Let's Recall,

- We were predicting the price of sold cars for CARS24.

We already know about Sklearn's linear regression.

Introduction to Statsmodel

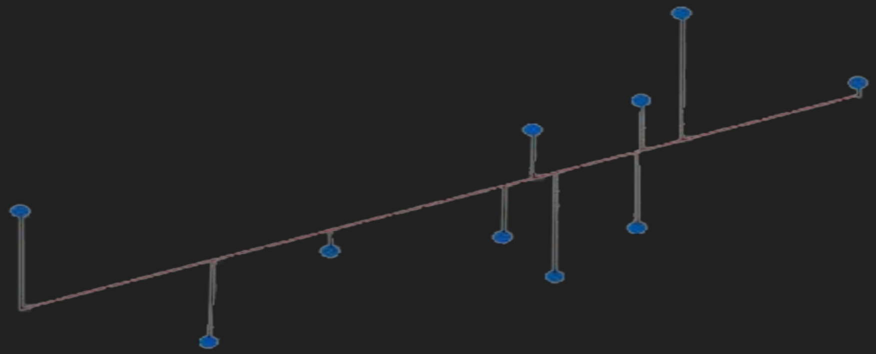
What is statsmodel?

- A python module with statistical functionalities.

Stats model Library

OLS (Ordinary Least Squares)

- OLR refers to a method of estimating the parameters of a linear regression by minimizing the sum of square residuals.



How is OLS different from sklearn’s Linear Regression?

OLS

- Focuses solely on estimating the parameters of a linear regression model

Sklearn

- Offers additional features and functionalities like :
  - Feature scaling
  - Regularization
  - Cross validation
  - Evaluation metrics

Assumptions of Linear Regression

- **Assumption of Linearity**
- **No Multicollinearity**
- **Normality of Residuals**
- **No Heteroskedasticity**
- **No Autocorrelation**

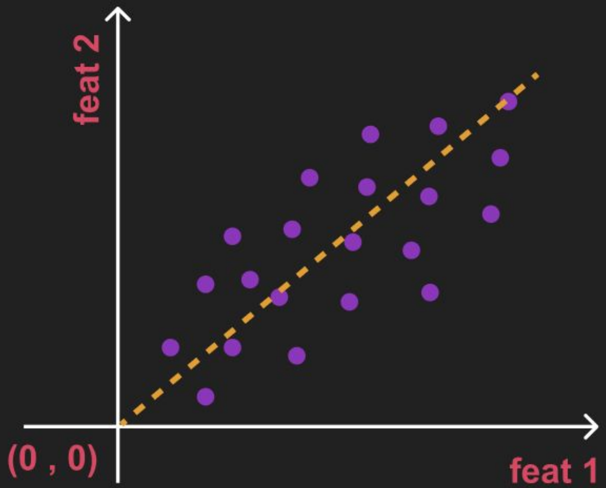
Assumptions of Linearity

Assumption of linearity

There should be linear relationship between :

- Independent Variables {X}
- Dependent Variables {y}

Straight line fit between variables



No Multicollinearity

What is collinearity?

Say we have two features : f_1 and f_2

If, $\boxed{f_1 = \alpha f_2 + \square}$

Then f_1 and f_2 are collinear

What is multicollinearity then?

Collinearity between multiple features

Example : $f_i = f_1 , f_2 , f_3$

$\boxed{\text{S.T.} \quad f_1 = \alpha_1 + \alpha_2 f_2 + \alpha_3 f_3}$

Then f_1 , f_2 , f_3 are multicollinear

Why multicollinearity is a problem?

Say we found the optimal weights w* for a model with 3 features

$w^*=[1,2,3]$ (corresponding to w_1, w_2, w_3) and $w_0 = 5$

So, $\hat{Y} = w_1x_1 + w_2x_2 + w_3x_3 + w_0$

$= x_1 + 2x_2 + 3x_3 + 5$

Now, let's say x_1 and x_2 are collinear.

Then, $x_2 = 1.5x_1$

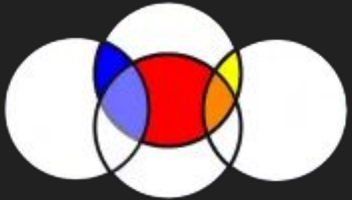
Hence, $\hat{Y} = 4x_1 + 3x_3 + 5$

$$\left.\begin{array}{l} \therefore\ w^* = < 4 , 0 , 3 > \\ \therefore\ w^* = < 1, 2, 3 > \end{array}\right\} \text{Same classifier}$$

This would mess up the weights and we won't be able to do feature importance.

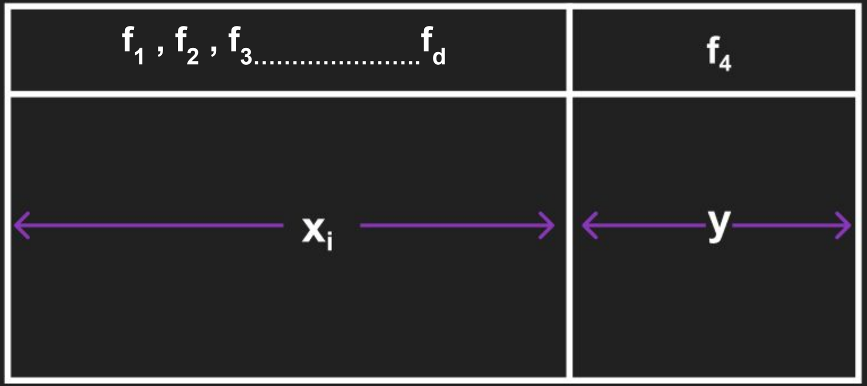
How to deal with multicollinearity?

We will use **Variance Inflation Factor (VIF)**



Say, we have 'd' features $< f_1 , f_2 , f_3 \ldots\ldots\ldots\ldots f_d >$

In, **(VIF)** we treat
- one feature as 'y'
- remaining features as 'x'



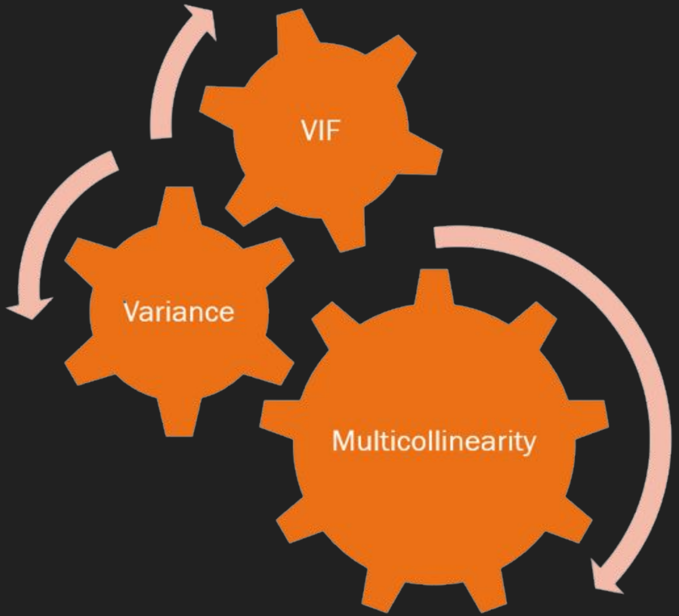
Now,

Train **linear regression** model with *(x_i , y)*

Find R^2 of the model

To Calculate VIF :

$$VIF = \frac{1}{1 - {R_j}^2}, {R_j}^2$$: R^2 for j^{th} feature



Step 1 : Start with a full regression model including all the independent variables.

Step 2 : Calculate VIF for each independent variable by regressing it against all other independent variables.

Step 3 : Check for variable with highest VIF, [Thumb rule on next slide]

Step 4 : Remove variable with highest VIF.

Step 5 : Re-fit the model without the removed variable.

Repeat steps 2 - 5 : Continue this until no variable has VIF above threshold.

Case 1

- If $R^2 \approx 1$

VIF = 1/1-1 = ∞

High R^2 means

⬇

Feature is highly collinear

⬇

Can be removed

Case 2

- If $R^2 \approx 0$

VIF = 1-1/0 = 1

Low R^2 means

⬇

Feature is not collinear

⬇

Don't remove

Thumb Rule

- ***VIF > 10: Very high multicollinearity, drop***
- ***5<=VIF<=10: High multicollinearity***
- ***VIF<5: Low multicollinearity***

**NOTE : We do this process for each feature. Calculate the VIF and based on that we keep / remove feature

Normality of Residuals

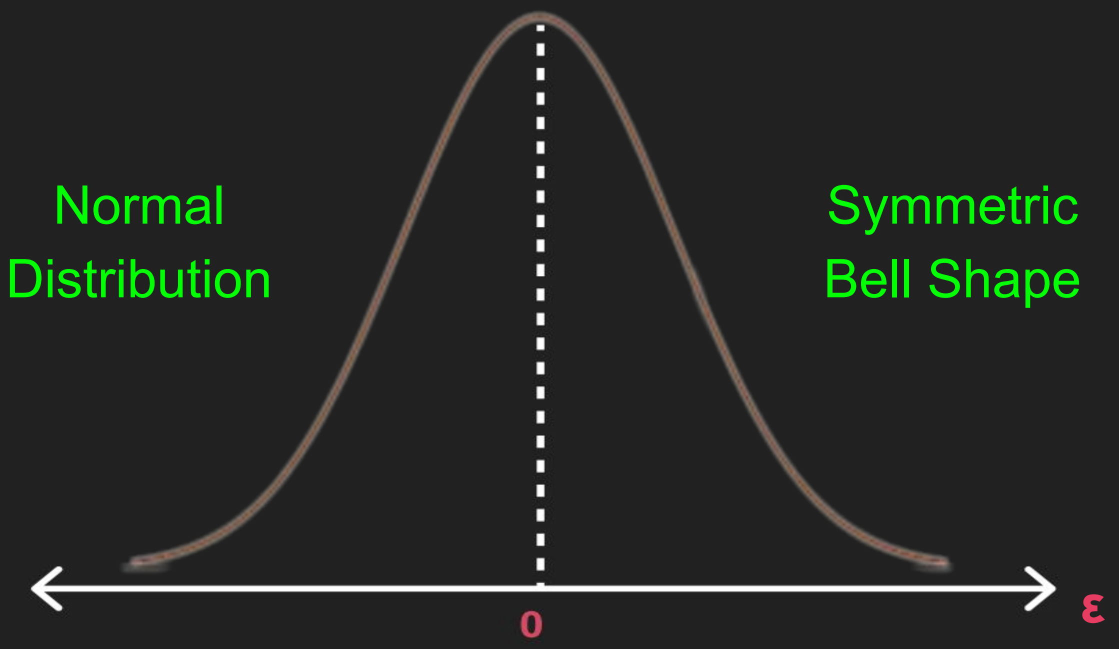
Residuals/ Errors follow multivariate normal distribution.

Every linear model has some error.

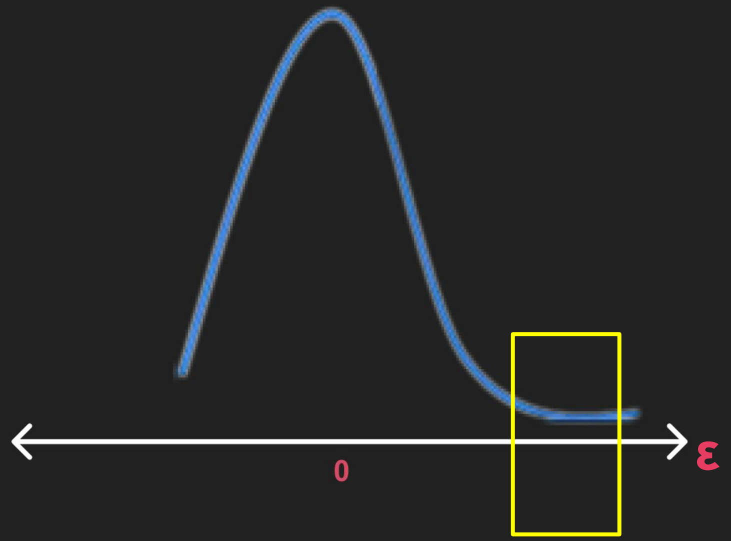
$$y^{i} = w_{o} + w^{t}_{x}{}^{(i)} + \varepsilon$$

$$\therefore \varepsilon^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

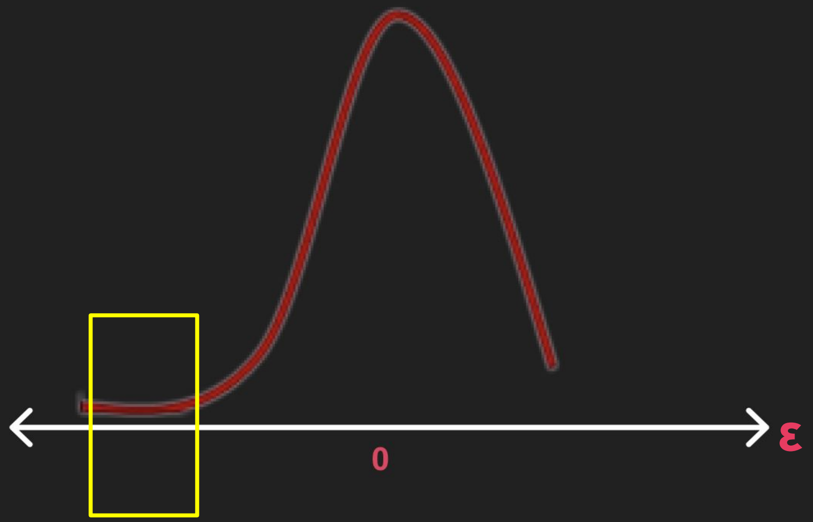
Plotting ‘ε’



On the other hand, if



Right Skewed



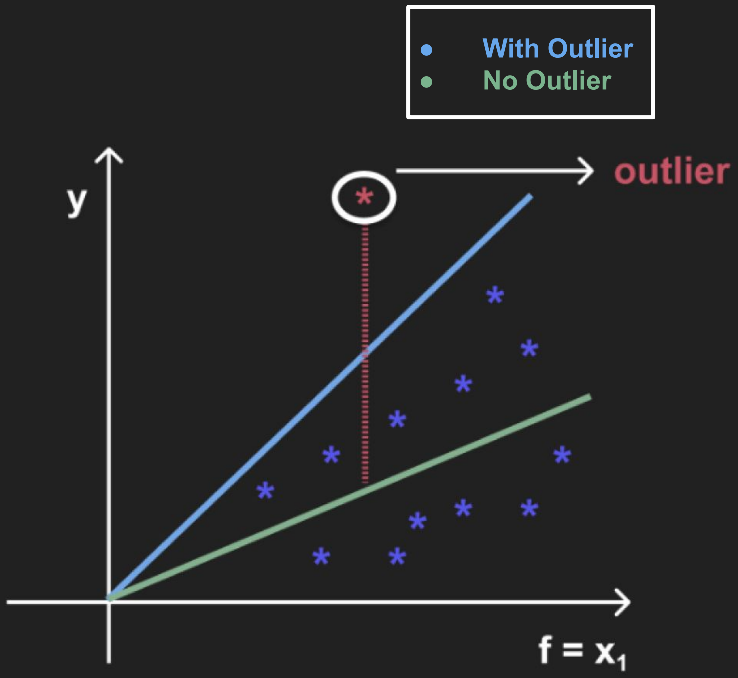
Left Skewed

ε is large, outliers are present.

What is the impact of outliers?

If we have outliers,

- The regression line gets pulled towards the outlier to minimize the squared loss.



Q : How to identify outliers ?

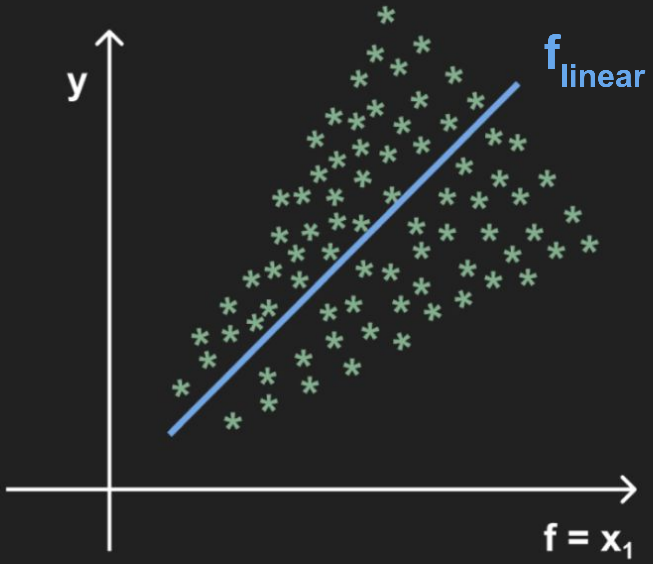
A: Outliers will have high error (ε).

Q : How to deal with outliers ?

A: Remove the points will high error as many as you want and fit the model again.

No Heteroskedasticity

When we plot the two features along with the regression line, notice :



As we go from left to right errors are increasing.

How to check Heteroskedasticity ?

By plotting,

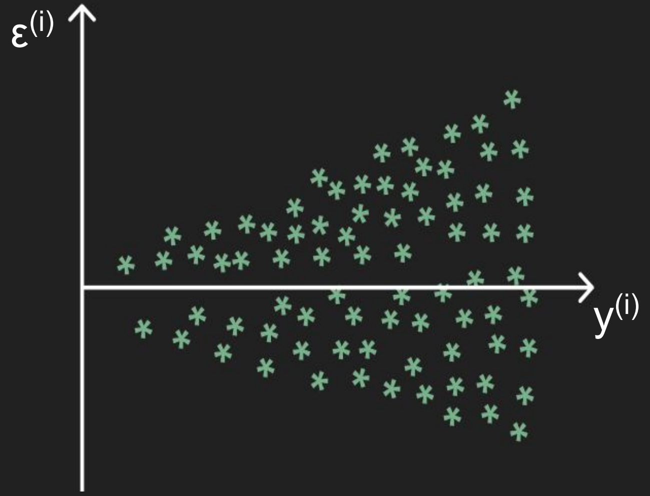
$$\boxed{y^{(i)} \text{ vs. } \varepsilon^{(i)}}$$

In maths/stats proof of linear regression, we assume the errors are normally distributed

$$\varepsilon^{(i)} \sim N(0, \sigma)$$

mean = 0

Std. Dev. is not constant



Spread of $\varepsilon^{(i)}$ is not same for all values of $y^{(i)}$, this is known as Heteroskedasticity.

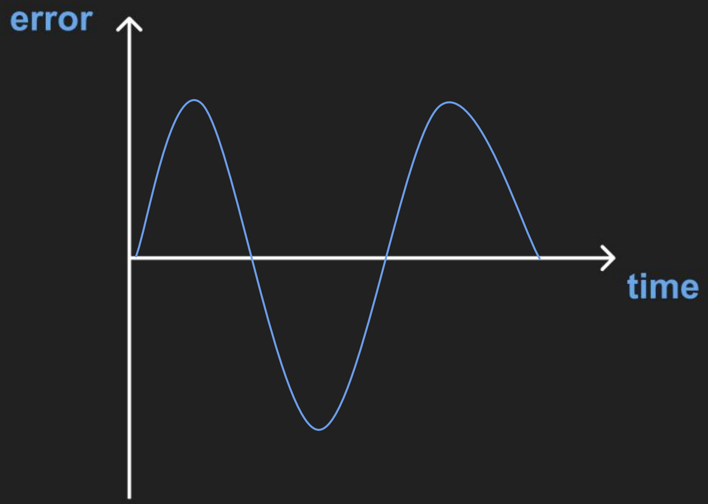
What does it tells us ?

Outliers in data

Linear model isn't right

No Autocorrelation

Autocorrelation plays a role only when **"Time Series"** data is involved



Example : Predicting sales on the same day

When we plot the error w.r.t. time, if some pattern is observed then autocorrelation exists.