Assumption of Linear Regression

Linear regression relies on several assumption to ensure the validity and reliablity of the estimate and inferences. The key assumption of linear regression are:

1. Linearity
2. Normality of Residuals
3. Homoscedasticity
4. No Autocorrelation
5. No or little Multicollinearity

1. Linearity

The Assumption

There is a linear relationship between the independent variables and the dependent Variable the model assume that changes in the independent Variable lead to propotional changes in the dependent variable.

What happens when this assumption is violated?

1. **Bias in parameter estimates:** When the true relationship is not linear, the estimated regression coefficient can be biased, leading to incorrect inference about the relationship between the independent and dependent variable.

2. **Reduced predictive accuracy:** A miss specified linear model may not accurately capture the underlying relationships, which can result in poor predictive performance. The model might underfit the data, missing important patterns and trends.

3. **Invalid hypothesis tests and confidence intervals:** The violation of the linearity assumption can affect the validity of hypothesis tests and confidence intervals, leading to incorrect inference about the significance of the independent variable and the effect size.

How to check this assumption

1. **Scatter plots**: Create scatter plots of the dependent variable against each independent variable. If the relationship appears to be linear, the linearity assump is likely satisfied. Non linear patterns or other trends may indicate that the assumption is violated.

2. **Residual plots**: Plot the residuals (the difference between the observed and predicted values) against the predicted values or against each independent variable. If the linearity assumption holds, the residual should be randomly scattered around zero, with no discernible pattern. Any trends, curvature, or heteroscedasticity in the residuals plots suggest that the linearity assumption may be violated.

3. **Polynomail terms**: Add polynomail term to your model and compare the model fit with the original linear model. If the new model with additional terms significance improves the fit, it may suggest that the linearity assumption is violated.

What to do when the assumption fails?

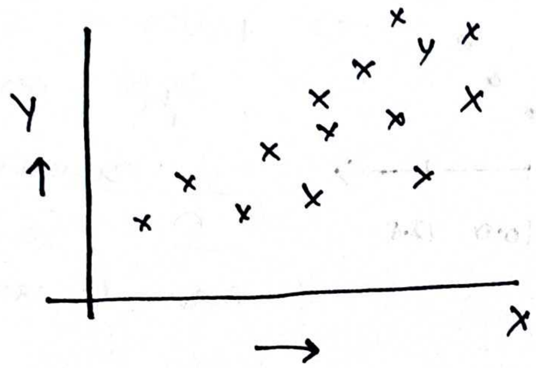
1. **Transformation:** Apply transformation to the dependent and / or Independent variables to make their relationship more linear. Common transformation including logarithmic, square root, and inverse transformations.

2. **Polynomail Regression** : Add polynomail terms of the Independent Variable to the model to capture non-linear relationship.

3. **Piecewise Regression** : Divide the range of the Independent variable into segments and fit seperate linear models to each segment.

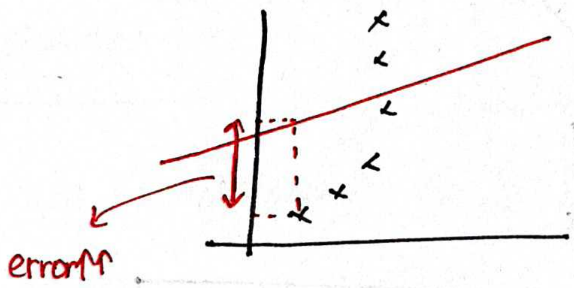
4. **Non-parametric or semi-parametric methods** : consider using non-parametric or semi-parametric methods that do not rely on the linearity assumption, such as generalized additive models (GAMs), splines or kernel regression.

X | Y



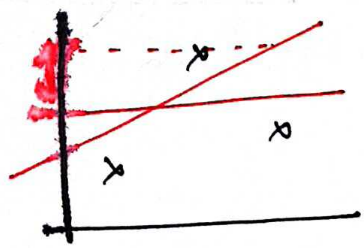
$y = \beta X$

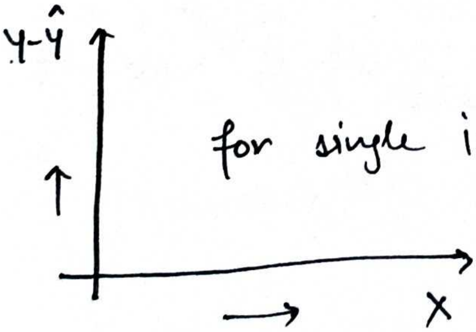
* Reduced predicted accuracy



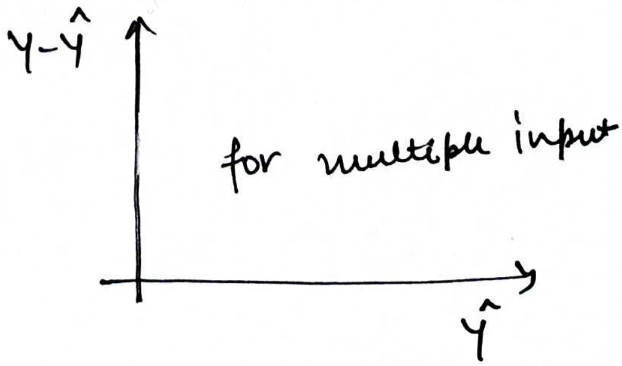
* Residual plots:

$X \mid Y \xrightarrow{\text{calculate}} \hat{Y}$

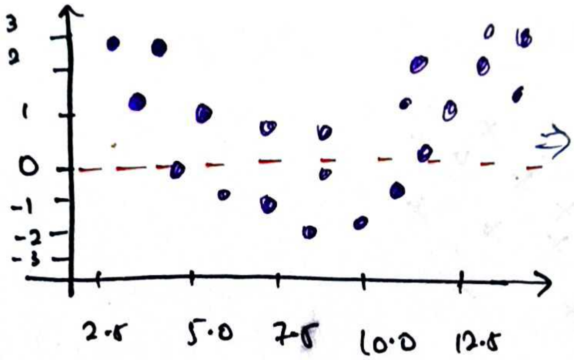




for single input



for multiple input



⇒ Pattern = Parabola type not a random at 0. So it is non-linearity

* Polynomail Regression (Polynomail term)

X	Y
2	3

Degree = 2

$Y = \beta_0 + \beta_1 x$

X^0	X^1	X^2	Y
1	2	4	3

$$\boxed{Y = \beta_0 + \beta_1 X^0 + \beta_2 X^1 + \beta_3 X^2}$$