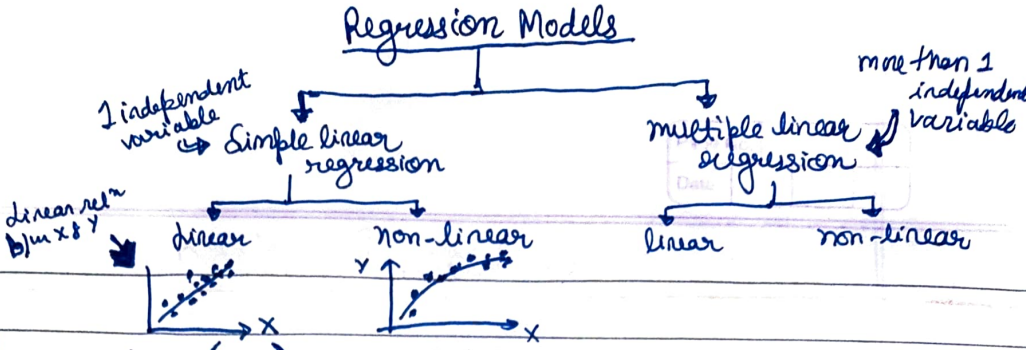
Unit-2

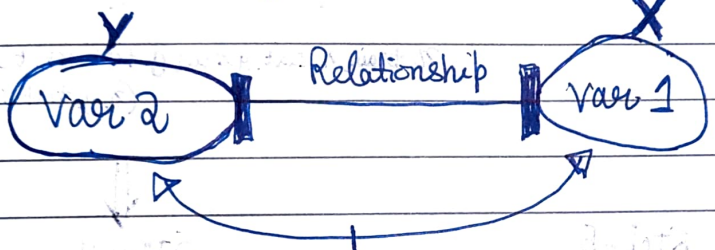


Simple Linear Regression (SLR)

↳ It is a statistical technique to find existence of an association relationship b/w a 'dependent variable' and an 'independent variable'.

⇒ Here, there is only one independent variable in the model.

★ Regression can only tell



does not mean that value of dependent variable (var 2 or Y) dependents on value of independent variable (var 1 or X)

- Def^n ⇒ It is a statistical model in which there is only one independent variable and the functional rel^n b/w the dependent variable and regression coeffient is linear.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Here,

Y_i = dependent variable in Sample / Response variable / outcome variable

X_i = independent variable in Sample / predictor variable / input or explanatory variable

β_0 and β_1 = Regression parameter or Regression coeffient
- β_0 = intercept of regression line (Y-intercept)
- β_1 = slope of the regression line

ε_i = random error (or residuals)

Simple Linear Regression Model Building

Step 1
Collect / Extract data

(on the dependent and independent variable. Time & resource consuming & expensive even with organisation having well-designed resource planning system (ERP) enterprise)

Eg: adut = [. . . . .]
sale = [. . . . .]
OR

df = pd.read_excel('___')
df.head()

Step 2
Pre-Process Data

- Ensure Quality of data
- Check for issues such as reliability, completness, usefulness, accuracy etc.
- All the steps required to give clean data)

Step 3
Dividing data into training and Validation datasets

- Data divided into two sets - training data & Test (validation) data.
- % of training data is usually 70% rest is testing data.
- Training data used for developing model, testing data to check & select the model

x_train, x_test, y_train, y_test = sk.train_test_split(x, y, test_size=0.2, random_state=

Step-4
Perform descriptive analysis of the data like boxplot, scatter plot (to identify any outliner)

- to reveal rel^n b/w the two variable
- useful to determine functional rel^n b/w dependent and independent variable.

Eg: df.boxplot("column name")
df.hist(df["___"])
df['column name'].median / mean etc.
(OR)
mean_value = np.mean(adut)
plt.boxplot(adut) etc.

Step-5
Build the Model

- Model is build using training data.
- OLS (ordinary least square) method used to estimate parameters in linear regression model.
- The goal of OLS is to find the best fit line through a set of data points by minimizing the sum of squares of vertical distances (residual) b/w the observed and value predicted by the model.

[Use SSR (sum of squared residual) = $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ where, $\hat{y}_i$ are predicted values.]

- Using this we find values of β_0 and β_1 and once these regression parameters are set ⇒ for a given x predicted value of y will be $\hat{y} = \beta_0 + \beta_1 X + \varepsilon$. Ie. our model is made ✓

Eg:
model = LinearRegression()
model.fit(x, y)
model.intercept_
model.coef_[0]

Step-6 ⇒ Perform model Dignostic

- Important to run dignostic test to see if the model runs correctly
- Important to ensure model runs consistently and somewhat accurately on the testing data as it did on training data.

Eg: y_predicted = model.predict(x)

Step-7 :- Validate the model and Measure model accuracy / plot the model

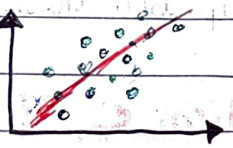
⇒ Ensure model runs accurately on the testing data and perform various test to see if the models fit perfectly.

r_square = r2_score (y, y_predicted)

mse = mean_squared_error (y, y_predicted)

Plot the model :-

```
plt.scatter (x, y, colour 'green')
plt.plot (x, y_predicted, colour='red'
plt.grid
plt.show
```



Step-8 : Decide on model deployment

⇒ Final step involves figuring out how to make statergic desciions like adjusting budgets, setting prices etc. based on the model output. It basically helps in decision making.

Regression Eq^n and its Analysis

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

β_0 → y-intercept, β_1 → slope

$$\hat{Y}_i = 61655.3553 + 3076.1774 X_i$$

(MBA stud salary vs % of marks in their class 10th)

So, if :-

Interpretation ⇒

If a student have 0% in class 10th still his/her salary = ₹ 61,555 approx

(X = 0, then, y = ? is y-intercept)

It means for every 1% increase in class 10th score the salary increases by ₹ 3077.1774 ...

★ Estimation of Parameters (using OLS)

⇒ A crucial step in regression model building is estimation of regression parameters.

⇒ Given an dependent variable (Y_i) and the corresponding independent variable values (X_i) and each subject to random error (ε_i), we have to find the best eqⁿ to represent the relationship b/w these two variables.

Assumption of OLS-based Linear Regression model

① **Linearity** :- Regression model must be linear in its parameters. meaning, β_0 and β_1 must be linear. [variables X_i etc. can or cannot be linear]

② **Variation in independent variable (X_i)** :- There should be sufficient variation in X_i values then only model can be correct Eg. Y = consumption level and X = income level. And the X i.e. income taken of every person is 210,000 ⇒ meaning no variation ⇒ so, model will not able to predict a ~~new~~ $\hat{Y}_i$ based on a new X.

③ **The explanatory variable X, is assumed to be non-stochastic (i.e. X is deterministic).** [⇒ Fixed value of independent variable]

⇒ This means, X is controlled and not subjected to random variation or influences ⇒ which will help predicting Y more straightforwardly.

Eg. X = Study hour
Y = marks in exam

} If we take into account the variation of students feeling like sometime they wanna study so before 1 day they study 18 hrs but before they distract due to their mood ⇒ So it makes determining y difficult. So we make 'X' more controlled and consistent so its easy to determine Y.

④ Residuals (ε_i) follow a normal distribution and have zero mean

⇒ This means, if we plot all residuals from our model, they would form the classic bell shaped curve of normal distribution.

⇒ Mean = 0, means on average the errors do not systematically overestimate or underestimate the dependent variable (Y).

★ Now Based on above 4 assumption :- In OLS, our objective is to find optimal value of β_0 and β_1 such that we minimize the Sum of Squared Errors (SSE) ie.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i})^2$$

on partial derivating it wrt. β_0 and β_1 we get :-

$$\beta_0 = Y_i - \beta_1 X_i \quad , \quad \beta_1 = \frac{cov(X,Y)}{var(X)} \text{ OR } r \times \frac{\sigma_x}{\sigma_y}$$

calculate β_1 and put value here.

(σ_x, σ_y are SD).

★ Imp

⑤ The variance (meaning how spread out the residuals are) of the residuals ⇒ $Var(\varepsilon_i | X_i)$ is constant for all values of X_i.

⇒ When the variance of residual is constant for different values of X_i its k/a ⇒ Homoscedasticity.

⇒ But when the variance of residual is non-constant its k/a ⇒ Heteroscedasticity

(we want Homoscedacity ✓ and we dont want heteroscedacity ✗)

Eg Whether its a cold day (less ice cream sold) or a hot day (more ice cream sold), the amount to which our prediction is wrong should have some spread. If the variance changed with temp ⇒ then Heteroscidacity which will violate our OLS assumption.

[Here, X_i = temp, Y_i = ice cream sold]

⑥ In case of time series data, residuals are uncorrelated that is $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

⇒ In time series data ⇒ data is collected over time ⇒ residual are uncorrelated means ⇒ residual/error at one point of time (eg. error in predicting electricity consumption on monday) doesnot tell anything about the residual (error in ,,,,, on Tuesday). They are independent of each other.

• Validation of Simple Linear Regression Model •

⇒ It's important to validate the regression model to ensure its validity and goodness of fit before it can be used for practical applications.

⇒ The following measures are used :- (+ Their interpretation on calculating them)

★ Coeff. of Determination (R^2) [R-square]

↳ objective of regression is to explain the variation in Y (or predict) based on knowledge of X.

⇒ R^2 ⇒ measures the % tage of variation in Y explained by the model ($\beta_0 + \beta_1 X_i$)

Simple Linear Regression model

⇓

Broken into —

Explained variation + Unexplained variation

$Y = \beta_0 + \beta_1 X_i + \varepsilon_i$

(Y: variation in Y)

($\beta_0 + \beta_1 X_i$: variation in Y explained by the model)

(ε_i: variation in Y not explained by the model)

So, Total variation in Y = Variation in Y explained by model + Variation in Y not explained by model.

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

Applying summation to this squares :-

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SSE}$$

SST (Sum of square of total variation)

SSR (Sum of square of variation explained by regression model)

SSE (~~Sum of squared of variation~~ Sum of square of error or unexplained variation)

Coeff. of Determination = R^2 = Explained variation / Total variation

$$= \frac{SSR}{SST} \quad \underline{OR} \quad 1 - \frac{SSE}{SST}$$

(Since, SSR = SST - SSE)

Interpretation

- R^2 lies b/w 0 and 1
- Higher the value of R^2 (meaning closer to 1) the better fit the model.
- R^2 less than 0.5 (or 50%.) ⇒ not good fit

Eg: $R^2 = 0.1563$ ⇒ means, grade 10 marks (Independent var) explain 15.63% of variation in the starting salary of MBA students (dependent var)

(X = grade 10 marks, Y = Salary)

⇓

(not good fit) → it should had ideally been able to explain 100% variability.

Analysis of variance (ANOVA) (F-static)

same context

★ Significance Test (p-value)

↳ In SLR we are trying to find a rel^n b/w X and Y.

⇒ p-value is used to determine weather X (independent var) has statistically significant relationship with Y (defendent var).

How it works in SLR :-

① H_0 (null Hypothesis) :- In SLR, null hypothesis states that there is "no" rel^n b/w X and Y. ⇒ means, $\beta_1 = 0$ (β_1 = coeff. of X).

② H_A (alternative Hypothesis) ⇒ There is rel^n b/w two ⇒ $\beta_1 \neq 0$.

③ Significance level (α) ⇒ we choose a significance level ($\alpha = 0.05$ mostly) before running regression analysis.

④ p-value ⇒ Fit SLR model into data, we get a p-value for β_1 (coeff of X)

⑤ Comparison ⇒

If, p-value $< \alpha$ ⇒ Reject H_0 and hence there is rel^n b/w two, Accept H_A

If, p-value $> \alpha$ ⇒ Accept H_0, there is no rel^n b/w two. Reject H_A

Eg. data set of House prices (Y) [trying to predict] based on size of house (X). ⇓

⇒ perform SLR and find p-value of the ~~coeff~~ β_1 (coeff of X_1), suppose it comes to be = 0.03.

⇒ our significance level = 0.05

Since, p-value $<$ 0.05 (α) ⇒ we reject H_0 and say there is rel^n b/w house price and size of house.

(p-value less than)

p-value for either test → t-test, f-test } same in simple linear regression.

so in other words,

p-value for t-test or f-test $< 0.05 \Rightarrow$ means rel^n b/w X and Y and model is statistically significant.

Interpretation

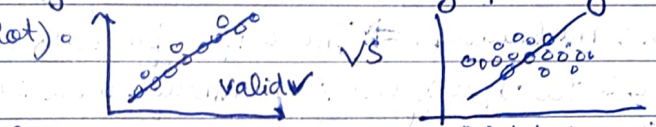
If, p-value less than 0.05 (α = level of significance) we reject H_0 meaning there is relationship b/w X (independent var) and Y (dependent var).

★ Residual Analysis

⇒ It is important to check if the assumptions of the SLR is being satisfied or not. It consist of checking the following four points:-

① The residual $(Y_i - \hat{Y}_i)$ is normally distributed or not.

↳ Easiest way to determine this is by plotting P-P plot (Probability-Probability plot).



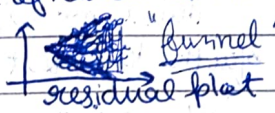
P-P plot of a regression model of eg^n $Y_i = \beta_0 + \beta_1 X_1$

P-P plot of regression model of eg^n $\Rightarrow Y_2 = \beta_2' + \beta_3' X_2$

dots close to diagonal ✓ ⇒ so residual follow normal distribution. And model is valid.

② Test of Homoscedasticity

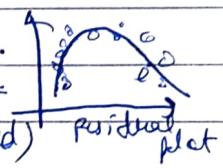
If there is heteroscedasticity (meaning variance of residual non-constant) there is funnel shape in residual plot.



③ Testing functional form of Regression model.

⇒ Residual plot shouldn't have any pattern (a parabola).

if it is there then ⟶ model is not valid. (incorrect functional form used) ⟸



④ Outlier Analysis

↳ outlier are those values whose value show a large deviation from mean value. Presence of outlier can have influence on regression coeff^n → changing the best fit line

★ The following distance measures are used to identify influencial observation/our outliners:-

① **Z-score** ⇒ It is the distance of observation from its mean value. $Z = \frac{\hat{Y}_i - \bar{Y}}{\sigma_Y}$ @ σ_Y and $\bar{Y}$ are SD and mean of dependent variable.

★ **Interpretation** ★

⇒ Any observation with Z-score greater than 3 will be classified as outliner / influencial obs. whose presence can alter regression coeffients.

② **Mahalanobis distance** ⇒ Its value greater than Chi-Square critical value for an obs. mean that that obs is outliner.

③ **Cooks distance** :- Cooks distance value more than 1 indicates highly influencial obs. / outliner.

Extra Point in R^2

⇒ Not all times, high R^2 value mean, that the model is better fit. It can sometimes be misleading. ⇒ This is k/a Spurious Regression :- meaning it shows there is rel^n b/w X and Y eventhough there is nt.

Eg:

	No. of facebook users (indep)	No. of people left Finland. (dep)
2004	1 lakh	1.2 lakh
2005	2 lakh	2.9 lakh
		etc.

⇒ on calculating $R^2 = 0.8$ it means they both are related / model is correct. our model eq^n is $Y = 1.996 + 2.1X$ ← means for every 1 person leaving finland 2.1 users ↑ss.

There is no relation b/w the two, but R^2 is high ⇒ hence mislead us so, R^2 is not always a correct measure to tell better fit.

Multiple Linear Regression

↳ It is statistical technique used to find existence of relationship b/w a dependent variable (Y) and several independent variable.

⇒ Functional form of MLR is given by :-

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \cdots \beta_n X_n + \varepsilon$$

here, $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ etc. are k/a 'Partial regression coeff'.

β_0 = y-intercept or constant term

ε = error term

★ Assumptions of MLR

↳ same as of that of SLR actually !!

MLR eq^n and its Analysis :-

⇒ ~~Consider, two independent variable :- promotion and rating point~~
~~dependent variable :- advt. revenue~~

⇒ Consider, dependent variable ⇒ House price
independent variable ⇒ size and Age of house

lets say,

$$Y = ₹\ 41,258 + 5931.85 \times \text{Size} + -1000 \times \text{Age of house}$$

⇓

- for every one unit ↑se in size ⇒ price of house ↑ses by ₹5931.85 (keeping age const.)
- for every one unit ↑se in age of House ⇒ price of house ↓se by ₹1000 (keeping size const.)

⇒ In MLR, it may not be possible to control a variable (i.e. keeping it constant) in many situations.

Working with categorical variables

↳ Those which represent characteristic like person gender, marital exam status. It can take numerical value. Eg. 1 for male, 2 for female etc.

↳ Eg. Y = Salary, X_1 = High school, X_2 = Bachelor, X_3 = Master, X_4 = Phd.

We take X_1 as refernce and calculate others in reference to it.

Multi-collinearity and VIF

↳ ⇒ It refers to presence of high correlation b/w two or more independent variables in a multiple linear regression model.

⇒ It can cause issue with the estimation of coeffs and effect the model.

⇒ It refers to ~~bad~~ presence of perfect linear relationship b/w few or all independent variables in a regression model.

Eg. There are n - independent variable then

$\alpha_1 X_1 + \alpha_2 X_2 \cdots\cdots \alpha_n X_n = 0$ (where α_n are constant)

↳ Perfect rel^n with each other.

⇒ One of the Assumptions of CLRM (classical linear regression model) is that there should be no multicollinearity among the independent variable.

⇒ As the no. of independent variable ↑ ⇒ problem of multicollinearity can also ↑.

⇒ Therefore we need to detect such multicollinearity in data and one such way is by calculating VIF.

◉ Variance Inflation Factor ◉

↳ ⇒ It is used to measure identifying existence of multi-collinearity

For Eg. consider a regression model,

$Y = B_0 + B_1 X_1 + B_2 X_2$ and thier R^2 value be R_{12} then,

$$VIF_{value} = \frac{1}{1 - R_{12}^2}$$ Formula

⇒ VIF greater than 5 requires investigation.

Interpretation ★

VIF = 1 ⇒ no correlation b/w independent variables
1 < VIF < 5 ⇒ moderate multicollinearity
VIF > 5 ⇒ High degree multicollinearity, problematic

⇒ To fix high multicollinearity we can do :-
1. Increase Sample Size.
2. Drop non-essential independent variables.

Outliers Analysis ⇒ Discussed in SLR.

★ Autocorrelation

↳ also K/a serial correlation. It happens when there is a rel" b/w current obs. and one or more past obs. It occurs when the residual of a regression model exhibit a pattern.

Eg. Measuring temp. every day ⇒ If since past 2-3 days temp is high ⇒ then this has chances that temp. tomorrow will also be high. In other words, autocorrelation occurs when past values in a data series influence future values.

⇒ Durbin-Watson test is a statistical test used to detect the presence of auto correlation. It is based on estimated residuals.

★ Interpretation : Its Range 0 to 4

- A value of 2.0 ⇒ no auto correlation
- Value less than 2.0 ⇒ +ve auto correlation (High stock yesterday → High stock today / or High temp yesterday → High temp tommorow)
- Value greater than 2.0 ⇒ -ve " " (rare, Eg. value = 3.5 means sales this quarter can go down if for past quarter were up. Sale High (Past) → Low now.

Validation of MLR model

⇒ To validate a MLR model we can follow the given measures and perform below test to check it.

① Coeff. of multiple Determination (R-square) and adjusted R-Square

$R^2 = 1 - \frac{SSE}{SST}$ (R^2 = portion of variance in Y explained by the model)

It explain how well your model explain variation in the dependent variable.

⇒ we use adjusted R^2 more bcoz in MLR, normal R^2 can be manufilated as it can be increased artificially by adding more variables, but adjusted R^2 takes care of that aspect.
⇒ Adjusted R^2 considered more reliable than R^2.

★ Interpretation

◉ R^2 ⇒ R^2 range 0 to 1. R-square = 0 ⇒ means model doesn't explain any variability in

⇒ more its near 1, the better fit the model with data is.

Eg: R-Squred = 70% or 0.70 means 70% of variation in dependent variable (Y) is explained by the independents variable.

◉ Adjusted R-square ⇒ ◉ Also, b/w 0 to 1
◉ less than or equal to R-square
◉ More near to 1 the better fit the model is.

② T-test, F-test ↓

H_0 = no rel^n b/w X & Y
H_A = rel^n b/w X & Y

Calculate p-value ⇒ If, p value $\overset{(\alpha)}{<} 0.05$ ⇒ H_0 reject, H_A accept ; rel^n b/w X & Y ✓

p value > 0.05 ⇒ opposite of above.

③ Check for multicollinearity, if present take app. steps.

④ And, check for autocorrelation (in case of timeseries data)

⑤ Residual Analysis (same as that of SLR ⇒ check if Homoscedacity is followed or not).

Unit-3

★ Logistic and Multinomial Regression

★ Logistic Regression

↳ It is used when the dependent variable has two possible outcomes ie. binary (Y)

⇒ Consider it answering a Yes or No question based on some info.

Eg. A doctor says a person has a disease (Yes or No) based on their symptoms (like cough and cold). simplified

⇒ Logistic Regression will help in calculating probability the patient has disease based on system. It outputs value 0 and 1, which can be interpreted as probabilities.

★ Multinomial Regression

↳ It is extention of Logistic Regression, but it is used when dependent variable has more than ~~one~~ two categories.

Eg: In above case based on symptoms (like cold, cough, fever) a newely appointed doctor said that the patient ⇒ Yes (Have this normal viral fever), No (he doesn't have its simply due to weather), Maybe.

⇒ So, your dependent variable is no longer binary and has more categories.

Eg Based on customer age, region, income level ⇒ The ice cream company has three flavours (= Y (dependent variable) :- Chocolate, Vanilla, Strawberry.....

⇓

Instead of binary outcome, the outcome/dependent variable has multiple categories.

★ Logistic Fnx (Sigmoid Fnx)

↳ The logistic regression model (or binary regression model) is given by; this sigmoid fnx :-

$$f(z) = \frac{1}{1+e^{-z}}$$

$$P(Y=1) = \frac{e^z}{1+e^z}$$

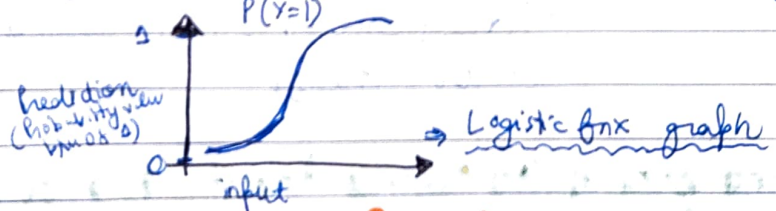
where,

$P(Y=1)$ - probability of obs labelled as 1

$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$

↳ X_1, X_2, X_3 are independent var.

⇒ The logistic fnx is a S-shaped curve (as it is a probability fnx).



Estimation of Parameters in Logistic Regression

↳ One of the major assumption of MLR was, residuals follow normal distribution, but here, in logistic regression residuals don't follow normal distribution, so it is difficult to follow a normal distribution ⇒ Thus we cannot use OLS to calculate parameters here. They are calculated using maximum likelihood estimates (**MLE**).

↳ using which β_0 & β_1 are calculated.

Interpretation of Parameters of Logistic Regr.

⇒ An easy interpretation of the regression coeff, β_1 is that :-

① +ve coeff (β_1 = +ve) ⇒ means probability of event occuring ($P(Y=1)$) also ↑es

② -ve coeff (β_1 = -ve) ⇒ means, probability of event occuring also ↓se.

Eg. Predict student pass ($Y=1$) or fail ($Y=0$) based on no. of hours they study (X)

⇒ If, $\beta_1 = 0.8$ ⇒ means, each additional hour of him/her studying will ↑se the probability of them passing.

$\beta_1 = -0.5$ ⇒ means, each additional hour of them studying will ↓se probability of them passing. (counter-intuitive consider X = stress).

Estimating probability using logistic Regression

⇒ It involves fitting data into the model, and then using this model to predict the probability of an event occuring.

Example :- Estimating Probability of getting a Loan Approved.

↳ Consider you are working for a bank and want to predict whether a person gets loan approved ✓ or not ✗ based on their age & income.

Step-1 :- Collect Data

① Gather past loan applicants data including their age, income & whether they got their loan approved or not.

Step-2 :- Fit into Regression model.

- Use collected data to fit a regression model. The regression model would have an eqn :-

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \times \text{Income} + \beta_2 \times \text{Credit score} \quad \text{—— (1)}$$

⇒ Here, $P(Y=1)$:- probability of loan approval, b_0 is intercept, b_1 and b_2 are coeff of income and credit score.

Step-3 :- Use model for prediction

⇒ once model is fitted, we can predict whether applicant loan will be approved or not.

Eg. If age = 25 and income = ₹ 50,000 (new data)

we put into the above eqn (1) to get log odds

↓

the convert that log odd to get the probability using logistic fnx :-

$$P(Y=1) = \frac{e^{\text{log-odds}}}{1 + e^{\text{log-odds}}}$$

Step-4 :- Interpret the results

↳ The result will be b/w 0 to 1 Eg. If its comes 0.7 or 70%. Meaning there is 70% chance that his/her loan will be approved.

Logistic Model Dignostic

↳ We perform dignostic test before a binary / logistic regression model is accepted for deployment.

4 test :-

- Omnibus Test
- Wald Test
- Hosmer demshow Test
- Pseudo R-Square

Omnibus :- ◉ To check overall significance of a logistic regression
◉ Checks weather models with predictor (ie. X or independent var) is significantly better at explaing the variance than a model with no predictors (null model)

Formula ⇒ omnibus = $\frac{\text{deviance reduced} - \text{deviance full}}{\text{deviance reduced}}$ > we obtain these value by SPSS (Statistical soft.)

Interpretation :-

⇒ After performing this test → then doing hypothesis test → If we get p-value < 0.05 ⇒ then it means, our predictors (X_1, X_2 etc.) were able to explain variance and have a significant impact on Y.

Eg. model :- predict admission ✓ or X based on GPA (X_1) and test-score (X_2). This test will show whether they both together significantly predict admission chances than a model that predicts admission without these X_1 & X_2 or predictors.

Wald Test :- ◉ Check significance of individual coeffs in the model (Eg. checking significance of β_1, β_2 etc.)

Formula ⇒ wald test = $W = \frac{(\text{Estimated coef})^2}{(\text{standard error of coef})^2}$

Interpretation :-

↳ If the p-value calculated after Wald test < 0.05 then, the wald test statistic for a coeff is a significant contributer to model.

Eg. ↳ p-value for β_1 of X_1 = 0.205 > 0.05 ⇒ this means X_1 may not have a significant impact on Y.

Hosmer Lemeshow Test ◉ It is a goodness-of-fit test for logistic regression model.
◉ It evaluate the models performance in predicting outcome based on observed or predicted probabilities.

Interpretation :-

↳ If p-value high ⇒ model a good fit
p-value low ⇒ mode not a good fit

Pseudo-R-square :- ⊙ Similar to what R^2 measure in LSR, but not directly comparable to it hence N/a.
('Pseudo R-Square' [since method use OLS here in Logistic]

⊙ It provides indication of how well predictor variable (X) explain the variance in dependent variable (Y) in logistic regression.

Interpretation → There are 3 types of pseudo R-square → McFadden, Cox and Snell, Nagelkerke ⇒ Range 0 to 1 closer to 1 the better

↳ In general,
Higher values indicate a better fit.

Eg. Cox and Snell R-square = 0.596 ⇒ 59% ⇒ meaning that around 59% of variance in the dependent/outcome variable (Y) is explained by the model which is moderate here.

Model Performance ⇒ The followings things help in determing the models performance.

★ Classification Table (or confusion matrix)

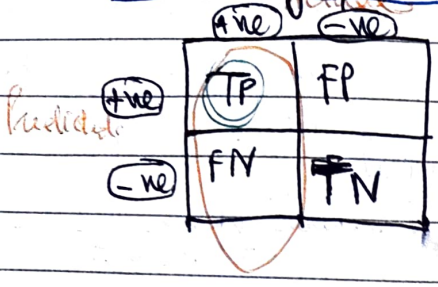
⇓

- helps to determine the performance of a model.

Predicted (हम अनुमान लगा रहे हैं)	Actual (Reality मे है) +ve	Actual (Reality मे है) −ve
+ve	TP	FP (Type I error)
−ve	FN (Type II error)	TN

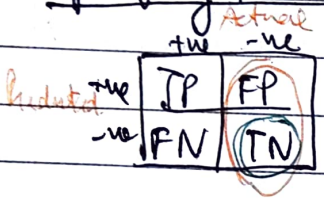
(True = correctly
False = incorrectly)

① Sensitivity / True Positive Rate / Recall



$$\text{Sensitivity} = \frac{\text{Predict bhi sahi kiya jo actual mai +ve}}{\text{actual mai +ve}} = \frac{TP}{TP + FN}$$

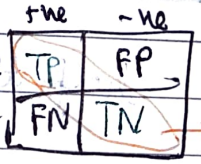
② Specificity / True Negative Rate



$$\text{Specificity} = \frac{\text{jin unko predict bhi -ve kiya hai actual mai -ve hai}}{\text{actual mai -ve hai}} = \frac{TN}{FP + TN}$$

③ Accuracy :- Accurately diagnosed to the total no. of patients

$$\text{Acuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$



④ Precision :- ~~correctly predicted~~

~~Correct~~ Precision $= \frac{TP}{TP + FP}$ → actual maibhi +ve / predicted +ve

⑤ F-score (or F1 score) :- It is the harmonic mean of Recall and Precision,

$$\text{F-score} = \frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \Rightarrow \text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Q.

Prediction	Observed: Correct	Observed: Incorrect
Correct	17 (TP)	3 (FP)
Incorrect	0 (FN)	4 (TN)

① Precision $= \frac{17}{20}$ ② Recall $= \frac{TP}{TP+FN} = \frac{17}{17+0} = 1$ ③ F score $= \frac{2 \times \frac{17}{20} \times 1}{\frac{17}{20} + 1}$

④ Specificity $= \frac{4}{4+3} = \frac{4}{7}$

Gini Coeff / Gini Index

- It has a value b/w 0 and 1.
- Performance of the model improves with ↑ing Gini coeff. (more near to 1 ⇒ better the model)
- Gini coef can be calculated from AUC of ROC.

Gini coefficient $= 2 \times AUC - 1$

ROC (Receiver operating characteristic)

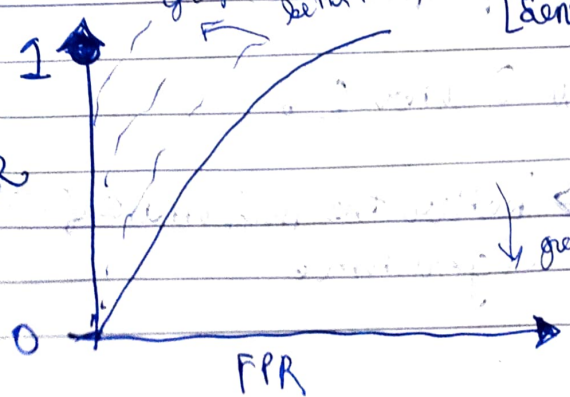
- The performance of logistic regression model can be assessed graphically by :- ROC ⇒ It is the graph b/w TPR and FPR

graph more near to 1 ⇒ better the model

(True +ve Rate) ⇓ [Sensitivity]

(False +ve rate) → [1 − True -ve Rate (i.e. specificity)]

TP	FP
FN	TN



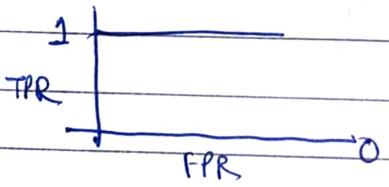
↓ graph more near to FPR, worst the model.

⇒ Basically we plot TPR and FPR at various threshold points then ⇒ ROC curve is formed.

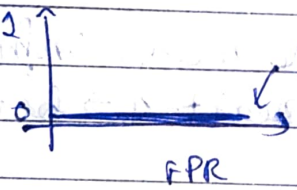
$$TPR = \frac{TP}{TP + FN} \quad , \quad FPR = \frac{FP}{FP + TN}$$

⇓ K/a

ⓔ with a TPR = 1 and FPR = 0, A perfect classifier so, ROC curve would look like :-



ⓔ With TPR = 0 and FPR = 1, worst model, ROC curve look like :-



★ Area Under Curve (AUC)

↳ When making a ROC ⇒ AUC is used to access its effectiveness.

⇒ Calculating AUC means calculating the area under ROC curve.

⇒ no formula ⇒ done using software.

⇒ AUC tells (more the AUC ⇒ tells that model is able to ~~explain all the~~ give better performance.

⇒ AUC values lie b/w 0 and 1.

⇓

More the AUC ⇒ better the performance of the algorithm.