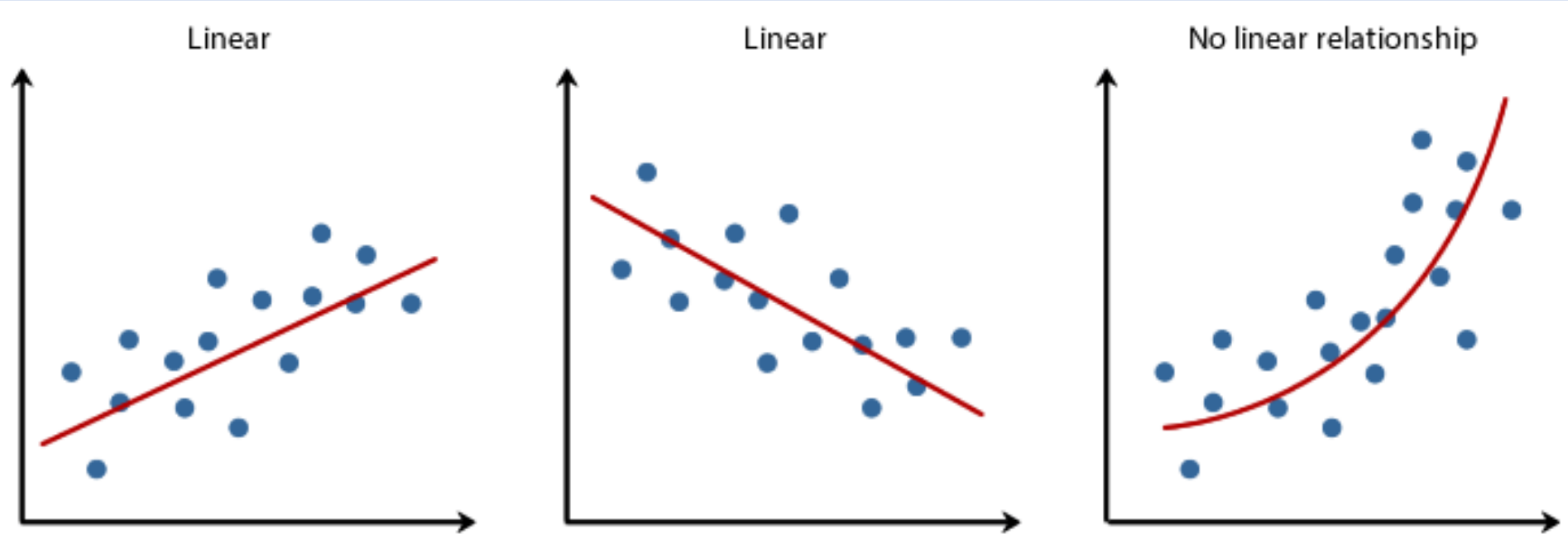
Linear Regression

Assumptions

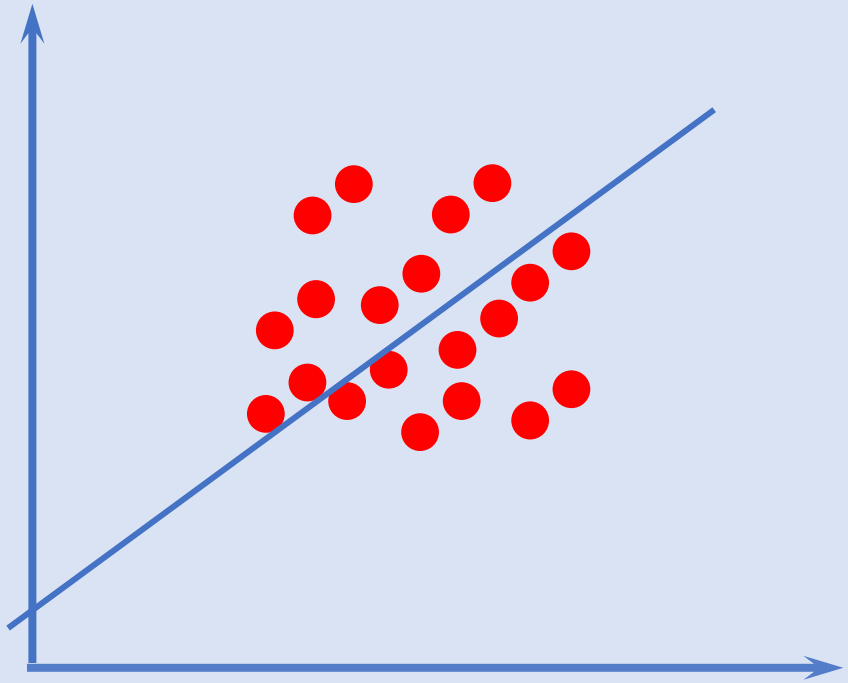
Linear Regression Assumptions

1. Linear Relationship Between input and output
2. No Multicollinearity
3. Normality of Residual
4. Homoscedasticity
5. No Autocorrelation in the errors

Linear Relationship



Linear Relationship



Non Linear

How to test Linearity of the data?

- The linear correlation coefficient is a number calculated from given data that measures the strength of the linear relationship between two variables: x and y
- Coefficient of correlation is also known as Pearson Correlation Coefficient or R-Value
- R-Value ranges from -1 to +1

How to compute R-Value/Correlation

$$R = Corr = \frac{Covariance(x, y)}{Product\ of\ Standard\ Deviation\ of\ x\ and\ y}$$

$$R = Corr = \frac{Covariance(x, y)}{\sqrt{Variance\ x} \times \sqrt{Variance\ y}}$$

How to compute Covariance?

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{N - 1}$$

$x_i = Values\ of\ x$

$y_i = Values\ of\ y$

$\bar{x} = Mean\ of\ x$

$\bar{y} = Mean\ of\ y$

$N = Number\ of\ data\ points$

Correlation Types

- When the R value is positive the correlation is positive
- When R value is negative the correlation is negative

- Good Predictor is when : $R \geq 0.7$ or $R \leq -0.7$
- Bad Predictor is when : $-0.3 < R < +0.3$

Multicollinearity

- Variance Inflation Factor

$$R^2 = 1 - \frac{MSE(Model)}{MSE(Baseline)} \qquad VIF = \frac{1}{1 - R^2}$$

- There should not be multicollinearity in the features of data
- Features having Variance inflation factor more than 5 is considered to have multicollinearity and should be dropped

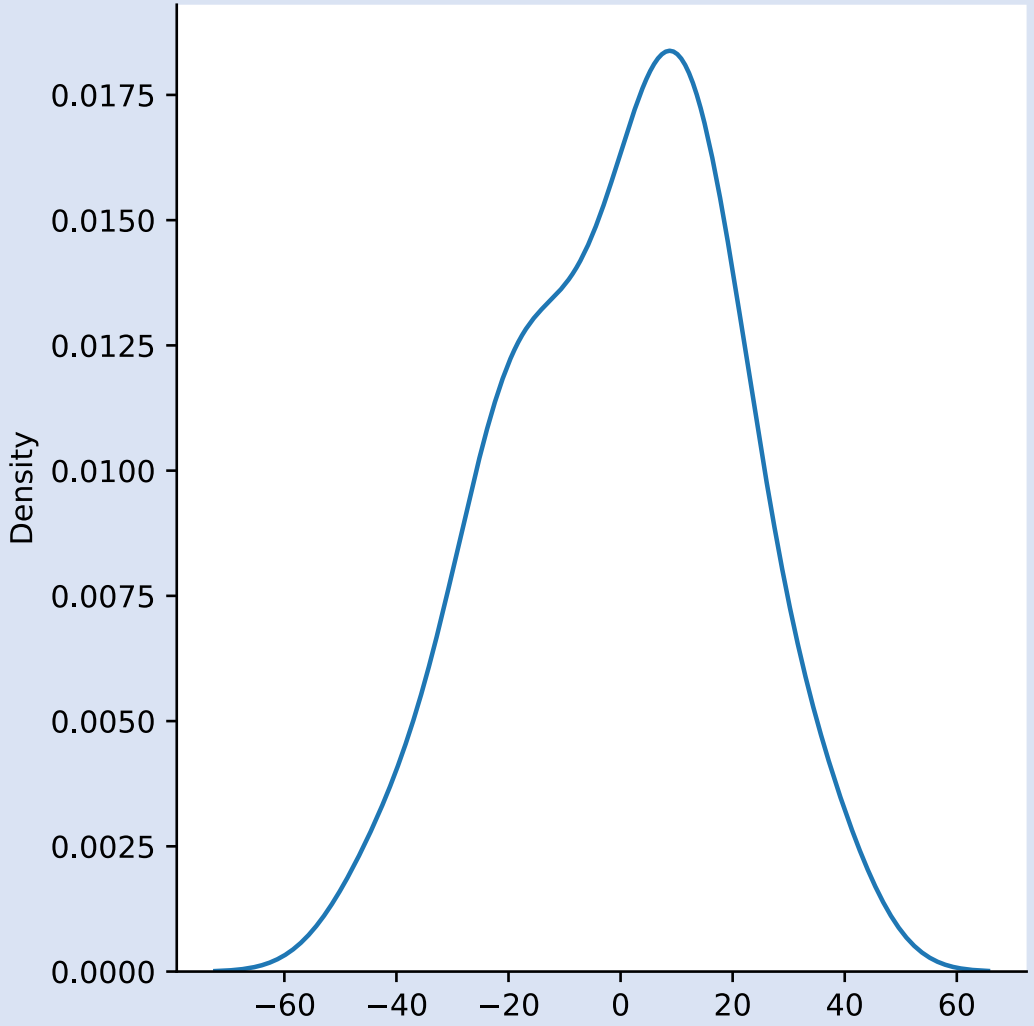
Normal Distribution of Residual

$$Residual = y - \hat{y}$$

$$Residual = y_{actual} - y_{predicted}$$

$$Residual = y_{test} - y_{pred}$$

Sns.distplot(residual)

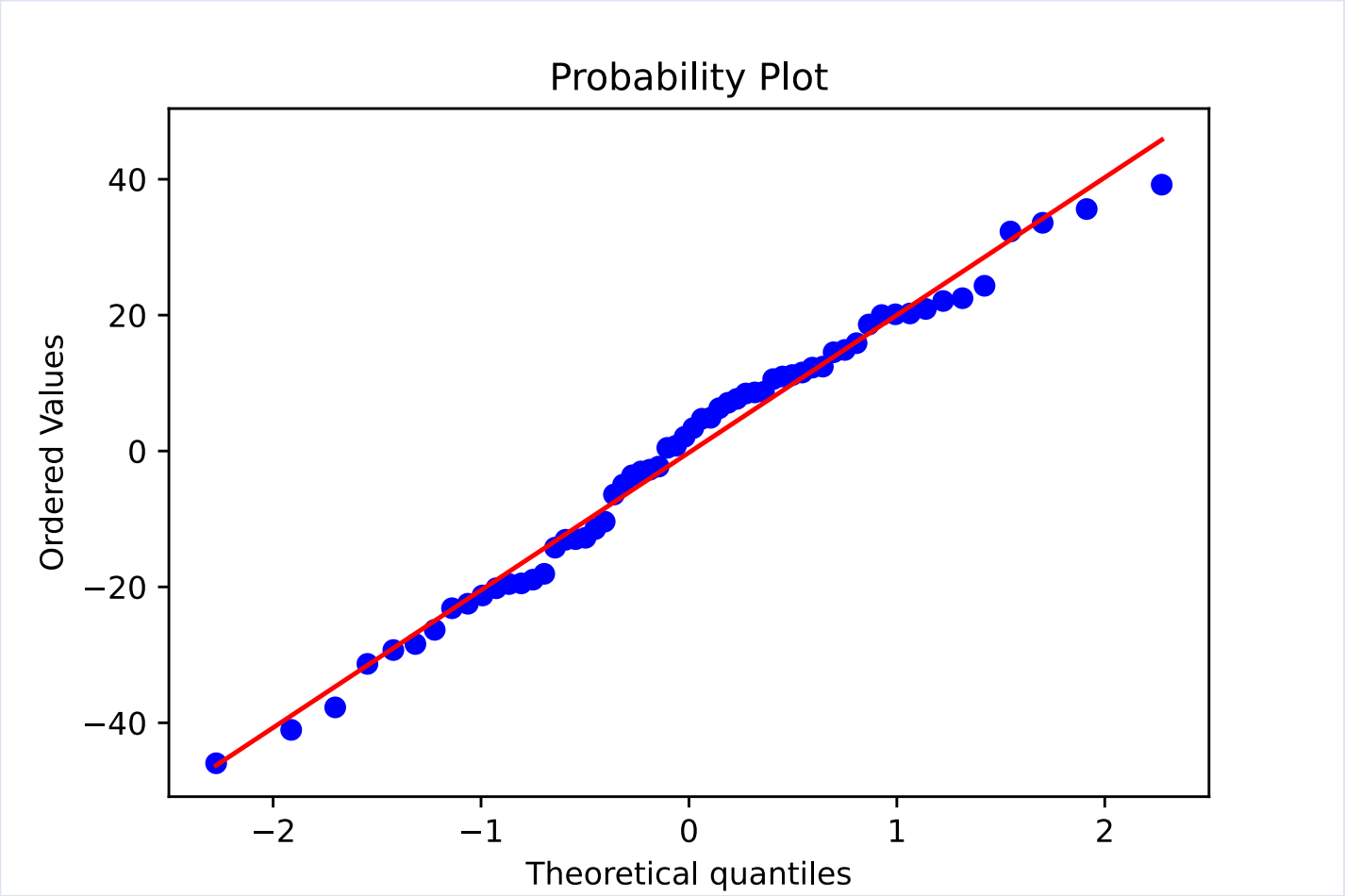


Normal Distribution of Residual

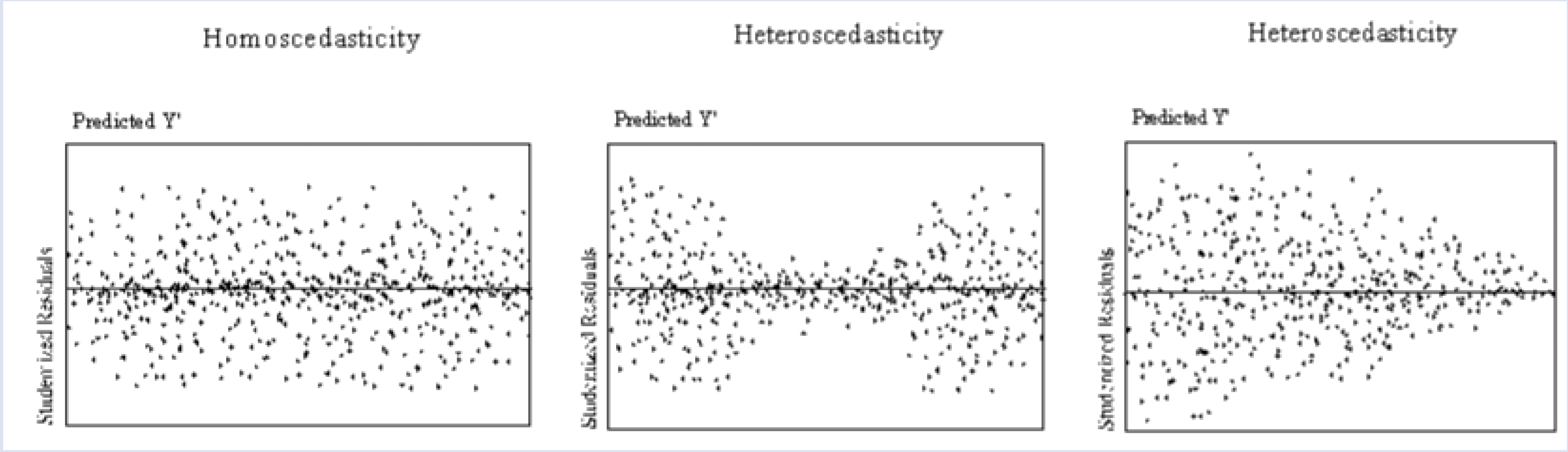
$$Residual = y - \hat{y}$$

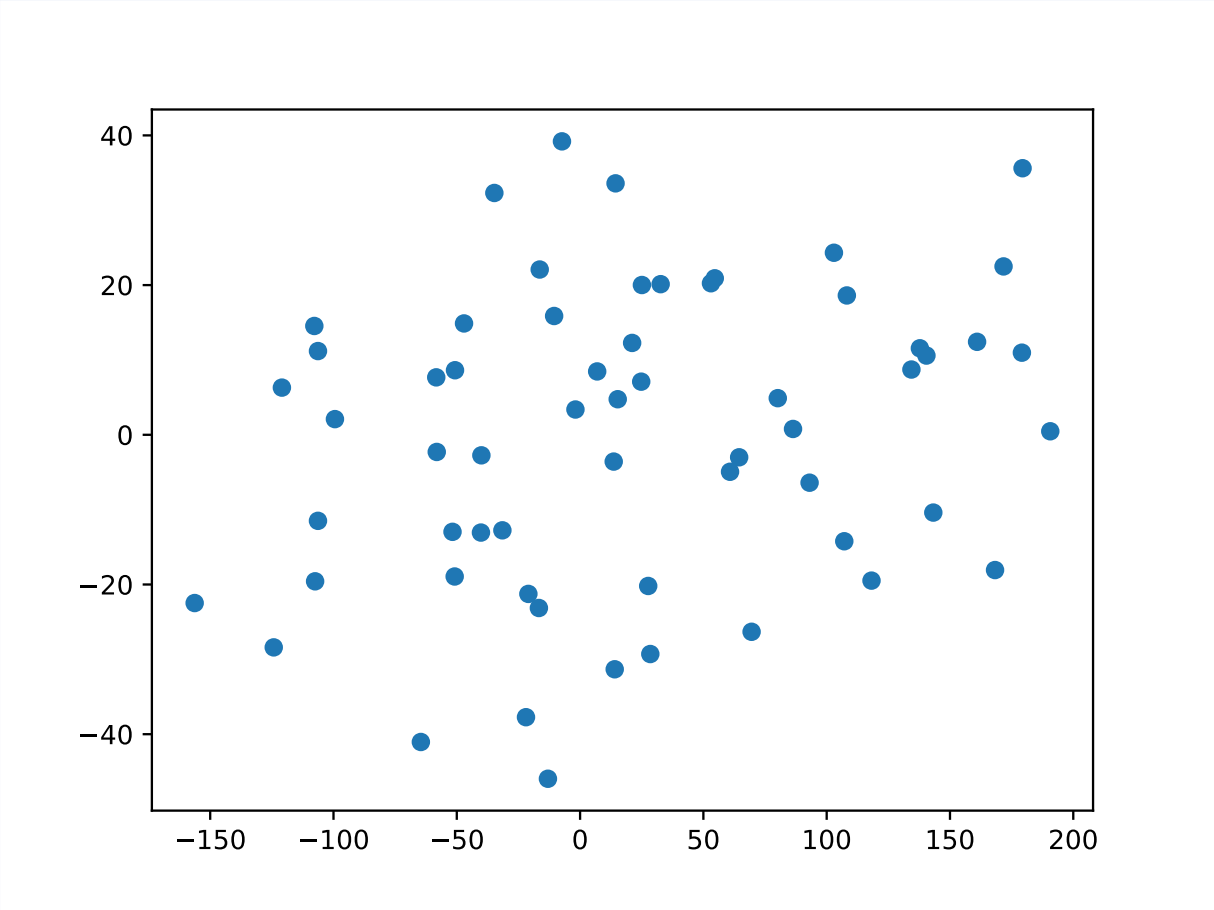
$$Residual = y_{actual} - y_{predicted}$$

$$Residual = y_{test} - y_{pred}$$



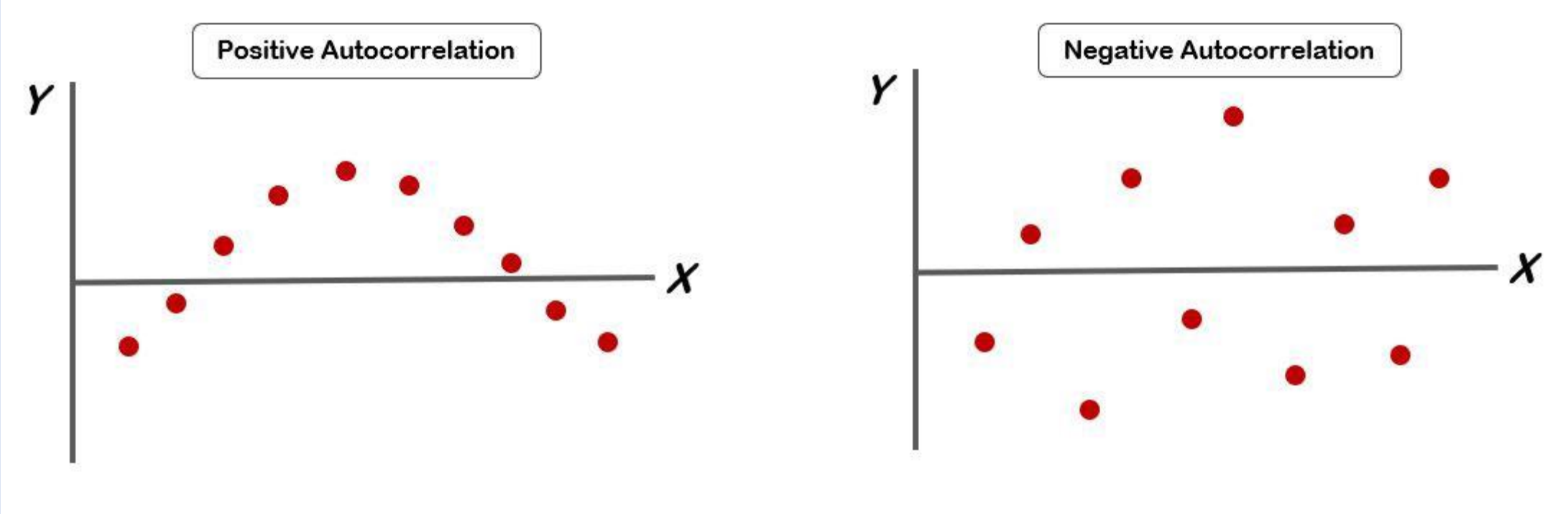
Homoscedasticity




40
20
0
−20
−40
−150
−100
−50
0
50
100
150
200

No Autocorrelation of error

- There should not be any pattern formation



No Autocorrelation of error

- There should not be any pattern formation