Multicollinearity
Studenmund Chapter 8

Definition
Multicollinearity is a violation of the assumption that no independent variable is a linear function of one or more other independent variables.

Basically, this means that two or more of your explanatory variables are very closely related and, as a result, you can’t distinguish the effect of one from the effect of the other.

For example, if you do some analysis to look at athletic performance, you might try to separate out the effects of heat and humidity. However, if in your observations, heat and humidity are highly correlated (that is, if the hot days also tended to be humid) then you really can’t tell how much of the impact is due to the heat and how much is due to the humidity. You can tell what the combined impact is, but you can’t separate out the effects of the two.

According to Studenmund, “The word collinearity describes a linear correlation between two independent variables, and multicollinearity indicates that more than two independent variables are involved. In common usage, multicollinearity is used to apply to both cases, and so we’ll typically use that term in this text, even though many of the examples and techniques discussed relate, strictly speaking, to collinearity.”

I find it a bit disturbing that these terms come up in “common usage.”

To distinguish multicollinearity from endogeneity, consider that endogeneity presumes some functional relationship between explanatory variables and/or the dependent variable in addition to the relationship that is to be estimated. Multicollinearity suggests that one or more of the explanatory variables are linearly related in your sample, but that there is no real causal relationship between them.

Multicollinearity may be perfect or imperfect.
Perfect multicollinearity means that one explanatory variable is an exact linear function of one or more explanatory variables with no error term. Imperfect multicollinearity means that there is a linear relationship between the variables, but there is some error in that relationship.

If collinearity is perfect, the mathematics that underlie regression analysis will fail because the matrix $X'X$ will not be invertible.

In the case of perfect or near perfect collinearity or multicollinearity, one of the involved variables may be automatically excluded from the analysis by the software package you are using. If you find that your software is kicking out one of your explanatory variables, some sort of collinearity may be to blame.

**Consequences of Multicollinearity**

1. Estimates are unbiased
   Unlike the omitted variable problem, coefficient estimates will be unbiased even if there is multicollinearity. This means that the expected value of the estimator of $\beta_K$ is equal to $\beta_K$:
   $$E[\hat{\beta}_K] = \beta_K$$
   So, biased estimates are not a problem with multicollinearity.

2. Standard errors of estimates will increase/t-stats will decrease
   This means that variables that play a significant role in explaining the value of the dependent variable won't have estimated coefficients that are significantly different from zero.

3. Estimates will be very sensitive to changes in specification of the model
   Dropping a variable or even excluding some observations may lead to large changes in estimated coefficients, leading to an important lack of confidence in the robustness of estimates. This effect is due to the high correlation between explanatory variables. Thinking back to the effects of excluded variable bias, the bias will be large when explanatory variables are correlated.

4. Explanatory power will be unaffected
   The nature of the problem with multicollinearity is that it becomes difficult or impossible to separate the effects of changes in individual variables. It is, however, not a problem to come up with good predictions given the complete set of explanatory variables, even if their effects cannot be separated.

**Detection of Multicollinearity**

1. High $R^2$ or adjusted $R^2$ and insignificant t-scores
This is the most evident sign. You have a model with good explanatory power but very few or no significant estimated coefficients. This situation should scream “Multicollinearity!” to you.

2. High correlation coefficients between explanatory variables
As we’ve discussed, you might want to look at a table of correlation coefficients to discover if multicollinearity might be a problem. This will mostly detect linear relationships between pairs of variables.

Studenmund offers 0.80 as a limiting value. If the correlation coefficient between two variables is much less than this, there’s not a problem. As it gets closer to 0.80 (or -0.80) then you need to be more concerned that multicollinearity might be a problem.

You're all intelligent enough to understand that there's no exact point at which multicollinearity becomes a problem. Rather it is more or less of a problem depending on the exactness of the linear relationship between explanatory variables.

3. Variance Inflation Factors (VIFs)
These are pretty cool. They're equal to the $1/(1-R^2)$ where the $R^2$ is that resulting from a regression of one of the explanatory variables on the other explanatory variables. Each explanatory variable will have its own VIF resulting from regressing it on the other independent variables.

To be clear about this, to calculate a VIF for one of the explanatory variables, $X_1$, you need to do the regression:

$$X_{1i} = \alpha_0 + \alpha_1X_{2i} + \alpha_2X_{3i} + \alpha_3X_{4i} + \ldots + \zeta_i$$

Then take the $R^2$ from this regression to calculate the VIF $= \frac{1}{1-R^2}$ You can calculate a VIF for every explanatory variable.

To get VIFs in SPSS, choose Statistics/ColinearityDiagnostics. You will get some additional output, but in the table with the estimated coefficients and t-stats, you'll get a column with VIFs for each explanatory variable. A VIF of greater than maybe 5 or 6 suggests that that variable exhibits some sort of collinearity.

The advantage of VIFs over correlation coefficients is that VIFs allow the collinearity to be between more than two explanatory variables while correlation coefficients allow only for bivariate (two variable) relationships.

**Remedies for Multicollinearity**
1. Do nothing
Your equation will have good explanatory power and suffer no problems other than artificially high standard errors if the only problem is multicollinearity. Coefficient estimates will be unbiased.
If you drop a variable, you may run into the problems associated with omitted variables. If you try to focus on the effect of one variable by removing another with which it is correlated, you will artificially inflate its effects.

2. Drop one or more of the collinear variables
If you do this, you run the risk of having an omitted variable bias. To review, omitted variables may make the estimated coefficients biased and the standard errors too small, and this problem is worse than collinearity. As a result, you shouldn’t do this.

3. Transform the variables
Studenmund describes this. It needs to be done carefully.

4. Get more observations for which the explanatory factors are not correlated
For the example of separating out the effects of heat and humidity on athletic performance, the problem of heat and humidity being collinear could be addressed by getting observations from hot, dry days and from cold, damp days.

**Conclusion**
Multicollinearity makes it impossible to distinguish the differential effects of correlated explanatory variables.

The estimated models can have good predictive power, even though no estimated coefficients are significantly different from zero.

Attempts to correct multicollinearity may often make things worse by introducing biases.