REGRESSION

3rd Sem, MCA

Contents

- Linear Regression
  - Model Specification, Cost functions
  - Gradient Descent, Batch Gradient Descent
  - Maximum Likelihood Estimation, Model Selection
  - Union and Chernoff bounds, VC dimensions.
  - Estimators- sampling distribution,
  - Bayes Risk, Desirable Properties,
- Logistic Regression
  - Model Specification
  - Model Fitting
- Generalized Linear Models
  - The exponential Family

Basic Maths

$$\frac{d}{dx}x^3 = 3x^{3-1} = \mathbf{3x^2}$$

$$\frac{d}{dx}x^{-1} = -1x^{-1-1}$$

$$= -x^{-2}$$

$$= \frac{-1}{x^2}$$

$$\frac{\text{Change in Y}}{\text{Change in X}} = \frac{\Delta y}{\Delta x}$$

$$\frac{dy}{dx} = lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$$

$f(x) = 5x^2 - 2x + 6$

$d/dx\ f(x) = d/dx\ (5x^2 - 2x + 6)$

$d/dx\ f(x) = d/dx\ (5x^2) - d/dx\ (2x) + d/dx\ (6)$

$= 5(2x) - 2(1) + 0 = 10x - 2$

- d/dx (k) = 0, where k is any constant
- d/dx(x) = 1
- $d/dx(x^n) = nx^{n-1}$
- d/dx (kx) = k, where k is any constant
- $d/dx\ (\sqrt{x}) = 1/2\sqrt{x}$
- $d/dx\ (1/x) = -1/x^2$
- d/dx (log x) = 1/x, x > 0
- $d/dx\ (e^x) = e^x$
- $d/dx\ (a^x) = a^x \log a$

Basic Maths

$$\int_0^3 x^2 dx = \left(\frac{x^3}{3}\right)_0^3 = \left(\frac{3^3}{3}\right) - \left(\frac{0^3}{3}\right) = 9$$

$$\int 1\, dx = x + C$$

$$\int a\, dx = ax + C$$

$$\int x^n\, dx = \frac{x^{n+1}}{n+1} + C;\ n \neq -1$$

$$\int (x^2-1)(4+3x)dx = \int 4x^2+3x^3-3x-4\ dx$$

$$= 4(x^3/3) + 3(x^4/4) - 3(x^2/2) - 4x + C$$

$$\int \frac{1}{x}\, dx = \ln|x| + C$$

$$\int e^x\, dx = e^x + C$$

$$\int a^x\, dx = \frac{a^x}{\ln a} + C;\ a > 0, a \neq 1$$

$$_0\int^2 x^2 - 2x\ dx = [x^3/3 - x^2]_0^2$$

$$= 4/3$$

Basic Maths

$$\iint_R f(x,y)\,dA = \int_{\theta_1}^{\theta_2}\int_{r_1}^{r_2} f(r\cos\theta, r\sin\theta)\,r\,dr\,d\theta$$

$$\int_a^b\int_c^d f(x,y)+g(x,y)dxdy = \int_a^b\int_c^d f(x,y)dxdy + \int_a^b\int_c^d g(x,y)dxdy$$

$$\int_a^b\int_c^d Cf(x,y)dxdy = C\int_a^b\int_c^d f(x,y)dxdy$$

$$f(x,y) = h(x)g(y)$$

$$\int_a^b\int_c^d f(x,y)dxdy = \int_c^d h(x)dx + \int_a^b g(y)dy$$

$$\int_0^3\int_1^5 5xydxdy = 5\int_0^3\int_1^5 xydxdy$$

$$\int_1^5 xydx = y\frac{x^2}{2}|_1^5$$

$$y\frac{x^2}{2}|_1^5 = y(\frac{5^2}{2} - \frac{1^2}{2})$$

$$y(\frac{5^2}{2} - \frac{1^2}{2}) = 12y$$

$$5\cdot 12\int_0^3 ydy = 60\frac{y^2}{2}|_0^3$$

$$60\frac{y^2}{2}|_0^3 = 60\frac{3^2}{2}$$

$$60\frac{3^2}{2} = 60\frac{9}{2} = 270$$

$$\int 1\,dx = x + C$$

$$\int a\,dx = ax + C$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C;\ n \neq -1$$

$$\int \frac{1}{x}\,dx = \ln|x| + C$$

$$\int e^x\,dx = e^x + C$$

$$\int a^x\,dx = \frac{a^x}{\ln a} + C;\ a > 0, a \neq 1$$

Basic Maths

$$\iint_R f(x,y)\,dA = \int_{\theta_1}^{\theta_2}\int_{r_1}^{r_2} f(r\cos\theta, r\sin\theta)\,r\,dr\,d\theta$$

$I = \int[\int(x^2+y^2)dx]dy$

$I = \int[x^3/3+y^2x]dy$

$I = x^3y/3+xy^3/3$

$I = [xy(x^2+y^2)]/3 + C$

$I = \int[\int(x+y)dx]dy$

$I = \int[x^2/2+yx]dy$

$I = x^2y/2+xy^2/2$

$I = (xy/2)(x+y) + C$

$$= \int_1^2 \left(\int_4^6 \frac{x}{y^2}\,dy\right) dx$$

$$= \int_1^2 \left(\left[-\frac{x}{y}\right]_{y=4}^{y=6}\right) dx$$

$$= \int_1^2 \left(\frac{x}{4} - \frac{x}{6}\right) dx$$

$$= \int_1^2 \frac{x}{12}\,dx$$

$$= \left[\frac{x^2}{24}\right]_{x=1}^{x=2}$$

$= (2)^2/24 - (1)^2/24$

$= (4 - 1)/24$

$= 3/24$

$= 1/8$

$$\int 1\,dx = x + C$$

$$\int a\,dx = ax + C$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C;\ n \neq -1$$

$$\int \frac{1}{x}\,dx = \ln|x| + C$$

$$\int e^x\,dx = e^x + C$$

$$\int a^x\,dx = \frac{a^x}{\ln a} + C;\ a > 0, a \neq 1$$

$$\int_a^b \int_c^d Cf(x,y)dxdy = C\int_a^b \int_c^d f(x,y)dxdy$$

$$f(x,y) = h(x)g(y)$$

$$\int_a^b \int_c^d f(x,y)dxdy = \int_c^d h(x)dx + \int_a^b g(y)dy$$

$$\int_a^b \int_c^d f(x,y) + g(x,y)dxdy = \int_a^b \int_c^d f(x,y)dxdy + \int_a^b \int_c^d g(x,y)dxdy$$

Basic Maths

$$\int_0^9 \int_6^8 3x^2 - 2xydxdy$$

$$\int_0^9 \int_6^8 3x^2 - 2xydxdy = \int_0^9 \int_6^8 3x^2 dxdy - \int_0^9 \int_6^8 2xydxdy$$

$$\int_6^8 3x^2 dx = x^3|_6^8 = 8^3 - 6^3 = 296$$

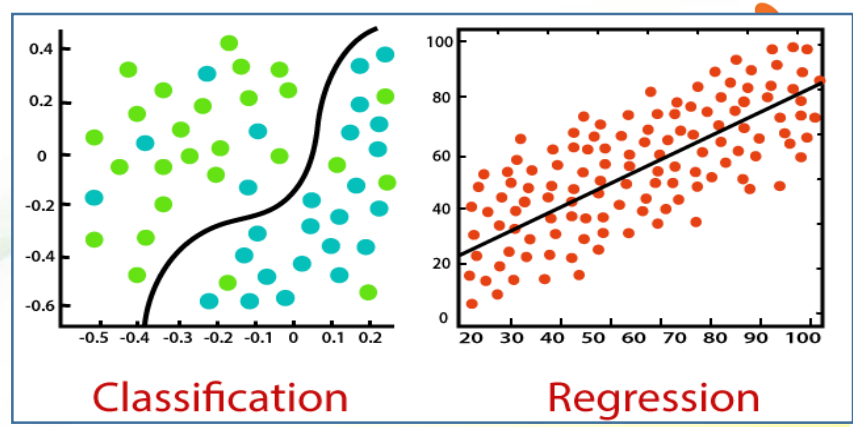
$$\int_6^8 2xydx = x^2y|_6^8 = y(8^2 - 6^2) = 28y$$

$$\int_0^9 296dy = 296y|_0^9 = 296(9 - 0) = 2664$$

$$\int_0^9 -28ydy = -14y^2|0^9 = -14(9^2 - 0^2) = -252$$

$$2664 - 252 = 2412$$

Machine Learning



Classification: Process of finding a model/function which helps in separating data into multiple categorical classes.

- Data is categorized under different labels according to some parameters given in input and then labels are predicted for the data.

Regression: Process of finding a model for distinguishing data into continuous real values instead of using classes.

- Helps to identify distribution movement depending on historical data.

- Regression algorithms are used to **predict continuous** values such as price, salary, age, etc.
- Classification algorithms are used to **predict/Classify discrete values** (True or False, Spam or Not Spam, etc.)

Clustering: Process data to find a structure/pattern/cluster in a collection of uncategorized data.

- Hierarchical clustering, K-means clustering

Association: Discovering exciting relationships between variables in large databases.

- Example: Shopping-cart recommendation, video streaming recommendation.

Machine Learning

Classification Algorithms:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Naïve Bayes
- Decision Tree Classification
- Random Forest Classification

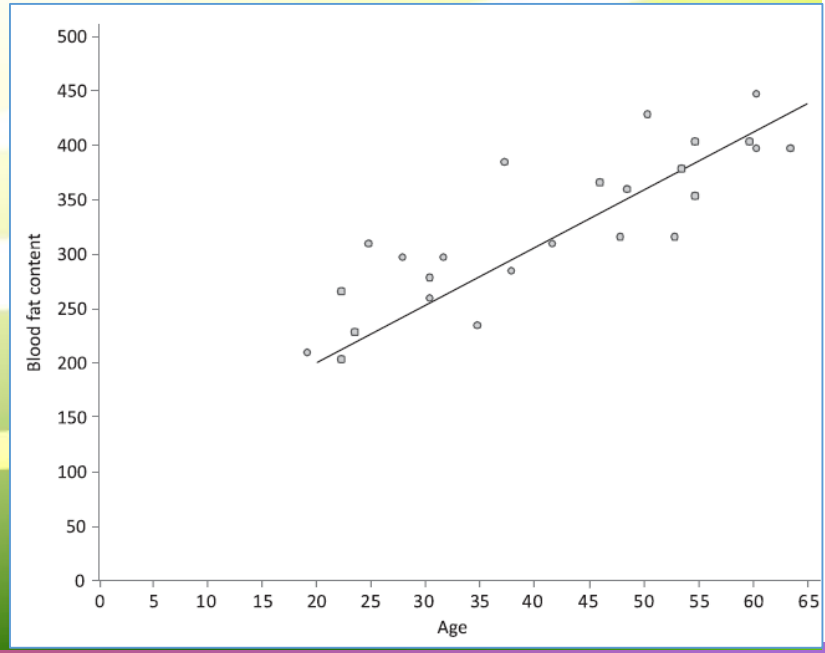
Regression Algorithms:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

Regression

Regression: build a function/model that describe relationship between one/more independent variables and a single response variable.

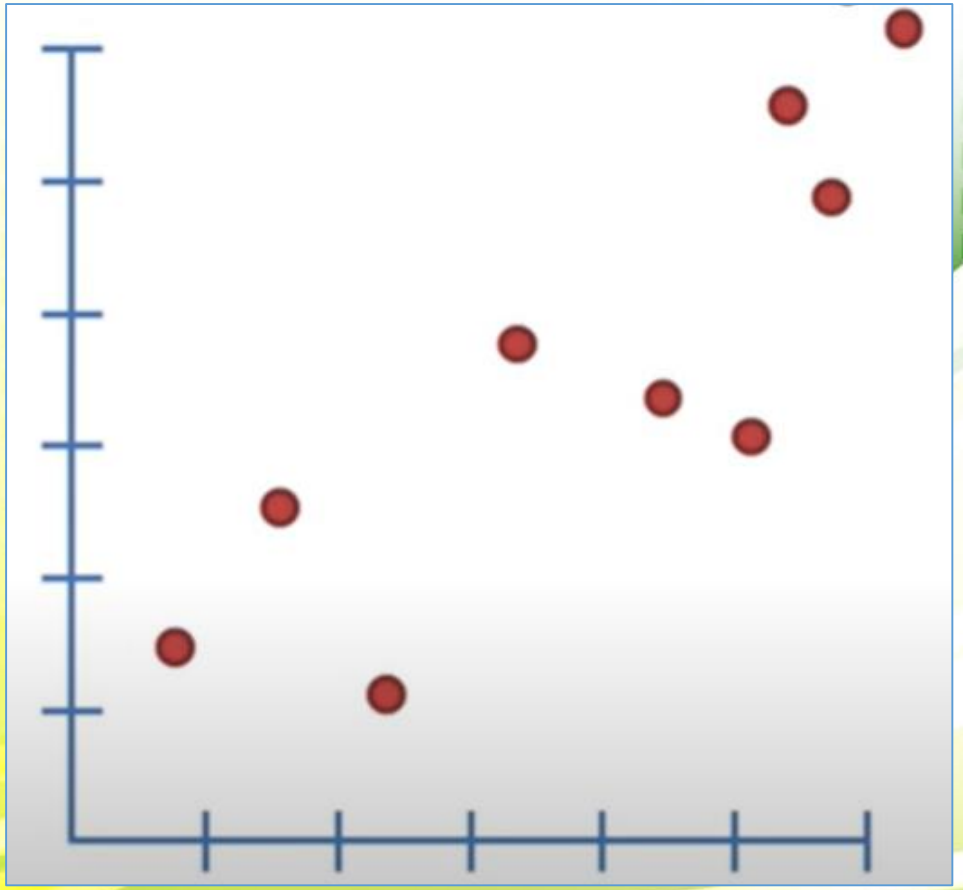
- Plot a graph between the variables which best fit the given data points.
- *Regression shows a line or curve that passes through all data points on a target-predictor graph in such a way that distance between data points and regression line is minimum.*

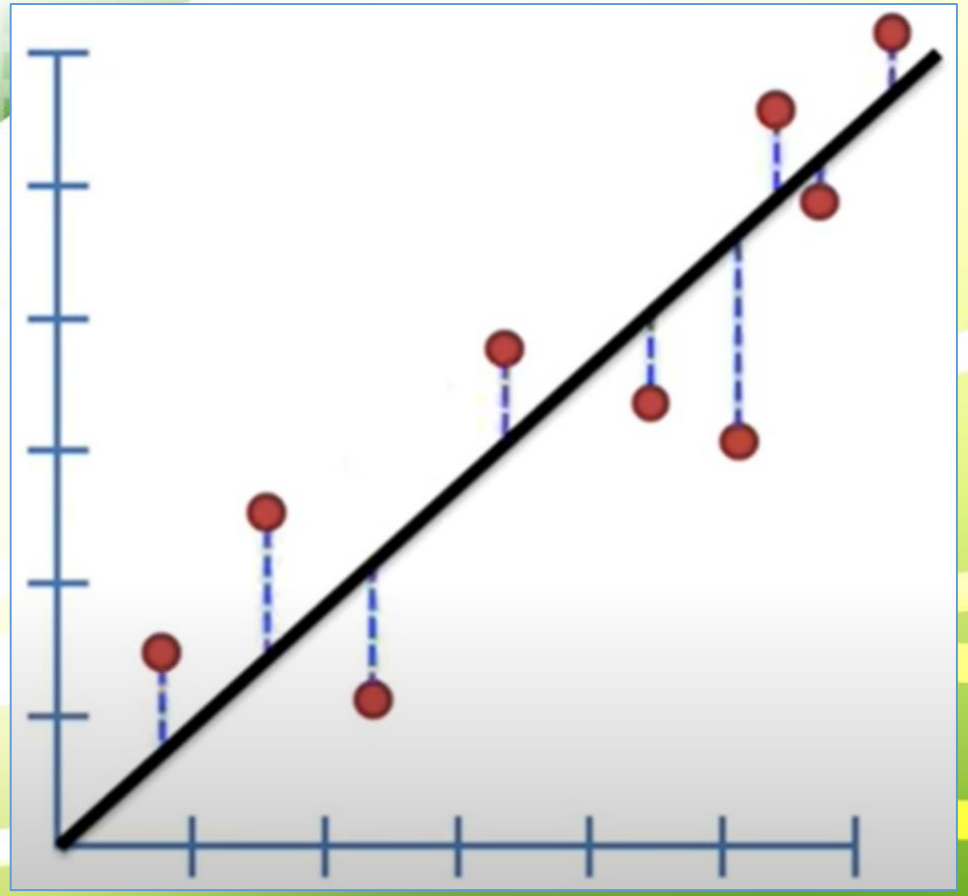


Types of Regression models

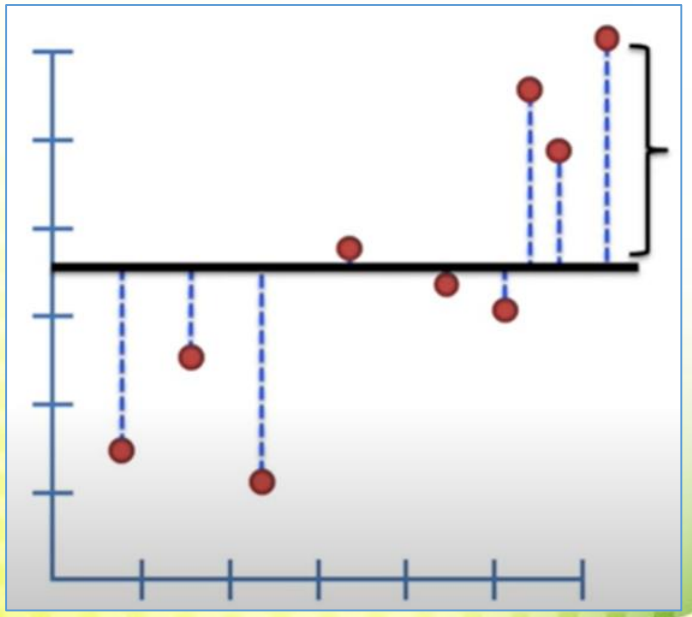
- Linear Regression
- Polynomial Regre6ssion
- Logistic Regression

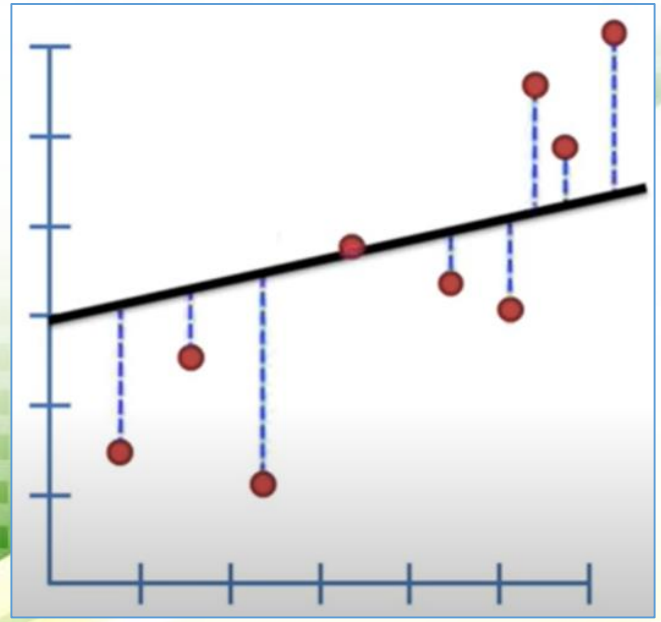
Linear Regression





Linear Regression

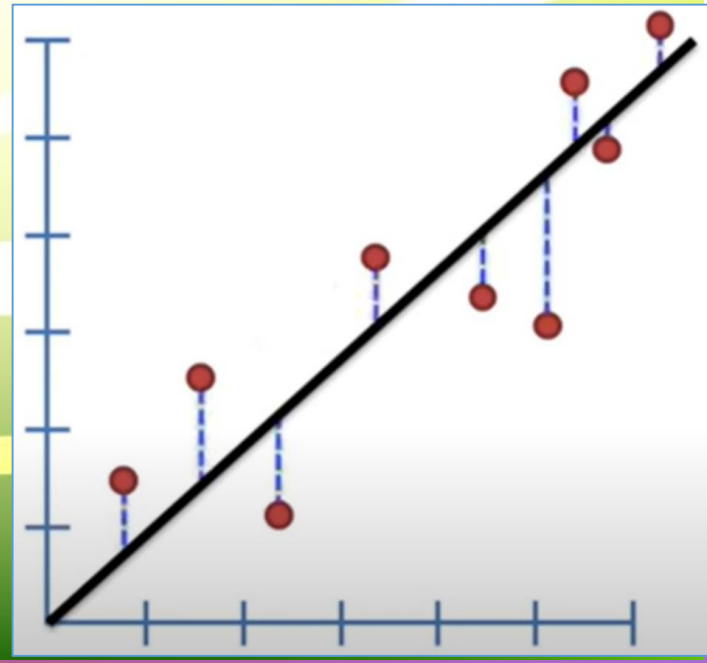




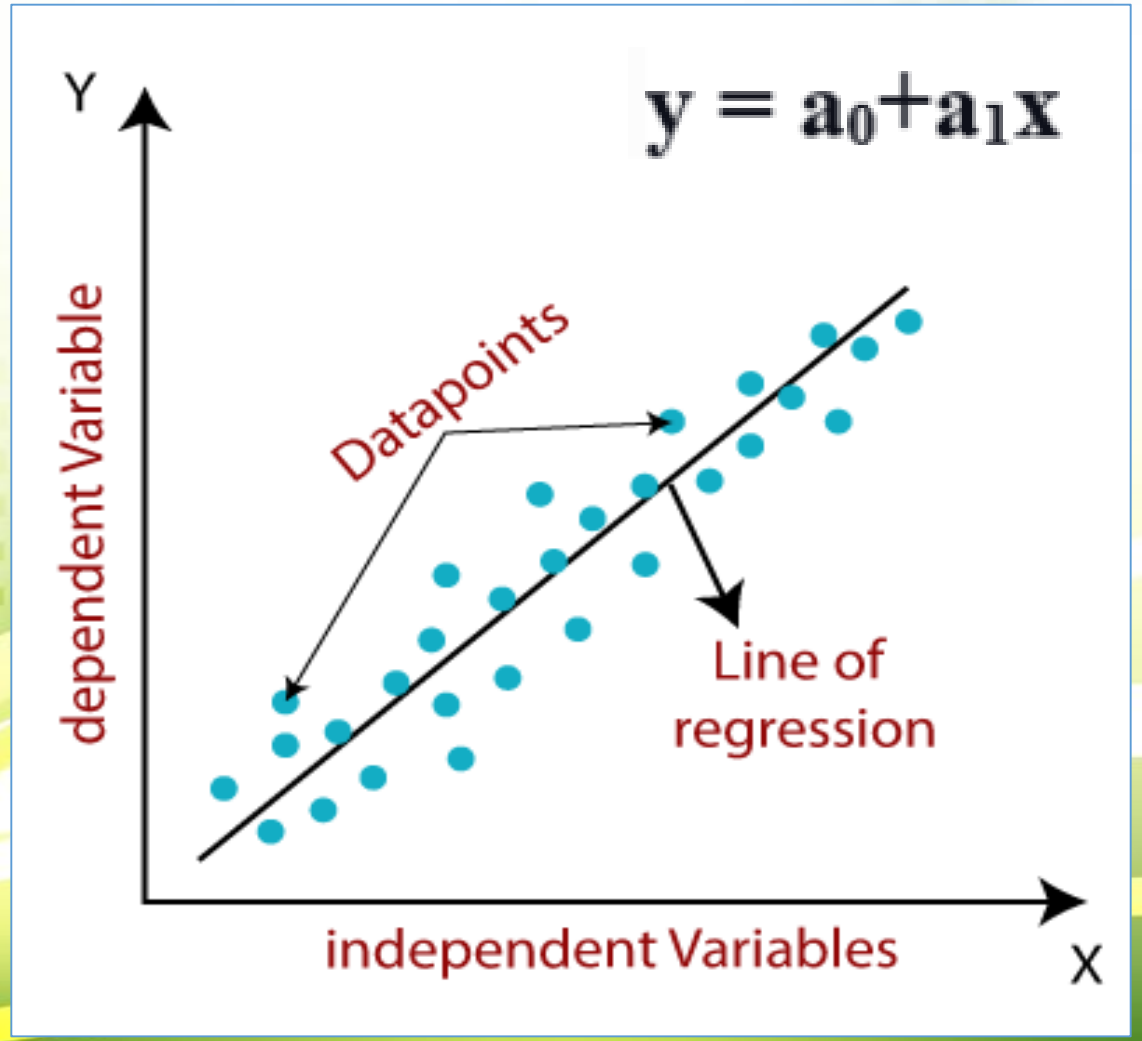
- Rotate the line aiming to reduce error.
- Find each Distance → square → add.

- Draw a random line through the points.
- Calculate distance between data point & corresponding point on line → Residual.
- Square of least distance.
- Sum of square.

- Regression line
- Lease square distance.



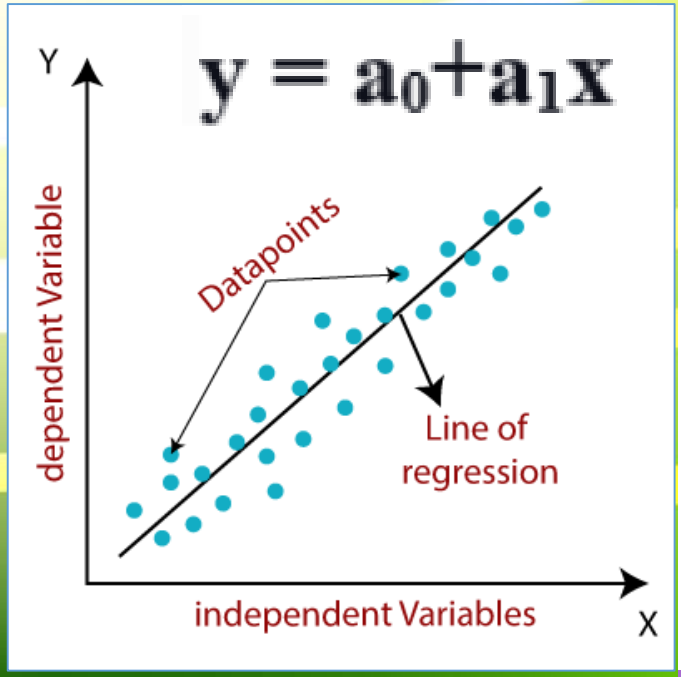
Linear Regression



Linear Regression

- Shows a linear relationship between a dependent (y) and one or more independent (x) variables.
- Finds how value of dependent variable is changing according to value of independent variable.
- Makes predictions for continuous/real or numeric variables; **sales, salary, age, product price,** etc.
- Provides a sloped straight line representing relationship between variables.
- Supervised learning algorithm.

- Goal of linear regression algorithm is to get best values for a_0 and a_1 to find best fit line.
- Best fit line should have least error → error between predicted values and actual values should be minimized.



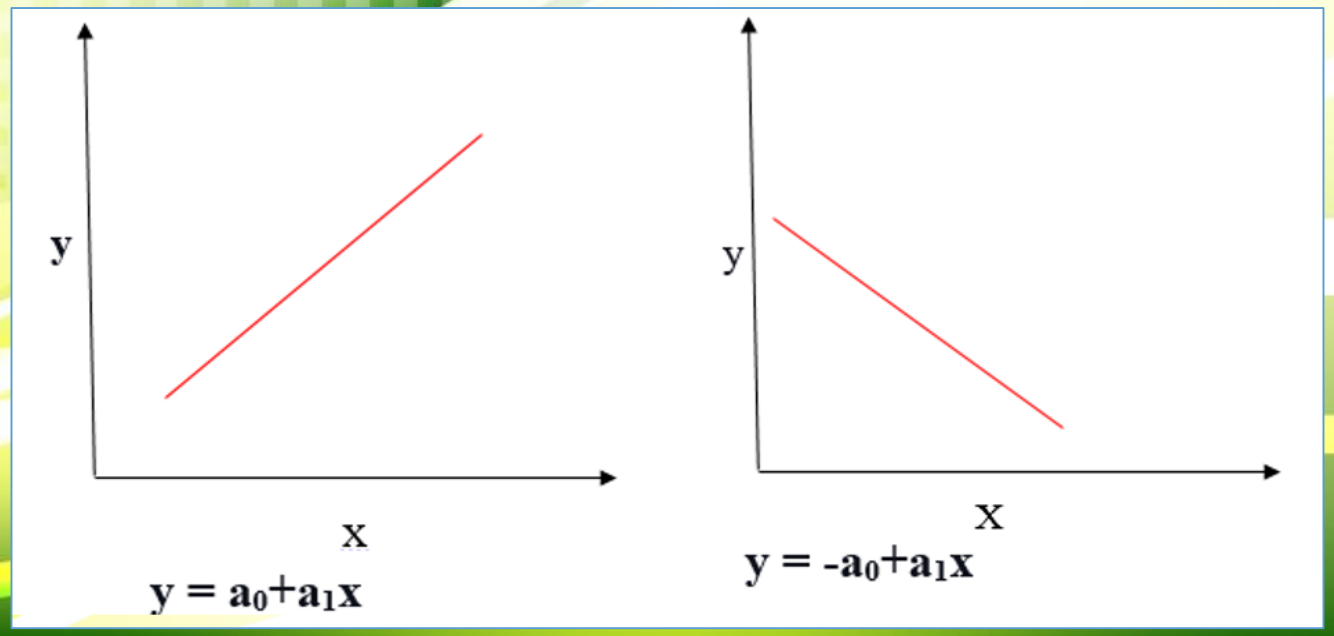
Linear Regression

Positive Linear Relationship

- If dependent variable expands on Y-axis given independent variable progress on X-axis.

Negative Linear Relationship

- If dependent variable decreases on Y-axis given independent variable increases on X-axis.



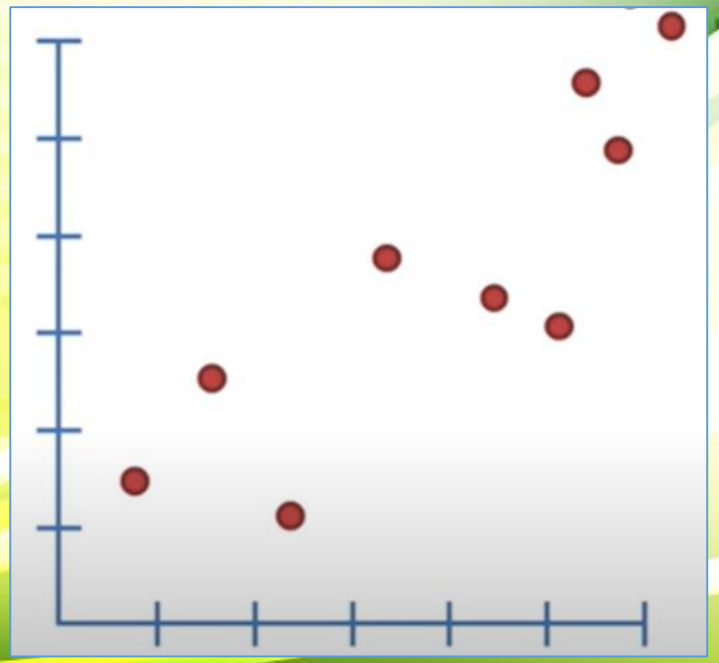
Linear Regression

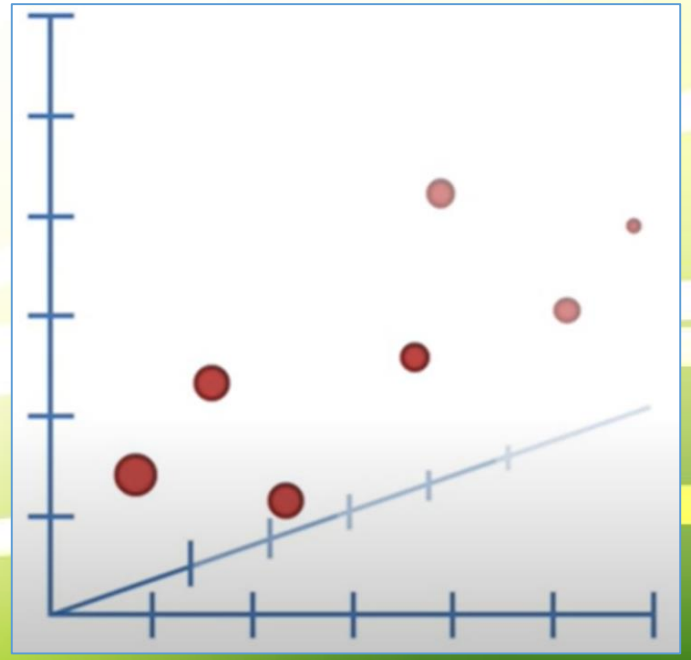
Simple Linear Regression:

- Single independent variable is used to predict value of a numerical dependent variable.

Multiple Linear regression:

- More than one independent variable is used to predict value of a numerical dependent variable.





Linear Regression

- **Simple Linear Regression:**
  - Single independent variable is used to predict value of a numerical dependent variable.
  - $Y = mx + c$
- **Multiple Linear regression:**
  - More than one independent variable is used to predict value of a numerical dependent variable.
  - $Y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + c$

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_k$$

- a_1, a_2, a_3, a_4 - Regression Coefficient
- capped → estimated coefficient

- $Y = 9 + 3x_1 + 0.9x_2 + 4x_3 + 1.4x_4$
- Regression Coefficient tells which feature(s) have more impact on dependent variable.

Linear Regression

- Size of coefficient for each independent variable tells the impact size of that variable on dependent variable.
- Sign on coefficient (+ or -) tells the direction of effect.

Response

Independent variables

Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_k$$

$$y = f(x_i)$$

Model

Linear Regression

Simple Linear Regression:

$$y = b_0 + b_1 x$$

- b_1 is slope
- b_0 is intercept.

Ordinary Least Squares (OLS) Estimator for Regression coefficient

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

e OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

$$\hat{u}_i = Y_i - \hat{Y}_i.$$

Linear Regression

Simple Linear Regression:

mean of x = 39.12
mean of y = 310.72

$$y = b_0 + b_1 x$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$Slope\,(b_1) = 19{,}157.84/3{,}600.64$

$Slope\,(b_1) = 5.32$

$Intercept\,(b_0) = 310.72 - (5.32 \times 39.12)$

$Intercept\,(b_0) = 102.6$

Blood fat content* = 102.6 + 5.32 × *Age

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
46	354	6.88	43.28	297.7664	47.3344
20	190	−19.12	−120.72	2,308.1664	365.5744
52	405	12.88	94.28	1,214.3264	165.8944
30	263	−9.12	−47.72	435.2064	83.1744
57	451	17.88	140.28	2,508.2064	319.6944
25	302	−14.12	−8.72	123.1264	199.3744
28	288	−11.12	−22.72	252.6464	123.6544
36	385	−3.12	74.28	−231.7536	9.7344
57	402	17.88	91.28	1,632.0864	319.6944
44	365	4.88	54.28	264.8864	23.8144
24	209	−15.12	−101.72	1,538.0064	228.6144
31	290	−8.12	−20.72	168.2464	65.9344
52	346	12.88	35.28	454.4064	165.8944
23	254	−16.12	−56.72	914.3264	259.8544
60	395	20.88	84.28	1,759.7664	435.9744
48	434	8.88	123.28	1,094.7264	78.8544
34	220	−5.12	−90.72	464.4864	26.2144
51	374	11.88	63.28	751.7664	141.1344
50	308	10.88	−2.72	−29.5936	118.3744
34	220	−5.12	−90.72	464.4864	26.2144
46	311	6.88	0.28	1.9264	47.3344
23	181	−16.12	−129.72	2,091.0864	259.8544
37	274	−2.12	−36.72	77.8464	4.4944
40	303	0.88	−7.72	−6.7936	0.7744
30	244	−9.12	−66.72	608.4864	83.1744
			Sum	19,157.84	3,600.64

Linear Regression

Example:

Build the simple linear regression model/function for the data given below.

Age (x)	Sugar Level (Y)
46	354
20	190
52	405
30	263
57	451

$$y = b_0 + b_1 x$$

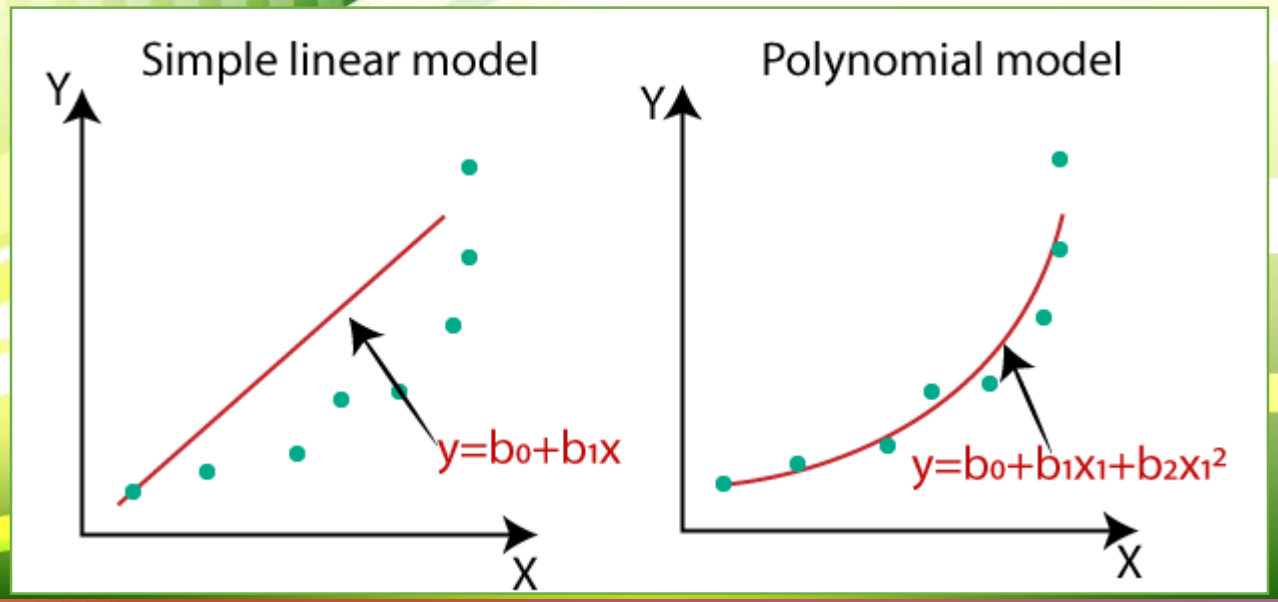
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Using this model predict the Sugar level for a new patient of age 39.

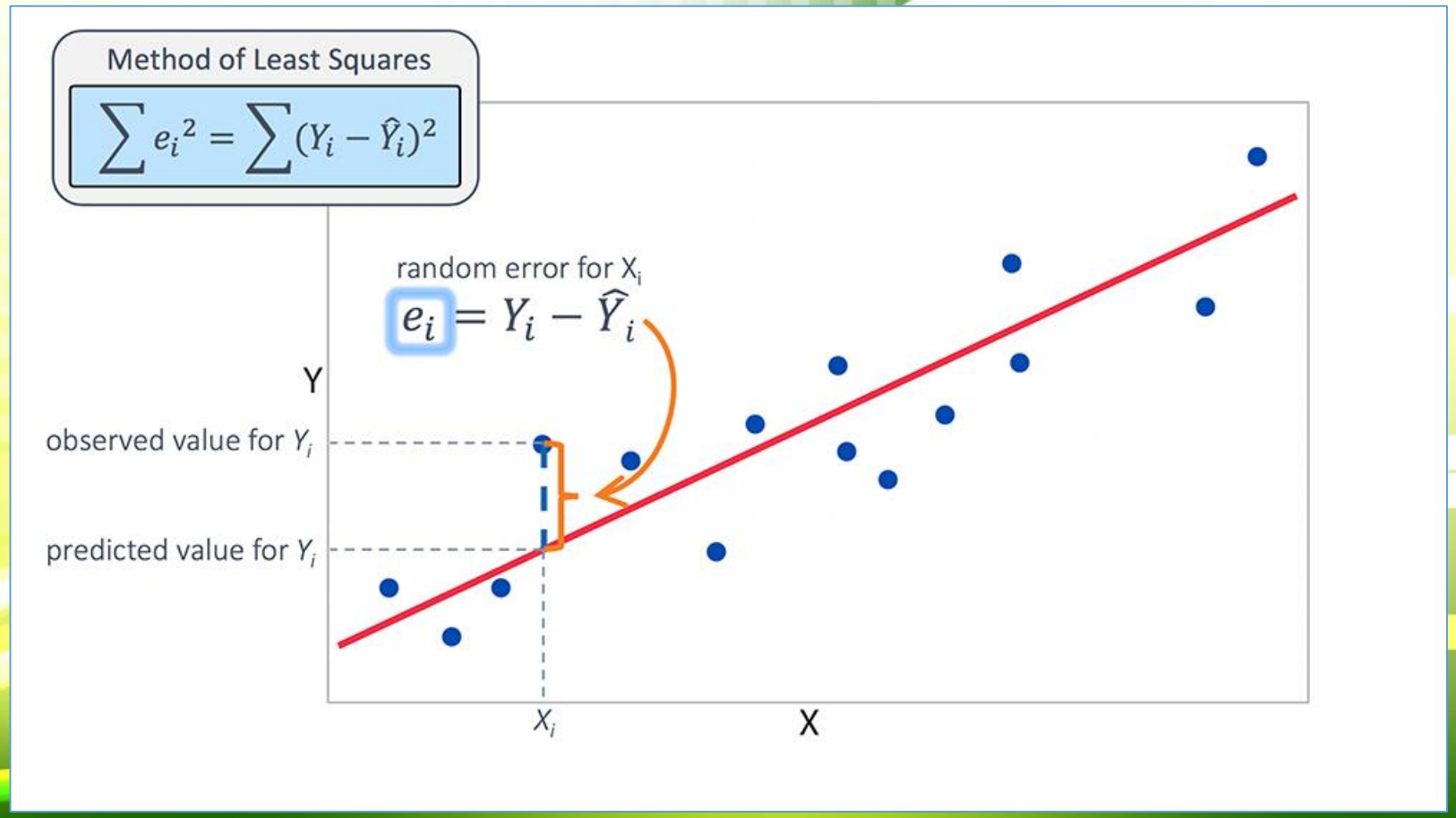
Polynomial Regression

- If data points clearly will not fit linear regression (straight line through all data points), it might be ideal for polynomial regression.
- (~~Linear~~)Relationship between variables x and y to find best way to draw a line through data points.
- Special case of Multiple Linear Regression, by adding some polynomial terms to it.
- Relationship between a dependent(y) and independent variable(x) as n^{th} degree polynomial.



Regression Model – Measures of Fit

- ***Residual***: error term representing difference between observed value (*y*) and predicted value.
- Residual analysis helps to better understand how well model is performing.



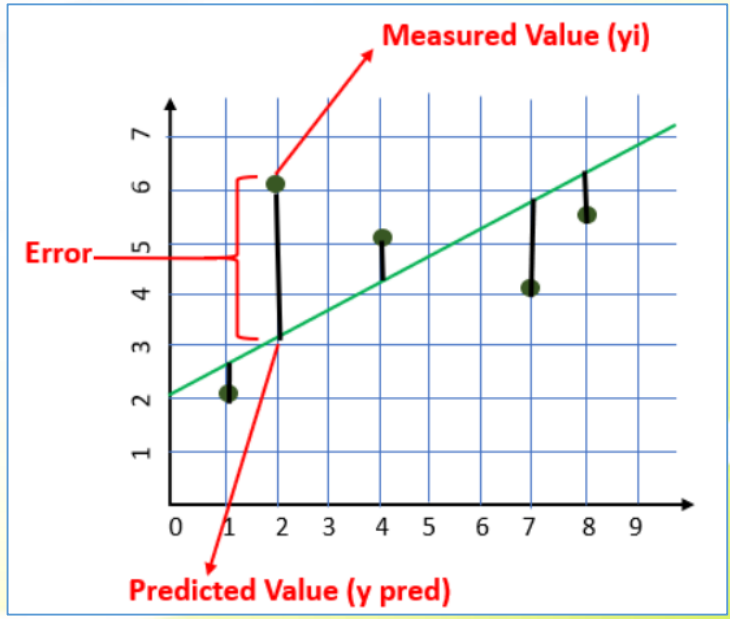
$$\hat{e} = y - \hat{y}$$

Regression Model - Measures of Fit

- **Loss function**: squared error (says how bad the fit is)

$$\mathcal{L}(y,t) = \frac{1}{2}(y-t)^2$$

  - **y – t** is the residual, and aim is to make this as small in magnitude
  - **½ factor** is just to make the calculations convenient.



- **Cost function: l**oss function averaged over all training examples
  - helps to reach optimal solution; technique of evaluating performance of model.
  - Takes both predicted outputs and actual outputs to calculate how much wrong model was in prediction.

$$\mathcal{J}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2$$

Regression Model - Measures of Fit

- ***standard error of the estimate*** ($S_{y,x}$) : measure of variation of *y*-values about regression line.
  - interpreted in a similar manner to standard deviation.
  - indicates model's accuracy: larger the value for standard error of estimate, lower the precision.

$$s_{y.x} = \sqrt{\frac{SSE}{n-2}}$$

- t-Test, F-Test performed to assess variable dependencies and model performance.

- **Mean Absolute Error (MAE):** simple metric; Not preferred where outliers are prominent.
- **Mean Squared Error (MSE):** most common metric for regression models.
- **Root Mean Squared Error (RMSE):** square root MSE.
  - RMSE penalizes large errors..

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

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

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

Regression Model – Measures of Fit

- ***Sum of squares total* (*SST*)** : measure of variation of *y*-values about their mean.
- ***Sum of squares due to regression* (*SSR*):** differences between predicted/regression values and average y-value.
- ***Sum of squares* of *error* (*SSE*) or Residual Sum of Squares (RSS):** differences between actual *y*-values and predicted *y*-values.

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

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

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

$$SST = SSR + SSE$$

$$TSS = ESS + SSR$$

Regression Model - Measures of Fit

- ***Coefficient of determination*** **(R^2)**: proportion of variation.

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

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

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- R^2 values vary between 0 and 1.
- R^2 closer to 1 → more accurate model predictions (models have a *closer fit).*
- In multiple linear regression, *adjusted* R^2 value (R^2 adj) is usually considered to better account for the multiple independent variables used in analysis as well as sample size.

$$R^2_{\text{adj}} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

n : number of observations
k : number of independent variables.

- correlation coefficient r is directly related to the coefficient of determination R^2

$$R^2 = r^2_{X,Y\cdot}$$
$$R^2 = r^2_{Y,\hat{Y}\cdot}$$

Regression Model - Measures of Fit

Pearson 'r' correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

r **= Pearson r correlation coefficient between x and y**
n **= number of observations**
x_i **= value of x (for ith observation)**
y_i **= value of y (for ith observation)**

S_x, S_y **= S.D. for x and y**

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Regression Model - Measures of Fit

'r' correlation coefficient:

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Population Correlation Coefficient

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where, $\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation

$\sigma_{xy} \rightarrow$ Population Covariance

$\bar{X}, \bar{Y} \rightarrow$ Population Mean

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)} \cdot \sqrt{\mathrm{var}(y)}}$$

$$\mathrm{Cov}\,(X, Y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\sigma^2 = \frac{\sum (xi - \bar{x})^2}{N}$$

Regression Model - Least Squares Assumptions

- **Least Squares Regression Line**: vertical distance from data points to regression line is as small as possible.
  - Its called “least squares” because the best line of fit is one that minimizes variance (sum of squares of errors).

- **Least squares fitting** (least squares estimation): way to find best fit line for a set of points.
  - sum of the squares of offsets (residuals) are used to estimate best fit line instead of absolute values of the offsets.

- (Least Square) Linear regression models are based on a series of assumptions.
- If a data set does not conform to these assumptions then either the model needs to be adjusted (by applying mathematical **transformation to data**) or particular linear regression not be suitable for modeling that data set.

Regression Model - Least Squares Assumptions

Least Squares Assumptions

- Model should have linear parameters.
  - relationship between independent & response variable should be linear
- data should be a random sample from the population.
- independent variables should not be strongly collinear.
- residuals have homogeneous variance.
  - variation of error/residual across each independent variables should remain constant (as a function of predicted value).
- independent variables have been measured accurately
  - Otherwise, small errors in measurement could result in huge errors for OLS regression.
  - Large Outliers are Unlikely
- ***independence of errors***. no trend in residuals based on order in which observations were collected.
- residuals follow a normal distribution.

Regression Model

- important to generate simplest possible model that contains only necessary independent variables.
- Ideally number of independent variables should be small and include at least **10 observations** in training set *for every independent variable* included in model. Example:
  - Dataset has 25 independent variables with 300 records.
  - Hypothesis Tests shows 13 important variables.
  - Training set should have minimum 130 records.
- Important to *perform exploratory data analysis t*o inspect relationships between variables.
- Perform *transformations* on potential independent variables.
- *Dummy, derived, or composite variables can be generated*.
- Continuous variables may need to be *transformed into a categorical variable*.
- If relationship between a potential independent and response variable needs to be converted from nonlinear to linear, suitable transformations can be used for same.
- Multiple combinations of different independent variables can be used to build set of models from which best performing, most plausible, and simplest model is selected.
- *Standard error*, *t-stat*, and *p-value are* calculated, which can be used to help in selection of independent variables.

Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Issues :

- **Multicollinearity** happens when independent variables in regression model are highly correlated to each other.
  - It makes it hard to interpret of model and also creates an overfitting problem.
  - test before selecting the variables into regression model.
- **Omitted variable bias (OBV)** occurs when a relevant explanatory variable is not included in a regression model, which can cause the coefficient of explanatory variables in the model to be biased.

Multiple Linear Regression

- **Omitted variable bias** in simple linear regression is bias in OLS estimator that arises when regressor X is correlated with an *omitted variable*.

- For omitted variable bias to occur, two conditions must be fulfilled:
  - X is correlated with the omitted variable.
  - Omitted variable is a determinant of dependent variable Y.

- An omitted variable is often left out of a regression model for one of two reasons:
  - Data for the variable is simply not available.
  - Effect of explanatory variable on response variable is unknown.

Multiple Linear Regression

- Similarly to SLR model, coefficients of multiple regression model can be estimated using OLS.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n.$$

- **Minimize** the sum of squared mistakes by choosing estimates $b_0, b_1, \ldots, b_k$ for the coefficients $\beta_0, \beta_1, \ldots, \beta_k$

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \cdots - b_k X_{ki})^2$$

difference between Y_i and its predicted value $\hat{Y}_i$ is called the OLS residual of observation i: $\hat{u} = Y_i - \hat{Y}_i$.

Regression - Specification

- Multiple Linear Regression model holds causal relationship between >1 independent variables and a dependent variable.
- **Model specification** is the process of determining which independent variables to include and exclude from a regression equation.
  - Specification error is when independent variables and their functional form inaccurately portray the real relationship present in data.
  - Specification error can cause bias (exaggerate, understate, or hide presence of underlying relationships).
- Regression analysis involves three distinct stages:
  - specification of a model,
  - estimation of the parameters of this model, and
  - interpretation of these parameters.

Regression - Specification

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

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

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- R^2 values vary between 0 and 1.
- R^2 closer to 1 → more accurate model predictions (models have a *closer fit).*
- In multiple linear regression, *adjusted* R^2 value (R^2 adj) is usually considered to better account for the multiple independent variables used in analysis as well as sample size.

$$R^2_{\text{adj}} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

n : number of observations
k : number of independent variables.

- correlation coefficient r is directly related to the coefficient of determination R^2

$$R^2 = r^2_{X,Y\cdot}$$
$$R^2 = r^2_{Y,\hat{Y}\cdot}$$

Regression - Specification

- Underfitting: omission of relevant variable(s)
  - Lack of data availability,
  - Oversighting,
  - Ignorance.
  - Leads to omitted variable bias: specification bias

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n.$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \epsilon_i$$

$$\epsilon_i = u_i + \beta_3 X_{3i}$$

- Overfitting: inclusion of irrelevant variable(s)
  - Wrong judgement/assumption of variable important,
  - Include everything available in data.
  - Leads to incorrect magnitudes of coefficients & efficiency loss.

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + v_i$$

$$v_i = u_i - \beta_3 X_{3i}$$

Regression - Specification

Commonly used decision methods to select explanatory variable:

- T-test, f-test
- Information criteria

Akaike: $AIC = \log(s^2) + \frac{2k}{n}$

Bayes: $BIC = \log(s^2) + \frac{k \log n}{n}$

- Out-of-sample prediction

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

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

S: squared standard error due to regression

K: number of variables, n: sample size

Regression – Optimization

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y^{(i)}} - y^{(i)} \right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

which is basically $\frac{1}{2}\bar{x}$ where $\bar{x}$ is the mean of squares of $h_\theta(x^{(i)}) - y^{(i)}$, or the difference between the predicted value and the actual value.

- Where
  - $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$
  - $(x^{(i)}, y^{(i)})$ is the i^{th} training data
  - m is the number of training example
  - $\frac{1}{2}$ is a constant that helps cancel 2 in derivative of the function when doing calculations for gradient descent

Model **objective is to minimize the cost function** → find best parameters to fit dataset i.e. **choose** θ_0 **and** θ_1 **so that** $h_\theta(x)$ **is close to y for training examples (x, y)**.

$$minimize_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

$$minimize_{\theta_0, \theta_1} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

- Least square method,
- Direct/analytical/closed form solution / OLS method,
- Gradient Descent

Regression – Optimization

$$\hat{Y} = \beta_0 + \beta_1 X$$

least squares estimates of β_0 and β_1 are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\mathbf{m} = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$\mathbf{b} = \frac{\Sigma y - m\,\Sigma x}{N}$$

- Least Square Estimation: find the β0 and β1 parameter estimates that minimize the error sum of squares.

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Regression Model – Ordinary Least Squares (OLS)

- OLS method is a linear regression technique that is used to estimate unknown parameters in a model by minimizing the sum of squared residuals between the actual and predicted values.
- OLS estimator used for unknown parameters in a linear regression model.
- best estimator is refer to as **BLUE** (best linear unbiased estimator)
- Estimates of β coefficients are values that minimize sum of squared errors for sample.
- Letter b is used to represent a sample estimate of a β coefficient… b_0 and b_1 are estimators of β_0 and β_1.
- Then the sum of squared estimation mistakes can be expressed as;

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$$

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \\ Y_N &= \beta_0 + \beta_1 X_N + \epsilon_N \end{cases}$$

Linear Regression – OLS Matrix method

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \\ Y_N &= \beta_0 + \beta_1 X_N + \epsilon_N \end{cases}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = [Y_1, \ldots, Y_N]^\top$, $\boldsymbol{\varepsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$, $\boldsymbol{\beta} = [\beta_0, \beta_1]^\top$ and $\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{bmatrix}$

Least squares estimates in matrix notation

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

- $(X'X)^{-1}$ is the **inverse** of the $X'X$ matrix, and
- X' is the **transpose** of the X matrix.

Linear Regression – OLS Matrix method

$$A = \begin{bmatrix} 1 & 5 \\ 4 & 8 \\ 7 & 9 \end{bmatrix}$$

$$A' = A^T = \begin{bmatrix} 1 & 4 & 7 \\ 5 & 8 & 9 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & x_n \\ 1 & \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

- $(X'X)^{-1}$ is the **inverse** of the $X'X$ matrix, and
- X' is the **transpose** of the X matrix.

$$A^{-1} = \frac{adj(A)}{|A|}; |A| \neq 0$$

$$adj\,A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

Change sign Interchange

If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of A — Determinant of A — Adjoint of A

Linear Regression – OLS Matrix method

Calculate OLS estimates for the given date.

x_i soap	y_i suds	$x_i \times y_i$ so*su	x_i^2 soap²
4.0	33	132.0	16.00
4.5	42	189.0	20.25
5.0	45	225.0	25.00
5.5	51	280.5	30.25
6.0	53	318.0	36.00
6.5	61	396.5	42.25
7.0	62	434.0	49.00
---	---	-----	-----
38.5	347	1975.0	218.75

$$X'X = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.75 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} 347 \\ 1975 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix}$$

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & x_n \\ 1 & \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

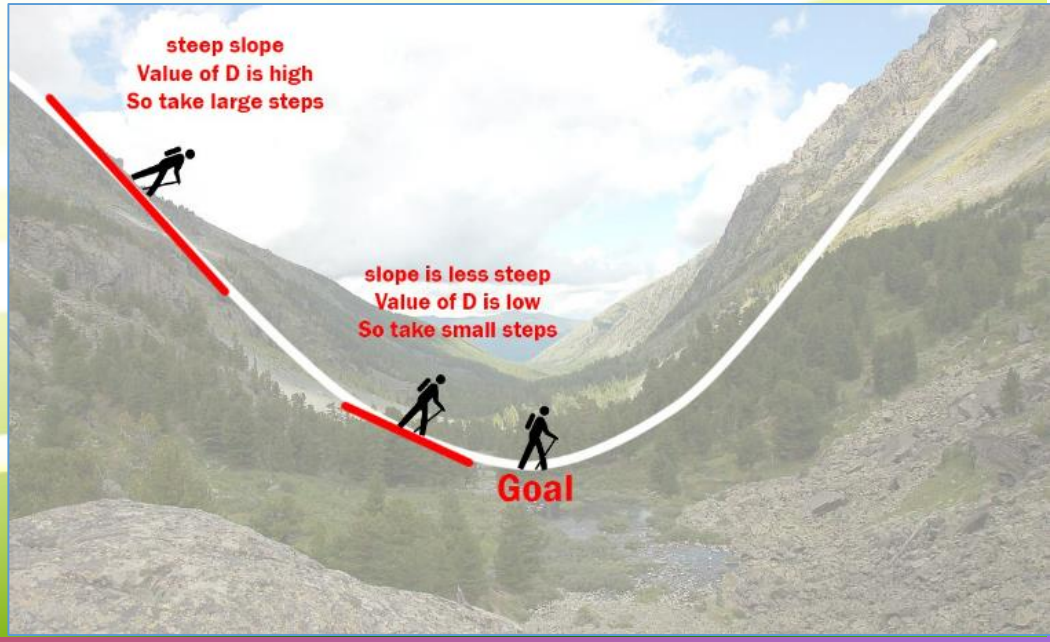
$$b = (X'X)^{-1}X'Y = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix} \begin{bmatrix} 347 \\ 1975 \end{bmatrix} = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

OLS estimated b_0 = -2.67 and b_1 = 9.51

Regression – Gradient Descent

- Gradient descent is an iterative optimization algorithm to find the minimum of a function.

- Imagine a valley and a person with no sense of direction who wants to get to bottom of the valley.
- He goes down the slope and takes large steps when slope is steep and small steps when slope is less steep.
- He decides his next position based on his current position.
- He stops when he gets to bottom of the valley (goal).



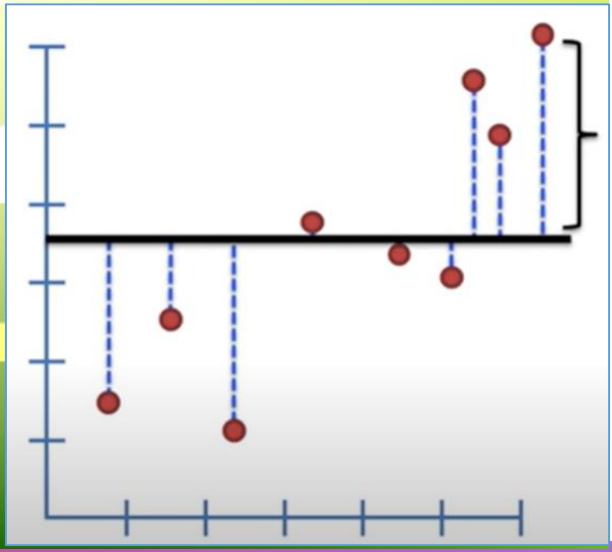
Regression – Gradient Descent

- Gradient descent is an iterative optimization algorithm to find the minimum of a function.
- model targets to minimize the cost function.
- To minimize cost function, model needs to have best value of θ1 and θ2.
- Initially model selects θ1 and θ2 values randomly and then iteratively update these value in order to minimize cost function until it reaches the minimum.
- By the time model achieves minimum cost function, it will have best θ1 and θ2 values.
- Using these finally updated θ1 and θ2, model predicts in best manner.

Linear Regression Cost Function

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$



Regression – Gradient Descent

- Gradient descent is an iterative optimization algorithm to find the minimum of a function.

Cost Function

$$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y_i]^2$$

Predicted Value; True Value

Gradient Descent

$$\Theta_j = \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta_0, \Theta_1)$$

Learning Rate

Now,

$$\frac{\partial}{\partial\Theta}J_\Theta = \frac{\partial}{\partial\Theta}\frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y]^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x_i) - y)\frac{\partial}{\partial\Theta_j}(\Theta x_i - y)$$

$$= \frac{1}{m}(h_\Theta(x_i) - y)x_i$$

Therefore,

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

θ_j : Weights of the hypothesis.

$h_{\theta(xi)}$: predicted y value for i^{th} input.

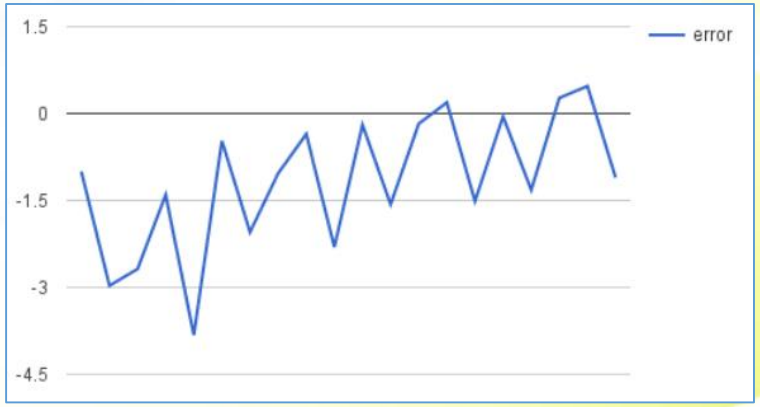
j : Feature index number (can be 0, 1, 2, ..., n).

α : Learning Rate of Gradient Descent.

Regression – Gradient Descent

Linear Regression Cost Function

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$



<u>Gradient descent steps</u>:

1. Initially let m = 0, c = 0, and L be the learning rate.
   - This controls how much the value of **m** changes with each step.
   - L could be a small value like 0.0001 for good accuracy.
2. Calculate partial derivative of loss function with respect to m, and plug in current values of x, y, m and c in it to obtain the derivative value **D**.

$$D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$$

3. Now update current value of m and c using following equation:

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

4. repeat this process until loss function is a very small value (ideally 0).
   - Current value of m and c will be optimum values.

$$D_c = \frac{-2}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)$$

Regression – Gradient Descent

$$D_m = \frac{\partial(Cost\ Function)}{\partial m} = \frac{\partial}{\partial m}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i^2 + m^2x_i^2 + c^2 + 2mx_ic - 2y_imx_i - 2y_ic)\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}x_i(y_i - (mx_i + c))$$

$$= \frac{-2}{n}\sum_{i=0}^{n}x_i(y_i - y_{i\,pred})$$

$$D_c = \frac{\partial(Cost\ Function)}{\partial c} = \frac{\partial}{\partial c}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

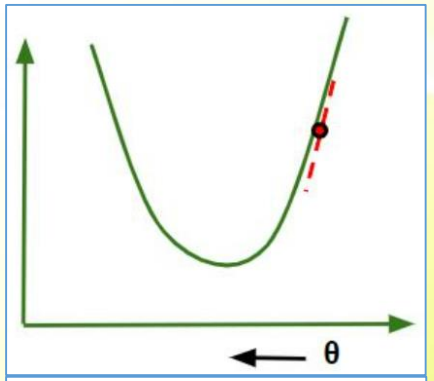
$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

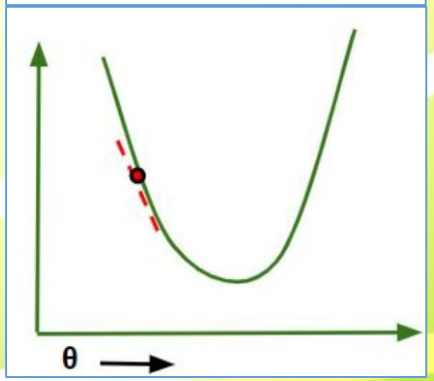
$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i^2 + m^2x_i^2 + c^2 + 2mx_ic - 2y_imx_i - 2y_ic)\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))$$

$$\frac{-2}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})$$

Regression – Gradient Descent

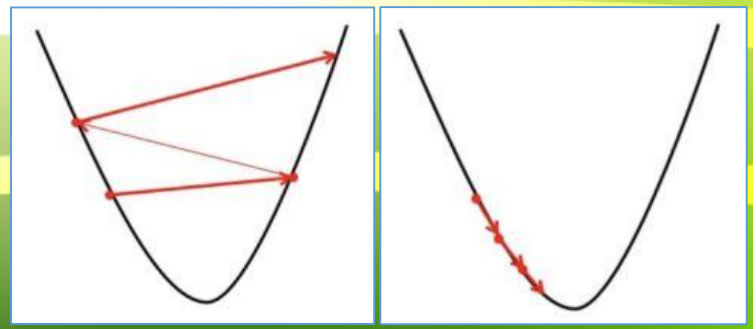




- Gradient Descent step-downs cost function in the direction of the steepest descent.
- size of each step is determined by parameter α known as Learning Rate.
  - **If slope is +ve** : $\theta_j = \theta_j$ – (+ve value). Hence value of θ_j decreases.
  - **If slope is -ve** : $\theta_j = \theta_j$ – (-ve value). Hence value of θ_j increases.

Choice of correct learning rate is very important as it ensures that Gradient Descent converges in a reasonable time. :

- If α chosen to be very large, Gradient Descent can overshoot the minimum.
  - Takes more time. It may fail to converge or even diverge.
- If α chosen to be very small, Gradient Descent will take small steps to reach local minima and will take a longer time to reach minima.



Regression – Gradient Descent

Convexity – In linear regression problem, there used to be only one minimum. The error surface is convex.

- Regardless of where we started, we would eventually arrive at the absolute minimum.
- Not always. It's possible to have a problem with **local minima** that gradient search can get stuck in.

Performance – Usually vanilla gradient descent used with a learning rate of 0.0005 (runs for 100's/1000's iterations).

- Different approaches can reduce the number of iterations required.
- Example, line search reduces number of iterations to arrive at a reasonable solution from several thousand to around few hundreds' (even < 100).

Regression – Gradient Descent

Gradient descent algorithms could be implemented in different ways:

Batch gradient descent: weight update is calculated based on all examples in training dataset.

- computationally highly expensive (not recommended) when number of training samples is huge.

Stochastic gradient descent (SGD): each iteration only analyses one training example.

- weight update is calculated incrementally after each training example.
- Large datasets often can't be held in RAM. Rather, each sample must be loaded, worked with, the results stored, and so on → lot faster than batch gradient descent.
- if number of training instances is huge, it will only process one of them, which will add to system's overhead because huge number of iterations.

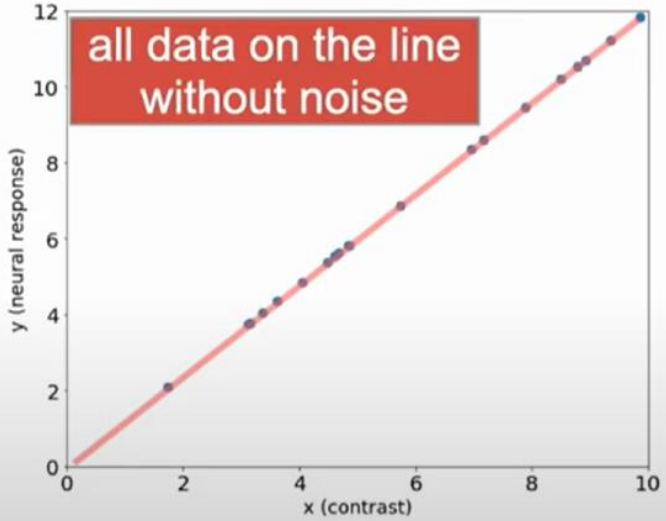
Mini Batch gradient descent: faster than both batch and stochastic gradient descent methods.

- Even if there are huge number of training examples, they are handled in 'm' batches of 'b' training examples at a time.

Linear Regression - Maximum Likelihood Estimation

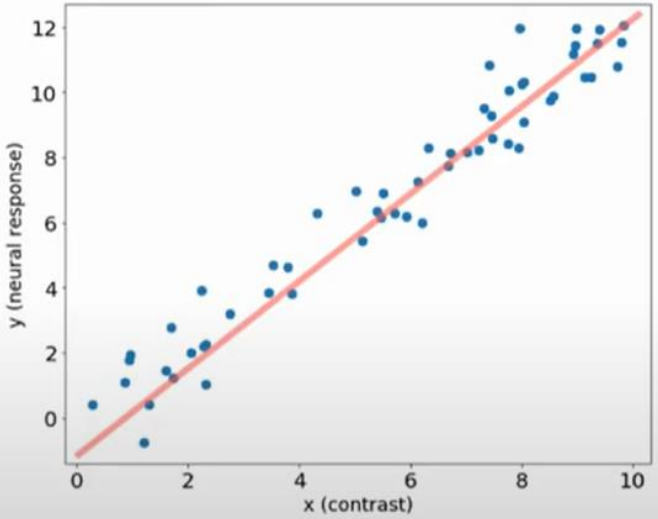
If we generate y from x using

$$y = \theta \times x$$



Model data with noise.

$$y = \theta \times x + \eta \leftarrow \text{Noise}$$



what we don't care about
(e.g., deviation from mean firing rate)

what we can't control
(e.g., measurement noise)

$$\eta \sim \mathcal{N}(0, \sigma^2)$$

$$\downarrow$$

$$y \sim \mathcal{N}(\theta x, \sigma^2)$$

$$y = \theta \times x + \eta$$

neural response | linear weight | contrast | Gaussian noise $\eta \sim \mathcal{N}(0, \sigma^2)$

$$p(y|x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\theta x)^2}$$

$$\text{mean}(y) = \theta x$$

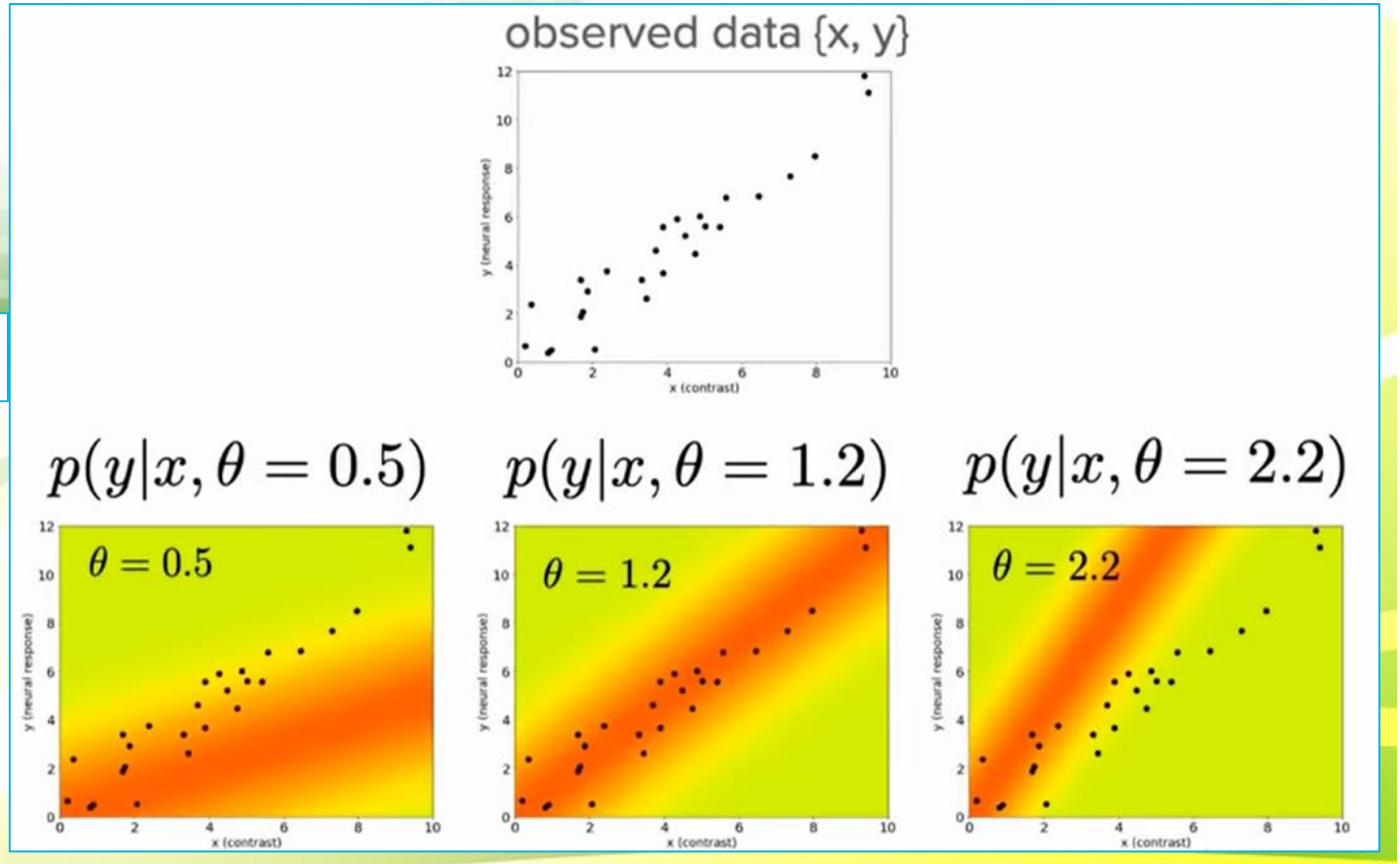
$$\text{var}(y) = \sigma^2$$

Linear Regression - Maximum Likelihood Estimation

- **Maximum likelihood estimation** (MLE): mechanism for finding optimal parameters of statistical models, that maximize joint density. $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Which θ will lead the gaussian distribution to most likely to match data

$$\mathcal{L}(\theta|x,y) = p(y|x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\theta x)^2}$$

- Calculate optimal θ by maximizing the likelihood.



$$\mathcal{L}(\theta=1.2) > \mathcal{L}(\theta=0.5) > \mathcal{L}(\theta=2.2)$$

Linear Regression - Maximum Likelihood Estimation

- Maximizing log likelihood is same as minimizing MSE.

$$\log \mathcal{L}(\theta|x,y) = \log \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i-\theta x_i)^2}$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i-\theta x_i)^2}$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sigma\sqrt{2\pi}} + \sum_{i=1}^{N} \log e^{-\frac{1}{2\sigma^2}(y_i-\theta x_i)^2}$$

$$= -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \theta x_i)^2$$

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

Linear Regression - Maximum Likelihood Estimation

Example: A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

For a given value of p, probability of getting 55 heads in this experiment is binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}$$

P(x:n,p) = ${}^{n}C_{x}\ p^{x}\ (1\text{-}p)^{n-x}$

probability of getting 55 heads depends on value of p (conditional probability):

$$P(55 \text{ heads} \mid p) = \binom{100}{55} p^{55}(1-p)^{45}$$

Find MLE $\hat{p}$ by taking derivative of likelihood function and setting it to 0.

$$\frac{d}{dp} P(\text{data} \mid p) = \binom{100}{55} (55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0$$

$$55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44}$$

$$55(1-p) = 45p$$

$$55 = 100p$$

$$\text{the MLE is } \hat{p} = .55$$

Linear Regression - Maximum Likelihood Estimation

Example: A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss. Use log likelihood estimator.

Likelihood $P(55 \text{ heads} \mid p) = \binom{100}{55} p^{55}(1-p)^{45}$

Log likelihood $\ln(P(55 \text{ heads} \mid p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)$

Maximizing likelihood is the same as maximizing log likelihoo

$$\begin{aligned}\frac{d}{dp}(\text{log likelihood}) &= \frac{d}{dp}\left[\ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)\right] \\ &= \frac{55}{p} - \frac{45}{1-p} = 0 \\ &\Rightarrow 55(1-p) = 45p \\ &\Rightarrow \hat{p} = .55\end{aligned}$$

Linear Regression - Maximum Likelihood Estimation

Example: Suppose that the lifetime of Badger brand light bulbs is modeled by an exponential distribution with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

Let X_i be the lifetime of i^{th} bulb.

Each X_i has pdf $f_{X_i}(x_i) = \lambda e^{-\lambda x_i}$

Assuming lifetimes of bulbs are independent, joint pdf:

$$f(x_1, x_2, x_3, x_4, x_5 \mid \lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}$$

Likelihood and log likelihood functions:

$$f(2, 3, 1, 3, 4 \mid \lambda) = \lambda^5 e^{-13\lambda}, \quad \ln(f(2, 3, 1, 3, 4 \mid \lambda) = 5 \ln(\lambda) - 13\lambda$$

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \Rightarrow \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Linear Regression - Model Selection

- Model selection is the process of selecting a model from a set of candidate models.
- Selecting best coefficient:
  - Least square method,
  - Direct/analytical/closed form solution / OLS method,
  - Gradient Descent
- Sometimes, dimension ‘p’ is too huge.
- Aim is to select a subset of predictor variables to perform regression, e.g. to choose k predicting variables from the total of p variables yielding minimum RSS.
- If model includes too many predictors → may lead to overfitting.
  - model gives good predictions to training data but performs much worse with new data.
  - Overfitting model normally gives low bias but high variance.
- Conversely, when there are too few predictors in model → underfitting may show up.
  - gives poor predictions to data used or not used for model fitting.
  - Underfitting models normally have low variance but high bias.

Linear Regression - Model Selection

- Two main approaches for model (variable) selection
  - Testing based approaches,
  - Criterion-based approaches.
- Testing-based approaches variables are selected based on whether they are significant or not when they are added/removed
  - *Example; include backward elimination, forward selection, stepwise regression, etc.*
- For criterion-based approaches, we have some idea about the purpose for which a model is intended, so we might propose some measures of how well a given model meets that purpose.
  - Then a model that optimizes a criterion which balances goodness-of-fit will be chosen.
  - *Examples of criterion-based approaches include Akaike information criterion(AIC) and Bayesian information criterion(BIC), adjusted R2, Mallow's cp, etc.*

Linear Regression - Model Selection

Univariate Analysis method:

- The initial step is to check each independent variable with dependent variable.
- It is to eliminate some independent variables which are not related to dependent variable at all.
- Check their individual model R^2 and p-value of their coefficient to test whether coefficient is significantly different from zero.
    - *P-value is a statistical number to conclude if there is a relationship between two variables.*
- Select predictor with Higher R^2 for given p-value.
- Not advisable; with 1000 predictors in regression model, It is memory intensive to run regression model 1000 times to produce R^2 of each variable.

Linear Regression - Model Selection

- In **Forward selection method**, build a regression model by adding predictors step-by-step, until the pre-set significance level is met for all predictors.
- Forward selection starts with a null model.
- It adds variables to the model one-by-one.
- The criteria of which variable to include each time is as following:
  1) test each variable that is not already in the model,
  2) check the significance of all of them to see if their P-value is below certain level, and
  3) choose the one that is the most significant.

<u>Drawbacks:</u>

- every time one variable is added into the model, it ignores the fact that new variable may render some of the existing variables to be non-significant.

Linear Regression - Model Selection

- In **backward elimination method**, build a regression model by removing predictors step-by-step, until the pre-set significance level is met for all predictors.
- Model starts with all variables.
- It removes variables from the model one-by-one.
- The criteria of which variable to include each time is as following:
    1. check the significance of each variable,
    2. drop one with the least significance each time until all the variables remained are statistically significant.

Linear Regression - Model Selection

- **STEPWISE selection algorithm** is a combination of backward and forward selection.
- build regression model by adding/removing predictors step-by-step, until pre-set significance level is met for all predictors.
  - In forward stepwise regression, variable which would add largest increment to R^2 *(variable with largest semi-partial correlation)* is added next *(provided it is statistically significant)*.
  - In backwards stepwise regression, variable which would produce the smallest decrease in R^2 *(variable with smallest semi-partial correlation)* is dropped next *(provided it is not statistically significant)*.

Linear Regression - Model Selection

- **Adjusted R-Square** penalizes the model for inclusion of each additional variable.
  - Adjusted R-square would increase only if the variable included in model is significant.
  - model with larger adjusted R-square value is considered to be better model.

- **Mallows' Cp Statistic** helps detect model biasness *(refers to either underfitting or overfitting model)*.

Mallows Cp = (SSE/MSE) – (n – 2p)

  - SSE is Sum of Squared Error
  - MSE is Mean Squared Error with all independent variables in model
  - N is sample size
  - p is number of estimates/predictor variables in model

- A final model should be selected based on the following two criterias -
  - First Step : Cp is less than or equal to p
  - Second Step : fewest parameters exist.

Predictor Variables	P+1	Mallows' Cp
Hours	2	45.5
Prep exams	2	31.4
GPA	2	29.3
Hours, Prep exams	3	3.4
Hours, GPA	3	2.9
Prep exams, GPA	3	2.7
Hours, Prep exams, GPA	4	4

Linear Regression - Model Selection

Akaike information criterion (AIC) & Bayes Information Criterion (BIC)

$$AIC = n \ln(\frac{RSS}{n}) + 2(p+1)$$

$$BIC = n \ln(\frac{RSS}{n}) + (p+1) \ln n$$

n is number of training data
p is number of parameters in model.

- It tells nothing about absolute quality of a model, rather only quality relative to other models.
- From a collection models, the 'best' (or 'least bad') one can be chosen by seeing which has **lowest AIC or BIC.**
- AIC is asymptotically equivalent to leave-one-out cross-validation.
- AIC tends to overfit models.
- BIC tends to favor simpler models than those chosen by AIC.
- AIC is generally considered better.

Linear Regression – Probability bound

- **Probability bounds** are inequalities that are usually applicable to a general scenario.
- Calculating exact value of probability might be difficult
  - unknown parameters; Not enough information to calculate a desired quantity,
  - problem might be complicated,
  - exact calculation might be very difficult.
- No need to actually find the desired probability exactly, rather an related inequality.
  - Example; exact mark of student (probability) might be difficult / not required.
  - Probability of marks < 40 is enough to classify as ‘Fail’.

Linear Regression – Probability bound

- **Union bound or Boole's inequality** is applicable to show that the probability of union of some events is less than some value.
- For any two events A and B;

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B).$$

- Similarly, for three events A, B, and C;

$$P(A \cup B \cup C) = P\big((A \cup B) \cup C\big) \leq P(A \cup B) + P(C) \leq P(A) + P(B) + P(C)$$

Union bound

For any events $A_1, A_2, \ldots, A_n$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

Linear Regression – Probability bound

- Generalization of Union Bound: **Bonferroni Inequalities**;

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B).$$

For any events $A_1, A_2, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right)$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i);$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j);$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k).$$

Linear Regression – Probability bound

For a random variable X;

For a > 0

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f_X(x)dx \\ &= \int_{0}^{\infty} x f_X(x)dx \\ &\geq \int_{a}^{\infty} x f_X(x)dx \\ &\geq \int_{a}^{\infty} a f_X(x)dx \\ &= a\int_{a}^{\infty} f_X(x)dx \\ &= aP(X \geq a). \end{aligned}$$

Markov's inequality

$$P(X \geq a) \leq \frac{EX}{a}, \qquad \text{for any } a > 0.$$

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} x f(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

$$Y = (X - EX)^2$$

$$P(Y \geq b^2) \leq \frac{EY}{b^2}$$

$$\begin{aligned} EY &= E(X - EX)^2 = Var(X), \\ P(Y \geq b^2) &= P\left((X - EX)^2 \geq b^2\right) = P\left(|X - EX| \geq b\right) \end{aligned}$$

Chebyshev's inequality

$$P\left(|X - EX| \geq b\right) \leq \frac{Var(X)}{b^2}$$

Linear Regression – Probability bound

For a random variable X;

$$P(X \geq a) = P(e^{sX} \geq e^{sa}), \quad \text{for } s > 0,$$
$$P(X \leq a) = P(e^{sX} \geq e^{sa}), \quad \text{for } s < 0.$$

$a, s \in \mathbb{R}$

For s > 0

$$P(X \geq a) = P(e^{sX} \geq e^{sa}) \leq \frac{E[e^{sX}]}{e^{sa}}, \quad \text{by Markov's inequality.}$$

For s < 0

$$P(X \leq a) = P(e^{sX} \geq e^{sa}) \leq \frac{E[e^{sX}]}{e^{sa}}.$$

$$P(X \geq a) \leq \frac{EX}{a}, \quad \text{for any } a > 0.$$

$E[e^{sX}]$ can be represented in terms of **moment generating function**, $M_X(s)$

Chernoff Bounds:

$$P(X \geq a) \leq e^{-sa} M_X(s), \quad \text{for all } s > 0,$$
$$P(X \leq a) \leq e^{-sa} M_X(s), \quad \text{for all } s < 0$$

Linear Regression - Probability bound

Example: Let X∼Binomial(n,p). Using markov's inequality, find an upper bound on P(X ≥ αn), where p<α<1. Evaluate the bound for p=1/2 and α=3/4.

$$P(X \geq a) \leq \frac{EX}{a}, \qquad \text{for any } a > 0.$$

Binomial: **EX=np**

$$P(X \geq \alpha n) \leq \frac{EX}{\alpha n} = \frac{pn}{\alpha n} = \frac{p}{\alpha}$$

for p=1/2 and α=3/4

$$P(X \geq \frac{3n}{4}) \leq \frac{2}{3}$$

Linear Regression - Probability bound

Example: Let X∼Binomial(n,p). Using Chebyshev's inequality, find an upper bound on P(X ≥ αn), where p<α<1. Evaluate the bound for p=1/2 and α=3/4.

$$P(|X - EX| \geq b) \leq \frac{Var(X)}{b^2}$$

Binomial: **EX=np**

$$\begin{aligned} P(X \geq \alpha n) &= P(X - np \geq \alpha n - np) \\ &\leq P(|X - np| \geq n\alpha - np) \\ &\leq \frac{Var(X)}{(n\alpha - np)^2} \\ &= \frac{p(1-p)}{n(\alpha - p)^2}. \end{aligned}$$

for p=1/2 and α=3/4

$$P(X \geq \frac{3n}{4}) \leq \frac{4}{n}$$

Linear Regression – Probability bound

For a random variable X;

Markov's inequality

$$P(X \geq a) \leq \frac{EX}{a}, \qquad \text{for any } a > 0.$$

Chebyshev's inequality

$$P(|X - EX| \geq b) \leq \frac{Var(X)}{b^2}$$

Chernoff Bounds:

$$P(X \geq a) \leq e^{-sa} M_X(s), \qquad \text{for all } s > 0,$$
$$P(X \leq a) \leq e^{-sa} M_X(s), \qquad \text{for all } s < 0$$

$$P(X \geq \frac{3n}{4}) \leq \frac{2}{3}$$
$$P(X \geq \frac{3n}{4}) \leq \frac{4}{n}$$
$$P(X \geq \frac{3n}{4}) \leq (\frac{16}{27})^{\frac{n}{4}}$$

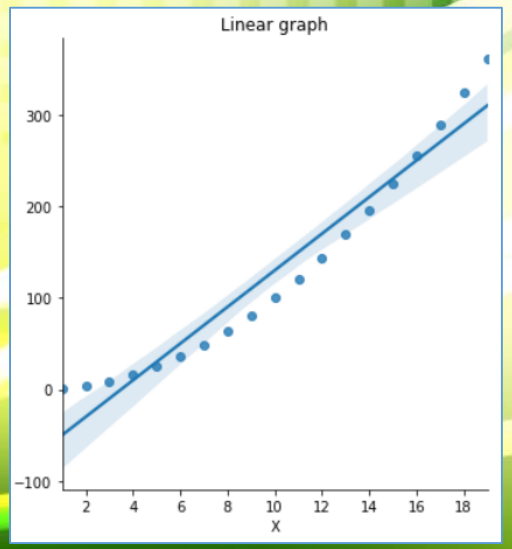
Example: Let X∼Binomial(n,p).

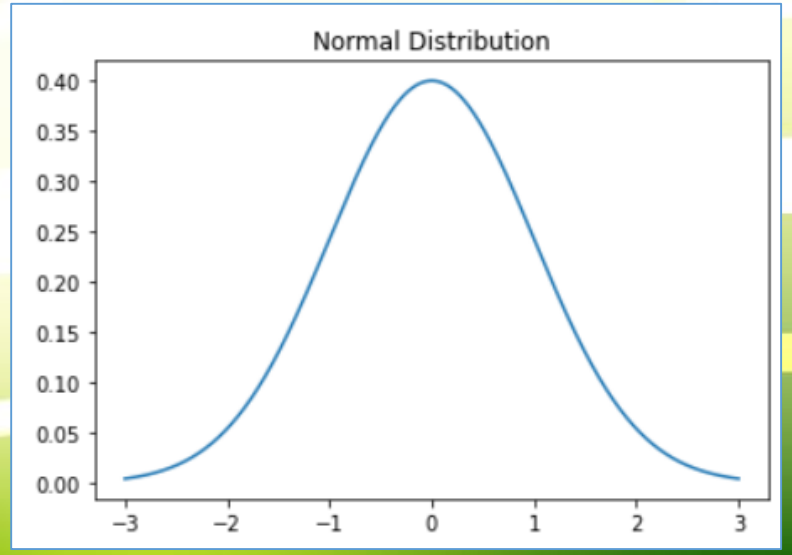
Using different inequality, upper bound on P(X ≥ αn), where p<α<1 for p=1/2 and α=3/4.

- **Bound given by Markov is "weakest".**
  - It is constant and does not change as n increases.
- **Bound given by Chebyshev's inequality is "stronger" than Markov's.**
  - 4/n goes to zero as n goes to infinity.
- **Strongest bound is Chernoff bound.**
  - It goes to zero exponentially fast.

Generalized Linear Model (GLM)

- GLM is an advanced statistical modelling technique.
- It is an umbrella term that encompasses many other models, Linear Regression, Logistic Regression, and Poisson Regression, etc.
- Allows response variable y to have an error distribution other than a normal distribution.
- In Linear Regression Model, response variable ‘y’ is expressed as linear function/combination/relation of all predictors ‘X’ (can simply visualize in form of straight line).
- Also, the error distribution of response variable should be normally distributed.





Generalized Linear Model (GLM)

- GLM models allow to build linear relationship between response and predictors, even though their underlying relationship is not linear.
- This is made possible by using a link function, which links the response variable to a linear model.
- Unlike Linear Regression models, error distribution of response variable need not be normally distributed.
- Errors in response variable are assumed to follow an exponential family of distribution (i.e. normal, binomial, Poisson, or gamma distributions).
- Aims to generalize a linear regression model that can also be applied in these cases → Generalized Linear Models.

Non-linear Regression

- All previous models have linear parameters.
- Nonlinear regression model:

$$Y = f(\mathbf{X}, \beta) + \epsilon$$

  - where X is a vector of p predictors,
  - β is a vector of k parameters,
  - f(·) is some known regression function, and
  - ϵ is an error term whose distribution may or may not be normal.
- No longer necessary to have dimension of parameter vector simply one greater than number of predictors.
- Some examples of nonlinear regression models:

$$y_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + \epsilon_i$$

$$y_i = \frac{\beta_0+\beta_1 x_i}{1+\beta_2 e^{\beta_3 x_i}} + \epsilon_i$$

$$y_i = \beta_0 + (0.4-\beta_0)e^{-\beta_1(x_i-5)} + \epsilon_i$$

- *Intrinsically non-linear models can be transformed into linear form.*

$$Y = \frac{\beta_0 X}{\beta_1 + X}$$

Non-linear Regression

- Nonlinear regression model: $Y = f(\mathbf{X}, \beta) + \epsilon,$
- Nonlinear Least Square;

$$Q = \sum_{i=1}^{n} (y_i - f(\mathbf{X}_i, \beta))^2$$

$$\hat{\beta} = \arg\min_{\beta} Q,$$

Algorithms for nonlinear least squares estimation include:

- **Newton's method**, classical method based on gradient approach.
- **Gauss-Newton algorithm**, modification of Newton's method.
- **Levenberg-Marquardt method**, takes care of computational difficulties arising with other methods.

Non-linear Regression

Argmax & argmin

- The argmax function returns the argument or arguments (arg) for the target function that returns the maximum (max) value from the target function.
  - *Example, given a function g() that takes the argument x,*
  - *the argmax operation of that function would be described as : result = argmax(g(x))*
  - $g(x) = x^2$, where x is integers from 1 to 5.
  - argmax(g(x)) = 5
  - argmin(g(x)) = 1

g(1) = 1^2 = 1
g(2) = 2^2 = 4
g(3) = 3^2 = 9
g(4) = 4^2 = 16
g(5) = 5^2 = 25

Logistic Regression

- Supervised classification algorithm.
- In linear regression problem, target variable(output) y can take only continuous values for a given set of features(inputs) X.
- Logistic regression is popular approach to building models where response variable is categorical.
- Just like Linear regression, it assumes that the data follows a linear function.
- Logistic Regression in its base form is a *Binary Classifier*.
  - Target vector may only take the form of one of two values.
  - Model builds a regression model to predict probability that a given data entry belongs to the category numbered as '1'.
  - A Linear Model, $\beta_0 + \beta_1 x$, is integrated into a Logistic Function (Sigmoid Function).

Sugar Level (X)	Diabetes (Y)
354	1
190	0
405	1
263	0
451	1

Logistic Regression

Binary Logistic Regression:

- Used when response is binary (two possible outcomes).
- Example; passing or failing a test, responding yes or no on a survey, and having high or low blood pressure.

Nominal Logistic Regression:

- Used when three or more categories with no natural ordering to levels.
- Example; departments at business (e.g., marketing, sales, HR), type of search engine used (e.g., Google, Yahoo!, MSN), and color (black, red, blue, orange).

Ordinal Logistic Regression:

- Used when three or more categories with a natural ordering to levels, but ranking of levels do not necessarily mean the intervals between them are equal.
- Example; how students rate effectiveness of a college course (e.g., good, medium, poor), levels of flavors for hot wings, and medical condition (e.g., good, stable, serious, critical).

Logistic Regression

- standard linear regression formula would compute values outside 0-1 range (not useful)
- Logistic function ensures prediction in 0-1 range

Linear regression/line function

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_k$$

logistic function for response = 1

$$P(y = 1) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

b_0 is a constant, and b_1 b_k are coefficients to k independent variables (x_1 x_k).

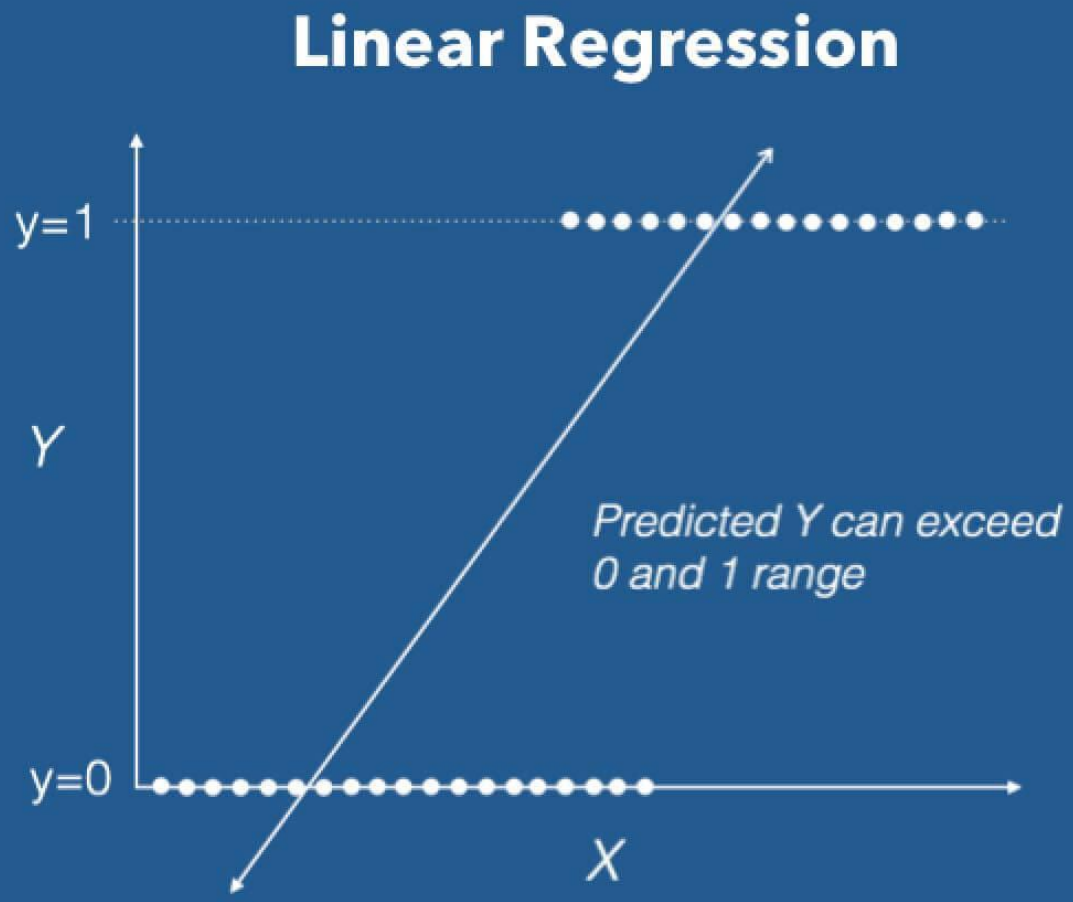
Logistic Regression

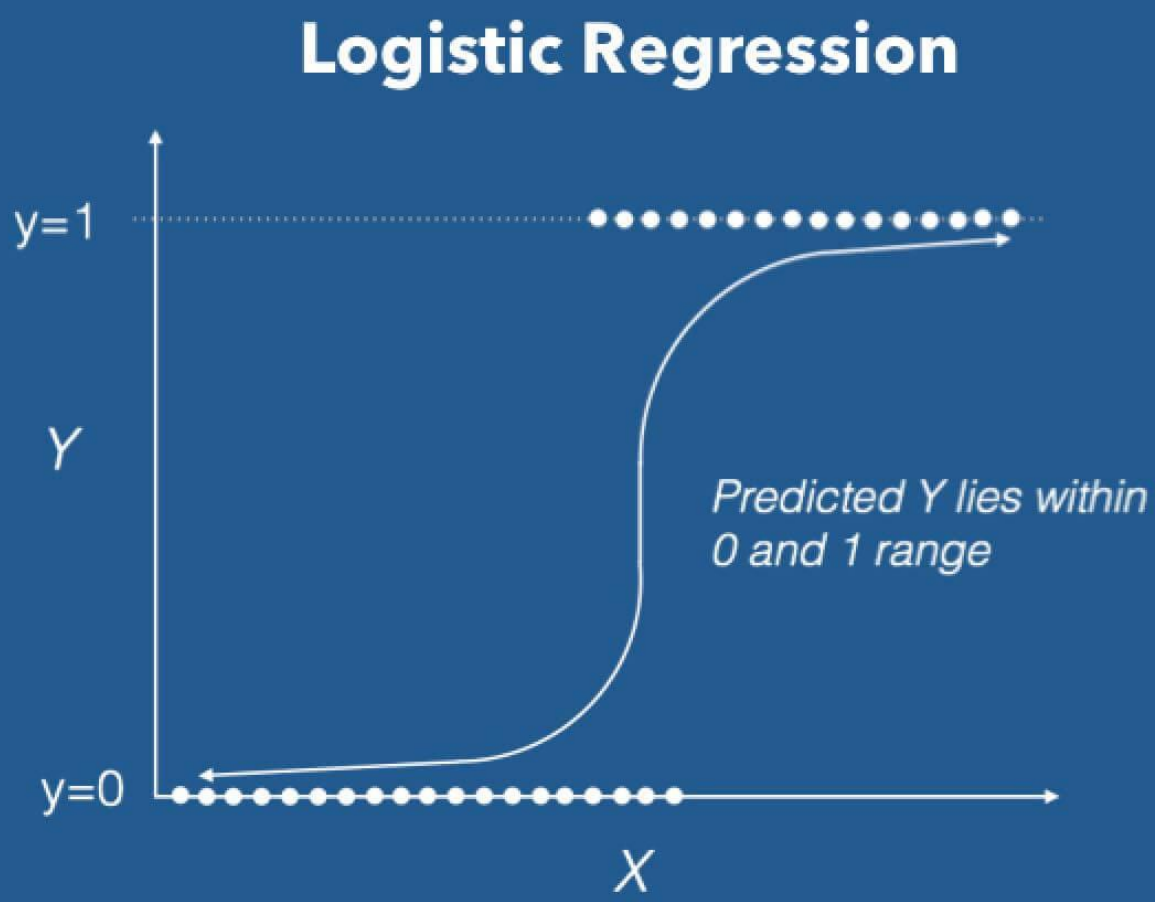
- Multiple binary logistic regression model:

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)} = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)} = \frac{1}{1 + \exp(-\mathbf{X}\beta)},$$

- π denotes a probability (not 3.14)
- π is the probability that an observation is in a specified category of the binary Y variable (success probability).
- estimates of π will always be between 0 and 1, because;
  - numerator $\exp(\beta_0+\beta_1X_1+\ldots+\beta_kX_k)$ is positive, because it is a power of positive value.
  - denominator of model is (1 + numerator), so answer will always be less than 1.
- With one X variable, theoretical model for π has an elongated "S" shape (or sigmoidal shape) with asymptotes at 0 and 1.

Logistic Regression

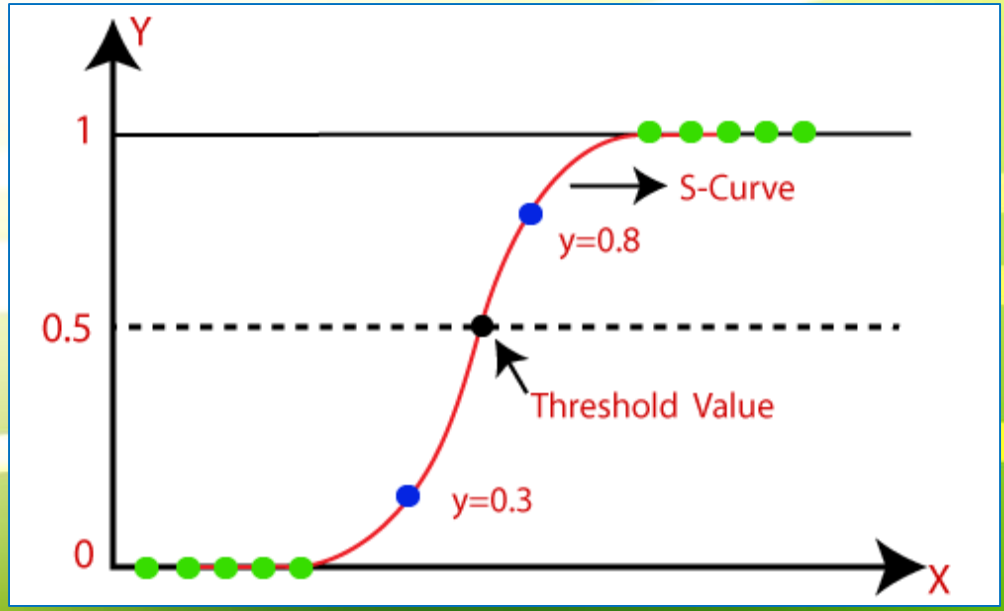




Logistic Regression

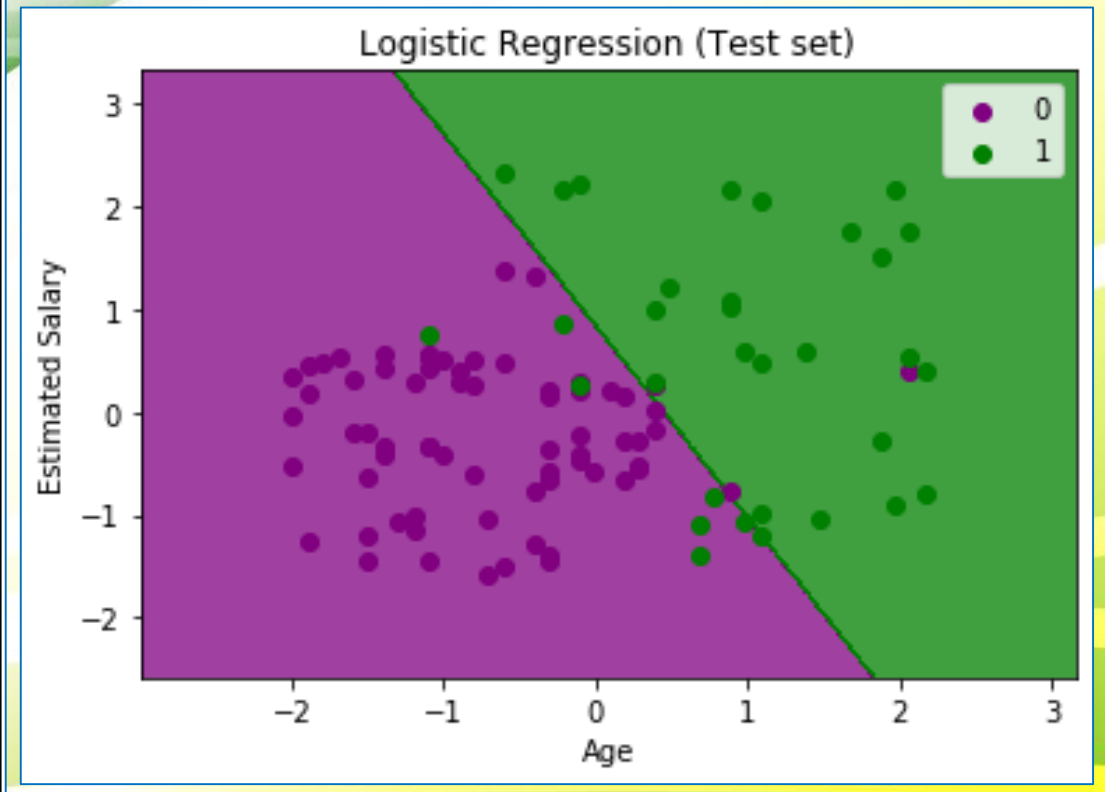
Logistic Function (Sigmoid Function):

- A mathematical function used to map the predicted values to probabilities.
- Maps any real value into another value within a range of 0 and 1.
- Value of logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- S-form curve is called Sigmoid function or logistic function.
- A threshold value defines probability of either 0 or 1.
    - values above threshold value tends to 1,
    - value below the threshold values tends to 0.



Logistic Regression

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1



Logistic Regression

- **Odds:** ratio of probability of an event occurring to probability of event not occurring.
- **Example,** let's assume that probability of winning a game is 0.02.

$$odds = \frac{P}{1-P}$$

  - Then, probability of not winning is 1- 0.02 = 0.98.
  - odds of winning game= (Probability of winning)/(probability of not winning) = 0.02/0.98
  - odds of winning the game are 1 : 49, and odds of not winning the game are 49 : 1.
- In a racetrack there is 80% chance that a certain horse will win the race.
- then his winning odds are 0.80 / (1 - 0.80) = 4, or 4:1.

Logistic Regression

- Transform the model from linear regression to logistic regression using the logistic function.

ln (odd)=bo+b1x

odds ratio (θ) between odds for two sets of predictors ($X_{(1)}$ and $X_{(2)}$):

$$\theta = \frac{(\pi/(1-\pi))|_{\mathbf{x}=\mathbf{x}_{(1)}}}{(\pi/(1-\pi))|_{\mathbf{x}=\mathbf{x}_{(2)}}}.$$

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

$$\frac{P}{1-P} = e^{b_0 + b_1 x}$$

$$P = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k),$$

$$\frac{\pi}{1-\pi} = \exp(\mathbf{X}\beta).$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

Logistic Regression

Sample proportion of women who are Instagram users is given as 61.08%, and sample proportion for men is 43.98%.

$$x = \begin{cases} 1 & \text{if the person is a woman} \\ 0 & \text{if the person is a man} \end{cases}$$

For women

$$\text{odds} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.6108}{1-0.6108} = 1.5694$$

y = log(odds) = log(1.5694) = 0.4507

For men

$$\text{odds} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.4398}{1-0.4398} = 0.7851$$

y = log(odds) = log(0.7851) = -0.2419

logistic regression model for women

$$\log\left(\frac{p_{\text{women}}}{1-p_{\text{women}}}\right) = \beta_0 + \beta_1$$

b_0 = - 0.2419

b_1 = 0.4507 - (-0.2419) = 0.6926

log(odds) = -0.2419 + 0.6926x

$$\log\left(\frac{p_{\text{men}}}{1-p_{\text{men}}}\right) = \beta_0$$

Logistic Regression

For sample of size n, likelihood for a binary logistic regression is :

$$L(\beta; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$
$$= \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i \beta)}{1+\exp(\mathbf{X}_i \beta)} \right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{X}_i \beta)} \right)^{1-y_i}.$$

log likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i) \log(1-\pi_i)]$$
$$= \sum_{i=1}^{n} [y_i \mathbf{X}_i \beta - \log(1+\exp(\mathbf{X}_i \beta))].$$

Logistic Regression – Goodness of Fit

Logistic regression model performance evaluation methods:

- Classification tables
- ROC curves
- Hosmer-Lemeshow tests
- Chi-square goodness of fit tests
- Deviance test
- Pearson test
- Logistic regression R^2

Logistic Regression– Goodness of Fit

Chi-square test:

Residual = $y_i - \hat{y}_i$

$$\hat{y}_i = \hat{\pi}_i(x_i) = \frac{\exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + \exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1 - \hat{y}_i)}}$$

$$X^2 = \sum_{i=1}^{n} r_i^2$$

$\chi^2 = \sum(O_i - E_i)^2/E_i$

$$\chi^2 = \sum \frac{(\text{Observed value - Expected value})^2}{\text{Expected value}}$$

Logistic Regression – Goodness of Fit

- Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

- **Question**: Are gender and education level dependent at 95% level of significance?
- In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

Logistic Regression – Goodness of Fit

Actual Data

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Expected Data

E_i = row total*column total/sample size

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

$$\chi^2 = \frac{(60 - 50.886)^2}{50.886} + \cdots + \frac{(57 - 48.132)^2}{48.132} = 8.006$$

- H0: There is no relationship between X and Y variable.
- H1: There is a relationship between X and Y variable.

- Critical value of χ2 with 3 degree of freedom is 7.815.
- 8.006 > 7.815 → reject the null hypothesis.
- Education level depends on gender at a 95% level of significance.

Logistic Regression– Goodness of Fit

Example: Suppose that we roll a die 30 times and observe the following table showing the number of times each face ends up on top. Calculate the chi-square coefficient.

$X^2 = \sum(O_i - E_i)^2/E_i$

Face	Count
1	3
2	7
3	5
4	10
5	2
6	3
Total	30

expected counts for each cell: $E_j = 30/6 = 5$

$$X^2 = \frac{(3-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(10-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(3-5)^2}{5} = 9.2$$

Logistic Regression – Goodness of Fit

Classification Tables

- 2 x 2 contingency table of observed and predicted results.
- model is used to classify each record (ranging between 0 and 1).
- Data records with probabilities greater than 0.5 *(50% as **cutoff value**)* are classified as 1.
- Those less than 0.5 are assigned a value 0.

	Observed positive	Observed negative
Predicted positive (above cutoff)	a	b
Predicted negative (below cutoff)	c	d

- Higher sensitivity and specificity indicate a better fit of the model.

$$\text{sensitivity} = \frac{a}{a+c}$$

TP/(TP + FN)

$$\text{specificity} = \frac{d}{b+d}$$

FP/(TN + FP)

Logistic Regression – Goodness of Fit

1. Consider two models—A and B—that each evaluate the same dataset. Which one of the following statements is true?

A. If model A has better precision and better recall than model B, then model A is probably better.

B. If model A has better recall than model B, then model A is better.

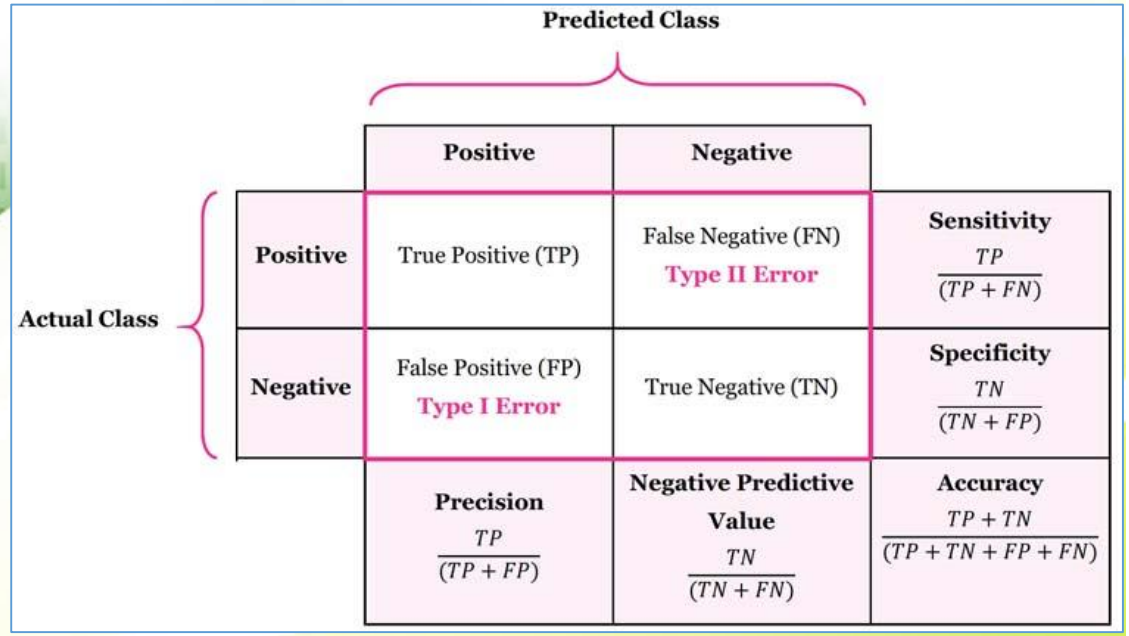
C. If Model A has better precision than model B, then model A is better.

2. Consider a classification model that separates email into two categories: "spam" or "not spam." If you raise the classification threshold, what will happen to recall?

A. Always decrease or stay the same.

B. Always increase.

C. Always stay constant.



3. Consider a classification model that separates email into two categories: "spam" or "not spam." If you raise the classification threshold, what will happen to precision?

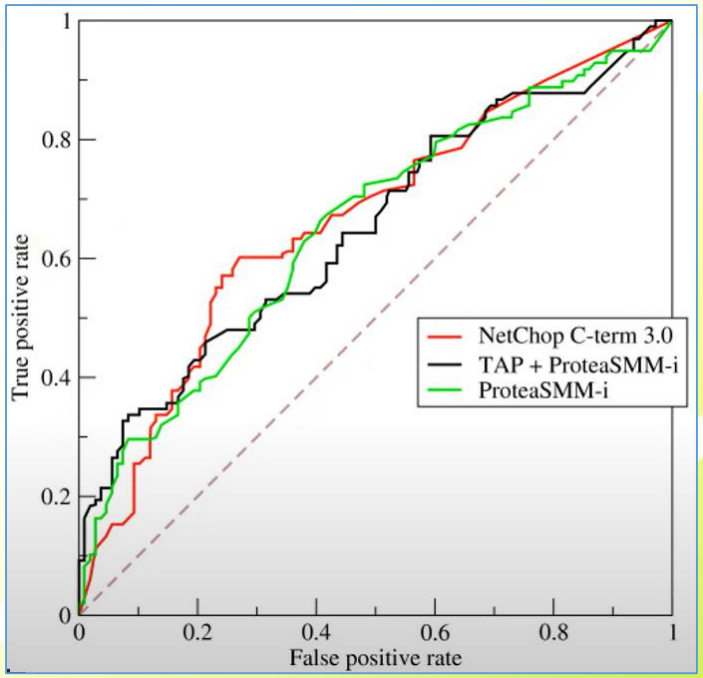
A. Definitely increase.

B. Definitely decrease.

C. Probably increase.

D. Probably decrease.

Logistic Regression – Goodness of Fit

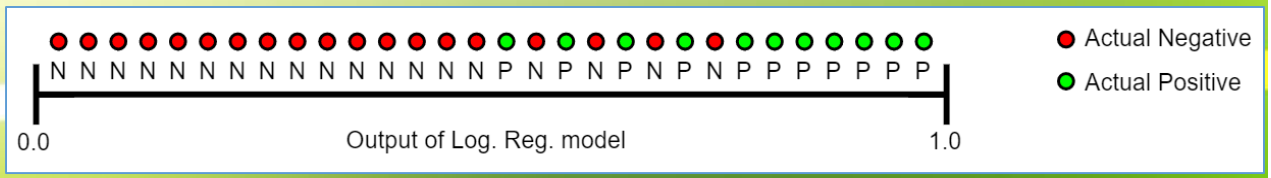


ROC curve (Receiver Operating Characteristic)

- Extending the classification (two-by-two) table idea, rather than selecting a single cutoff, we can examine the full range of cutoff values from 0 to 1.
- For each possible cutoff value, form a two-by-two table.
- ROC curve is plot between pairs of sensitivity *(true positive rate/TPR)* & specificities *(false positive rate/FPR)*.
- ROC curves helps determining best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1).
- ROC shows the performance of a classification model at all classification thresholds.
- This curve plots two parameters:
  - **True Positive Rate (TPR) / recall**
  - **False Positive Rate (FPR)**

$$TPR = \frac{TP}{TP + FN}$$

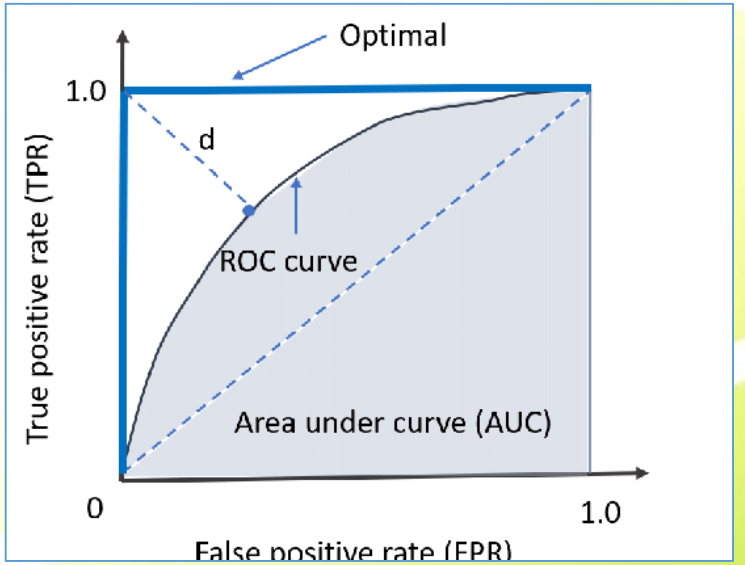
$$FPR = \frac{FP}{FP + TN}$$

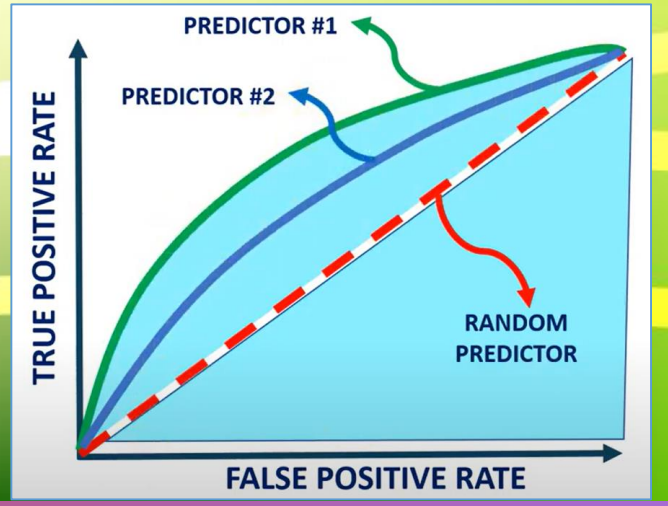


Logistic Regression – Goodness of Fit

AUC (Area under the ROC Curve)

- AUC provides overall measure of fit of the model.
- An excellent model has AUC near to the 1 which means it has a good measure of separability.
- A poor model has an AUC near 0 which means it has the worst measure of separability.
  - In fact, it means it is reciprocating the result.
  - It is predicting 0s as 1s and 1s as 0s.
- And when AUC is 0.5, it means the model has no specific class separation capacity whatsoever.

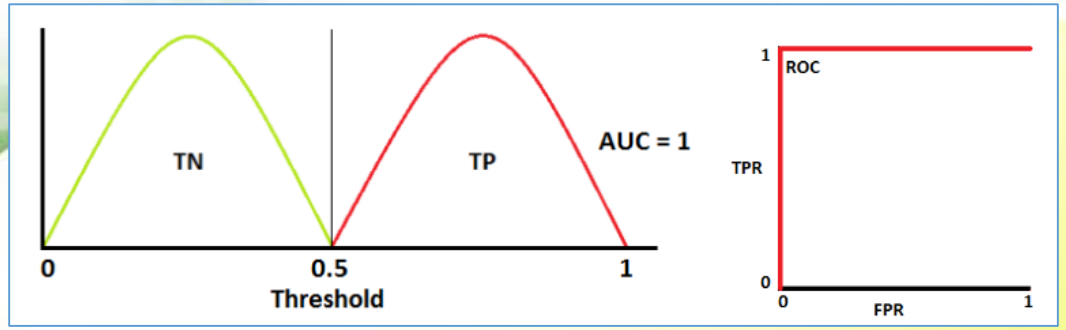


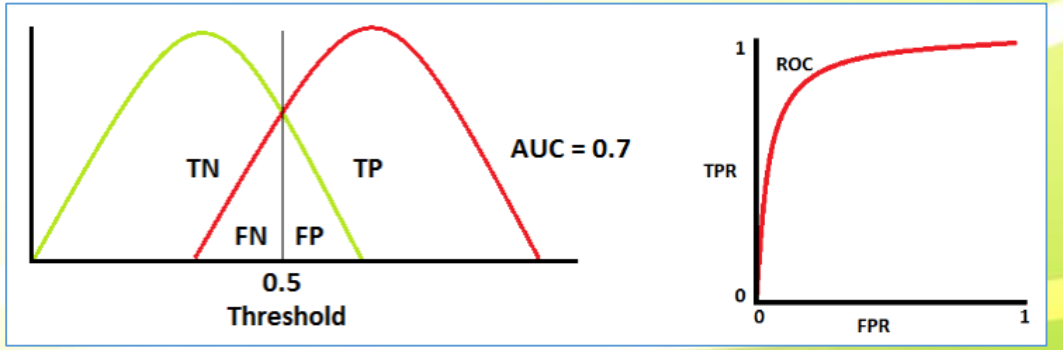


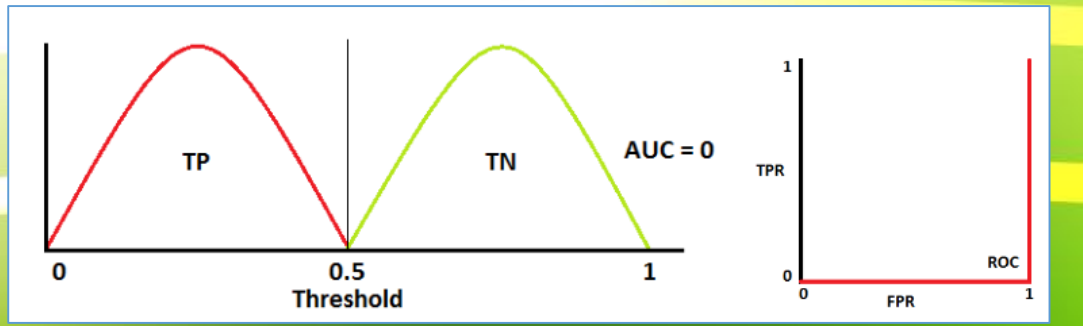
Logistic Regression – Goodness of Fit

AUC & ROC

- Red distribution curve is of positive class (patients with disease) and green distribution curve is of negative class (no disease).
- AUC = 1 When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.
- If two distributions overlap, we introduce type-1 & type-2 error.
- Depending upon threshold, we can minimize or maximize them.
- AUC = 0.7 means there is a 70% chance that the model will be able to distinguish between positive class and negative class.
- AUC = 0 means model is actually reciprocating classes → model is predicting a negative class as positive class and vice versa.







Information Theory

- Information theory revolves around quantifying how much information is present in a signal.
- The intuition behind quantifying information is the idea of measuring how much surprise there is in an event.
  - Learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.
- Those events that are rare (low probability) are more surprising and therefore have more information than those events that are common (high probability).
  - ***Low Probability Event****: High Information (surprising).*
  - ***High Probability Event****: Low Information (unsurprising).*
- Calculate the amount of information there is in an event using probability of the event; called "*Shannon information*," "*self-information*," or "*information*" :

information(x) = - log(p(x))

log() is the base-2 logarithm

p(x) is the probability of event x.

Information Theory

- For any random variable that follows a probability distribution with a probability density function (p.d.f.) or a probability mass function (p.m.f.) p(x), expected amount of information is measured through entropy (or Shannon entropy)
- **Entropy**: measure of uncertainty of occurrence of certain event, given partial information about the system.

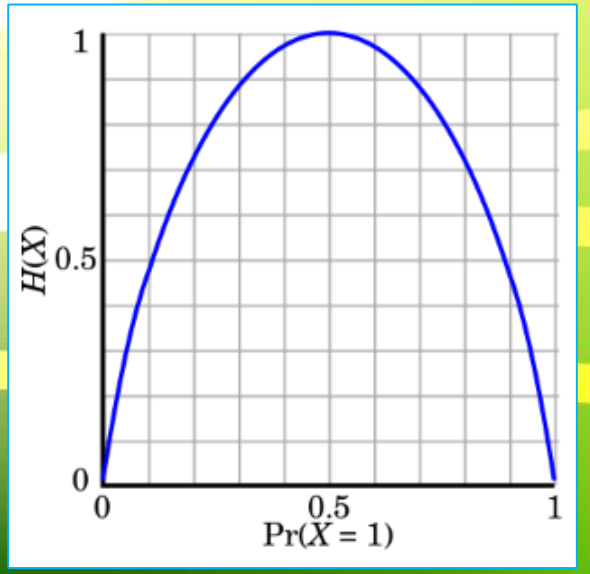
if X is discrete

$$H(X) = -\sum_i p_i \log p_i, \text{ where } p_i = P(X_i).$$

if X is continuous

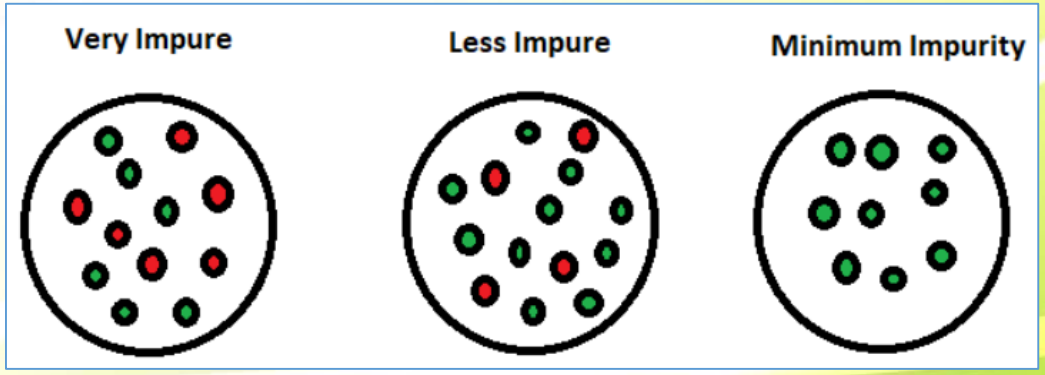
$$H(X) = -\int_x p(x) \log p(x)\, dx.$$

- this function is maximized when the two possible outcomes are equally probable.
- *example, unbiased coin toss*



Information Theory

- "Goodness" of splitting criteria is determined by how clean each split is.
- **Impurity**: proportion of different categories of response variable.
- Cleaner splits result in lower scores.
- As the tree is being generated, it is desirable to decrease level of impurity until ideally there is only one category at a terminal node (a node with no children).
- Three primary methods for calculating impurity:
  - Misclassification,
  - Gini,
  - Entropy.
    - S: set of observations.
    - Pi : fraction of observations that belong to a particular value
    - C : number of different possible values of response variable.



$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Information Theory

Split a

Entropy (N1) = $-(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$
Entropy (N2) = $-(10/10)\log_2(10/10) - (0/10)\log_2(0/10) = 0$
Entropy (N3) = $-(0/10)\log_2(0/10) - (10/10)\log_2(10/10) = 0$

Split b

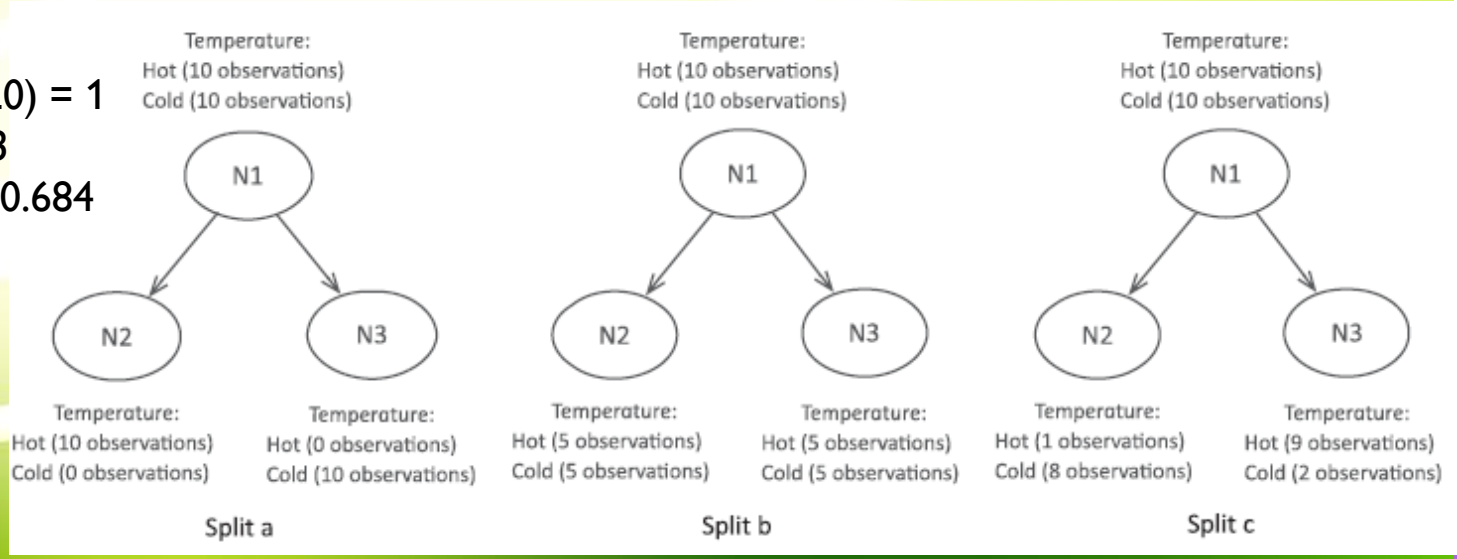
Entropy (N1) = $-(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$
Entropy (N2) = $-(5/10)\log_2(5/10) - (5/10)\log_2(5/10) = 1$
Entropy (N3) = $-(5/10)\log_2(5/10) - (5/10)\log_2(5/10) = 1$

Split c

Entropy (N1) = $-(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$
Entropy (N2) = $-(1/9)\log_2(1/9) - (8/9)\log_2(8/9) = 0.503$
Entropy (N3) = $-(9/11)\log_2(9/11) - (2/11)\log_2(2/11) = 0.684$

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

S: set of observations.
Pi : fraction of observations that belong to a particular value
C : number of different possible values of response variable



Information Theory

Information Gain is a measure of this change in entropy.

$$\text{Gain} = \text{Entropy(parent)} - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$

$$Gain = E_{parent} - E_{children}$$

- *N:* number of observations in parent node,
- *K:* number of possible resulting nodes,
- *N(vj)* : number of observations for each of the *j* child nodes,
- *vj :* set of observations for the j^{th} node.

- Gain formula can be used with other impurity metrics by replacing the entropy calculation.

  *Gain(Split a) = 1 − (((10/20) *0) + ((10/20) *0)) = 1*

  *Gain(Split b) = 1 − (((10/20) *1) + ((10/20) *1)) = 0*

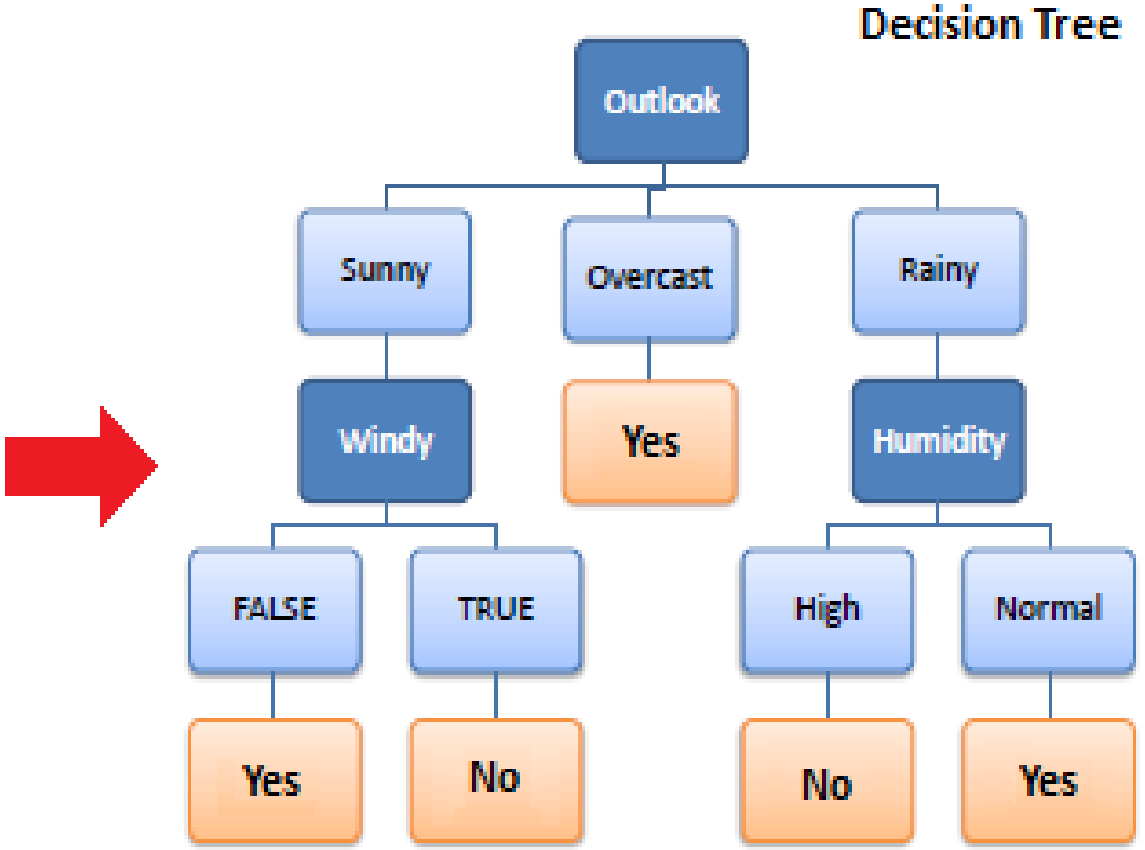
  *Gain(Split c) = 1 − (((9/20) *0.503) + ((11/20)* 0.684)) = 0.397*

- Condition used in **Split a** is selected as best splitting criteria.
- During tree generation process, algorithm examines all possible splitting values for all splitting variables, calculates a gain function, and selects best splitting criterion.

Information Theory

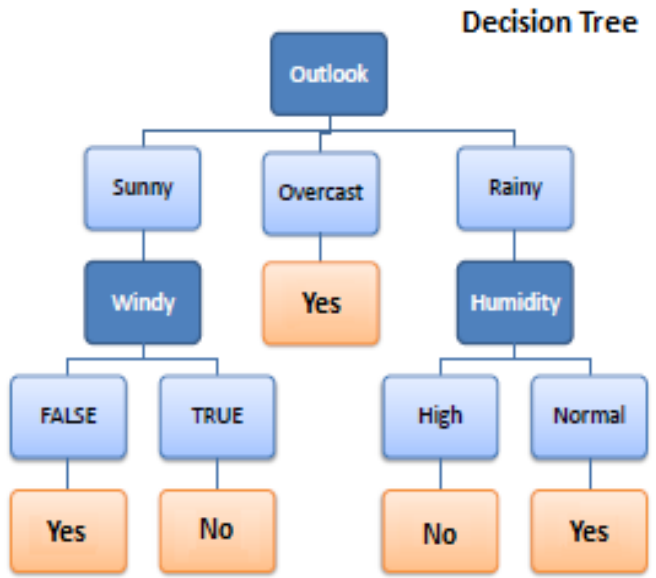
Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Information Theory

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

Information Theory

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

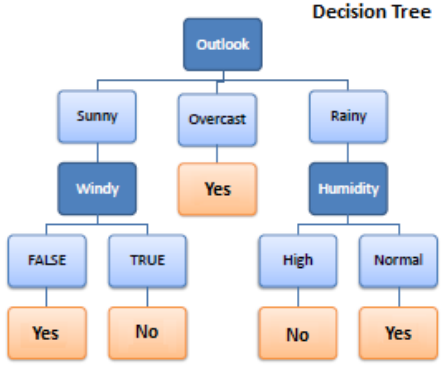
$$Gain = E_{parent} - E_{children}$$

$$\text{Gain} = \text{Entropy(parent)} - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Decision Tree



$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)*__E__(3,2) + **P**(Overcast)*__E__(4,0) + **P**(Rainy)*__E__(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

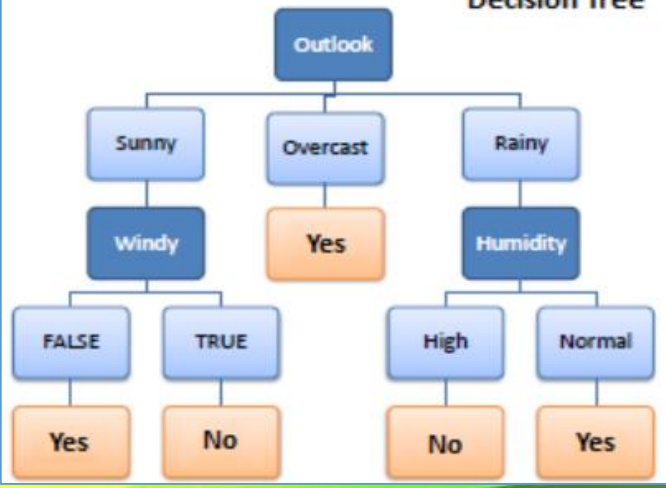
= 0.693

Information Theory

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Decision Tree



		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

Information Theory

- Joint Entropy:

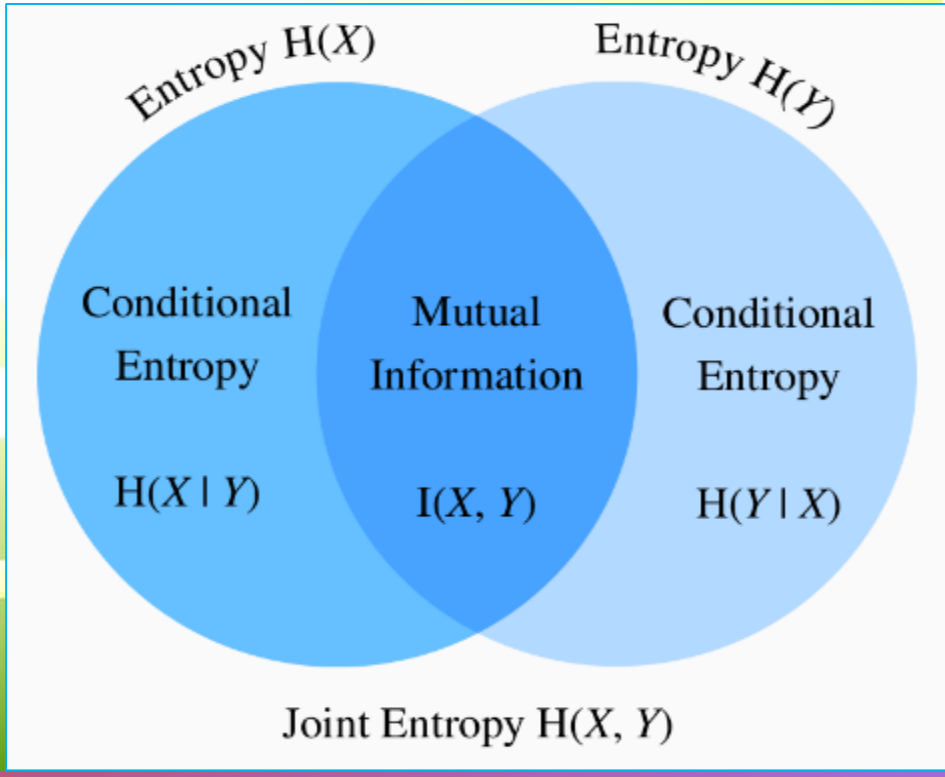
$$H(X,Y) = \mathbb{E}_{X,Y}[-\log p(x,y)] = -\sum_{x,y} p(x,y) \log p(x,y)$$

$$H(X,Y) = -\sum_{x} \sum_{y} p_{X,Y}(x,y) \log p_{X,Y}(x,y)$$

$$H(X,Y) = -\int_{x,y} p_{X,Y}(x,y) \ \log p_{X,Y}(x,y) \ dx \ dy$$

- Joint entropy is equal to sum of individual entropies of X and Y when they are independent.

$$H(X), H(Y) \leq H(X,Y) \leq H(X) + H(Y)$$



Information Theory

- Conditional Entropy:

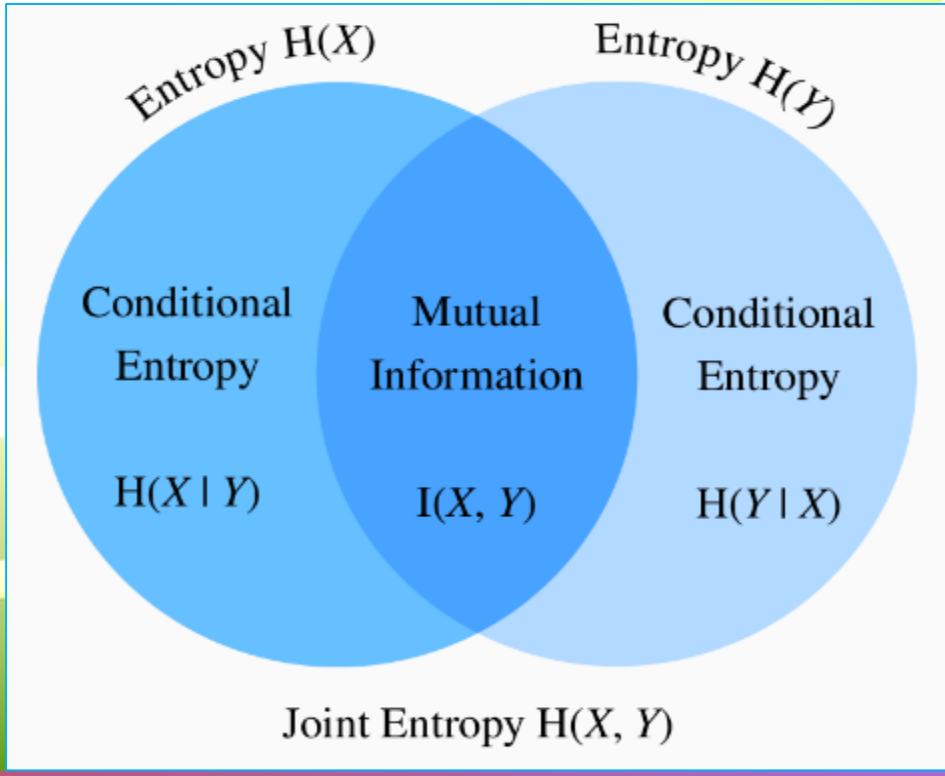
$$H(Y \mid X) = -E_{(x,y)\sim P}[\log p(y \mid x)]$$

$$H(Y \mid X) = -\sum_x \sum_y p(x,y) \log p(y \mid x)$$

$$H(Y \mid X) = -\int_x \int_y p(x,y)\ \log p(y \mid x)\ dx\ dy$$

- Conditional entropy is equal to joint entropy minus self entropy.

$$H(Y \mid X) = H(X,Y) - H(X)$$



Information Theory

- Mutual Information:

$$I(X,Y) = H(X,Y) - H(Y \mid X) - H(X \mid Y)$$

$$H(X) - H(X \mid Y)$$
$$H(Y) - H(Y \mid X)$$

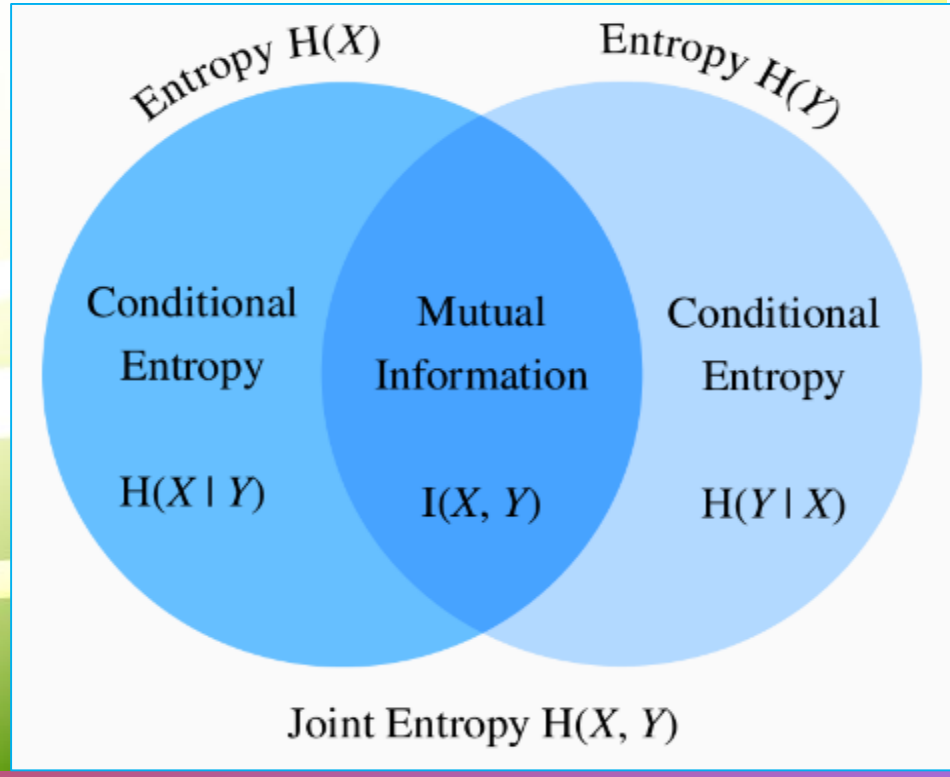
$$I(X,Y) = E_x E_y \left\{ p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right\}$$

- Mutual Information is symmetric, non-negative.

$$I(X,Y) = I(Y,X).$$

$$I(X,Y) \geq 0$$

- Mutual Information = 0, only if X & Y are independent.



Information Theory

- **KL divergence (relative entropy and I-divergence)** is a type of statistical distance: a measure of how one probability distribution P is different from another potential distribution Q.
- KL divergence of P from Q is the expected excess surprise from using Q as a model when the actual/better distribution is P.
- KL divergence = 0 indicates two distributions in question have identical quantities of information.
- Relative entropy is a non-symmetric, nonnegative function of two distributions or measures.

$$D_{KL}(p||q) = E[\log p(x) - \log q(x)]$$

$$D_{\mathrm{KL}}(P \parallel Q) \geq 0$$

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$$

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot log \frac{p(x_i)}{q(x_i)}$$

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Information Theory

Example: Calculate the KL divergence for the following probabilities across two difference distribution.

x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$D_{KL}(p||q) = E[\log p(x) - \log q(x)]$$

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot log \frac{p(x_i)}{q(x_i)}$$

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \\ &= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln\left(\frac{4/25}{1/3}\right) \\ &= \frac{1}{25}\left(32\ln(2) + 55\ln(3) - 50\ln(5)\right) \approx 0.0852996 \end{aligned}$$

Information Theory

- H(P,Q): cross entropy of P and Q.
- H(P): entropy of P (cross-entropy of P with P itself).

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$
$$= \mathrm{H}(P, Q) - \mathrm{H}(P)$$

Cross-entropy is a measure of difference between two probability distributions for a given random variable or set of events.

$$H(p, q) = -\mathrm{E}_p[\log q]$$

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

END !!