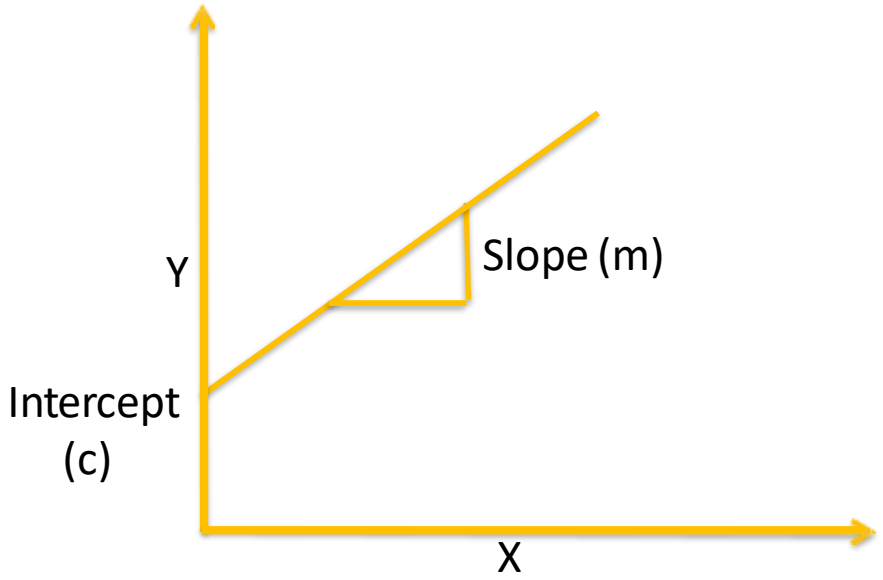
Regression - Extended

Linear Regression

- **What is Linear Regression ?**
  - In simple terms linear regression is the study of relation between dependent(Y) variable and independent(X) variable.
  - Simple linear regression equation is can be written as follows
    - **Y = mX+c where, m – slope and c – Y intercept**
  - Regression gives a best fit line for given data, which can be used to predict the dependent variable.



Linear Regression

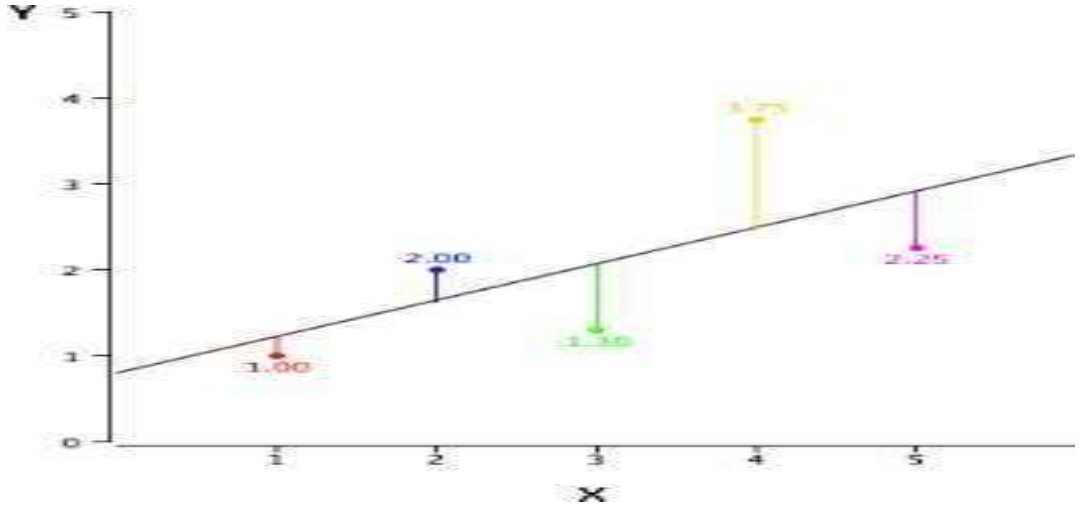
Example: We are predicting the salary of the new employees based on their educational qualification using the history data of other employees. Consider the following data where salary is in Y*1000 INR per month and Education in number of years.

Salary (Y)	Education (X_1)
82	14
48	10
60	14
85	16
72	10
62	10
90	14
101	16

- Regression equation for the data is:
  - Y (Salary) = m * Education(X) + c

Linear Regression

- **Calculation of slope and intercept:**
  - Slope(m) = 5.1
  - Intercept (c) = 8.9
- **So now regression equation becomes:**
  - **Y (Salary) = 5.1 * Education(X) + 8.9**



- **Obtaining best fit :** We use **LSM**(least square method) which gives the vertical distance between data point and the regression line.
- Using equation **min|Y – Xi|² where I = 1 to n.** The figure depicts the **LSM.**

Linear Regression

- **Multiple Linear Regression:** It is same as simple linear regression but there more than one predictor/independent variables to predict the target variable.

- **Formula:** $Y = m_0 + m_1X_1+m_2X_2+...+m_5X_5$
- **Example:** Consider the example of employee salary prediction, if we add features like **Education, Experience** and **Designation** of the employee to predict the salary of new employee then it becomes a multiple linear regression problem.

- **Limitations of Linear Regression:**
  - Very sensitive to outliers.
  - More susceptible for overfitting
  - Mainly linear regression is used to find out the linear relationship between variables, if the variables are non linear then resulting model will be very poor.

Stepwise Regression

- Helps us to build our regression model by adding or removing independent/predictor variables one at a time in stepwise to obtain better model.
- The list of independent variables obtained at the end stepwise adding/removing process should contain all the variables which predicts the target variable.
- Usually stepwise regression is performed on large datasets where regression on all independent variables is not suitable to get a good accuracy model. To avoid this we can only select the significant independent variables.
- **Alpha-to-enter (α_E) / Alpha-to-remove(α_R)**: We set a threshold to add/remove a predictor based on their significant values. By default **α_E or α_R =0.15**
- There are 3 types of stepwise regressions they are:
    1. Standard
    2. Forward
    3. Backward

Stepwise Regression

- **Standard Stepwise Regression**: In Standard Stepwise Regression we add/remove the independent variables based on our requirement.
- Example: Consider the example of predicting employee salary based on various independent variables.

Salary (LPA)	Experience (Years)	Education (Years)	Designation	Age	Address
7	3	16	Software Engineer	26	Place-A
8	2	18	Business Analyst	25	Place-B
3	4	12	Clerk	28	Place-C
12	6	16	Manager	30	Place-D
2	5	10	Security	31	Place-E
3	6	12	Clerk	32	Place-F

Stepwise Regression

- In the example table we have the independent variables which we can connect with salary of a particular employee.
- If we apply standard stepwise regression to add are remove variables stepwise.
- At the end we remove Age and Address because they are less significant.
- **Forward Stepwise Regression:** In this type of regression we first choose a single variable which has highest significant relation and in subsequent steps we add variables which are relatively less significant than previous variable but are essential for building good model.
- **Example:** In the employee salary prediction example we first choose the predictor variable **Education** and then in subsequent steps we add variables in the order **Experience, Designation, Age**.

Stepwise Regression

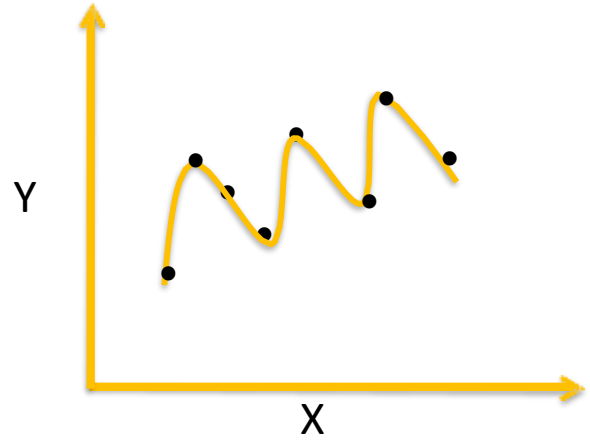
- **Backward Stepwise Regression: This is reverse of forward stepwise regression.** In this type of regression we first choose all independent/predictor variables and in subsequent steps we remove variables which are less significant in order to build a good model.
- **Example**: Consider the same example of salary prediction. Here we first choose all the predictor variables in first step and in subsequent steps we remove the variable **Address** and **Age** which are less significant to predict the salary.

Regularization

- To over come the problem of over-fitting we use the technique of regularization by introducing a new term called penalty in the equation which finds the coefficient of the independent variables.
- Consider a linear regression equation which predicts the value of **Y** using the predictors $X_0, X_1, X_2, .., X_5$. The equation looks like below.
  - $Y = m_0 + m_1X_1 + m_2X_2 + \ldots + m_5X_5$
- In above equation $m_0, m_1, m_2, \ldots m_5$ are called as coefficients of $X_0, X_1, X_2, .., X_5$ and we can calculate the coefficients using the Least Square Estimation.
  - $\text{Min } |Y_i - f(X_i)|^2$
- In the regression equation there are too many predictor variables i.e. This may leads to over-fitting.

Regularization

- **Over-fitting:** In a modeling when we try to fit a curve very closely in the presence of very less data points. **Following figure shows the over-fitted curve.**



- To avoid over-fitting, we make the changes to the Least Square estimator function as follows.
  - **Min $|Y_i - f(X_i)|^2 + \text{Penalty}$**
- The penalty may be of 2 types:
  - Ridge Regression (L2 Regularization)
  - Lasso Regression (L1 Regularization)

Mathematically

- In case of Ridge Regression, the objective function (or the cost) is to minimize the RSS plus the sum of square of the magnitude of weights. This can be depicted mathematically as:

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

Ridge Regression

- It is a L2 regularization technique.
- Here Penalty = (Magnitude of the Coefficient)^2 . Let us see the equation for getting coefficients after penalizing using Ridge.
  - **Min $|Y_i - f(X_i)|^2 + \lambda|f|^2$**
  - Here the term lambda can be changed to get type of fit.
  - If **lambda = 0** Coefficients will be produced same as linear regression. i.e. **Overfitting** still remains as it is.
  - If **0 < lambda < ∞** in this case the magnitude of the lambda decides shape of the fit. We need to choose lambda wisely in between 0 and infinity.
  - If **lambda = ∞** it will brings down the regression line to mean resulting in **underfitting**.

Ridge Regression Usage

- The prime nature of ridge regression is, it considers or includes all or some of the predictor variables. Because of this nature of Ridge it is effective in shrinking the coefficients and reducing the complexity of the model.
- Ridge is mainly used to counter the over-fitting. It acts fairly good with datasets containingrelatively smaller number of features.
- It is not preferred for dataset with large number features because it includes all the features, which is susceptible for computational challenges.
- Ridge handles multi-collinearty. It includes all correlated features but with different coefficients.

Lasso Regression

- It is a L1 regularization technique.
- Here **Penalty = mod|Magnitude of the Coefficient|** . Let us see the equation for getting coefficients after penalizing using Lasso.
    - **Min $|Y_i - f(X_i)|^2 + \lambda|f|$**
    - Similar to Ridge Here also the term lambda can be changed to get type of fit.
    - If **lambda = 0** Coefficients will be produced same as linear regression. i.e. **Overfitting** still remains as it is.
    - If **0 < lambda < ∞** in this case the magnitude of the lambda decides shape of the fit. We need to choose lambda wisely in between 0 and infinity.
    - If **lambda = ∞** it will brings down the regression line to mean resulting in **underfitting**.

Lasso Regression Usage

- LASSO (Least absolute shrinkage and selection operator) does both parameter shrinkage and variable selection automatically.
- It drops some of the features which are correlated by zeroing the coefficient.
- LASSO is the model of choice for the problems which include large number of features to deal with. Lasso produces sparse solutions by eliminating features zero coefficients.
- LASSO handles multicollinearity by selecting the a single feature among the highly correlated features and coefficients of the rest all features are reduced to zero.

Difference between Ridge and LASSO

proschool
An ims Initiative

1. The most critical difference

- **Ridge:** It includes all (or none) of the features in the model. Thus, the major advantage of ridge regression is coefficient shrinkage and reducing model complexity.
- **Lasso:** Along with shrinking coefficients, lasso performs feature selection as well. As we observed earlier, some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model.

2. Sparse and Non-sparse

- **Ridge:** It is majorly used to *prevent over fitting*. Since it includes all the features, it is not very useful in case of exorbitantly high #features, say in millions, as it will pose computational challenges.
- **Lasso:** Since it provides *sparse solutions*, it is generally the model of choice (or some variant of this concept) for modeling cases where no. of features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored.

3. High Correlated Features

- **Ridge:** It generally works well even in presence of highly correlated features as it will include all of them in the model but the coefficients will be distributed among them depending on the correlation.
- **Lasso:** It arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero. Also, the chosen variable changes randomly with change in model parameters. This generally doesn't work that well as compared to ridge regression.

Polynomial Regression

- If we want a single curve then we choose a quadratic equation, cubic equation is needed for curves with two bends, for curves with multiple bends we use equations with higher order.
- **Significance tests:** Explains how variance(r^2) changes with respect to change in order of the polynomial.
- Usually variance increases with the increasing order of polynomial. As the order approaches to **n-1**, equation gives the **perfect fit,** but it may result in **overfit**.
- It is better fit a curve based on nature of the problem. It is advisable to stay within the order range of **quadratic to cubic.**

KNN Regression

- K Nearest Neighbor is an algorithm which stores all the different cases of a problem and predicts the numerical value of a target variable using distance functions which helps to measure the similarity between the cases available and target value.
- To measure the similarities between there are many distance formulae in use.
- For numerical and continuous variables we use following distance formulae Euclidean, Manhattan, Minkowski.
  - Euclidean distance: In 2-dimentional It is a great measure of shortest distance between two points.
  - Formula - $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
  - Example: Consider 2 points p = (0,0), q = (5,12) then Euclidean distance is calculated as **sqrt [(5-0)² + (12 – 0)²] = 13**

KNN Regression

- **Manhattan Distance**: In 2-dimentional It is a great measure of shortest path between two points when we are restricted move horizontally or vertically.
- Formula - $$\sum_{i=1}^{k} |x_i - y_i|$$
- Example: Consider 2 points p = (0,0), q = (5,12) then Manhattan distance is calculated as **abs(5-0) + abs(12-0) = 17**
- **Minkowski Distance: It calculates the distance between two points in a normed vector space. It is usually known as mixture of both Euclidean and Manhattan distance.**
- **Formula –** $$\left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q}$$
- Where if **q=1** then it's a **Manhattan** distance else **q = 2** it's **Euclidean** distance

KNN Regression

- **Hamming Distance**: In order to find the distance between categorical variable we use Hamming distance. It measures the distance between 2 strings of equal length by finding the number of positions where the two strings differs.
- **Formula:** $D_H = \sum_{i=1}^{k} |x_i - y_i|$ where if **X=Y** then **D=0** else **X ≠ Y** then **D=1**
- **Example:** Distance between "KRISHNA" and "KRISHNA" is 0 and KRISHNA and KRUSHNA is 1
- **Example of kNN regression**: To predict the tomorrow's temperature in Bangalore by using the historical data of previous 10 years. In this we approximately get 3650 data points and we get 10 nearest neighbor weeks containing 10 values for tomorrow in each year. By taking the **weighted average** of **$tomorrow_0$ + $tomorrow_1$ +……$tomorrow_9$** we can predict tomorrow'stemperature.

Poisson Regression

- **Poisson Regression:** Regression analysis methodology used to model count data (Non negative integer values) and contingency tables(table used to study correlation).
- **How it Works?** :Poisson regression assumes the response variable *Y* has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters.
- When Poisson Distribution is used to model contingency tables (table used to study correlation) then it is also called as log-linear model.
- **Formula:**

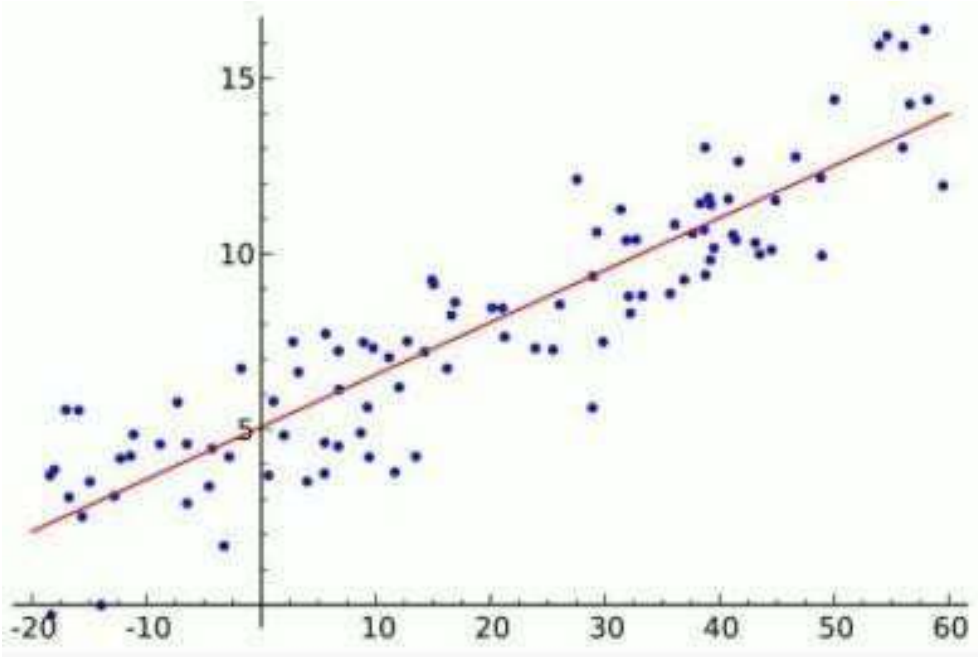
$$\Pr[Y = y] = \frac{e^{\lambda}\lambda^{y}}{y!}, y = 0,1,2..$$

  - Where,
  - λ : Average number of occurrences in a specified interval.
  - y : Actual number of successes that occur in a specified region.
  - e : A constant equal to approximately 2.71828.

Poisson Regression

- **Example:** Problem of estimating number of cars in the parking lot of a hotel at a given interval of time. Here predictors include any special day (like festivals, weekends, New Year), any special offers (like discounts) or any special invitees(Celebrity visit, concerts, etc).

- Following figure shows the poisson regression line for data points



Multinomial Regression

- Prior to understanding Multinomial Regression first let us learn about Logistic Regression.
- **Logistic Regression:** Logistic Regression used to predict the dependent variable, which has binary outputs (1/0, TRUE/FALSE, YES/NO).
- Logistic Regression finds the probability of success and failure of an event.
- **Example Logistic Regression:** Predicting the **win/loss** of a cricket team based on the historical data of the team's performance.
- **Multinomial Regression:** It generalizes the logistic regression to multiple classes. i.e. Classifies the dependent variable into more than 2 classes.
- Multinomial regression predicts the probabilities of the different possible outcomes of a categorically distributed target/dependent variable, given a set of predictor/independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

Multinomial Regression

- **Example of Multinomial Logistic Regression:** If you classifying each state of India based on amount of rain fall every year, we classify sates into 3 classes like **HIGH** rain fall, **MEDIUM** rain fall and **LOW** rain fall states based on various independent variables.
- Formula: In the following formula there are dependent variable has **k-categories, β_k** represents the set of regression coefficients, **X_i** represents set of predictor variables for i^{th} observation. **Yi** represents target category for i^{th} observation.

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

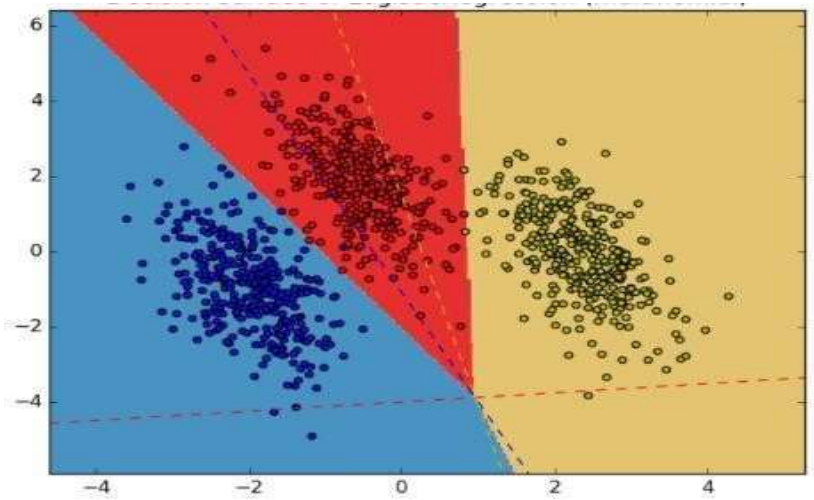
$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K-1) = \frac{e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

Important Points

- As it is a classification model, it expects large sample sizes.
- The dependent variable need not to have a linear relation with independent variables as in linear regression.
- Logistic regression handles various types of relations by using log function on predicted outcome which is non-linear in nature.
- **No Multicolinearity:** The independent variables must not be correlated.
- Following figure is the example of how output of the model divided into 3 categories(k).



Packages used in Case Studies

- **datetime** - In Python, date, time and datetime classes provides a number of function to deal with dates, times and time intervals. Date and datetime are an object in Python, so when you manipulate them, you are actually manipulating objects and not string or timestamps. Whenever you manipulate dates or time, you need to import datetime function.

- **Sklearn -** scikit-learn is a collection of Python modules relevant to machine/statistical learning and data mining.

- **Joblib -** Joblib is a set of tools to provide lightweight pipelining in Python. In particular, joblib offers: transparent disk-caching of the output values and lazy re-evaluation (memoize pattern).

- **Pandas_profiling** - Generates profile reports from a pandas DataFrame. The pandas df.describe() function is great but a little basic for serious exploratory data analysis. pandas_profiling extends the pandas DataFrame with df.profile_report() for quick data analysis.

- **Seaborn** - Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- **Warnings** - Warning are typically issued in situations where it is useful to alert the user of some condition in a program, where that condition (normally) doesn't warrant raising an exception and terminating the program.

Thank You.