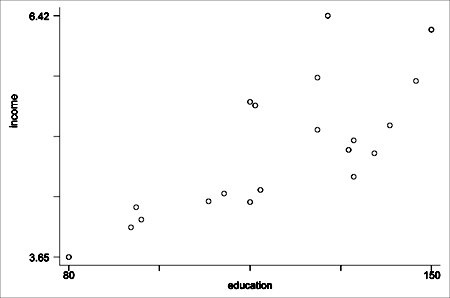
**Multicollinearity**– A Brief Research Paper by Ali A Nathani – Data Science & Application Course – Metro College of Technology – Instructor: Vijay Kumar

Multicollinearity is when there are very high correlations between two or more predictor variables, or when “one independent variable can be linearly predicted from one or multiple other independent variables with a substantial degree of certainty”1. The result of this occurrence would be redundant information that would skew the results in a regression model, due to unreliable statistical inferences.   
  
To understand multicollinearity, we need to understand the reasons why it occurs, how to measure it, what are some results and possibly some examples that help support the explanation of the definition.

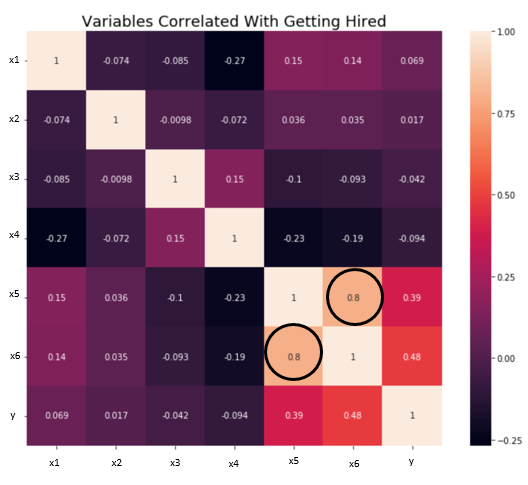
Causes:  
“**There are certain reasons why multicollinearity occurs**:

It is caused by an inaccurate use of dummy variables.  
It is caused by the inclusion of a variable which is computed from other variables in the data set.  
Multicollinearity can also result from the repetition of the same kind of variable.  
Generally occurs when the variables are highly correlated to each other.”2  
Finally, it can be caused by insufficient data.

What are the problems that can result when one variable can accurately predict another variable or are highly correlated and both variables are used in calculating regression of yet another variable?

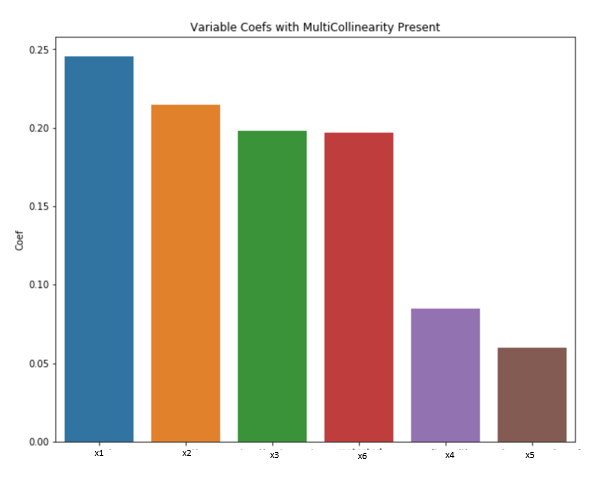
  
**“Multicollinearity can adversely affect your regression results.”3.**  
  
Some of the Problems that can occur with multicollinearity are:  
Inaccurate results when calculating partial regression coefficient2;  
A change in the signs as well as the magnitudes of the partial regression coefficients from one sample to another2;  
The multicollinear variables together can cause a variation in the predictor variable, but it is difficult to explain the variation caused by these dependent variables due to their multicollinearity;  
“It becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data under study”2.

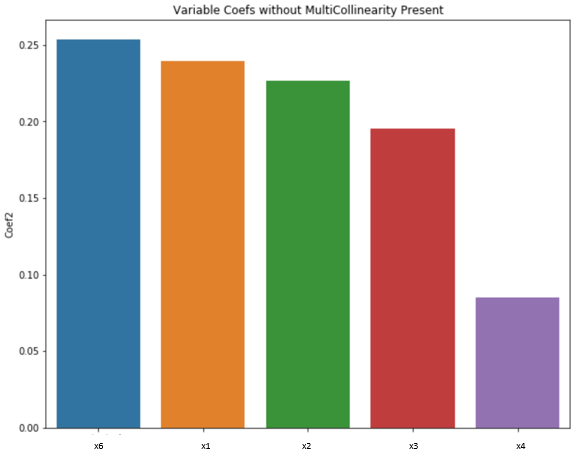
In other words, it reduces “the quality of the interpretation of the independent variables”, making the “explanation of how the model takes inputs to produce the outputs”1, unreliable.  
  
 Detecting or Measuring Multicollinearity:

In order to detect multicollinearity, one can calculate the correlation coefficients for all the pairs of the predictor variables and “if the correlation coefficient, r, is exactly +1 or -1, this is called perfect multicollinearity. If r is close to or exactly -1 or +1, one of the variables should be removed from the model if at all possible.”3 Using VIF, Variance Inflation Factor as a measure of correlation, if there is no correlation VIF will be 1 and if the predictors are correlated, then the variance of an estimated regression coefficient increases.   
Also, a visual method to detect multicollienarity is a correlation matrix or heat maps, with a scale of 0 – 1 with 1 being perfectly correlated1:  
 Correlation Heat Map

Solutions:  
1. Feature Engineering: Combine 2 correlated variables using feature engineering or aggregate to turn them into one variable.

2. Drop One of the Variables that is highly correlated, as a general rule, by dropping the one that is not as strongly correlated with the target variable.

Examples  
  
Feature Coefficients with all variables present1

  
Feature Coefficients after removing Multicollinearity1

In the above example: after you drop one of the variables (the one least correlated with the target), the coefficients do in fact change.  
  
**Conclusion:**

Multicollinearity may not affect the accuracy of the model or the prediction. However, it will affect the interpretation or analysis of how the prediction was achieved. Hence, it is better to discover if multicollinearity exists and to deal with it using the appropriate methods.