**Readings on Multicollinearity**

* **What is multicollinearity?**
* Collinearity (or multicollinearity) is the undesirable situation where the correlations among the independent variables are strong.

In some cases, multiple regression results may seem paradoxical. For instance, the model may fit the data well (high F-Test), even though none of the X variables has a statistically significant impact on explaining Y. How is this possible? When two X variables are highly correlated, they both convey essentially the same information.  When this happens, the X variables are *collinear* and the results show *multicollinearity*.

* To help you assess multicollinearity, SPSS tells you the Variance Inflation Factor (VIF) that measures multicollinearity in the model.
* **Why is multicollinearity a problem?**

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0, whereas without multicollinearity and with lower standard errors, these same coefficients might have been found to be significant and the researcher may not have come to null findings in the first place.

**In other words**, multicollinearity misleadingly inflates the standard errors. Thus, it makes some variables statistically insignificant while they should be otherwise significant.

**It is like two or more people singing loudly at the same time. One cannot discern which is which. They offset each other.**

**How to detect multicollinearity?**

*Formally, variance inflation factors* (VIF) measure how much the variance of the estimated coefficients are increased over the case of no correlation among the X variables. If no two X variables are correlated, then all the VIFs will be 1.

If VIF for one of the variables is around or greater than 5, there is collinearity associated with that variable.

The easy solution is: If there are two or more variables that will have a VIF around or greater than 5, one of these variables must be removed from the regression model.

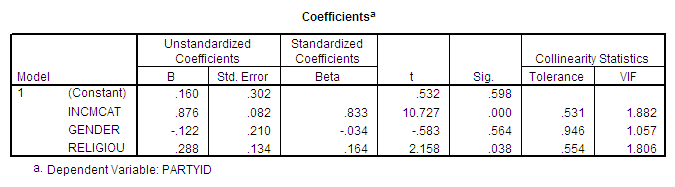
To determine the best one to remove, remove each one individually. Select the regression equation that explains the most variance (R2 the highest).

How to get VIF:

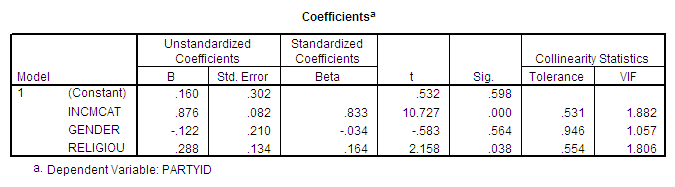
In SPSS Regression dialogue box: Select statistics button;

Check collinearity diagnostics in window.   
Example: Download the following file: [collinearity.sav](http://www.chsbs.cmich.edu/fattah/courses/msa600/data/collinearity.sav).

Let us assume that the variable "RELIGIO2" is another measurement of religiosity beside the one that was already there "RELIGIOU." When one puts both of them together in the same model, none of them is statistically significant. The VIF is above 5 which means that multicollinearity inflated the standard errors which lowers the T test below 2 which means that the significance level becomes above .05.

****

**Remedy:** Let us delete  "RELIGIO2" from the model. Religiosity becomes statistically significant.

****

Other informal signs of multicollinearity are:

* Regression coefficients change drastically when adding or deleting an X variable.
* A regression coefficient is negative when theoretically Y should increase with increasing values of that X variable, or the regression coefficient is positive when theoretically Y should decrease with increasing values of that X variable.
* None of the individual coefficients has a significant t statistic, but the overall F test for fit is significant.
* A regression coefficient has a nonsignificant t statistic, even though on theoretical grounds that X variable should provide substantial information about Y.
* High pairwise correlations between the X variables. (But three or more X variables can be multicollinear together without having high pairwise correlations.)

* **What can be done to handle multicollinearity?**
* 1.     Increasing the sample size is a common first step since when sample size is increased, standard error decreases (all other things equal). This partially offsets the problem that high multicollinearity leads to high standard errors of the b and beta coefficients.
* 2.      **The easiest solution:** Remove the most intercorrelated variable(s) from analysis. This method is misguided if the variables were there due to the theory of the model, which they should have been.
* 3.     Combine variables into a composite variable through building indexes such as the one we did for religiosity through factor analysis. Remember: creating an index theoretical and empirical reasons to justify this action.
* 4.     Use **centering**: transform the offending independents by subtracting the mean from each case. The resulting centered data may well display considerably lower multicollinearity. You should have a theoretical justification for this consistent with the fact that a zero b coefficient will now correspond to the independent being at its mean, not at zero, and interpretations of b and beta must be changed accordingly.
* 5.     Drop the intercorrelated variables from analysis but substitute their crossproduct as an interaction term, or in some other way combine the intercorrelated variables. This is equivalent to respecifying the model by conceptualizing the correlated variables as indicators of a single latent variable. Note: if a correlated variable is a dummy variable, other dummies in that set should also be included in the combined variable in order to keep the set of dummies conceptually together.
* 6.     Leave one intercorrelated variable as is but then remove the variance in its covariates by regressing them on that variable and using the residuals.
* 7.     Assign the common variance to each of the covariates by some probably arbitrary procedure.
* 8.     Treat the common variance as a separate variable and decontaminate each covariate by regressing them on the others and using the residuals. That is, analyze the common variance as a separate variable.