Twitter Thread by Prashant

Prashant
@capeandcode

Checking for multicollinearity is a routine task while solving a data problem.

Most people rely on pairwise correlations for detecting multicollinearity which is a faulty approach in many scenarios

So what to do? Check this ↓

1/11

Multicollinearity creates a problem in the interpretation of the model when some predictors explain predictors.

Because then we are unable to understand the effect of each predictor in isolation towards the target variable and our coefficients become less useful.

2/11

Variance Inflation Factor or VIF is an efficient method to check for multicollinearity.

Variance Inflation refers to the inflation in the variance of the estimated coefficient of the independent variable because of the presence of multicollinearity.

3/11

We could have low pairwise correlations, but still have high VIF and vice-versa.

A strong relationship is possible between a predictor and other variables together combined, even though there's no high correlation individually.

4/11

Hence we prefer VIF

VIF performs a set of multivariate regression analyses to check the dependence among the independent variables by fitting multiple regression models on the dependent variables.

5/11

In each fit, one of the variables is treated as a target while the rest of them act as regressors.

If we have regressors A,B,C,D, then VIF would fit models like,

A <- B,C,D
B <- A,C,D
C <- A,B,D
D <- A,B,C

and since we are looking for explained variance, we would use...

6/11

...R2 as a metric for each model.

We will calculate the VIF for each predictor by using the R2 given by the model where the predictor has been used as the target.

Putting the values in the following formula we will get our VIF values:

7/11

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

The high R2 would show that other predictors explain most of the variance in the current predictor (current model target).

And in the formula, high R2 would yield high VIF values as well, suggesting multicollinearity.

8/11

Now that we got the calculated VIF values what to do with them?

VIF values start from 1 and do not have an upper limit.

1 suggests that there's no multicollinearity among the variables,
as a rule-of-thumb, values > 5 or > 10 indicate high multicollinearity but not always.

9/11

Once we know the correlated variables, for removing the multi-collinearity we can,

• either remove all but one of the variables with high VIF or
• combine those variables into a single variable.

and we are good to go!

10/11

That's one good way to check for multicollinearity!

11/11

If you do find a mistake or have any questions, drop them below ↓