REGRESSION → Linear regression is a linear model that assume a linear relationship between input variables (x) and output variable (y). In other word, y can be calculated from a linear combination of input variable (x). [y=mx+b]

Simple linear regression – Single input variable, Multiple linear regression → More than 1 i/p var

Example of LR – Estimation of Bad loans in a bank. Bank has disbursed 100 loans in quarter. Additionally, we had 2.5% of bad rate so build regression model for estimating bad rate in next quarter.

Assumption in regression → MANHL, Multicollinearity, autocorrelation, normality, homoscedasticity, linear

MULTICOLLINEARITY

Collinearity – When one regressor is highly correlated with another regressor or in other word when one regressor is highly correlated with linear combination of other regressor. Eg – DOB and age.

Why collinearity is a problem – Each regressor is trying to "tell a story" about dependent variable. So if the regressor are correlated the story will be same. So if the story is same, but the regressor are different then model estimate will be confuse. In other word, p value becomes irrelevant and coefficients becomes very sensitive, can swing widly to small changes.

Multicollinearity → When multiple regressor are themselves correlated with each other.

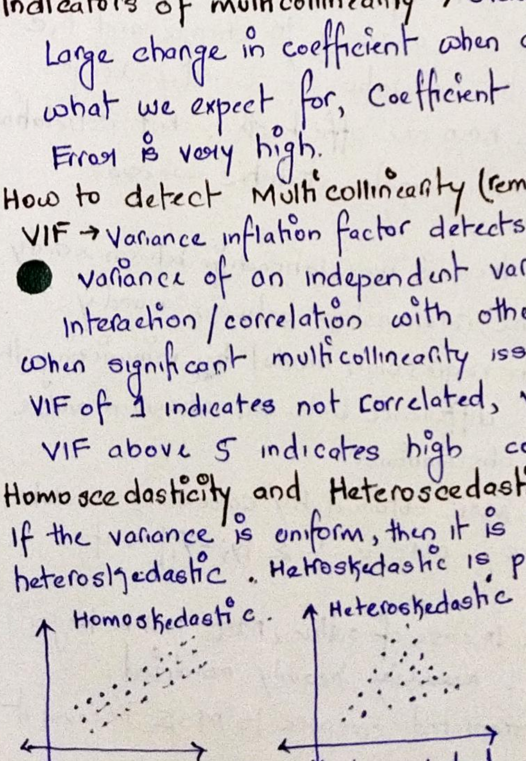
Main causes of multicollinearity → Two identical variables, Combination of two variable, Dummy variable may be incorrectly used.

Indicators of multicollinearity → Overall model is significant but none of the coefficient are. Coefficient have opposite sign what we expect for, Large change in coefficient when adding predictors, Coefficient on different samples are wildly different, Standard Error is very high.

How to detect Multicollinearity (remove) → Correlation and VIF.

VIF → Variance inflation factor detects degree of multicollinearity. It measure behaviour / variance of an independent variable., how much it will be inflated/influenced by its interaction / correlation with other independent variables.
when significant multicollinearity issue exist, VIF will be very large for variables.
VIF of 1 indicates not correlated, VIF between 1 and 5 indicates moderate collinearity, VIF above 5 indicates high correlation.

Homoscedasticity and Heteroscedasticity –
If the variance is uniform, then it is homoskedastic. If the variance is not uniform, it is heteroskedastic. Hetroskedastic is present when size of error term is different across values of an independent variable.



Scatterplot of Residual against predicted value of dependent variable across time.

Eg – Imagine relationship between family income and spending on luxury items. We found through regression, there is a strong, positive association between income and spending. When we observe residual are very small for low value but greater variation of residuals for wealthier families.

OLS gives equal weight to all observations, but when heteroskedasticity is present, larger disturbances lead to large residuals and leads to high standard error.
In these cases, weighted least square regression is more suitable.

Normality – When the sample size is sufficiently large (> 200), normality assumptions is not needed as Central limit theorem ensures that the distributions will be approx to normality. For small samples, sample stability become issue and in that case, use more conservative p values (0.01 rather than 0.05) for conducting significance test. And spread of error should be normally distributed

Linearity – Functional form should follow Linear equation i.e, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$
If the funchonal form is incorrect, both coefficient & standard error become unreliable

Autocorrelation – Auto correlation refers to the degree of correlation between values of the same variable across different observations in the data. It is the similarity between observations as a function of time lag between them.

Eg – One might expect sale of FD to be high during first two days of a month (suppose) compare to 30/31st end of the month. But if the values that occured are farther away and similar, then the data will be autocorrelated.

Problem → If we are attempting simple linear regression, but observed relationship is non-linear (follow curved / U shape / any shape) then residuals will be autocorrelated

Overall assumptions in Regression – MANBHL.

i) Multicollinearity – No or little multicollinearity. Independence variable should not be correlated.

ii) Autocorrelation – No correlation between residual (error) terms.

iii) Normality – Error term must be normally distributed.

iv) Homoscedasticity – Error term must have constant variance.

v) Linear – Linear relationship between independent and dependent variable.

Correlations vs Regression – Regression establishes how x causes y to change and the result will change if x and y are interchanged. With correlation, if x and y changes to see how one affect other but correlation result will be same. Regression allow us show relationship between two variables. Correlation is single statistic whereas regression is entire equation.

Null hypothesis of LR – if $y = \beta_0 + \beta_1 x$ null is $\beta_1 = 0$, there is no relationship between x and y.
Alternate → $\beta_1 \neq 0$, different value from 0 shows some relationship between x and y.

OLS – Ordinary least squares, estimates parameter in regression model by minimizing the sum of squared residuals. Residuals means difference between observed value and mean values that model predicts for that observations.

Model evaluation metrics – i) MAE (Mean Absolute error) – MAE obtained by calculating absolute differences between model prediction & actual values. $MAE = \frac{1}{n} \Sigma |y_i - \hat{y}_i|$. If MAE = 0, the model is perfect.

ii) MSE (Mean Square error) – $MSE = \frac{1}{n} \Sigma (y_i - \hat{y}_i)^2$. In case of outlier, MSE will be larger. Since error is squared any predicting error is being heavily penalized.

iii) RMSE (Root mean square error) – RMSE is easily intreprated compare to MSE because it match unit of output. $RMSE = \sqrt{\frac{1}{n} \Sigma (y_i - \hat{y}_i)^2}$

iv) MAPE (Mean Absolute Percentage error) – MAE ranges from 0 to ∞, so MAPE provides error in %.

$$MAPE = \frac{100\%}{n} \Sigma \left| \frac{y_i - \hat{y}}{y_i} \right|$$