows:

1. Choose a significance level to enter and exit the model (e.g. SL_in = 0.05 and SL_out = 0.05 with 95% confidence).
2. Perform the next step of forward selection (newly added feature must have p-value < SL_in to enter).
3. Perform all steps of backward elimination (any previously added feature with p-value>SL_out is ready to exit the model).
4. Repeat steps 2 and 3 until we get a final optimal set of features.