Multicollinearity

**Introduction:**

Multicollinearity (also collinearity) is a phenomenon in which two or more predictor [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) (Independent variables) in a [regression](https://en.wikipedia.org/wiki/Multiple_regression) model are highly [correlated](https://en.wikipedia.org/wiki/Correlation_and_dependence), meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the [coefficient estimates](https://en.wikipedia.org/wiki/Regression_coefficient) of the multiple regression may change erratically in response to small changes in the model or the data. Collinearity is a linear association between two [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable). Two variables are perfectly collinear if there is an exact linear relationship between them.

**Types of multicollinearity:**

There are two types of multicollinearity:

1. Structural multicollinearity is a mathematical artifact caused by creating new predictors from other predictors — such as, creating the predictor *x*2 from the predictor *x*.
2. Data-based multicollinearity, on the other hand, is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

**Detect Multicollinearity:**

Indicators that multicollinearity may be present in a model include the following:

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted
2. Insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero (using an [F-test](https://en.wikipedia.org/wiki/F-test))
3. If a multivariable regression finds an insignificant coefficient of a particular explanator, yet a [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression) of the explained variable on this explanatory variable shows its coefficient to be significantly different from zero, this situation indicates multicollinearity in the multivariable regression.
4. VIF (Variance inflation factor) can be used to detect multicollinearity in the regression model{\displaystyle \mathrm {tolerance} =1-R\_{j}^{2},\quad \mathrm {VIF} ={\frac {1}{\mathrm {tolerance} }},}

**Why is this problem?**

* Collinearity tends to inflate the variance of at least one estimated regression coefficient.
* This can cause at least some regression coefficients to have the wrong sign.

**Ways of dealing with collinearity**

* Ignore it. If prediction of y values is the object of your study, then collinearity is not a problem.
* Get rid of the redundant variables by using variable sélection technique.

There are multiple techniques to select variables which are less correlated with high importance

1. Correlation method
2. PCA (Principal Component Analysis)
3. SVD (Singular value Decomposition)
4. Machine learning algorithms (Random Forest, Decision trees)

## Remedies for multicollinearity

1. Drop one of the variables. An explanatory variable may be dropped to produce a model with significant coefficients. However, you lose information (because you've dropped a variable). Omission of a relevant variable results in biased coefficient estimates for the remaining explanatory variables that are correlated with the dropped variable.
2. Obtain more data, if possible. This is the preferred solution. More data can produce more precise parameter estimates (with lower standard errors), as seen from the formula in [variance inflation factor](https://en.wikipedia.org/wiki/Variance_inflation_factor) for the variance of the estimate of a regression coefficient in terms of the sample size and the degree of multicollinearity.
3. Try seeing what happens if you use independent subsets of your data for estimation and apply those estimates to the whole data set. Theoretically you should obtain somewhat higher variance from the smaller datasets used for estimation, but the expectation of the coefficient values should be the same. Naturally, the observed coefficient values will vary, but look at how much they vary.
4. Standardize your independent variables. This may help reduce a false flagging of a condition index above 30.
5. It has also been suggested that using the [Shapley value](https://en.wikipedia.org/wiki/Shapley_value), a game theory tool, the model could account for the effects of multicollinearity. The Shapley value assigns a value for each predictor and assesses all possible combinations of importance.
6. If the correlated explanators are different lagged values of the same underlying explanator, then a [distributed lag](https://en.wikipedia.org/wiki/Distributed_lag) technique can be used, imposing a general structure on the relative values of the coefficients to be estimated.

Interview Questions:

# What is multicollinearity and what techniques would you use to solve for it in a model?

1. What is VIF?