Multiple Linear Regression

$$Y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + b + \varepsilon$$

(i) x_1, x_2, x_3, x_4 are independent of each other

(ii) Plot , correlation

(iii) $VIF = \frac{1}{1-R^2}$

regressions:

$x_1 \sim x_2 + x_3 + x_4$

$x_2 \sim x_1 + x_3 + x_4$

$x_3 \sim x_1 + x_2 + x_4$

$x_4 \sim x_1 + x_2 + x_3$

R^2

if VLF calculated from each regression 20, the R^2 are not coming from normal.

VIF = Variance inflation factor

If VIF calculated > 20, that variable is dependent on others

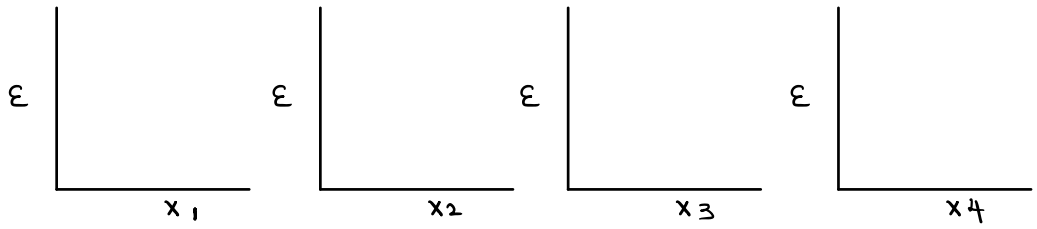
→ this dependency is called collinearity.

correlation : $$\frac{\Sigma(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(z_i - \bar{z})^2}}$$

If all VIF < 20

Second assumption : x_1, x_2, x_3 and x_4 $\Rightarrow$ Error is independent

calculate :



$\Rightarrow$ There shouldn't be any pattern in these plots. That is there shouldn't be any dependency.

If there is dependency found then we need to perform transformation on that data to remove the dependency structure.

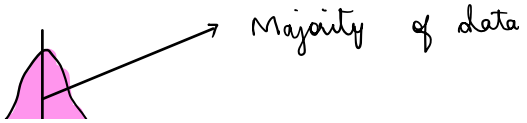
Assumptions

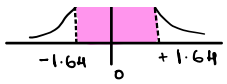
i) We take variables are Indepentent to remove multi-dependent-representation

ii) the variables are independent of error

iii) Errors are coming from normal with mean zero.

majority of data is close to zero.





→ How to prove error is coming from Normal distribution

$x_1, x_2, \ldots, x_n$ $\quad$ $y_1, y_2, \ldots, y_n$ $\quad$ Pred-y $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ $\longrightarrow$ $\varepsilon_1 \ldots \varepsilon_n \longrightarrow N(0, \sigma^2)$

$\Downarrow$ normalize

$\frac{\varepsilon_1 - 0}{\sigma^2}$

$\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots \tilde{\varepsilon}_n \rightarrow N(0,1) \rightarrow$ predicted

$z_1, z_2 \ldots z_n \rightarrow N(0,1) \rightarrow$ unpredicted

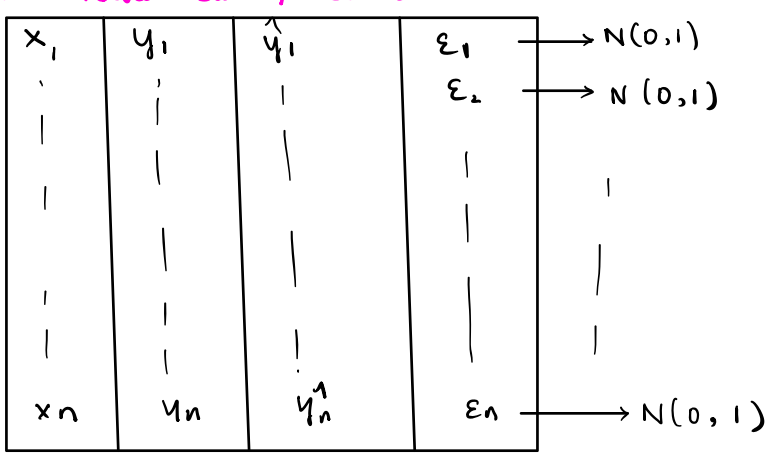
Calculate the quantiles and they should be same.

Transformation of Data ⇒ Box - Cox transformation.

5

Constant Variance (Homoscedasticity)

Its a property that all calculated Errors come from normal distributions and have same / constant Variance.



i.e when ε vs $\hat{y}_i$ is plotted, The graph should be random.

4) Partial Regression ⇒ Independent of each other

LR- Model & Assumptions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon$$

assumption

- Linear
  - Linear in parameters
- Errors
  - I I D Normal
  - Zero mean
  - Constant variance (Homoscedasticity)
  - Independent of each other
- Outlier detection
  - Box plot { 1-D data }
  - Plotting
    - Scatter plot ⇒ intuition
- Cooks plot
  - → remove a row at a time a plot
  - → plot all these, clearly an outlier will be visible.