Identification of Multicollinearity: VIF and Condition Number
Note by Hubert Gatignon
July 7, 2013; updated March 4, 2014

Multicollinearity is practically detected by the presence of some of its effects. These are presented in Chap. 5 of "Statistical Analysis of Management Data" (Gatignon 2014, pp158 and 163). In particular, due to the small determinant, the consequence of near-collinearity is that the variances of the estimated coefficients are very large while the R-squared can be reasonably high.

Because multicollinearity involves more than two variables, the bivariate correlations provide typically poor guidelines to identify its presence. Mason and Perreault (1991) show that "bivariate correlations as high as .95 have virtually no effect on the ability to recover "true" coefficients and to draw the correct inferences if the sample size is 250 and the R^2 is at least .75" (p. 269) and conclude that "collinearity per se is of less concern than is often implied in the literature" (p. 280).

Two major indices are used to identify collinearity: the variance inflation factor (VIF) and the condition number.

Variance Inflation Factor

To compute the VIF, the auxiliary regressions of each independent variable on all the other *K*-1 independent variables are performed. For the variable *j*, the R-squared is R_j^2. Then the VIF for independent variable *j* is defined as:

$$VIFj = \frac{1}{1 - R_j^2} \qquad (1)$$

Examples of values showing the correspondence between the R-squared and the VIF are given in Table 1:

Table 1 – Correspondence between R-square and VIF

R^2	VIF
0.2	1.25
0.9	10
0.99	100

The average VIF is the average VIF_j across the *K* independent variables.

As a rule of thumb, collinearity is potentially a problem for values of VIF>10.

Condition Number

The condition number of a data matrix is $\kappa(\mathbf{X})$ and measures the sensitivity of the parameter estimates to small changes in the data matrix (Belsley, Kuh and Welsh 1980, Belsley 1982). It is calculated by taking the ratio of the largest to the smallest singular values from the singular value decomposition of **X**.

A condition number above 30 is considered to be indicative of collinearity.

While parameter estimates do not change whether mean centering or not, the collinearity measures (VIF and condition number) decrease dramatically. This illustrates that these measures are, in fact, inadequate to identify collinearity (Belsley 1984). This is especially the case in the context of moderated regression since mean centering is often proposed as a way to reduce collinearity (Aiken and West 1991). However, Echambadi and Hess (2007) prove that the transformation has no effect on collinearity or the estimation.

The VIF and condition number can be obtained in STATA using the "collin" command. Building on the example in Chap. 5 shown in Fig. 5.3 (STATA input example) and Fig.5.5 (STATA output), Table 2 shows the additional instructions to request after the regression commands. The command "collin" in used (it can be installed after searching it using the "findit collin" command), followed by the list of independent variable in the regression. The input "if !missing(lms)" removes from the analysis all observations on the independent variable for which the dependent variable is missing, so as to use the same observations as used in the regression.

Table 2 – VIF and condition number in STATA

```
. regress lms brand2 brand3 brand4 ldist lprice ldist2 lprice2 ldist3 lprice3 ldist4 lprice4

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F( 11,    16) =  462.83
       Model |   47.198075    11  4.29073409           Prob > F      =  0.0000
    Residual |  .148329751    16  .009270609           R-squared     =  0.9969
-------------+------------------------------           Adj R-squared =  0.9947
       Total |  47.3464047    27  1.75357054           Root MSE      =  .09628

------------------------------------------------------------------------------
         lms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      brand2 |  -2.231837   .0516191   -43.24   0.000    -2.341265    -2.12241
      brand3 |  -1.014442   .0515121   -19.69   0.000    -1.123643   -.9052411
      brand4 |   1.264971   .0515001    24.56   0.000     1.155796    1.374147
       ldist |   .9553852   .5182457     1.84   0.084    -.1432467    2.054017
      lprice |   .2487778    .805241     0.31   0.761    -1.458257    1.955812
      ldist2 |    .106274   .5530961     0.19   0.850    -1.066237    1.278785
     lprice2 |  -1.855945   .9255221    -2.01   0.062    -3.817964    .1060746
      ldist3 |  -.0349298   .7525606    -0.05   0.964    -1.630287    1.560427
     lprice3 |  -.9055384   1.196263    -0.76   0.460    -3.441502    1.630425
      ldist4 |   .7047077   1.641841     0.43   0.673    -2.775839    4.185254
     lprice4 |  -1.104441     1.1231    -0.98   0.340    -3.485306    1.276425
       _cons |  -1.676908   .0364238   -46.04   0.000    -1.754123   -1.599693
------------------------------------------------------------------------------

. *Measures of collinearity
. collin brand2 brand3 brand4 ldist lprice ldist2 lprice2 ldist3 lprice3 ldist4 lprice4 if
!missing(lms)
(obs=28)

  Collinearity Diagnostics

                    SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
    brand2      1.51    1.23     0.6627      0.3373
    brand3      1.50    1.23     0.6655      0.3345
    brand4      1.50    1.23     0.6658      0.3342
     ldist     10.79    3.29     0.0927      0.9073
    lprice      7.84    2.80     0.1276      0.8724
    ldist2      8.80    2.97     0.1136      0.8864
   lprice2      4.39    2.10     0.2276      0.7724
    ldist3      3.19    1.79     0.3138      0.6862
```

```
   lprice3      3.06    1.75    0.3267      0.6733
    ldist4      1.83    1.35    0.5457      0.4543
   lprice4      3.40    1.84    0.2944      0.7056
----------------------------------------------------
  Mean VIF      4.35

                          Cond
        Eigenval          Index
---------------------------------
    1     2.6660          1.0000
    2     1.8582          1.1978
    3     1.7073          1.2496
    4     1.5410          1.3153
    5     1.1646          1.5130
    6     0.9976          1.6347
    7     0.9830          1.6469
    8     0.4191          2.5222
    9     0.3717          2.6780
    10     0.1356          4.4339
    11     0.1196          4.7207
    12     0.0363          8.5720
---------------------------------
 Condition Number         8.5720
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.0033
```

References

Aiken, L. S., and S. G. West (1991), *Multiple Regression: Testing and Interpreting Interactions*, Sage Publications, Newbury Park, CA.

Belsley, David A. (1982), "Assessing the Presence of Harmful Collinearity and Other Forms of Weak Data Through a Test for Signal-to-Noise," *Journal of Econometrics*, 20(2), 211–253.

Belsley, David A. (1984), "Demeaning Conditioning Diagnostics Through Centering," *The American Statistician*, 38(2), 73–77.

Belsley, David A., Edwin Kuh and Roy E. Welsh (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York, NY: John Wiley and Sons.

Echambadi, Raj, and James D. Hess (2007), "Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models," *Marketing Science*, 26(3), 438–445.

Mason, Charlotte H., and William D. Jr. Perreault (1991), "Collinearity, Power, and Interpretation of Multiple Regression Analysis," *Journal of Marketing Research*, 28(3), 268–280.