REGRESSION

The definitions of regression

- Regression is the process of predicting a continuous value using some other variables
- There are 2 types of variables in regression:
    - Dependent variable (predict value)
        - This variable can be seen as the state, target or final goal we study and try to predict
        - Shown notated by Y
    - Independent variable (predictor value)
        - This variable also known as explanatory variables can be seen as the cause of those states
        - Shown conventionally by X
- Regression model relates dependent variable to a function of independent variable
- The key point in the regression is that **dependent variable value should be continuous and cannot be a discrete value**, however the **independent variables can be measured on either a categorical or continuous measurement scale**
- There are 2 types of regression models:
    - Simple Regression
        - This regression is when 1 independent variable is used to estimate a dependent variable
        - Types of this regression
            - Simple Linear Regression
            - Simple Non-linear Regression
    - Multiple Regression
        - This regression is when more than 1 independent variable are used to estimate a dependent variable
        - Types of this regression
            - Multiple Linear Regression
            - Multiple Non-linear Regression
- There are many regression algorithms which each of them has its own importance and a specific condition to which their application is best suited
    - Ordinal regression
    - Poisson regression
    - Fast forest quantile regression
    - Linear, Polynomial, Lasso, Stepwise, Ridge regression
    - Bayesian linear regression
    - Neural network regression
    - Decision forest regression
    - Boosted decision tree regression
    - KNN (K-nearest neighbors)

- Linear Regression is the approximation of a linear model used to describe the relationship between 2 or more variables
- In this regression, there are 2 variables
    - Dependent variable
    - Independent variable
- The key point in this regression is that the dependent value should be continuous and cannot be a discrete value, however the independent variables can be measured on either a categorical or continuous measurement scale
- Linear regression topology
    - Simple Linear Regression
        - This regression is when 1 independent variable is used to estimate a dependent variable
    - Multiple Linear Regression
        - This regression is when more than 1 independent variable are used to estimate a dependent variable

Simple Linear Regression

$\hat{y} = \beta_0 + \beta_1 x_1$ is the math equation for Simple Linear Regression

- $\hat{\mathbf{y}}$ is the dependent variable of the predicted value
- $\mathbf{x_1}$ is the independent variable
- $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are also called **the coefficients** of the linear equation
    - $\boldsymbol{\beta_0}$ is known as **the intercept**
    - $\boldsymbol{\beta_1}$ is known as **the slope** or gradient of the fitting line
- we can interpret this equation as $\hat{\mathbf{y}}$ being a function of $\mathbf{x_1}$ or $\hat{\mathbf{y}}$ being dependent of $\mathbf{x_1}$
- these are how we determine which line fits best with the data (let say we already fitted the model to the data and need to be checked if is the model is good fit or bad fit):
    - Linear Regression estimates the coefficients of the line which means we must calculate $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ to fit the data
    - y is the actual value and ŷ is the predicted value
    - we can compare the actual value of the dependent variable with the predicted value and find the error value using this equation $\text{Error} = y - \hat{y}$
        - with this error value can see how accurate our model to predict a value
        - this error is called a **Residual Error**
        - the mean of all residual errors shows how poorly the line fits with the whole data set
        - mathematically, the mean of all residual errors can be shown by the equation Mean Squared Error (MSE) with this formula:
            - $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
        - Our objective is to find a line where the MSE is minimized
- The objective of Linear Regression is to minimize the MSE value and to minimize it we should find the best parameters of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ with this way based on mathematic approach:
    - It requires that we calculate the mean of the independent and dependent from the data set
    - Notice that all of the data must be available to traverse and calculate the parameters
    - Calculate the intercept ($\boldsymbol{\beta_1}$)
        - using this equation $\beta_1 = \frac{\sum_{i=1}^{S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{S}(x_i - \bar{x})^2}$
        - calculate this value to find the slope of a line based on the data
        - $\bar{x}$ and $\bar{y}$ are the average value for the independent the dependent variable in our data set
        - x_i and y_i in the equation refer to the fact that we need to repeat these calculations across all values in our data set
        - this value is the coefficient for dependent variable
    - calculate the slope ($\boldsymbol{\beta_0}$)
        - once we get the value of $\boldsymbol{\beta_1}$ then we use this equation $\beta_0 = \bar{y} - \beta_1\bar{x}$ to find the slope
        - this value is also called the bias coefficient
    - now we can write down the polynomial of the line so we know how to find the best fit for our data and its equation
- pros of Linear Regression
    - it is the most basic regression to use and understand
    - very fast
    - it doesn't require tuning of parameters
    - easy to understand and highly interpretable

---

the goal of regression is to build a model to accurately predict an unknown case. To this end, we have to perform **regression evaluation** after building the model. There are 2 types of evaluation approaches that can be used to achieve this goal:

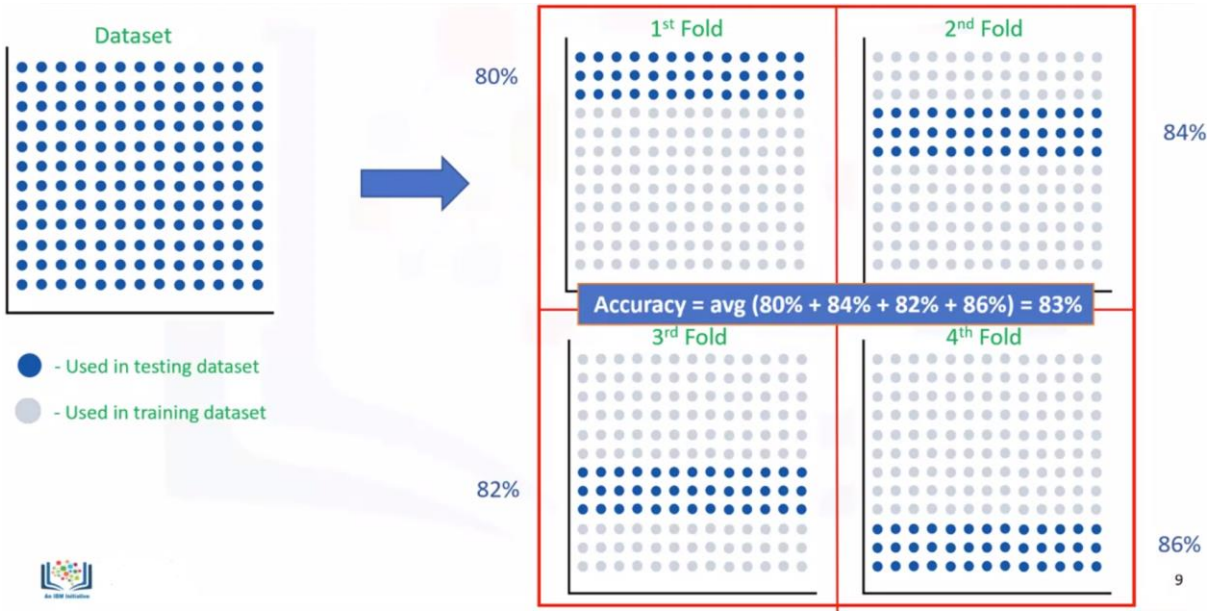
- Train and Test on the same Dataset
- Train/Test Split

When considering evaluation models, we clearly want to choose the one that will give us the most accurate results. We can calculate the accuracy of our model and determine how much we can trust this model for prediction of an unknown sample using a given dataset and having built a model such as linear regression with these solutions:

- (**Train and Test on the same Dataset**) Select a portion of our dataset for testing while we use the entire dataset for training and we build a model using this training set
    - We select a small portion of the dataset (about 10%-30% of our dataset) but without the labels (dependent value) and this set is called **Test Set** which has the labels but the labels are not used for prediction and is used only as ground truth, and the labels are called actual values of the test set
    - We pass the feature set of the testing portion to our built model and predict the target values
    - Finally, we compare the predicted values by our model with the actual values in the test set to indicate how accurate our model actually is
    - There are different metrics to report the accuracy of the model but most of them work generally based on the similarity of the predicted and actual values
        - This $Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$ equation is one of the simplest metrics to calculate the accuracy of our regression model
            - The error of the model is calculated as the average difference between the predicted and actual values for all the rows
    - When we test with a dataset in which we know the target value for each data point, we're able to obtain a percentage of accurate predictions for the model
    - Since the model knows all of the testing data points from the training, this evaluation approach would most likely have
        - **High "training accuracy"**
            - This is the percentage of correct predictions that the model makes when using the test dataset
            - However, this isn't necessarily a good thing
            - Having this may result in an over-fit of the data
            - **Over-fit** means the model is overly trained to the dataset which may capture noise and produce a non-generalized model
        - **Low "out-of-sample accuracy"**
            - This is the percentage of correct predictions that the model makes on data that model has not been trained on
            - Doing a train and test on the same dataset will most likely have low out-of-sample accuracy due to the likelihood of being over-fit
            - It's important that our models have a high out-of-sample accuracy because the purpose our model is to make correct predictions on unknow data
            - We can improve out-of-sample accuracy is to use another evaluation approach called Train/Test split
- (**Train/Test Split**) in this approach we split our dataset into 2 sets which are mutually exclusive, Train set and Test set
    - we select a portion of our dataset for training (about 70%-90% of our dataset) and the rest is used for testing which the model is built on the training set
    - the test feature set is passed to the model for prediction
    - Finally, the predicted values for the test set are compared with the actual values of the testing set
    - this evaluation approach would most likely have
        - more accurate evaluation on **out-of-sample accuracy** because the testing dataset is not part of the dataset that has been used to train the data which is more realistic for real-world problems

- the issue with this evaluation approach is that it's highly dependent on the datasets on which the data was trained and tested
- the variation of this causes Train/Test Split to have a better out-of-sample prediction than training and testing on the same dataset, but it still has some problems due to this dependency
    - another evaluation model (**K-fold Cross Validation**) resolves most of these issues

- (**K-fold Cross Validation**) this approach fixes high variation that results from a dependency

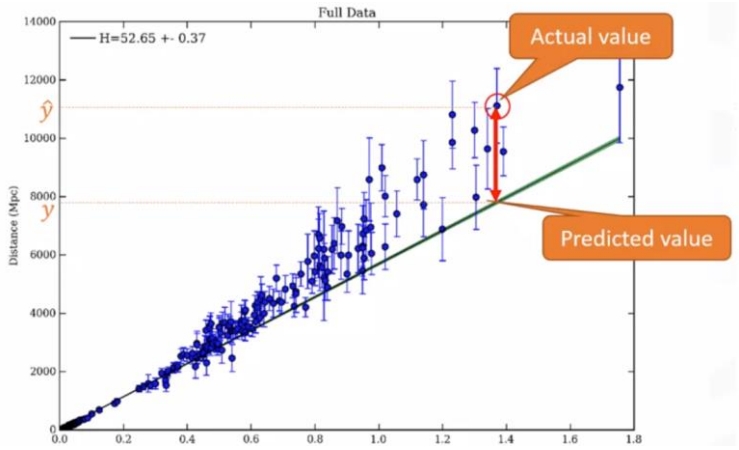


- The entire dataset is represented by the point in the image at the top left
- If we have K equals 4 folds, then we split up that dataset as shown at top right
- In the first fold for example, we use the first 25% of the dataset for testing and the rest for training
- The model is built using the training set and is evaluated using the test set
- Then in the second fold, the second 25% of the dataset is used for testing and the rest for training the model and the accuracy of the model is calculated
- We continue for all folds and finally the result of all four evaluations are averaged
- That is, the accuracy of each fold is then averaged, keeping in mind that each fold is distinct where no training data in one-fold is used in another
- This approach in its simplest form performs multiple Train/Test Splits using the same dataset where each split is different then the result is average to produce a more consistent out-of-sample accuracy

---

Evaluation Metrics are used to explain the performance of a model and we'll define more about the model evaluation metrics that are used for regression

- Basically, we can compare the actual values (y) and predicted values ($\hat{y}$) to calculate the accuracy of our regression model
- Evaluation metrics provide a key role in the development of a model as it provides insight to areas that require improvement



- In the context of regression we define **the error of the model is the difference between the data points and the trend line generated by the algorithm**, since there are multiple data points an error can be determined in multiple ways
- Here are some of model evaluation metrics that can be used for quantifying of our prediction:
    - **Mean Absolute Error (MAE)**
        - This is the equation of MAE $MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$
        - MAE is the mean of the absolute value of the errors (this is just the average error)
    - **Mean Squared Error (MSE)**
        - This is the equation of MSE $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
        - MSE is the mean of the squared error and this is more popular than MAE because the focus is geared more towards large errors
        - This is due to the squared term exponentially increasing larger errors in comparison to smaller ones
    - **Root Mean Squared Error (RMSE)**
        - This is the equation of RMSE $RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$
        - RMSE is the square root of the MSE which one of *the most popular of the evaluation metrics* because RMSE is interpretable in the same units as the response vector or $\boldsymbol{y}$ units making it easy to relate its information
    - **Relative Absolute Error (RAE)**
        - This is the equation of RAE $RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$
        - Also known as Residual Sum of Square, where $\bar{\boldsymbol{y}}$ is a mean value of $\boldsymbol{y}$, takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor
    - **Relative Squared Error (RSE)**
        - This is the equation of RSE $RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$
        - RSE is very similar to RAE but is *widely adopted* by the data science community as it is used for calculating R-Squared
    - **R-Squared (R^2)**
        - This is the equation of R-Squared $R^2 = 1 - RSE$
        - R^2 is not an error per SE but it is a popular metric for the accuracy of our model that represents how close the data values are to the fitted regression line
        - **The higher the R^2 the better the model fits our data**
- The choice of metric completely depends on the type of model, datatype, and domain of knowledge

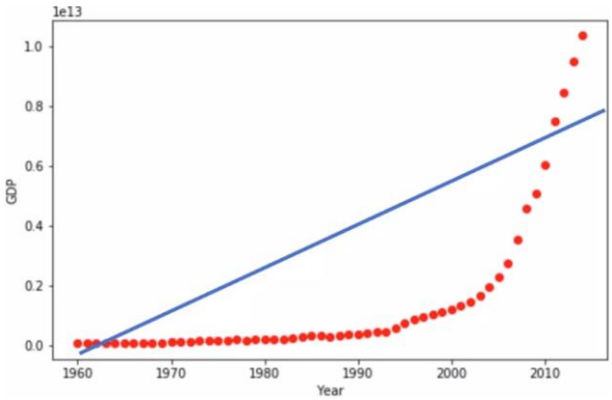
Multiple Linear Regression

- Multiple Linear Regression (**MLR**) is the extension of the Simple Linear Regression model
- MLR is a method of predicting a continuous variable that uses multiple variables called independent variables or predictors that best predict the value of the target variable which is also called the dependent variable
- Basically, there are 2 applications for MLR
  - It can be used when we would like to identify the strength of the effect that the independent variables have on the dependent variable
  - It can be used to predict the impact of changes, that is, to understand how the dependent variable changes when we change the independent variables
- In MLR, the target value ($\boldsymbol{y}$) is a linear combination of independent variables ($\boldsymbol{x}$)
- MLR is very useful because we can examine which variables are significant predictors of the outcome variable and also, we can find out how each feature impacts the outcome variable
- Generally, the model of MLR is of the form $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ or $\hat{y} = \beta^T X$
- In higher dimensions when we have more than 1 input ($\boldsymbol{x}$), the line is called a plane or a hyperplane and this is what we use for MLR
- The whole idea is to find the best fit hyperplane for our data and also we should estimate the values of intercept and slopes that best predict the value of the target field in each row
  - To achieve this goal, we have to minimize the error of the prediction
- **Optimized parameters** are the ones which lead to a model with the fewest errors
- Before we go to prediction phase, we need to find the optimized parameters for our model with these ways:
  - let's assume that we have already found the parameter vector of our model, it means we already know the values of $\boldsymbol{\beta^T}$
  - find the different error value between the predicted value and the actual value (Residual Error) by subtracting the actual value with the predicted value – this value we can get from a single observation, which means we need to find all the values from all observations data
    - as in the case in Linear Regression, we can say the error here is the distance from the data point to the fitted regression model
  - the mean of all residual errors shows how bad the model is representing the data set
    - it is called the MSE and this is one of the most popular ways to do so
  - the best model for our data set is the one with minimum error for all prediction values so **the objective of MLR is to minimize the MSE equation** and to minimize it we should find the best parameters $\boldsymbol{\beta^T}$ by estimating MLR parameters:
    - there are several ways to estimate the best parameters or coefficients for MLR
      - **Ordinary Least Squares (OLS)**
        - It tries to estimate the values of the coefficients by minimizing the MSE
        - This approach uses the *data as a matrix* and uses *linear algebra operations* to estimate the optimal values for $\boldsymbol{\beta^T}$
        - The problem with this technique is the time complexity of calculating matrix operations as it can take a very long time to finish – especially when we have large datasets (10K+ rows)
      - **Optimization algorithm**
        - We can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on our training data
        - **Gradient Descent** – starts optimization with random values for each coefficient then calculates the errors and tries to minimize it through y's changing of the coefficients in multiple iterations
        - Gradient Descent is a proper approach if we have a large data set
  - After we find the best parameters for our model then we can go to the prediction phase

Some concerns that we might already be having regarding MLR

- How to determine whether to use simple or multiple linear regression?
  - We can use multiple independent variables to predict a target value in MLR and it sometimes results in a better model compared to using a SLR which uses only one independent variable to predict the dependent variable
- How many independent variables should we use for the prediction?
  - Basically, adding too many independent variables without any theoretical justification may result in overfit model
  - An overfit model is a **real problem** because it is too complicated for our data set and not general enough to be used for prediction
  - **It is recommended to avoid using many variables for prediction**
  - There are different ways to avoid overfitting a model in regression
- Should independent variables be continuous?
  - Basically, categorical independent variables can be incorporated into regression model by converting them into numerical variables – for example OneHotEncoder method
- Remember that MLR is a specific type of Linear Regression so there needs to be a linear relationship between the dependent variable and each of independent variables. There are a number of ways to check for linear relationship
  - For example – we can use scatter plots and then visually checked for linearity, if the relationship displayed in scatter plot is not linear then we need to use non-linear regression

Non-Linear Regression

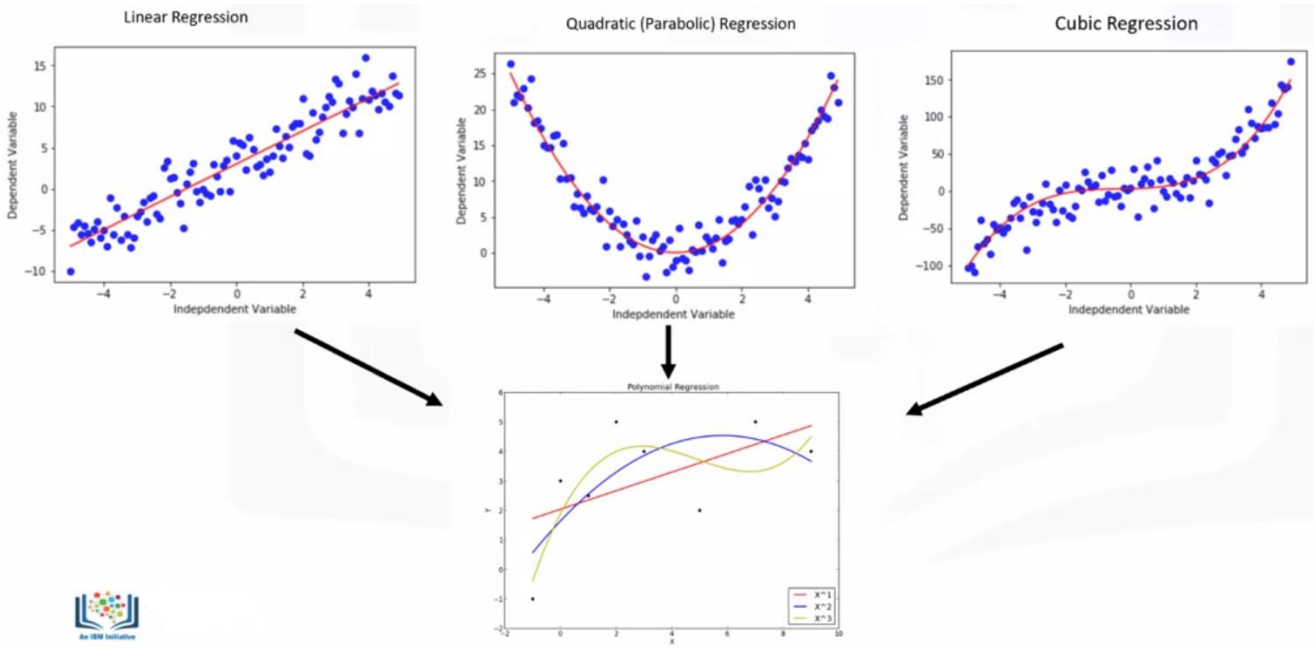


If the data shows a curvy trend (red dots), then linear regression (blue line) wouldn't produce very accurate results when compared to a non-linear regression. Simply because linear regression presumes that the data is linear like the scatterplot on the above

- The scatter plot shows that there seems to be strong relationship between "*GDP*" and "*Year*", but the relationship is not linear
- It looks like either a logistical or exponential function so it requires a special estimation method of the **Non-Linear Regression** procedure

$\hat{y} = \beta_0 + \beta_1 {\beta_2}^x$ is the main exponential function that we need to build a Non-Linear model

- Our job is to estimate the parameters of the model ($\boldsymbol{\beta}$) and use the fitted model to predict unknown or future cases
- Many different regressions exist that can be used to fit whatever the dataset looks like
  - Linear Regression – 1st degree of polynomial
  - Quadratic (parabolic) Regression – 2nd degree of polynomial
  - Cubic Regression – 3rd degree of polynomial
  - And it can go on and on to infinite degrees



- In essence, we can call all of those regression in the above as **Polynomial Regression**, where the relationship between the independent variable and the dependent variable is modelled as an N^{th} degree polynomial in X
- With many types of regression to choose from, there is a good chance that one will fit our dataset well

Polynomial Regression can be described as

- Some curvy data can be modelled by this regression and it fits a curve line to our data
- A simple example of polynomial with degree 3 is shown as $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ where ($\boldsymbol{\beta}$) are parameters to be estimated that makes the model fit perfectly to the underlying data
- Though the relationship between X and Y is non-linear here and Polynomial Regression can't fit them, a Polynomial Regression model can still be transformed into Linear Regression
  - Here is an example:
    - Given the 3rd degree polynomial equation by defining $\boldsymbol{x_1 = x}$, $\boldsymbol{x_2 = x^2}$ and $\boldsymbol{x_3 = x^3}$
    - The model is converter to a Simple Linear Regression with new variables as $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
    - Then this model is linear in the parameters to be estimated
- Therefore, this Polynomial Regression is considered to be a special case of traditional MLR so we can use the same mechanism as linear regression to solve such a problem
- Polynomial Regression models can fit using the model of **Least Squares**
  - Least Squares is a method for estimating the unknown parameters in a linear regression model by minimizing the sum of the squares of the differences between the observed dependent variable ($\boldsymbol{y}$) in the given dataset and those predicted ($\hat{\boldsymbol{y}}$) by the linear function

Non-Linear Regression can be explained as

- Is a method to model a non-linear relationship between the dependent variable and a set of independent variables
- For a model to be considered as non-linear, $\hat{\boldsymbol{y}}$ must be a non-linear function of the parameters $\boldsymbol{\beta}$, not necessarily the features x
- When it comes to non-linear equation, it can be the shape of exponential, logarithmic, logistic or many other types
  - $\hat{y} = \beta_0 + {\beta_2}^2 x$
  - $\hat{y} = \beta_0 + \beta_1 {\beta_2}^x$
  - $\hat{y} = log(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3)$
  - $\hat{y} = \frac{\beta_0}{1+{\beta_1}^{(x-\beta_2)}}$
- As you can see in all of those equations in above, the change of $\hat{\boldsymbol{y}}$ depends on changes in the parameters $\boldsymbol{\beta}$ not necessarily on $\boldsymbol{x}$ only. That is, in Non-Linear Regression, a model is non-linear by parameters
- In contrast to linear regression, we **can't use the Ordinary Least Squares (OLS) method to fit the data in Non-Linear Regression** and in general estimation of the parameters is **not easy**

Linear vs Non-Linear Regression with 2 important questions

- How can we know if a problem is linear or non-linear?
  - First, we need to visually figure out if the relation is linear or non-linear
    - It's best to plot **bivariate plots** of output variables with each input variable
    - Also, we can **calculate the correlation coefficient** between independent and dependent variables, and if for all variables it is 0.7 or higher there is a linear tendency and thus it's not appropriate to fit a Non-Linear Regression
  - Second, we use Non-Linear Regression instead of Linear Regression when **we can't accurately model the relationship with linear parameters**
- How should we model our data, if it displays non-linear on a scatter plot?
  - To address this question, we have to use either:
    - **Polynomial** Regression,
    - **Non-Linear** Regression model or
    - **Transform** our data