MULTICOLLINEARITY

Herman J. Bierens

Pennsylvania State University

Revised: April 5, 2007

1. *Introduction*

Consider the linear regression model

$$y_j = \theta_1 x_{1,j} + ... + \theta_k x_{k,j} + u_j,\ j = 1,...,n, \quad \textbf{(1)}$$

where y_j is the dependent variable, the $x_{i,j}$'s are the independent (or explanatory) variables, the u_j's are unobservable error terms, n is the sample size, and the θ_i's are the model parameters. If the model contains an intercept, then one of the $x_{i,j}$'s is equal to 1, say $x_{1,j}$. Throughout we assume that the errors are normally distributed:

$$u_j \sim N(0, \sigma^2), \quad \textbf{(2)}$$

Also, for the sake of the argument we assume that the x-variables are nonstochastic. This assumption is of course not very realistic, but it allow the discussion below to stay focused, and is harmless in that everything we are going to derive also holds conditional on the x variables if they are stochastic.

Denoting

$$y = \begin{pmatrix} y_1 \\ y_2 \\ : \\ y_n \end{pmatrix},\ X = \begin{pmatrix} x_{1,1} & ... & x_{k,1} \\ x_{1,2} & ... & x_{k,2} \\ : & ... & : \\ x_{1,n} & ... & x_{k,n} \end{pmatrix},\ u = \begin{pmatrix} u_1 \\ u_2 \\ : \\ u_n \end{pmatrix}, \quad \textbf{(3)}$$

we can cast model **(1)** in vector/matrix form as:

$$y = X\theta + u,\quad u \sim N_n(0, \sigma^2 I_n), \quad \textbf{(4)}$$

Exact multicollinearity occurs if the columns of the matrix X are linearly dependent: at least one of the columns can be written as a linear combination of the other columns. This case is always due to model specification errors, or to transformation errors. For example, if the there are dummy variables among the x variables like seasonal dummies that add up to 1, and if also an intercept is included, these dummy variables will be exactly multicollinear with the column on 1-s in the matrix X. The

solution is to delete either one of the dummy variables or the intercept itself. See Greene. Another common error is where you want to specify a quadratic model, for example $y_j = \theta_0 + \theta_1 x_j + \theta_2 x_j^2 + u_j$, and instead of making the variable x_j^2 by using the option "multiplicative transformation" you use the option "linear transformation", with coefficient 2. Then the transformed variable will be $2x_j$ rather than x_j^2, and obviously x_j and $2x_j$ are exactly multicollinear.

Note that in the case of exact multicollinearity the matrix X^TX is singular, hence we cannot compute the least squares estimator $\hat{\theta} = (X^TX)^{-1}X^Ty$.

If the matrix X^TX is nonsingular but the smallest eigenvalue is very small, we have a case of near-multicollinearity.

2. The effect of near-multicollinearity on the t-values

Denoting by e_i the i-th unit vector of length k, i.e., e_i is the i-th column of the unit matrix I_k, we can write the variance of the i-th component $\hat{\theta}_i$ of the least squares estimator $\hat{\theta}$ as

$$Var(\hat{\theta}_i) = \sigma^2 e_i^T (X^TX)^{-1} e_i. \qquad \textbf{(5)}$$

Since we can write

$$X^TX = Q\Lambda Q^T, \qquad \textbf{(6)}$$

where Λ is the diagonal matrix of eigenvalues of X^TX and Q is the orthogonal matrix of corresponding eigenvectors, we can write this variance also as

$$Var(\hat{\theta}_i) = \sigma^2 e_i^T Q\Lambda^{-1} Q^T e_i = \sigma^2 \sum_{m=1}^{k} \frac{q_{m,i}^2}{\lambda_m}, \qquad \textbf{(7)}$$

where $q_{m,i}$ is the (m,i)-th element of the matrix Q and the λ_m's are the eigenvalues (λ_m is the m-th diagonal element of Λ). Now if λ_1 is the smallest eigenvalue, and $q_{1,i} \neq 0$, then the smaller λ_1, the larger the variance involved. Therefore, near-multicollinearity may inflate all the variances and consequently deflate all the t-values. But how can we distinguish true insignificance from near-multicollinearity?

If the insignificant parameters are really zero, then the F-test of the joint hypothesis involved

should not reject. If it does, we have an indication that the low t-values are due to near-multicollinearity. In order to illustrate this, suppose that all the t-values are insignificant due to near-multicollinearity. Then the F-test of the (incorrect) null hypothesis $\theta = 0$ takes the form

$$\hat{F} = \frac{\hat{\theta}^T (X^T X)^{-1} \hat{\theta}/k}{s^2}, \qquad \textbf{(8)}$$

where s^2 is the well-known unbiased estimator of σ^2. It is also well-known that $\hat{\theta}$ is unbiased and independent of s^2. Denoting $\gamma = (\gamma_1, .., \gamma_k)^T = Q^T \theta$ we therefore have:

$$E(\hat{F} | s^2) = \frac{\theta^T (X^T X)^{-1} \theta/k}{s^2} = \frac{\theta^T Q \Lambda^{-1} Q^T \theta/k}{s^2} = \frac{\gamma^T \Lambda^{-1} \gamma}{k s^2} = \frac{1}{k s^2} \sum_{i=1}^{k} \gamma_i^2 \lambda_i^{-1}. \qquad \textbf{(9)}$$

Since $\gamma \neq 0$ as otherwise $\theta = 0$, we see that a very small λ_1 inflates the F- statistic rather than deflating it.

3. A cure for near-multicollinearity

The only cure for near-multicollinearity is to reduce the number of explanatory variables by imposing restrictions on the parameters. The best way of doing this is to impose restrictions that are prescribed by economic theory. Take for example the translog production function

$$\ln(Y) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(K) + \beta_3 \frac{(\ln(L))^2}{2} + \beta_4 \ln(L)\ln(K) + \beta_5 \frac{(\ln(K))^2}{2} + u, \qquad \textbf{(10)}$$

where Y is output, L is labor, and K is capital. If you suspect that $\ln(L)$ and $\ln(K)$ are near-multicollinear, then imposing the restriction of constant return to scale, which amounts to the restriction $\beta_1 + \beta_2 = 1$ and $\beta_3 = \beta_5 = -\beta_4$, will cure the problem, because then the model becomes

$$\ln(Y) - \ln(K) = \beta_0 + \beta_1 (\ln(L) - \ln(K)) + \beta_3 \frac{(\ln(L) - \ln(K))^2}{2} + u. \qquad \textbf{(11)}$$

If economic theory does not provide guidelines for parameter restrictions, the only other option is to delete the variables from the model that cause the problem. As we have seen, the t-values cannot be used for testing which set of variables should be deleted, as they are "polluted" by the near-singularity of the $X^T X$ matrix. The solution I propose is to orthogonalize the columns of the matrix

X similarly to the Gramm-Schmidt orthogonalization of a basis of a vector space, as follows.

First, determine which of the insignificant *x*-variables is the least important for your economic theory, because you don't want to start off throwing away the key-variables. Suppose it is the last one. Then we can partition the matrix *X* in:

$$X = (X_{k-1}, x_k), \tag{12}$$

where X_{k-1} is the matrix of the first *k*-1 columns of *X* and x_k is the last column. Next, regress x_k on X_{k-1}, and let the vector of residuals be z_k:

$$z_k = x_k - X_{k-1}\alpha_{k-1}, \text{ where } \alpha_{k-1} = (X_{k-1}^T X_{k-1})^{-1} X_{k-1}^T x_k . \tag{13}$$

Then

$$X_{k-1}^T z_k = 0 \tag{14}$$

(*exercise*: why?). Next, replace x_k in *X* by z_k. Partitioning the parameter vector θ as

$$\theta = \begin{pmatrix} \theta_{k-1}^* \\ \theta_k \end{pmatrix} \tag{15}$$

the new model is related to the old one as follows:

$$y = X\theta + u = X_{k-1}\theta_{k-1}^* + x_k\theta_k + u = X_{k-1}(\theta_{k-1}^* + \alpha_{k-1}\theta_k) + z_k\theta_k + u \tag{16}$$

hence

$$y = X_{k-1}\beta_{k-1} + z_k\theta_k + u, \tag{17}$$

where $\beta_{k-1} = \theta_{k-1}^* + \alpha_{k-1}$. Thus we only have reparametrized the model, but now z_k is orthogonal to the columns in X_{k-1}.

THEOREM 1: *The least square estimator and the t-value of the parameter* θ_k *in the reparametrized model* **(17)** *are the same as in the original model* **(4)**.

Proof: We have

$$\left(X^{T}X\right)^{-1} = \begin{pmatrix} X_{k-1}^{T}X_{k-1} & X_{k-1}^{T}x_k \\ x_k^{T}X_{k-1} & x_k^{T}x_k \end{pmatrix}^{-1} = \begin{pmatrix} (X_{k-1}^{T}X_{k-1})^{-1} + \dfrac{\alpha_{k-1}\alpha_{k-1}^{T}}{z_k^{T}z_k} & -\dfrac{\alpha_{k-1}}{z_k^{T}z_k} \\ -\dfrac{\alpha_{k-1}^{T}}{z_k^{T}z_k} & \dfrac{1}{z_k^{T}z_k} \end{pmatrix} \qquad \textbf{(18)}$$

(*exercise*: Verify this), and

$$X^{T}y = \begin{pmatrix} X_{k-1}^{T}y \\ x_k^{T}y \end{pmatrix} \qquad \textbf{(19)}$$

hence

$$\hat{\theta}_k = \frac{x_k^{T}y - \alpha_{k-1}^{T}X_{k-1}^{T}y}{z_k^{T}z_k} = \frac{z_k^{T}y}{z_k^{T}z_k} \qquad \textbf{(20)}$$

Moreover, the least squares estimator of the parameter vector $(\beta_k^{T}, \theta_k)^{T}$ in model **(17)** is

$$\begin{aligned} \left((X_{k-1}, z_k)^{T} (X_{k-1}, z_k) \right)^{-1} (X_{k-1}, z_k)^{T}y &= \begin{pmatrix} (X_{k-1}^{T}X_{k-1})^{-1} & 0 \\ 0 & \dfrac{1}{z_k^{T}z_k} \end{pmatrix} \begin{pmatrix} X_{k-1}^{T}y \\ z_k^{T}y \end{pmatrix} \\ &= \begin{pmatrix} X_{k-1}^{T}X_{k-1})^{-1}X_{k-1}^{T}y \\ \dfrac{z_k^{T}y}{z_k^{T}z_k} \end{pmatrix}. \end{aligned} \qquad \textbf{(21)}$$

Comparing the last two results, it follows that the least squares estimator of θ_k in both models is the same. Replacing *y* in **(20)** by the right-hand side of **(17)**, and using **(14)**, it easily follows that

$$\hat{\theta}_k - \theta_k = \frac{z_k^{T}u}{z_k^{T}z_k} \sim N\left(0, \frac{\sigma^2}{z_k^{T}z_k}\right). \qquad \textbf{(22)}$$

From this result it follows that also the t-values are the same, because the sum of squared residuals of both models is the same, and so is the estimator s^2 of σ^2. Q.E.D.

The t-values of the x-variables in the matrix X_{k-1}, however, will change is there is near-multicollinearity. If all the t-values of these variables are now significant, and some of them were not before, then you are done, in the sense that you now may blame the near-multicollinearity on x_k, and solve the problem by removing x_k from the model. If some t-values of variables in X_{k-1} are still insignificant, we may repeat the procedure, by selecting among the insignificant variables the one that is of the least importance (but do not choose the intercept! The intercept is important for other reasons.), and replace it by the residual of the regression of the variable involved on the remaining variables in X_{k-1}. If the last column, x_{k-1}, in the matrix X_{k-1} is this variable, and partitioning $X_{k-1} = (X_{k-2}, x_{k-1})$, then we replace x_{k-1} by the vector z_{k-1} of residuals of the regression of x_{k-1} on X_{k-2}, and run the regression

$$y = X_{k-2}\beta_{k-2} + z_{k-1}\beta_{k-1,k-1} + z_k\theta_k + u, \tag{23}$$

where $\beta_{k-1,k-1}$ is the last component of β_{k-1}. Again the t-value of $\beta_{k-1,k-1}$ will be the same as in model **(17)**, as follows easily from Theorem 1. Repeating this procedure until all the t-values of the remaining variables are significant, we end up with a model of the form

$$y = X_{k-m}\beta_{k-m} + \sum_{i=1}^{m-1} z_{k-i}\beta_{k-i,k-i} + z_k\theta_k + u, \tag{24}$$

where the t-values of the parameters in β_{k-m} are significant, except perhaps the intercept, and the coefficients of the residuals z_i are all insignificant. This model is still equivalent to the original model **(4)**, but we now have isolated the source of the problem, and since the residuals z_i are all insignificant, it is now clear which variables to remove from the model: the variables corresponding to the residuals in model **(24)**, so that the final model becomes

$$y = X_{k-m}\beta_{k-m} + u. \tag{25}$$

This approach has the advantage that in the selection process one can also weigh the theoretical importance of each x-variable.