Multicollinearity

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised January 13, 2015

Stata Example (See appendices for full example).

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/multicoll.dta, clear
. reg  y x1 x2, beta

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =    3.24
       Model |  55.7446181     2   27.872309           Prob > F      =  0.0436
    Residual |  835.255433    97  8.61088075           R-squared     =  0.0626
-------------+------------------------------           Adj R-squared =  0.0432
       Total |  891.000051    99  9.00000051           Root MSE      =  2.9344

------------------------------------------------------------------------------
           y |      Coef.  Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   .0153846   .1889008     0.08   0.935                  .025641
          x2 |   .1353847   .1889008     0.72   0.475                 .2256411
       _cons |   10.49231   .6655404    15.77   0.000                        .
------------------------------------------------------------------------------

. corr y x1 x2, means

(obs=100)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
           y |           12            3     4.899272     18.91652
          x1 |           10            5    -1.098596     23.10749
          x2 |           10            5    -.0284863     23.72392


             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.2400   1.0000
          x2 |   0.2500   0.9500   1.0000
```

Note that

- The t-statistics for the coefficients are not significant. Yet, the overall F is significant.
- Even though both IVs have the same standard deviations and almost identical correlations with Y, their estimated effects are radically different.
- X1 and X2 are very highly correlated (r_{12} = .95).
- The N is small

These are all indicators that *multicollinearity* might be a problem in these data. (See the appendices for more ways of detecting problems using Stata.)

What multicollinearity is. Let H = the set of all the X (independent) variables. Let G_k = the set of all the X variables *except* X_k. The formula for standard errors is then

$$s_{b_k} = \sqrt{\frac{1-R^2_{YH}}{(1-R^2_{X_kG_k})*(N-K-1)}} * \frac{s_y}{s_{X_k}}$$

$$= \sqrt{\frac{1-R^2_{YH}}{Tol_k*(N-K-1)}} * \frac{s_y}{s_{X_k}}$$

$$= \sqrt{Vif_k} * \sqrt{\frac{1-R^2_{YH}}{(N-K-1)}} * \frac{s_y}{s_{X_k}}$$

Questions: What happens to the standard errors as R^2_{YH} increases? As N increases? As the multiple correlation between one DV and the others increases? As K increases?

From the above formulas, it is apparent that

- The bigger R^2_{YH} is, the smaller the standard error will be.
- Larger sample sizes decrease standard errors (because the denominator gets bigger). This reflects the fact that larger samples will produce more precise estimates of regression coefficients.
- The bigger R^2_{XkGk} is (i.e. the more highly correlated X_k is with the other IVs in the model), the bigger the standard error will be. Indeed, if X_k is perfectly correlated with the other IVs, the standard error will equal infinity. This is referred to as the problem of *multicollinearity*. The problem is that, as the Xs become more highly correlated, it becomes more and more difficult to determine which X is actually producing the effect on Y.
    - Also, 1 - R^2_{XkGk} is referred to as the *Tolerance* of X_k. A tolerance close to 1 means there is little multicollinearity, whereas a value close to 0 suggests that multicollinearity may be a threat.
    - The reciprocal of the tolerance is known as the *Variance Inflation Factor (VIF)*. The VIF shows us how much the variance of the coefficient estimate is being inflated by multicollinearity. For example, if the VIF for a variable were 9, its standard error would be three times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 3 times as large to be statistically significant.
- Adding more variables to the equation can increase the size of standard errors, especially if the extra variables do not produce increases in R^2_{YH}. Adding more variables decreases the (N-K-1) part of the denominator. More variables can also decrease the tolerance of the variable and hence increase the standard error. In short, adding extraneous variables to a model tends to reduce the precision of all your estimates.

Causes of multicollinearity

- Improper use of dummy variables (e.g. failure to exclude one category)

- Including a variable that is computed from other variables in the equation (e.g. family income = husband's income + wife's income, and the regression includes all 3 income measures)
- In effect, including the same or almost the same variable twice (height in feet and height in inches; or, more commonly, two different operationalizations of the same identical concept)
- The above all imply some sort of error on the researcher's part. But, it may just be that variables really and truly are highly correlated.

Consequences of multicollinearity

- Even extreme multicollinearity (so long as it is not perfect) does not violate OLS assumptions. OLS estimates are still unbiased and BLUE (Best Linear Unbiased Estimators)
- Nevertheless, the greater the multicollinearity, the greater the standard errors. When high multicollinearity is present, confidence intervals for coefficients tend to be very wide and t-statistics tend to be very small. Coefficients will have to be larger in order to be statistically significant, i.e. it will be harder to reject the null when multicollinearity is present.
- Note, however, that large standard errors can be caused by things besides multicollinearity.
- When two IVs are highly and *positively* correlated, their slope coefficient estimators will tend to be highly and *negatively* correlated. When, for example, b_1 is greater than β_1, b_2 will tend to be less than β_2. Further, a different sample will likely produce the opposite result. In other words, if you overestimate the effect of one parameter, you will tend to underestimate the effect of the other. Hence, coefficient estimates tend to be very shaky from one sample to the next.

Detecting high multicollinearity. Multicollinearity is a matter of degree. There is no irrefutable test that it is or is not a problem. But, there are several warning signals:

- None of the t-ratios for the individual coefficients is statistically significant, yet the overall F statistic is. If there are several variables in the model, though, and not all are highly correlated with the other variables, this alone may not be enough. You could get a mix of significant and insignificant results, disguising the fact that some coefficients are insignificant because of multicollinearity.
- Check to see how stable coefficients are when different samples are used. For example, you might randomly divide your sample in two. If coefficients differ dramatically, multicollinearity may be a problem.
- Or, try a slightly different specification of a model using the same data. See if seemingly "innocuous" changes (adding a variable, dropping a variable, using a different operationalization of a variable) produce big shifts.
- In particular, as variables are added, look for changes in the signs of effects (e.g. switches from positive to negative) that seem theoretically questionable. Such changes may make sense if you believe suppressor effects are present, but otherwise they may indicate multicollinearity.
- Examine the bivariate correlations between the IVs, and look for "big" values, e.g. .80 and above. However, the problem with this is

- ✓ One IV may be a linear combination of several IVs, and yet not be highly correlated with any one of them
- ✓ Hard to decide on a cutoff point. The smaller the sample, the lower the cutoff point should probably be.
- ✓ Ergo, examining the tolerances or VIFs is probably superior to examining the bivariate correlations. Indeed, you may want to actually regress each X on all of the other X's, to help you pinpoint where the problem is. A commonly given rule of thumb is that VIFs of 10 or higher (or equivalently, tolerances of .10 or less) may be reason for concern. This is, however, just a rule of thumb; Allison says he gets concerned when the VIF is over 2.5 and the tolerance is under .40. In Stata you can use the `vif` command after running a regression, or you can use the `collin` command (written by Philip Ender at UCLA).

- Look at the correlations of the estimated coefficients (not the variables). High correlations between pairs of coefficients indicate possible collinearity problems. In Stata you get it by running the `vce, corr` command after a regression.
- Sometimes condition numbers are used (see the appendix). An informal rule of thumb is that if the condition number is 15, multicollinearity is a concern; if it is greater than 30 multicollinearity is a very serious concern. (But again, these are just informal rules of thumb.) In Stata you can use `collin`.

Dealing with multicollinearity

- Make sure you haven't made any flagrant errors, e.g. improper use of computed or dummy variables.
- Increase the sample size. This will usually decrease standard errors, and make it less likely that results are some sort of sampling "fluke."
- Use information from prior research. Suppose previous studies have shown that $\beta_1 = 2*\beta_2$. Then, create a new variable, $X_3 = 2X_1 + X_2$. Then, regress Y on X_3 instead of on X_1 and X_2. b_3 is then your estimate of β_2 and $2b_3$ is your estimate of β_1.
- Use factor analysis or some other means to create a scale from the X's. It might even be legitimate just to add variables together. In fact, you should do this anyway if you feel the X's are simply different operationalizations of the same concept (e.g. several measures might tap the same personality trait). In Stata relevant commands include `factor` and `alpha`.
- Use joint hypothesis tests—instead of doing t-tests for individual coefficients, do an F test for a group of coefficients (i.e. an incremental F test). So, if X1, X2, and X3 are highly correlated, do an F test of the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$.
- It is sometimes suggested that you "drop" the offending variable. If you originally added the variable "just to see what happens," dropping may be a fine idea. But, if the variable really belongs in the model, this can lead to specification error, which can be even worse than multicollinearity.
- It may be that the best thing to do is simply to realize that multicollinearity is present, and be aware of its consequences.

Appendix: Stata example.

We use the `corr`, `regress`, `vif`, `vce`, and `collin` commands.

```
. use https://www3.nd.edu/~rwilliam/statafiles/multicoll.dta, clear
. corr y x1 x2, means

(obs=100)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
           y |           12            3     4.899272     18.91652
          x1 |           10            5    -1.098596     23.10749
          x2 |           10            5    -.0284863     23.72392


             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.2400   1.0000
          x2 |   0.2500   0.9500   1.0000

. reg  y x1 x2, beta

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =    3.24
       Model |  55.7446181     2  27.872309           Prob > F      =  0.0436
    Residual |  835.255433    97  8.61088075           R-squared     =  0.0626
-------------+------------------------------           Adj R-squared =  0.0432
       Total |  891.000051    99  9.00000051           Root MSE      =  2.9344

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   .0153846   .1889008     0.08   0.935                  .025641
          x2 |   .1353847   .1889008     0.72   0.475                 .2256411
       _cons |   10.49231   .6655404    15.77   0.000                        .
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          x1 |     10.26    0.097500
          x2 |     10.26    0.097500
-------------+----------------------
    Mean VIF |     10.26

. vce, corr

             |       x1       x2    _cons
-------------+---------------------------
          x1 |   1.0000
          x2 |  -0.9500   1.0000
       _cons |  -0.1419  -0.1419   1.0000
```

Note that

- X1 and X2 are very highly correlated (r_{12} = .95). Of course, the tolerances for these variables are therefore also very low and the VIFs exceed our "rule of thumb" of 10.

- The t-statistics for the coefficients are not significant. Yet, the overall F is significant.
- Even though both IVs have the same standard deviations and almost identical correlations with Y, their estimated effects are radically different.
- The correlation between the coefficients for X1 and X2 is very high, -.95
- The sample size is fairly small (N = 100).
- The condition number (reported below) is 16.964. This falls within our "rule of thumb" range for concern. Again, this is based on the uncentered variables; if I thought centering was more appropriate I would just need to change the means of X1 and X2 to 0. (Doing so produces a condition number of 6.245, as Stata confirms below.)

All of these are warning signs of multicollinearity. A change of as little as one or two cases could completely reverse the estimates of the effects.

```
. * Use collin with uncentered data, the default. (Same as SPSS)
. collin x1 x2 if !missing(y)

  Collinearity Diagnostics

                        SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
        x1     10.26    3.20    0.0975      0.9025
        x2     10.26    3.20    0.0975      0.9025
----------------------------------------------------
  Mean VIF     10.26
                           Cond
        Eigenval          Index
---------------------------------
    1     2.8546          1.0000
    2     0.1355          4.5894
    3     0.0099         16.9635
---------------------------------
 Condition Number        16.9635
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.0975
```

```
. * Use collin with centered data using the corr option
. collin x1 x2 if !missing(y), corr

  Collinearity Diagnostics

                        SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
        x1     10.26    3.20    0.0975      0.9025
        x2     10.26    3.20    0.0975      0.9025
----------------------------------------------------
  Mean VIF     10.26

                           Cond
        Eigenval          Index
---------------------------------
    1     1.9500          1.0000
    2     0.0500          6.2450
---------------------------------
 Condition Number         6.2450
 Eigenvalues & Cond Index computed from deviation sscp (no intercept)
 Det(correlation matrix)    0.0975
```

`collin` is a user-written command; type `findit collin` to locate it and install it on your machine. Note that, with the `collin` command, you only give the names of the X variables, not the Y. If Y has missing data, you have to make sure that the same cases are analyzed by the `collin` command that were analyzed by the regress command. There are various ways of doing this. By adding the optional `if !missing(y)` I told Stata to only analyze those cases that were NOT missing on Y. By default, `collin` computed the condition number using the raw data (same as SPSS); adding the `corr` parameter makes it compute the condition number using centered data. [NOTE: `coldiag2` is yet another Stata routine that can give you even more information concerning eigenvalues, condition indices, etc.; type `findit coldiag2` to locate and install it.]

Incidentally, assuming X1 and X2 are measured the same way (e.g. years, dollars, whatever) a possible solution we might consider is to simply add X1 and X2 together. This would make even more sense if we felt X1 and X2 were alternative measures of the same thing. Adding them could be legitimate if (despite the large differences in their estimated effects) their two effects did not significantly differ from each other. In Stata, we can easily test this.

```
. test x1 = x2

 ( 1)  x1 - x2 = 0

       F(  1,    97) =    0.10
            Prob > F =    0.7484
```

Given that the effects do not significantly differ, we can do the following:

```
. gen x1plusx2 = x1 + x2
. reg y x1plusx2

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    6.43
       Model |  54.8536183     1  54.8536183           Prob > F      =  0.0128
    Residual |  836.146432    98  8.53210645           R-squared     =  0.0616
-------------+------------------------------           Adj R-squared =  0.0520
       Total |  891.000051    99  9.00000051           Root MSE      =   2.921

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    x1plusx2 |   .0753846   .0297309     2.54   0.013     .0163846    .1343846
       _cons |   10.49231   .6624892    15.84   0.000      9.17762      11.807
------------------------------------------------------------------------------
```

The multicollinearity problem is obviously gone (since we only have one IV). As noted before, factor analysis could be used for more complicated scale construction.

Appendix: Models with nonlinear or nonadditive (interaction) terms.

Sometimes models include variables that are nonlinear or nonadditive (interactive) functions of other variables in the model. For example, the IVS might include both X and X^2. Or, the model might include X1, X2, and X1*X2. We discuss the rationale for such models later in the semester. In such cases, the original variables and the variables computed from them can be highly correlated. Also, multicollinearity between nonlinear and nonadditive terms could make it difficult to determine whether there is multicollinearity involving other variables. Consider the following simple example, where X takes on the values 1 through five and is correlated with X^2.

```
. clear all
. set obs 5
obs was 0, now 5
. gen X = _n
. gen XSquare = X ^ 2
. list

     +-------------+
     | X   XSquare |
     |-------------|
  1. | 1         1 |
  2. | 2         4 |
  3. | 3         9 |
  4. | 4        16 |
  5. | 5        25 |
     +-------------+

. corr X XSquare, means
(obs=5)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
           X |            3     1.581139            1            5
     XSquare |           11      9.66954            1           25


             |        X  XSquare
-------------+------------------
           X |   1.0000
     XSquare |   0.9811   1.0000
```

As we see, the correlation between X and X^2 is very high (.9811). High correlations can likewise be found with interaction terms.

It is sometimes suggested that, with such models, the original IVs should be *centered* before computing other variables from them. You center a variable by subtracting the mean from every case. The mean of the centered variable is then zero (the standard deviation, of course, stays the same). The correlations between the IVs will then often be far smaller. For example, if we center X in the above problem by subtracting the mean of 3 from each case before squaring, we get

```
. replace X = X - 3
(5 real changes made)
. replace XSquare = X ^ 2
(5 real changes made)
```

```
. list

     +--------------+
     |  X   XSquare |
     |--------------|
  1. | -2         4 |
  2. | -1         1 |
  3. |  0         0 |
  4. |  1         1 |
  5. |  2         4 |
     +--------------+

. corr X XSquare, means
(obs=5)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
           X |            0     1.581139           -2            2
     XSquare |            2     1.870829            0            4


             |        X  XSquare
-------------+------------------
           X |   1.0000
     XSquare |   0.0000   1.0000
```

As you see, the extremely high correlation we had before drops to zero when the variable is centered before computing the squared term.

We'll discuss nonlinear and nonadditive models later in the semester. We'll also see other reasons why centering can be advantageous. In particular, centering can make it a lot easier to understand and interpret effects under certain conditions. The specific topic of centering is briefly discussed on pages 30-31 of Jaccard et al's Interaction Effects in Multiple Regression. Also see pp. 35-36 of Aiken and West's Multiple Regression: Testing and Interpreting Interactions. Note, incidentally, that centering is only recommended for the IVs; you generally do not need or want to center the DV.

Appendix: Multicollinearity in Non-OLS techniques

The examples above use OLS regression. As we will see, OLS regression is not an appropriate statistical technique for many sorts of problems. For example, if the dependent variable is a dichotomy (e.g. lived or died) logistic regression or probit models are generally better. However, as Menard notes in Applied Logistic Regression Analysis, much of the diagnostic information for multicollinearity (e.g. VIFs) can be obtained by calculating an OLS regression model using the same dependent and independent variables you are using in your logistic regression model. "Because the concern is with the relationship among the independent variables, the functional form of the model for the dependent variable is irrelevant to the estimation of collinearity." (Menard 2002, p. 76). In other words, you could run an OLS regression, and ignore most of the results but still use the information that pertained to multicollinearity. Even more simply, in Stata, the `collin` command can generally be used regardless of whether the ultimate analysis will be done with OLS regression, logistic regression, or whatever.

In short, multicollinearity is not a problem that is unique to OLS regression, and the various diagnostic procedures and remedies described here are not limited to OLS.

Appendix: Condition Numbers (Optional)

Warning: This is a weak area of mine so if you really want to understand these you should do a lot more reading.

Sometimes *eigenvalues*, *condition indices* and the *condition number* will be referred to when examining multicollinearity. While all have their uses, I will focus on the condition number. The condition number (κ) is the condition index with the largest value; it equals the square root of the largest eigenvalue ($\lambda_{\max}$) divided by the smallest eigenvalue ($\lambda_{\min}$), i.e.

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

When there is no collinearity at all, the eigenvalues, condition indices and condition number will all equal one. As collinearity increases, eigenvalues will be both greater and smaller than 1 (eigenvalues close to zero indicate a multicollinearity problem), and the condition indices and the condition number will increase. An informal rule of thumb is that if the condition number is 15, multicollinearity is a concern; if it is greater than 30 multicollinearity is a very serious concern. (But again, these are just informal rules of thumb.) In SPSS, you get these values by adding the `COLLIN` parameter to the `Regression` command; in Stata you can use `collin`.

CAUTION: There are different ways of computing eigenvalues, and they lead to different results. One common approach is to center the IVs first, i.e. subtract the mean from each variable. (Equivalent approaches analyze the standardized variables or the correlation matrix.) In other instances, the variables are left uncentered. SPSS takes the uncentered approach, whereas Stata's `collin` can do it both ways. If you center the variables yourself, then both approaches will yield identical results. If your variables have ratio-level measurement (i.e. have a true zero point) then not centering may make sense; if they don't have ratio-level measurement, then I think it makes more sense to center. In any event, be aware that authors handle this in different ways, and there is sometimes controversy over which approach is most appropriate.

I have to admit that I don't fully understand all these issues myself; and I have not seen the condition number and related statistics widely used in Sociology, although they might enjoy wider use in other fields. See Belsley, Kuh and Welsch's Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (1980) for an in-depth discussion.

Appendix: SPSS Example (Optional)

Consider the following hypothetical example:

```
MATRIX DATA VARIABLES = Rowtype_ Y X1 X2/ FORMAT = FREE full
                        /FILE = INLINE / N = 100.

BEGIN DATA.
MEAN     12.00     10.00   10.00
STDDEV   3.00     5.00   5.00
CORR     1.00      .24    .25
CORR      .24     1.00   0.95
CORR     0.25     0.95   1.00
END DATA.

REGRESSION   matrix = in(*)
            /VARIABLES Y X1 X2
            /DESCRIPTIVES
            /STATISTICS DEF TOL BCOV COLLIN TOL
            /DEPENDENT Y
            /method ENTER X1 X2 .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Y	12.000000	3.0000000	100
X1	10.000000	5.0000000	100
X2	10.000000	5.0000000	100

Correlations

		Y	X1	X2
Pearson Correlation	Y	1.000	.240	.250
	X1	.240	1.000	.950
	X2	.250	.950	1.000

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X2, X1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.250[a]	.063	.043	2.9344301

a. Predictors: (Constant), X2, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	55.745	2	27.872	3.237	.044[a]
	Residual	835.255	97	8.611		
	Total	891.000	99			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	10.492	.666		15.765	.000		
	X1	.015	.189	.026	.081	.935	.098	10.256
	X2	.135	.189	.226	.717	.475	.098	10.256

a. Dependent Variable: Y

Coefficient Correlations[a]

Model			X2	X1
1	Correlations	X2	1.000	-.950
		X1	-.950	1.000
	Covariances	X2	.036	-.034
		X1	-.034	.036

a. Dependent Variable: Y

Collinearity Diagnostics[a]

				Variance Proportions		
Model	Dimension	Eigenvalue	Condition Index	(Constant)	X1	X2
1	1	2.855	1.000	.02	.00	.00
	2	.136	4.589	.98	.02	.02
	3	.010	16.964	.00	.98	.98

a. Dependent Variable: Y