INTERVIEW DAY - CORRELATION, VIF (~~Variation~~ Variance Inflation factor)

Q) How to remove correlation among variables?

Ans - If they are correlated, they are correlated. We cannot remove a correlation.

To handle this, we can handle in two ways -

I) One is to choose one variable from each highly correlated pair. For example, age OR experience. Choose based on which one is more logically connected to what we are trying to predict (For example, suppose we want in domain experience, therefore total experience may be more but in domain experience may be less, so in domain experience is more logically connected) or else, go with the one that correlates most strongly with outcome variable.

II) The other option is to create a new variable by combining them. First we need to convert two variable (age and experience) onto similar scale then summing them and create dummy variable.

Q) What is VIF / Variation Inflation Factor?

Ans - VIF detects multicollinearity in regression analysis.

- Multicollinearity is when there's correlation between predictors (independent variables) in a model, its presence can adversely affect regression result.
- VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

Q) ~~How the VIF is computed?~~ How to estimate Standard Error in regression?

Ans - The standard error of an estimate in a linear regression is determined by 4 things -

I) Overall amount of noise (error). More the noise in data, higher standard error.

II) Variance of the associated predictor variable. Greater the variance of predictor, smaller the standard error (Scaling effect).

III) Sampling mechanism used to obtain the data. For example, smaller the sample size with a simple random sample, bigger the standard error.

IV) Extend to which a predictor is correlated with other predictor in a model.

Q) How VIF is computed?

Ans - The extend to which a predictor is correlated with other predictor variables in a linear regression can be quantified as R squared statistic of the regression where the predictor of interest (target variable) is predicted by all other predictor variable. (Independent variable)

Therefore VIF is computed as $\boxed{VIF = \frac{1}{1-R^2}}$

Q) How to interpret the VIF?

Ans - VIF can be computed for each predictor in a predictive model.

- Higher the value, greater the correlation of variable with other variable.

VIF value = 1, means predictor is not correlated with other variable.

VIF = Between 1 and 5 = moderately correlated.

VIF = Greater than 5 = highly correlated

These numbers are just thumb rules, in some context even VIF=2 could be problem.

If one variable has a high VIF it means other variable must also have high VIFs. In simplest cases, two variable will be highly correlated & each will have same high VIF.

- Higher the VIF, more the Standard Error is inflated and larger the confidence interval and smaller the chance that a coefficient is determined to be statistically significant.

Q) What are the remedies of VIF?

Ans - When VIF is too high for variable, solutions are -

I) Obtain more data to reduce standard error.

II) Recode the predictor in a way that reduce correlation (choose one variable/ create a dummy variable).