Assumptions of LR -

1). Linearity
2) Normality of residuals
3) Homoscedasticity
4) No auto correlation
5) No or little multicollinearity

1) Linearity.

↳ There is linear relationship b/w dependent & independent var

→ what if assumption fails.

- Bias in parameter estimates
- Reduce predictive accuracy
- Invalid hypothesis test & confidence interval

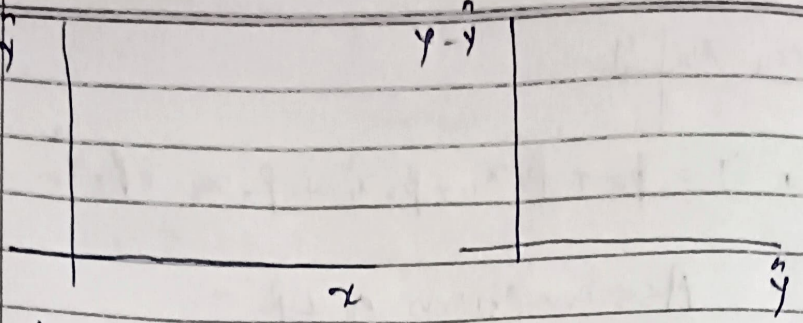
- How to check assumptions -

1) Scatter plot
   ↳ if 2D or 3D then we can see but if 4D or more then make scatter plot individually

x_1	x_2	x_3	Y

$x_1 \rightarrow Y$, $x_2 \rightarrow Y$, $x_3 \rightarrow Y$

2) Residual plot
   ↳ Residuals $Y - \hat{Y}$



↳ If linearity holds then there will be random scatter and if there is any kind of pattern then linearity fails.

3) Polynomial Regression -
↳ apply linear Reg and then check R^2 score
↳ apply polynomial . . . . .
if there is significant improvement means non-linearity

⇒ What if assumption fails -
i) Apply transformation
↳ on dep. & indep. variables
ii) Polynomial Reg -
↳ apply on indep. variable & check R^2 score.

2) Normality of Residuals -
↳ The error terms (residuals) are assumed to follow normal distribut^n with mean zero & constant variance

$$e_i \sim N(0, \sigma^2)$$

• what if assumptn fails
1) Inaccurate hypothesis test -
↳ F-tests & t-tests assume that residuals follow normal dist.

- If it doesn't follow then hypothesis tests may be inaccurate

2) Invalid confidence interval
3) Bad model performance

- How to check assumptions -

1) Plot histogram or kde plot for residual
   $X, Y, \hat{Y} \Rightarrow Y - \hat{Y}$
   ↳ hist() or kde
2) Draw Q-Q plot
3) Statistical test
   ↳ Shapiro wilk, omnibus, Jarque-Bera.

→ Omnibus test:-

1) H_0: residuals are normally dist
   H_a: " " not "

2) Fit linear Reg model
3) Calculate Residuals $(Y - \hat{Y})$
4) Calculate skewness
5) " kurtosis
6) $k^2 = n\left[\frac{(\text{skewness})^2}{6} + \frac{(\text{kurtosis})^2}{24}\right]$

n → no. of observers

k^2 follows χ^2 dist. So plot it in χ^2 graph (df = 2)

→ compare with p-value & decide whether reject or accept hypothesis

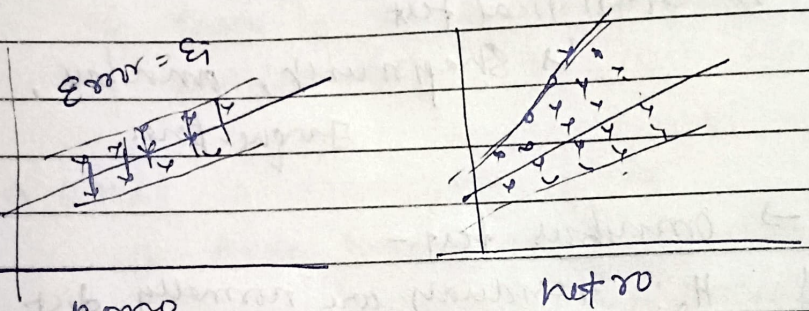
→ What to do when assumption fails -

- ↳ feature selection
- ↳ Robust regression
- ↳ Use Non-parametric methods
- ↳ Use bootstraping

[No need to worry if sample size > 30]

3) Homoscedasticity :-

↳ The spread of error terms should be constant across all level of indep. variable



if $Var(\varepsilon_i) = \sigma^2$ ↳ homo

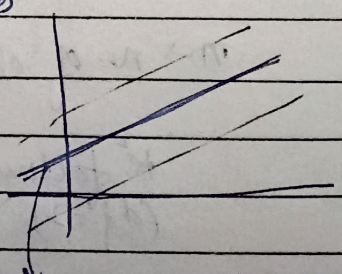
$Var(\varepsilon_i) = f(x)$ ↳ hetro

→ what happens if this fails -

- SE, b_0, b_1 will be unreliable
- t-test, f-test fails
- Invalid confidence interval.

Standard error → Standard dev of the sample means

$$SE = \frac{\sigma}{\sqrt{n}}$$



This line can be anywhere b/w the two lines