Regression Analysis

Regression analysis is a statistical method used to examine the relationship between one dependent variable and one more independent variables. The goal of regression analysis is to understand how the dependent variable changes when one or more independent variables are altered, and to create a model that can predict the value of the dependent variable based on the values of the independent variables.

1. **Define the research question**: Identify the dependent variable (the variable you want to predict or explain) and the independent variable(s) (the variables that you think influence the dependent variable).

2. **Collect and prepare data**: Gather data for the dependent and independent variables. The data should be organized in a tabular format, with each row representative an observation and each column representing a variable. It's essential to clean and pre-process the data to handle missing value, outliers, and other potential issue that may affect the analysis.

3. **Visualize the data**: Before fitting a linear regression model, it's helpful to create scatter plots to visualize the relationship between the dependent variable and each independent variable. This can help you identify trends, outliers and any potential issue with the data.

4. **Check assumption**:- Linear regression model has some underlying assumption, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. You can use diagnostic plots and statistical test to check whether these assumption hold your data.

Regression Analysis

5. Fit the Linear Regression model.

6. **Interpet the model**: Analyse the estimated regression coefficient their standard errors, t-value, and p-value to determine the statistical significance of the relationship between the dependent and independent variables. The R squared and adjusted R-squared value can provide insight into the goodness-of-fit of the model and the propotion of variation in the dependent variable explained by the independence variables.

7. **Validating the model :-** If you have a sufficient large dataset, you can split it into a training and testing set. Fit the linear regression model to the training set, and then use the model to predict the dependent variable in the testing set. Calculate the mean squared error, root mean squared error, or another performance metric to access the predictive accuracy of the model.

8. **Report results :** Summarize the findings of the linear regression analysis in a clear and concise manner, including the estimated coefficient their interpretation, and any limitations or assumption that may impact the results.

Why ML problems are a Statistical Inference Problems? [with Example]

India → avg salary
(impossible)

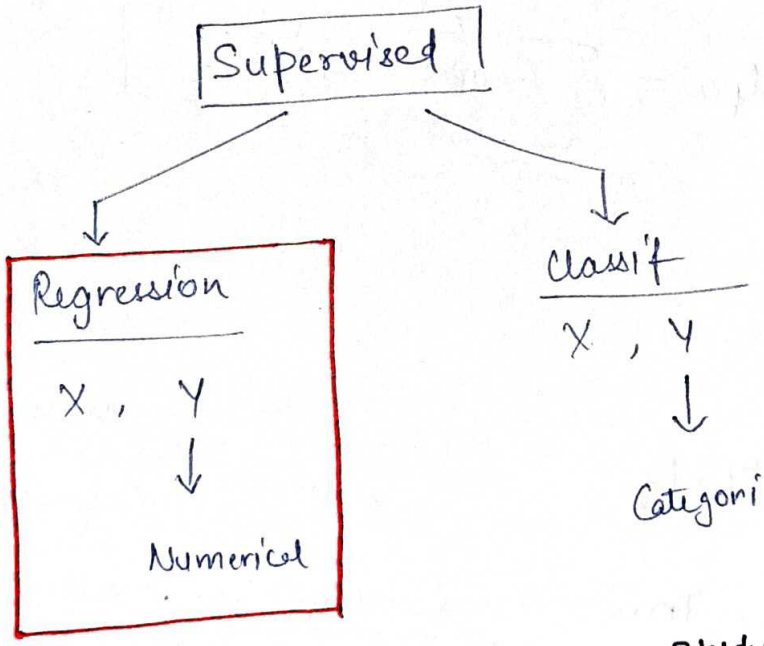
Sample → 1000 ← Statistical problem

Pop

$\bar{x}$ → M → pop parameter

$\bar{x}$: Sample mean

estimate → Statistical Inference



100 data

cgpa	iq	lpa
–	–	–
–	–	–
–	–	–

input | output

machine learing problem

* Study 100 data of Students and predict for all the student and understant the mathematical relation

$lpa = f(cgpa, iq)$

$$lpa = f(cgpa, iq) + \varepsilon$$

ε → error → irreducible error

→ Why add error?

lpa is not directly depend on cgpa and iq. Sometimes Other factors are also involved like (cgpa and iq is good but student not feeling well during interview so he got low lpa). That's Why add error.

* If we predict ~~exactly~~ perfectly ~~same~~ ~~true~~ lpa but error add in prediction and become not perfectly.

True relationship

$$lpa = f(cgpa, iq) + \varepsilon$$

Parameter

$$lpa = \beta_0 + \beta_1 cgpa + \beta_2 iq$$

non-parameter

decision Tree

* Let assume, True relationship is really ~~True~~ linear and relationship is $lpa = \beta_0 + \beta_1 cgpa + \beta_2 iq$

apply ~~rotato~~ Linear Regression and get coefficient is b_0 b_1 b_2.

which is not exactly equal to population coefficient

b_0 b_1 b_2 $\neq$ β_0 β_1 β_2

because if dataset change, then coefficient is also change. Parameter doesn't match with current dataset's parameter.

Try to close the perfect output and we call

$f'(cgpa, iq) \longrightarrow b_0 \quad b_1 \quad b_2$ — b_{c2} we have sample not whole dataset

↳ f dash

True coefficient → $\beta_0 \quad \beta_1 \quad \beta_2$

$f() - f'()$ → reducible error

$lpa = f'(cgpa, iq) + \text{reducible} + \underset{\text{irreducible}}{\varepsilon}$

estimate of x and y based on given data

$f'() \simeq f()$

↑

True of x, y for pop

Regression - Analytical.ipynb

Inference Vs Prediction

[why regression Analysis is required]

cgpa	iq	lpa
–	–	–
–	–	–
–	–	–
–	–	–

linear

β_0 β_1 β_2

$$lpa = \beta_0 + \beta_1 cgpa + \beta_2 iq$$

inference (relationship study)

lpa → cgpa, iq

Y X

relationship?

Individual

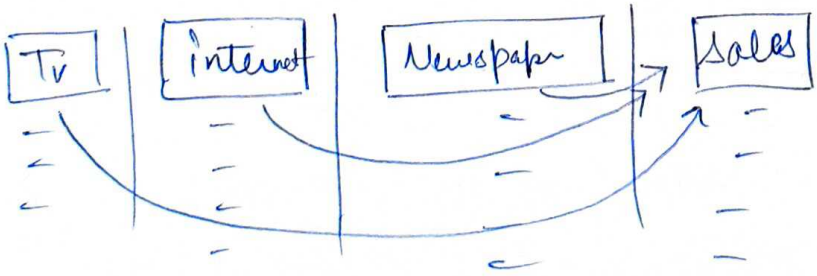
lpa → cgpa

lpa → iq

relationship?

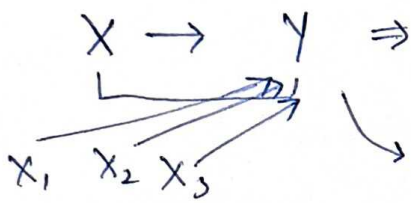
example:-

ads → sales



* Not predicting just find relationship individually so, all these relationship helps for future investment in ads.

Statsmodel Linear Regression



Hypothesis test → F- test for overall significance (ANOVA)

this test help to find relationship betn X and Y.

Tss, Rss and Ess

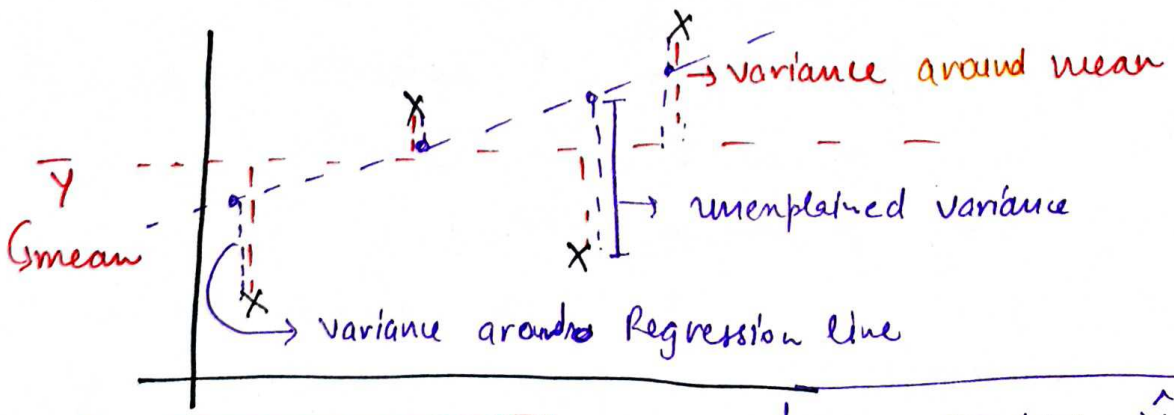
Tss → Total sum of squared

Rss → Residual sum of square

Ess → Explained sum of square

example :

experience	salary
X	Y



$$Tss = \sum (Y_i - \bar{Y})^2$$

$$Rss = \sum (Y_i - \hat{Y}_i)^2$$

predict

TSS - RSS = Explained Variance