Advanced Data Science

Regression

Session 5

Kiran Waghmare
Program Manager
C-DAC Mumbai

Agenda

- Regression
- Types of Regression
  - Linear
  - Logistic

What is Linear Regression (LR)?

- Linear regression (LR) models the linear relationship between the one independent (x) variable with that of the dependent variable (y). If there are multiple independent variables in a model, it is called as multiple linear regression.
- For example, how the likelihood of blood pressure is influenced by a person's age and weight. This relationship can be explained using linear regression.
- In LR, the y variable should be continuous, whereas the x variable can be continuous or categorical. If both x and y are continuous, the linear relationship can be estimated using correlation coefficient (r) or the coefficient of determination (R-Squared)

- LR is useful if the relationships between the `x` and `y` variables are linear
- LR is helpful to predict the value of `y` based on the value of the `x` variable

Note: Dependent variable also called a response, outcome, regressand, criterion, or endogenous variable.

Independent variable also called explanatory, covariates, predictor, regressor, exogenous, manipulated, or feature (mostly in machine learning) variable.

Types of Linear Regression (LR)?

- Univariate LR: Linear relationships between `y` and `x` variables can be explained by a single `x` variable

$$y = a + bX + \epsilon$$

Where, a = y-intercept, b = slope of the regression line (unbiased estimate) and ϵ = error term (residuals)

- Multiple LR: Linear relationships between `y` and `x` variables can be explained by multiple `x` variables

$$y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n + \epsilon$$

Where, a = y-intercept, b = slope of the regression line (unbiased estimate) and ϵ = error term (residuals)

- The y-intercept (a) is a constant and slope (b) of the regression line is a regression coefficient.

9.Monitoring and Maintenance:

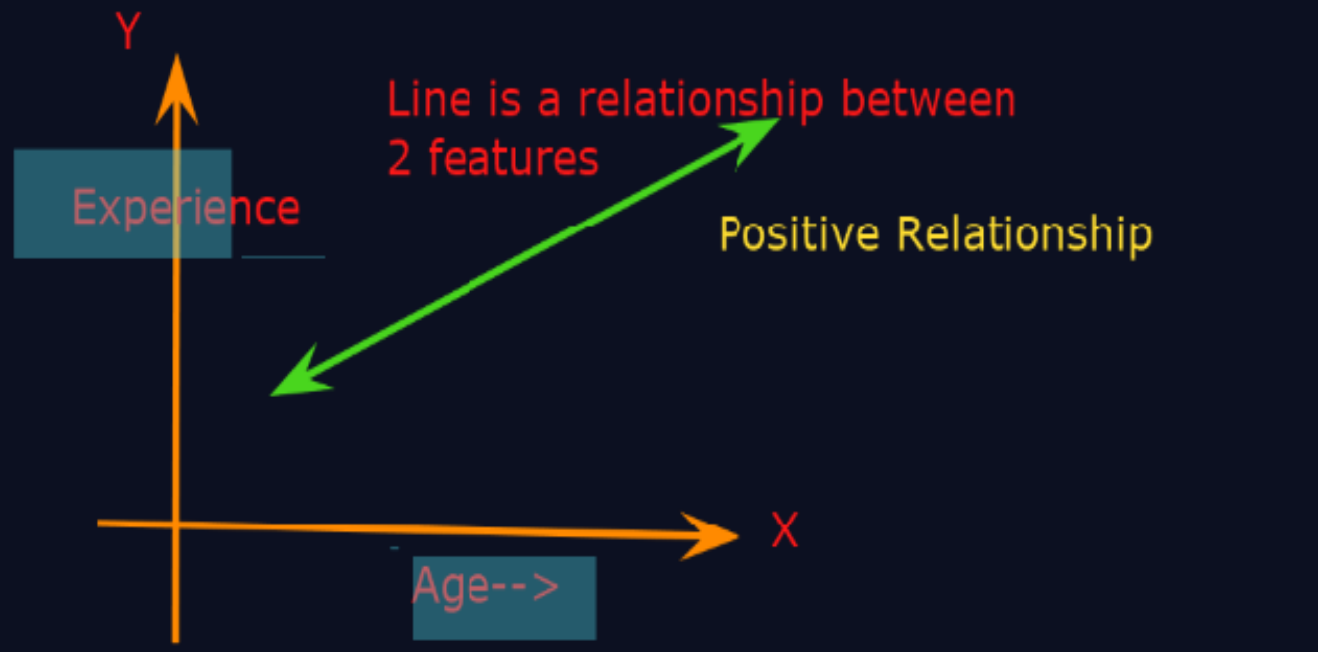
Model types:

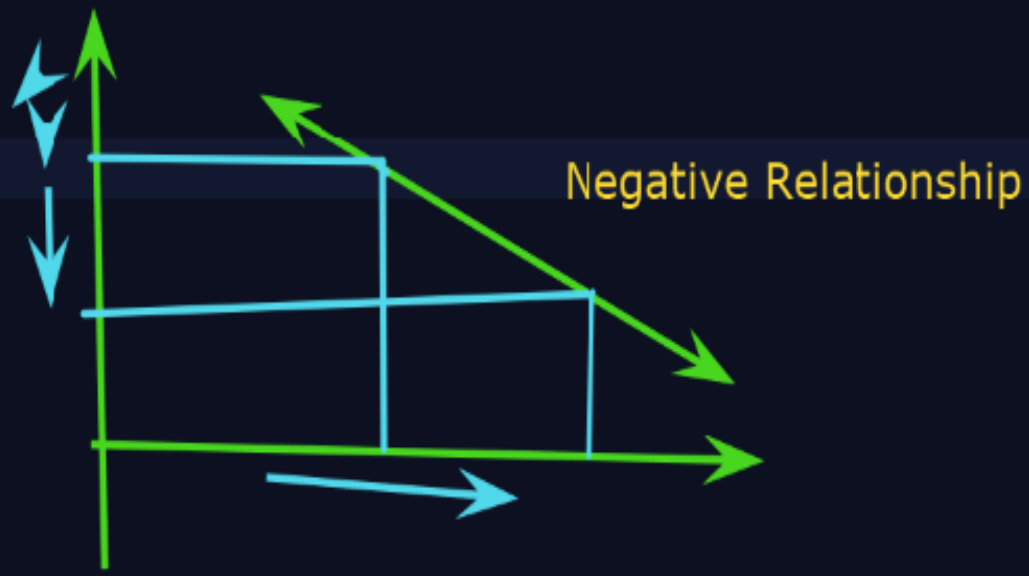
---

1. Regression Mode
    1.Linear Regression

    2.Multi linear REgression
    3.Polynomial Regression
    4.popular)Lasso Regression (
    5.Ridge Regression
    6.Elasticnet Regression

2. Classification Model
3. Unsupervised Model : Clustering





Model types:

---

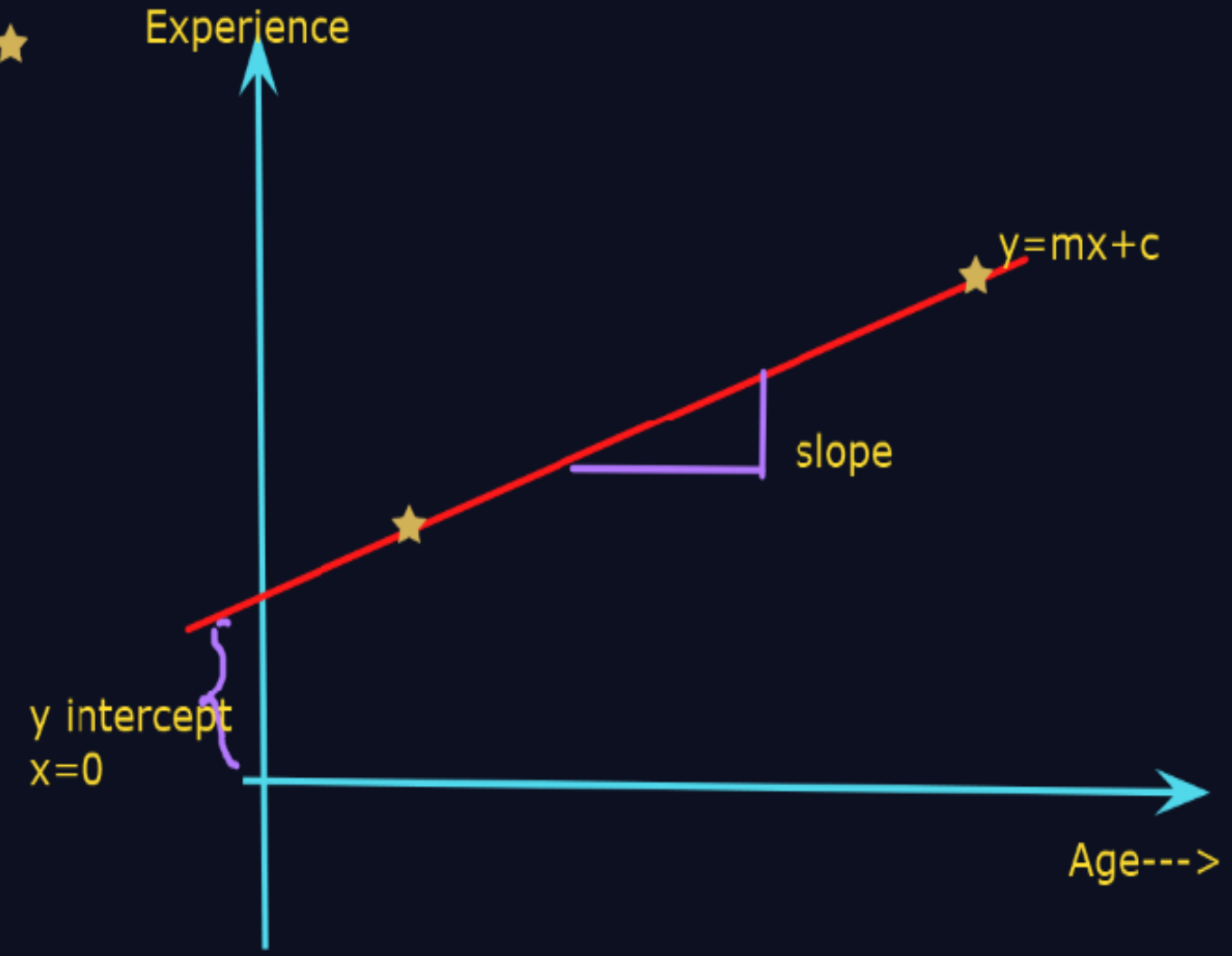
1. Regression Model
    1.Linear Regression

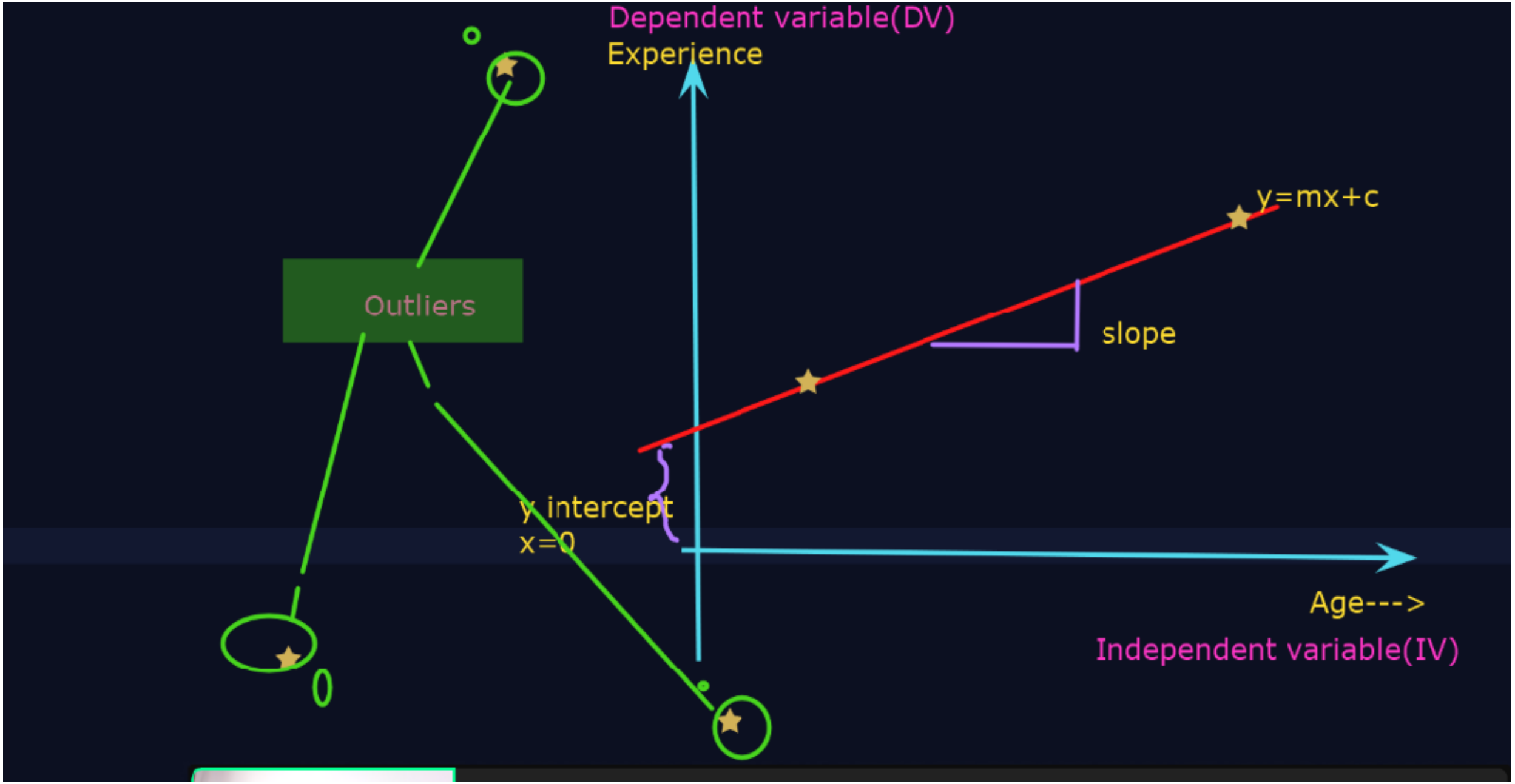
    2.Multi linear REgression
    3.Polynomial Regression
    4.popular)Lasso Regression (
    5.Ridge Regression
    6.Elasticnet Regression

    Equation for line: y=mx+c
        m=slope of line =3
        c= constant (y-intercept) =5
        X=2

        y = mx + C
          =3*2 + 5 = 11

2. Classification Model
3. Unsupervised Model : Clustering




Dependent variable(DV)
Experience
Outliers
y=mx+c
slope
y intercept
x=0
Age--->
Independent variable(IV)
0

Terminologies:

- Dependent variable:Target variable to be predicted
- Independent cariable:Predictor affecting the dependent variable
- Outliers:Extreme values tha t can skew the results
- Multicollinearity:High coorelation among independent varaibles
- Underfitting:
- Overfitting:

jupyter SLR Last Checkpoint: 5 minutes ago

Edit View Run Kernel Settings Help

Code JupyterLab

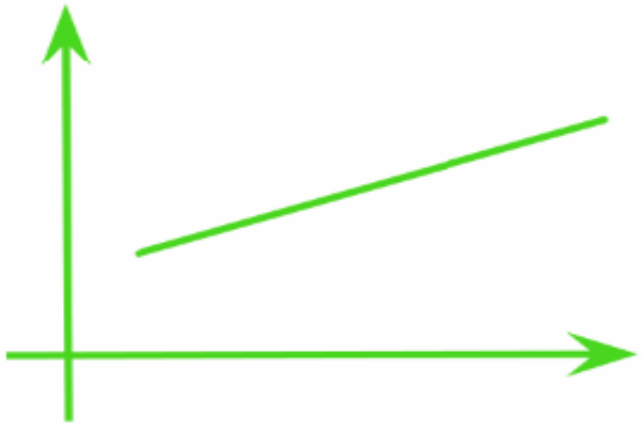
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_24772\2668718341.py:2: SyntaxWarning: invalid escape sequence
  dataset = pd.read_csv('D:\Test\Salary_Data.csv')
```

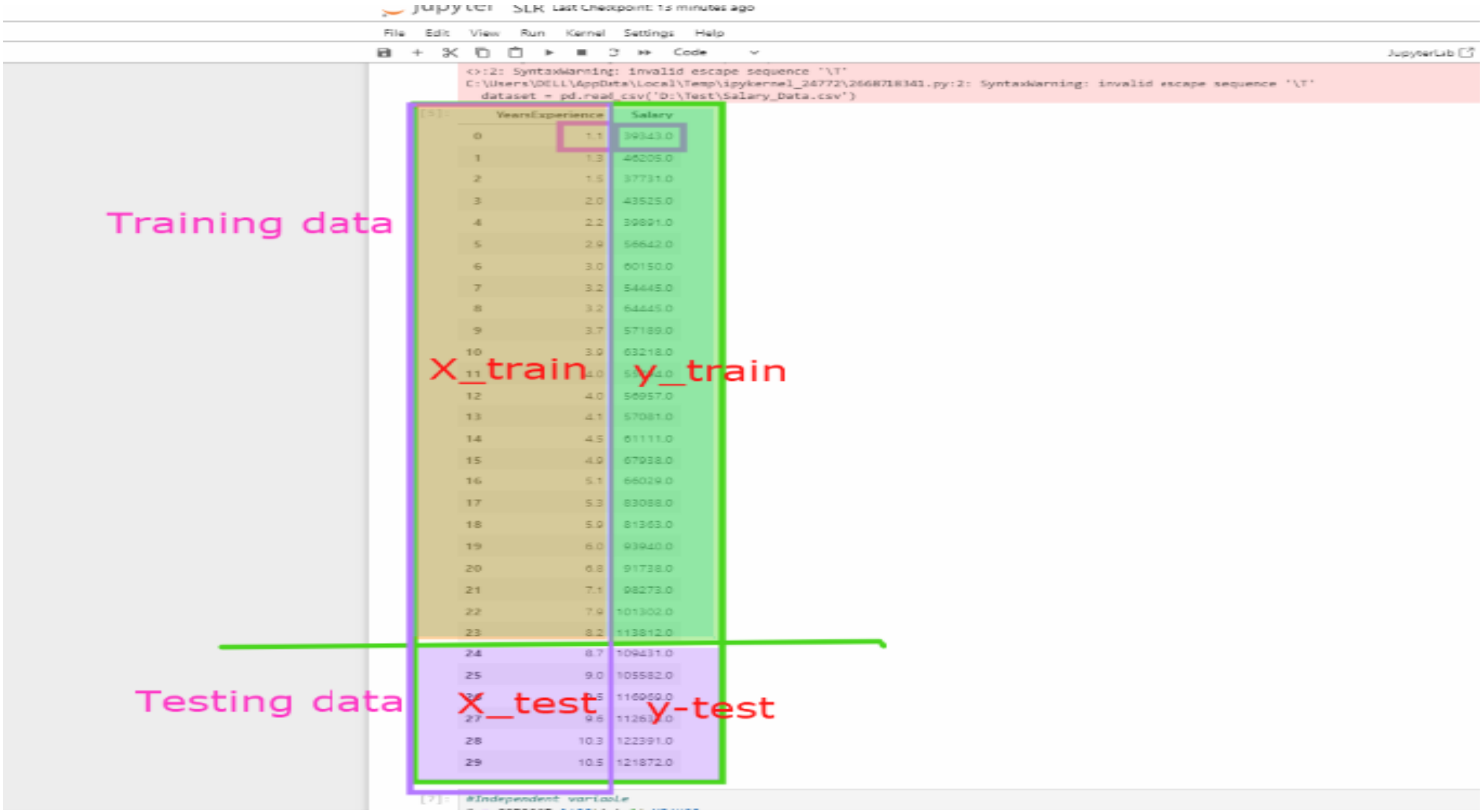
[5]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0
5	2.9	56642.0
6	3.0	60150.0
7	3.2	54445.0
8	3.2	64445.0
9	3.7	57189.0

y= Salary (Dependent variable)

x= Yrs of Exp (Independent variable)




Jupyter SLR Last Checkpoint: 13 minutes ago
File Edit View Run Kernel Settings Help
Code
JupyterLab
<>:2: SyntaxWarning: invalid escape sequence '\T'
C:\Users\DELL\AppData\Local\Temp\ipykernel_24772\2668718341.py:2: SyntaxWarning: invalid escape sequence '\T'
dataset = pd.read_csv('D:\Test\Salary_Data.csv')
YearsExperience Salary
0 1.1 39343.0
1 1.3 46205.0
2 1.5 37731.0
3 2.0 43525.0
4 2.2 39891.0
5 2.9 56642.0
6 3.0 60150.0
7 3.2 54445.0
8 3.2 64445.0
9 3.7 57189.0
10 3.9 63218.0
11 4.0 55794.0
12 4.0 56957.0
13 4.1 57081.0
14 4.5 61111.0
15 4.9 67938.0
16 5.1 66029.0
17 5.3 83088.0
18 5.9 81363.0
19 6.0 93940.0
20 6.8 91738.0
21 7.1 98273.0
22 7.9 101302.0
23 8.2 113812.0
24 8.7 109431.0
25 9.0 105582.0
26 9.5 116969.0
27 9.6 112635.0
28 10.3 122391.0
29 10.5 121872.0
#Independent variable
Training data
X_train
y_train
Testing data
X_test
y-test

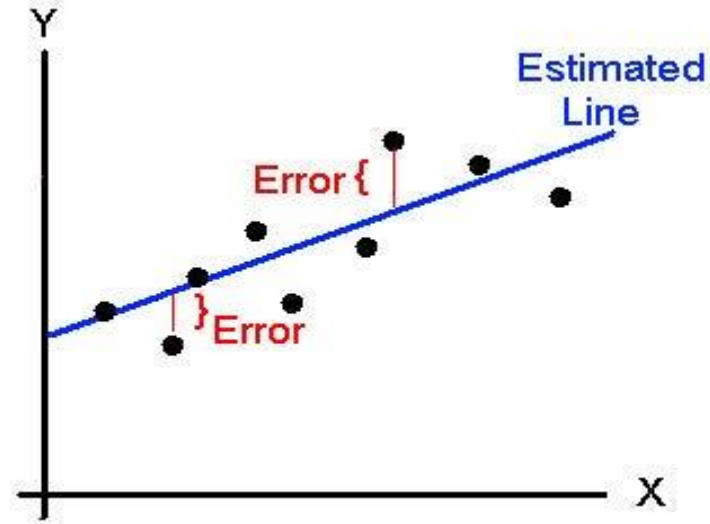
Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$



SLR Program Implementation

1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Residuals (regression error)

- Residuals or error in regression represents the distance of the observed data points from the predicted regression line

$$residuals = actual\ y(y_i) - predicted\ y\ (\hat{y}_i)$$

Root Mean Square Error (RMSE)

- RMSE represents the standard deviation of the residuals. It gives an estimate of the spread of observed data points across the predicted regression line.