Linear Regression in Python – SKlearn and Statsmodels

Pre-Requisites

1. High school Maths - Line of Best Fit
2. Derivatives - Gradient Descent
3. High School Algebra - Matrices and Vectors
4. Numpy and Matplotlib experience/knowledge
5. Some Probability (Gaussian / Normal Distribution)
6. What is Mean, Variance, Standard Deviation etc.

Content

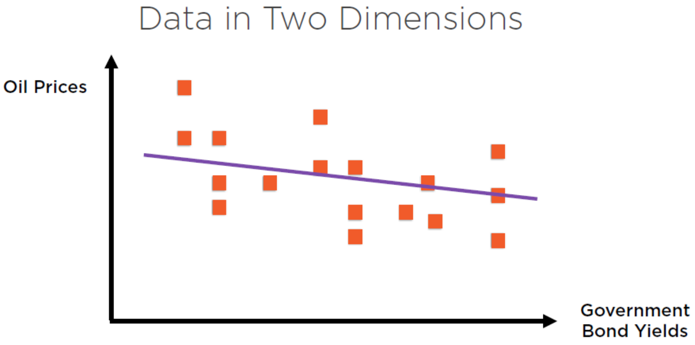
1. Machine Learning - Role of Linear Regression
2. Visualization of regression problem and it's situation
3. Extending the problem by adding more input variables - Multiple Linear Regression

Machine Learning – Role of Linear Regression

1. Types of Machine Learning (ppt)

Role of Linear Regression

1. (Simple) Linear Regression also known as "line of best fit"
2. Supervised (X → Y)
3. Regression output (Y) is a number



A straight line represents a linear relationship

Real world Examples

1. X = Government Bond Yields → Y = Oil prices
2. Credit Risk Scoring
    - (X = Customer Payment, Demographic, Financial Standing data vs Y = Credit Risk Score)
3. X = # of hours of studying → Y = grade

4. X = # of hours of exercise in a week → Y = Body Mass Index

X = Independent or Explanatory Variables (Cause)

Y = Dependent Variables (Effect)

→ = Direction of causality

It's important we determine the direction of causality. There can be some other variables that cause the effect.

QUIZ

1) What can linear regression be used for?

    a) Predicting a person’s height, given the amounts of certain substances in their diet

    b) Predicting whether or not a person likes cake, given the amounts of certain substances in their diet

Correct Answer:

Predicting a person’s height, given the amounts of certain substances in their diet

2) Are you guys ready to do some regression analysis?

    a) Yes

    b) No

Some Guidelines

1. Please ask questions. I'll try to answer most of these during the session, but if I can't I'll definitely make it a point to come back with one.

2. If you don't meet pre-requisites then please study these after the classes. You can open github issues in the MachineLearningSL repo.

   https://help.github.com/en/articles/creating-an-issue

3. Practice, Practice, Practice…

Linear Regression Model

Linear regression is a linear approximation of a causal relationship between two or more variables.

Regression models are highly valuable of establishing relationship.

Process

1) Get sample data
2) Design a model that works for that sample
3) Make predictions for the whole population

Dependent Variable (Y) → That are being predicted (Y)

Independent or Explanatory Variables (X) → That are predictors (x1, x2, x3,…, xn)

Y = F(x1, x2, x3,…, xn)

The dependent variable Y is the function of the independent variables x1 to xn.

Easiest model is Simple Linear Regression Model.

Y = B0 + B1.x1 + c

- Causality ------ Years of Education (x) → Income (Y)
  - Income = f(Years of Education (x))
- **What if it was reverse?**
  - **It would be faulty right?**

- B1 = Coefficient that quantifies the effect of education on income, let's say B0 = 99
  - Y = B0 + 99.x1 + c
  - So for every additional year of your education, the income is going to jump around 99 times
- B0 = Another coefficient, but constant one, let's say B0 = 1000
  - So it's like minimum guaranteed income, even if you did not have any education, you'll still be paid $1000 from let's say government.
  - Y = 1000 + 99.x1 + c
- c = Error coefficient
  - To correct the error that's generated from equation

- Since it’s a mathematical equation, there can be an error in the result so we need to adjust the output with an error coefficient

Quiz

- You’ve an ice cream shop. You notice relationship between the number of cones you order and the number of ice creams you sell. Is this a suitable situation for regression analysis?
  - ○ Yes
  - ○ No

Correct Answer: No

While it’s true that if you run out of cones you can’t sell more ice creams, this is not a regression analysis problem because there is a 1:1 relationship between ice cream and a cone (assuming you’re selling only cone ice creams ☺️).

ICECREAM # = # of CONES

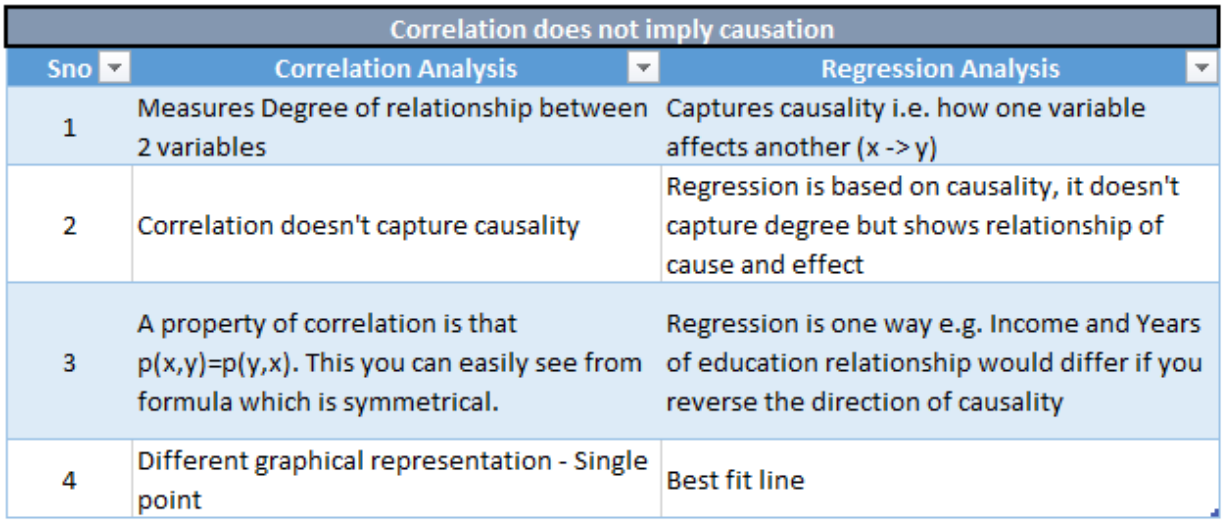
- You’re trying to predict the amount of beer consumed in India, depending on state, is this a problem that can be solved by regression?
  - ○ Yes
  - ○ No

# of Beer cans = a.(Weather) + b.(Finance) + c.(Demographic) + Dry

Correct Answer: Yes

Logic shows that in different states people drink different amounts of beer, Some states are warmer vs colder, While many more things will be part of this regression like Gender, income etc.

Correlation versus Regression



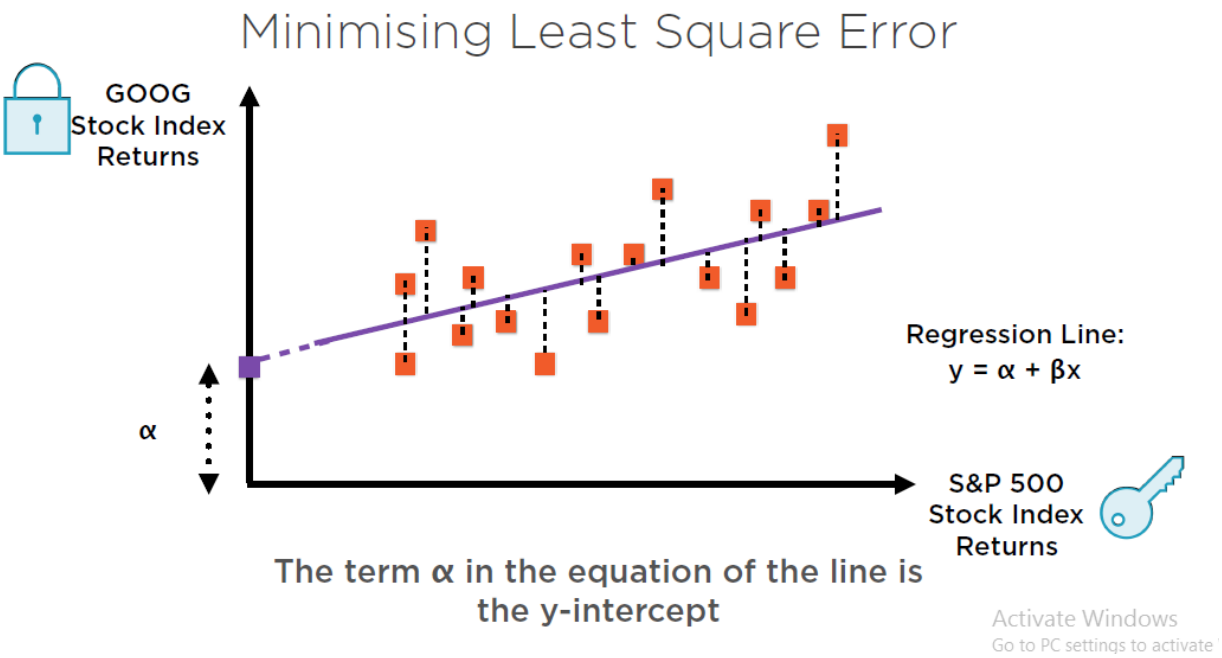
Correlation does not imply causation		
Sno	Correlation Analysis	Regression Analysis
1	Measures Degree of relationship between 2 variables	Captures causality i.e. how one variable affects another (x -> y)
2	Correlation doesn't capture causality	Regression is based on causality, it doesn't capture degree but shows relationship of cause and effect
3	A property of correlation is that p(x,y)=p(y,x). This you can easily see from formula which is symmetrical.	Regression is one way e.g. Income and Years of education relationship would differ if you reverse the direction of causality
4	Different graphical representation - Single point	Best fit line

Quiz

- Which statement is false?
  - ○ Correlation does not imply Causation.
  - ○ Correlation is symmetrical regarding both variables
  - ○ Correlation can be represented as a line.
  - ○ Correlation does not capture the direction of causal relationship.

Correct: Correlation can be represented as a line.

Geometrical Representation via Scatter Plot



Minimising Least Square Error

The term α in the equation of the line is the y-intercept

Red points = observed points

a = y-intercept

b = slope

e = perpendicular distance between the observed points and the best fit line is the estimator of the error.

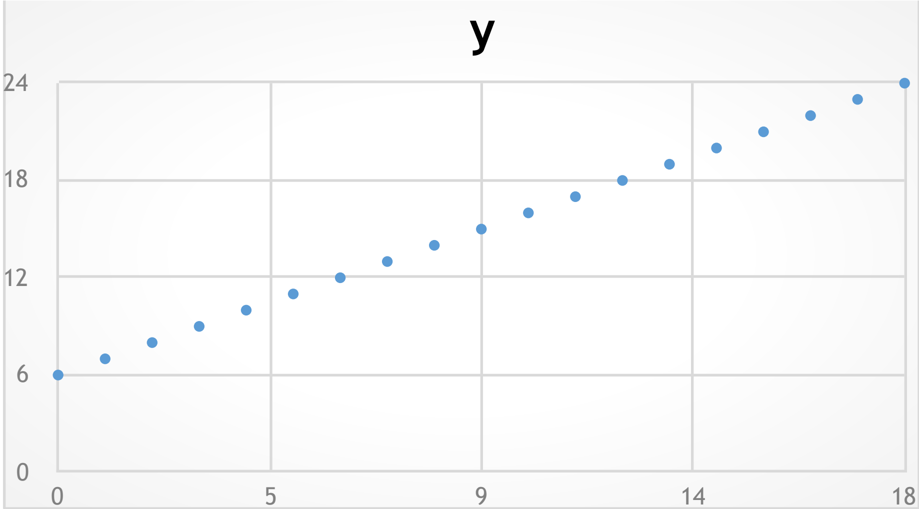
Quiz

- Assume you’ve the following sample regression line y=6+x. If we draw the regression line what would be the slope?
    - 1
    - 6
    - Xi
    - None of the above

Y = m.X + b

b = Y-Intercept

m = Slope



Coding – Packages

- Numpy
- Pandas
- Scipy
- Statsmodels -
    - built on top of numpy and scipy and integrates well with pandas. SM provides with very good summaries as we'll see in code shortly.
- Matplotlib
- Seaborn
- Sklearn -
    - Sklearn is the Real deal
    - Statsmodel makes it easier to understand

First regression Model using statsmodels and Understanding Statsmodels tables

Let's open Simple regression model.ipynb to create our first

Quiz

- What is predicted GPA of the students with a SAT score of 1850?

  (Assume that any co-efficient with a p-value greater than 0.05 is not significant)

  a) 3.420

  b) 3.060

c) 3.145

d) 3.230

Correct Answer: c) 3.145

- What does p-value of 0.503 suggest about the intercept coefficient?

  e) It is significantly different from 0

  f) It is not significantly different from 0

  g) It is equal to 0.503

  h) None of the above

Correct Answer: b) It is not significantly different from 0

- What does p-value of 0.000 suggest about the coefficient of x?

  i) It is significantly different from 0

  j) It is significantly not different from 0

  k) It tells nothing

  l) None of the above

Correct Answer: a) It is significantly different from 0

Determinants of good regression

- SST - Sum of squares total or Total sum of squares (TSS) – **Total Variability**

    m) Summation of Squared difference between observed dependent variable and it’s mean. Dispersion of observed variables around the mean.

    n) $\Sigma (yi – y)^2$

- SSR - Sum of squares regression or Explained sum of squares (ESS) – **Explained Variability**

    o) Summation of difference between predicted value and mean of dependent variable.

    p) How well your line fits the data

    q) If SSR = SST then it means your regression captures all the variability and is perfect.

- SSE - Sum of squares error or Residual sum of squares (RSS) – **Unexplained Variability**

    ○ Difference b/w observed and predicted value

    ○ Smaller the error the better the regression, idea is to optimize the error to a minimum

- Connection between these 3: SST = SSR + SSE

- Total Variability = Explained Variability + Unexplained Variability