Cardon Research Papers
in Agricultural and Resource Economics

Research Paper 2012-02
June 2012

A Note on the Harmful Effects of Multicollinearity

Lester D. Taylor
University of Arizona

Department of Agricultural and Resource Economics
College of Agriculture and Life Sciences
The University of Arizona

This paper is available online at http://ag.arizona.edu/arec/pubs/researchpapers.html

Tucson, AZ *April 2012*

A Note on The Harmful Effects of Multicollinearity

Lester D. Taylor[*]
Department of Economics
Department of Agricultural & Resource Economics
University of Arizona

ltaylor@email.arizona.edu

Abstract

Assessing the harmful effects of multicollinearity in a regression model with multiple predictors has always been one of the great problems in applied econometrics. As correlations amongst predictors are almost always present to some extent (especially in time-series data generated by natural experiments), the question is at what point does inter-correlation become harmful. Despite receiving quite a bit of attention in the 1960s and 1970s (but only limited since), a fully satisfactory answer to this question has yet to be developed. My own thoughts on the issue have always been that multicollinearity becomes "harmful" when there is an R^2 *in the predictor matrix that is of the same order of magnitude as the* R^2 *of the model overall. An empirical examination of this "rule-of-thumb", in a stylized Monte Carlo study, is the purpose of this communication.*

[*] I am grateful to Timothy Tardiff for comment and criticism.

A Note on The Harmful Effects of Multicollinearity

Lester D. Taylor
University of Arizona

I. INTRODUCTION

Assessing the harmful effects of multicollinearity in a regression model with multiple predictors has always been one of the great problems in applied econometrics. As correlations amongst predictors are almost always present to some extent (especially in time-series data generated by natural experiments), the question is at what point does inter-correlation become harmful. Despite receiving quite a bit of attention in the 1960s and 1970s (but only limited since), a fully satisfactory answer to this question has yet to be developed. My own thoughts on the issue have always been that multicollinearity becomes "harmful" when there is an R^2 in the predictor matrix that is of the same order of magnitude as the R^2 of the model overall. An empirical examination of this "rule-of-thumb", in a stylized Monte Carlo study, is the purpose of this communication.

II. BACKGROUND AND DESIGN

Once, many years ago, when at the University of Michigan I attended a lunch-time seminar run by one of my colleagues, whereby a professor in the political science department presented a multiple regression model in which two variables and their difference were specified as predictors. As politely as I could, I mentioned that this involved a problem, as the X'X matrix would be singular and none of the coefficients in the model could be estimated. The visitor responded that this was indeed correct, but that the computer regression program being used was able to get around the problem. At this point, I decided just to sit and listen.

A perfectly singular X'X matrix is, of course, the extreme of harmful multicollinearity, but is such that, in practice, is only encountered when the same variable is inadvertently (or otherwise) included twice, whether directly or as an exact linear combination of other independent variables. The best that can be done in this situation is to estimate a set of linear combinations of the original coefficients that are equal in number to the rank of the X'X matrix. Since this results in fewer equations than unknowns, at least some (if not all) of the original coefficients of the model cannot be identified.

The multicollinearity problem in practice is not a perfectly singular X'X matrix, but one that is nearly so -- or so it was thought to be the case in the days when "harmful" multicollinearity first became a topic of serious research. As a consequence, early investigations focused strictly on the structure of the X'X matrix as source of the problem. Farrar and Glauber (1964), for example, approached the problem in terms of departures from orthogonality of the columns of the X matrix, while Belsley, Kuh, and Welsch (1980) sought to pinpoint it via singular-value decomposition of the X'X matrix. Increasingly, however, it became clear that focus on just the X'X matrix is only part of the story, and that the strength of the relationship in the overall regression is a factor as well.

My own experience certainly attests to this, for I have estimated many models in which inter-correlations amongst the independent variables are extremely high, but because the R^2s for the estimated equations are even higher, “harmful” multicollinearity does not appear to be present. The communication that follows is essentially an exercise in examples that this is the case.

The model used in the investigation involves three independent variables x, w, and z and an error term η,

$$.y = \alpha + \beta x + \gamma w + \kappa z + \eta \,. \tag{1}$$

The variables x and w are orthogonal to one another by construction, while z is created as

$$z = x + w + \delta e \tag{2}$$

in one design and as

$$z = x + \delta e \tag{3}$$

in a second design. The first design investigates the effects of the “closeness” of z to the plane spanned by x and w, while the second design focuses on the effects of near co-linearity of z with x alone. The vector e represents realizations of a pseudo random variable generated from a 0-1 uniform distribution, with δ a parameter that can be varied to give desired correlations between z, x, and w.[1] The orthogonal variables x and w are constructed as principal components of household consumption data from a BLS Consumer Expenditure Survey.[2] The results are all based upon the model with assumed coefficients:

$$y = 10 + x + 2w + 2z + \Delta\varepsilon \,, \tag{4}$$

where ε is a vector of realizations of a pseudo unit-normal random variable and Δ is a parameter for adjusting the model’s overall R^2. The sample size is 100, with the same values for x, w, e, and ε in all of the estimations.

The results for a variety of values of δ and for “high”, “moderate”, and “low” R^2s for the overall fit of the model are tabulated in Tables 1 and 2. The first two columns in these tables describe the degree of co-linearity in the independent variables, while the columns on the right show the effects of this co-linearity on the coefficient estimates. The information given includes the R^2

---

[1] R^2s between the vectors e and x and e and w employed in the realizations are 0.0136 and 0.0025, respectively, and 0.0002 and 0.0096 for ε and x and ε and w.

[2] The data set used consists of expenditures for 14 exhaustive categories of consumption for 100 households from the fourth-quarter BLS survey of 1999. All calculations are done in SAS.

Table 1

Monte Carlo Multicollinearity Results

$$y = 10 + x + 2w + 2z + \Delta\varepsilon$$
$$z = x + w + \delta e$$

Co-linearity $z = x + w + \delta e$		High R^2 ($\Delta = 100$) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9995	19.72	1.13	2.73	0.88	3.86	1.24	0.27	0.09	0.9616
20	0.9981	19.72	1.13	1.86	1.19	2.99	1.92	1.13	0.73	0.9617
30	0.9957	19.72	1.13	1.57	1.51	2.70	2.59	1.42	1.38	0.9619
40	0.9924	19.72	1.13	1.43	1.81	2.56	3.24	1.56	2.02	0.9620
50	0.9882	19.72	1.13	1.34	2.12	2.47	3.87	1.65	2.67	0.9622
60	0.9832	19.72	1.13	1.28	2.41	2.42	4.49	1.71	3.32	0.9624
70	0.9774	19.72	1.13	1.24	2.71	2.37	5.09	1.75	3.96	0.9625
80	0.9708	19.72	1.13	1.21	2.99	2.34	5.67	1.78	4.61	0.9627
90	0.9636	19.72	1.13	1.28	3.27	2.31	6.22	1.81	5.25	0.9629
100	0.9556	19.72	1.13	1.17	3.55	2.30	6.75	1.83	5.90	0.9632
200	0.8503	19.72	1.13	1.08	5.89	2.21	10.77	1.91	12.36	0.9657
400	0.6117	19.72	1.13	1.03	8.71	2.17	14.21	1.96	25.28	0.9722
800	0.3230	19.72	1.13	1.02	10.74	2.15	15.87	1.98	51.13	0.9833
1600	0.1432	19.72	1.13	1.01	11.63	2.14	16.41	1.99	102.81	0.9932

Co-linearity $z = x + w + \delta e$		Moderate R^2 ($\Delta = 400$) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9995	48.89	0.70	7.91	0.64	9.44	0.76	-4.93	-0.40	0.6263
200	0.8503	48.89	0.70	1.32	1.81	2.85	3.47	1.65	2.67	0.6488
300	0.7268	48.89	0.70	1.21	2.17	2.74	4.07	1.77	4.29	0.6686
400	0.6117	48.89	0.70	1.15	2.42	2.68	4.39	1.82	5.90	0.6916
800	0.3230	48.89	0.70	1.06	2.81	2.59	4.79	1.91	12.36	0.7871
1600	0.1432	48.89	0.70	1.02	2.95	2.55	4.90	1.96	25.28	0.9025

Co-linearity $z = x + w + \delta e$		Low R^2 ($\Delta = 1000$) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9995	107.22	0.61	18.28	0.59	20.60	0.66	-15.33	-0.50	0.2370
400	0.6117	107.22	0.61	1.38	1.16	3.71	2.42	1.57	2.02	0.2778
800	0.3230	107.22	0.61	1.16	1.23	3.48	2.57	1.78	4.61	0.3751
1200	0.2017	107.22	0.61	1.09	1.23	3.41	2.59	1.86	7.19	0.4890
1600	0.1432	107.22	0.61	1.05	1.22	3.37	2.59	1.89	9.78	0.5927

Table 2

Monte Carlo Multicollinearity Results

$$y = 10 + x + 2w + 2z + \Delta\varepsilon$$
$$z = x + \delta e$$

Co-linearity $z = x + \delta e$		High R^2 (Δ = 100) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9993	19.72	1.13	2.73	0.88	2.13	16.64	0.27	0.09	0.9445
20	0.9973	19.72	1.13	1.86	1.19	2.13	16.64	1.13	0.73	0.9448
30	0.9939	19.72	1.13	1.57	1.51	2.13	16.64	1.42	1.38	0.9450
40	0.9893	19.72	1.13	1.43	1.81	2.13	16.64	1.56	2.02	0.9453
50	0.9835	19.72	1.13	1.34	2.12	2.13	16.64	1.65	2.67	0.9456
60	0.9766	19.72	1.13	1.28	2.41	2.13	16.64	1.71	3.32	0.9459
70	0.9685	19.72	1.13	1.24	2.71	2.13	16.64	1.75	3.96	0.9463
80	0.9595	19.72	1.13	1.21	2.99	2.13	16.64	1.78	4.61	0.9466
90	0.9496	19.72	1.13	1.28	3.27	2.13	16.64	1.81	5.25	0.9470
100	0.9389	19.72	1.13	1.17	3.55	2.13	16.64	1.83	5.90	0.9474
200	0.8030	19.72	1.13	1.08	5.89	2.13	16.64	1.91	12.36	0.9523
400	0.5334	19.72	1.13	1.03	8.71	2.13	16.64	1.96	25.28	0.9636
800	0.2614	19.72	1.13	1.02	10.74	2.13	16.64	1.98	51.13	0.9805
1600	0.1140	19.72	1.13	1.01	11.63	2.13	16.64	1.99	102.81	0.9929

Co-linearity $z = x + \delta e$		Moderate R^2 (Δ = 400) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9993	48.89	0.70	7.91	0.64	2.51	4.90	-4.93	-0.40	0.5272
200	0.8030	48.89	0.70	1.32	1.81	2.51	4.90	1.65	2.67	0.5602
300	0.6576	48.89	0.70	1.21	2.17	2.51	4.90	1.77	4.29	0.5897
400	0.5334	48.89	0.70	1.15	2.42	2.51	4.90	1.82	5.90	0.6234
800	0.2614	48.89	0.70	1.06	2.81	2.51	4.90	1.91	12.36	0.7551
1600	0.1140	48.89	0.70	1.02	2.95	2.51	4.90	1.96	25.28	0.8924

Co-linearity $z = x + \delta e$		Low R^2 (Δ = 1000) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9993	107.22	0.61	18.28	0.59	3.26	2.56	-15.33	-0.50	0.1680
400	0.5334	107.22	0.61	1.38	1.16	3.26	2.56	1.57	2.02	0.2137
800	0.2614	107.22	0.61	1.16	1.23	3.26	2.56	1.78	4.61	0.3257
1200	0.1607	107.22	0.61	1.09	1.23	3.26	2.56	1.86	7.19	0.4553
1600	0.1140	107.22	0.61	1.05	1.22	3.26	2.56	1.89	9.78	0.5708

for the regression of z on x and w (0.9995 for $\delta = 10$ in Table 1) and estimated coefficients, t-ratios, and R^2 for the model in expression (4). Table 1 shows results for z constructed according to expression (2) and Table 2 for z constructed according to expression (3).

The key features in Table 1 are as follows:

(1). When z lies very close to plane spanned by x and w (see the $\delta = 10$ lines in Table 1), regression coefficients are estimated with little precision. Note, however, that degradation is much greater in the moderate and low R^2 cases than when the R^2 is high. While both results are to be expected, it will be seen below that, even with z nearly lying in the x-w plane (per the R^2 of 0.9995), degradation can be "trumped" by the dependent variable lying even closer to its regression plane (i.e., by a model R^2 that is of the same order.

(2). Again, as is to be expected, precision of the estimates increases as z moves away from the x-w plane. Taking a t-ratio of 2 as a benchmark, this is reached for all three coefficients at $\delta = 50$ and $\delta = 300$ in the high and moderate R^2 cases and at $\delta = 400$ for the coefficients of w and z in the low R^2 case. However, the surprising thing is that, in all three cases, the R^2s of z on the x-w plane are greater than for the models overall: 0.9924 vs. 0.9620, 0.7268 vs. 0.6686, and 0.6117 vs. 0.2788, respectively.[3]

The only difference between the design underlying the results in Table 2 and the design underlying Table 1 is that the co-linearity of z is now in relation to x alone rather than with respect to the x-w plane. The data are otherwise all identical. As seen in the table, the effect of this change is to confine the ill-effects of multicollinearity to estimates of the coefficients for x and z. Since w is orthogonal to both x and e, and therefore to z, the estimates of the coefficient for w have large t-ratios and are invariant (for an R^2 regime) across realizations. This is simply a consequence of OLS estimation and orthogonality. That the estimated coefficients for x are identical for the two designs may seem strange, but this, too, is a straightforward consequence of OLS estimation in light of the orthogonality of x with both e and w. The final thing to note in Table 2 is that the apparent "harmful" effects co-linearity of z with x dissipate (using a t-ratio of 2 as a benchmark) at the same values of δ as in Table 1, which is to say, that the ill-effects of multicollinearity are independent of the form the co-linearity takes. What matters is the degree, not the form.

Next on the agenda is to investigate the effect of co-linearity when the R^2s of z and y with their respective regression planes (i.e., z on x and w, and y on x, w, and z) are both extremely close to 1. The design for this case has been to hold δ constant at 10 in the construction of $z = x + w + \delta e$ and then to vary Δ in the generation of $y = 10 + x + 2w + 2z + \Delta\varepsilon$. The results are presented in Table

---

[3] The invariance of the intercept and its t-ratio across different values of δ reflects the fact that the means of z and the dependent variable always change by the same amount.

Table 3

Monte Carlo Multicollinearity Results

$$y = 10 + x + 2w + 2z + \Delta\epsilon$$

$$z = x + w + 10e$$
$$R^2 = 0.9995$$

$$y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$$

Δ	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
5	1.09	6.99	2.09	13.48	1.91	12.36	0.9999
10	1.17	3.78	2.19	7.04	1.83	5.90	0.9996
20	1.36	2.17	2.37	3.82	1.65	2.67	0.9984
30	1.52	1.63	2.56	2.75	1.48	1.59	0.9964
40	1.69	1.36	2.74	2.21	1.31	1.06	0.9936
50	1.86	1.20	2.93	1.89	1.13	0.73	0.9900
100	2.73	0.88	3.86	1.24	0.27	0.09	0.9616

3.[4] The results are interesting in that they show, in line with the thesis of this communication, that multicollinearity, and whether it is harmful, is not an absolute concept, but depends upon the relationship between the largest R^2 amongst the regressors (where each predictor is regressed on all of the others) and the R^2 of the model. Table 3 shows this very clearly, where, despite an R^2 of 0.9995 in the regression of z on x and w, an R^2 for the model of the same (or even slightly lower) magnitude, estimated coefficients are seen to remain stable with t-ratios comfortably greater than 2.

As a check on the results presented in Tables 1 - 3, results from a from a second set of realizations for the vectors e and ε (keeping x and w the same) are presented in Tables 4 - 6.[5] While the results are obviously not the same, they clearly support the thesis that ill-effects of multicollinearity depend upon the highest R^2 amongst the independent variable in relation to the R^2 of the overall model.

III. CONCLUSION

Most earlier analyses of "harmful" multicollinearity in linear regression involving multiple predictors have focused on the structure of the X'X matrix without regard to the strength of the relationship between the dependent variable and the independent variables. The thesis in this communication has been that the ill-effcts of co-linearity (as reflected in unstable and imprecise

---

[4] Since the intercept is of little interest at this point, it is not included in this table.

[5] The R^2s between e and x and w are 0.0012 and 0.0021, respectively. The R^2s between ε and x and ε and w are 0.0030 and 0.0018, respectively.

Table 4

Monte Carlo Multicollinearity Results
Second Set of Error Vectors e and ϵ

$$y = 10 + x + 2w + 2z + \Delta\epsilon$$
$$z = x + w + \delta e$$

Co-linearity $z = x + w + \delta e$		High R^2 (Δ = 100) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9995	27.20	1.34	5.09	1.49	6.33	1.85	-2.14	-0.63	0.9547
20	0.9981	27.20	1.34	1.86	1.19	3.02	1.76	-0.07	-0.04	0.9546
30	0.9957	27.20	1.34	2.33	2.04	3.57	3.11	0.62	0.54	0.9545
40	0.9924	27.20	1.34	1.98	2.31	3.22	3.73	0.96	1.13	0.9545
50	0.9881	27.20	1.34	1.76	2.58	3.01	4.34	1.17	1.71	0.9545
60	0.9830	27.20	1.34	1.64	2.85	2.88	4.93	1.31	2.29	0.9544
70	0.9769	27.20	1.34	1.54	3.12	2.78	5.51	1.41	2.88	0.9544
80	0.9700	27.20	1.34	1.46	3.38	2.70	6.06	1.48	3.46	0.9545
90	0.9622	27.20	1.34	1.41	3.64	2.64	6.60	1.54	4.04	0.9545
100	0.9536	27.20	1.34	1.36	3.89	2.60	7.12	1.59	4.63	0.9545
200	0.8336	27.20	1.34	1.15	6.09	2.39	11.03	1.79	10.47	0.9558
400	0.5431	27.20	1.34	1.05	8.72	2.29	14.23	1.90	22.14	0.9615
800	0.2106	27.20	1.34	1.00	10.32	2.24	15.52	1.95	45.49	0.9757
1600	0.0501	27.20	1.34	0.97	10.77	2.21	15.797	1.99	92.19	0.9906

Co-linearity $z = x + w + \delta e$		Moderate R^2 (Δ = 400) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9995	78.80	0.97	17.36	1.27	19.31	1.41	-14.58	-1.06	0.5885
200	0.8336	78.80	0.97	1.61	2.13	3.57	4.11	1.17	1.71	0.5852
300	0.6845	78.80	0.97	1.34	2.37	3.29	4.65	1.45	3.17	0.5944
400	0.5431	78.80	0.97	1.20	2.49	3.15	4.90	1.59	4.63	0.6100
800	0.2106	78.80	0.97	0.99	2.56	2.95	5.11	1.79	10.47	0.7063
1600	0.0501	78.80	0.97	0.89	2.46	2.84	5.08	1.90	22.14	0.8616

Co-linearity $z = x + w + \delta e$		Low R^2 (Δ = 1000) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9995	182.01	0.90	41.90	1.22	45.28	1.32	-39.44	-1.15	0.2223
400	0.5491	182.01	0.90	1.50	1.24	4.88	3.04	0.96	1.13	0.2140
800	0.2106	182.01	0.90	0.98	1.01	4.36	3.03	1.48	3.06	0.2705
1200	0.0956	182.01	0.90	0.81	0.88	4.19	2.97	1.65	5.80	0.3684
1600	0.0501	182.01	0.90	0.72	0.80	4.10	2.93	1.74	8.13	0.4764

Table 5

Monte Carlo Multicollinearity Results
Second Set of Error Vectors e and ϵ

$$y = 10 + x + 2w + 2z + \Delta\epsilon$$
$$z = x + \delta e$$

Co-linearity $z = x + \delta e$		High R^2 (Δ = 100) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9993	27.20	1.34	5.09	1.49	2.18	15.71	-2.14	-0.63	0.9340
20	0.9973	27.20	1.34	1.86	1.19	2.18	15.71	-0.07	-0.04	0.9339
30	0.9940	27.20	1.34	2.33	2.04	2.18	15.71	0.62	0.54	0.9338
40	0.9893	27.20	1.34	1.98	2.31	2.18	15.71	0.96	1.13	0.9337
50	0.9834	27.20	1.34	1.76	2.58	2.18	15.71	1.17	1.71	0.9337
60	0.9761	27.20	1.34	1.64	2.85	2.18	15.71	1.31	2.29	0.9337
70	0.9677	27.20	1.34	1.54	3.12	2.18	15.71	1.41	2.88	0.9337
80	0.9582	27.20	1.34	1.46	3.38	2.18	15.71	1.48	3.46	0.9337
90	0.9475	27.20	1.34	1.41	3.64	2.18	15.71	1.54	4.04	0.9338
100	0.9359	27.20	1.34	1.36	3.89	2.18	15.71	1.59	4.63	0.9340
200	0.7308	27.20	1.34	1.15	6.09	2.18	15.71	1.79	10.47	0.9367
400	0.4590	27.20	1.34	1.05	8.72	2.18	15.71	1.90	22.14	0.9481
800	0.1613	27.20	1.34	1.00	10.32	2.18	15.71	1.95	45.49	0.9712
1600	0.0380	27.20	1.34	0.97	10.77	2.18	15.71	1.99	92.19	0.9900

Co-linearity $z = x + \delta e$		Moderate R^2 (Δ = 400) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9993	78.80	0.97	17.36	1.27	2.74	4.92	-14.58	-1.06	0.4796
200	0.7808	78.80	0.97	1.61	2.13	2.74	4.92	1.17	1.71	0.4763
300	0.6071	78.80	0.97	1.34	2.37	2.74	4.92	1.45	3.17	0.4918
400	0.4190	78.80	0.97	1.20	2.49	2.74	4.92	1.59	4.63	0.5170
800	0.2106	78.80	0.97	0.99	2.56	2.74	4.92	1.79	10.47	0.6583
1600	0.0380	78.80	0.97	0.89	2.46	2.74	4.92	1.90	22.14	0.8534

Co-linearity $z = x + \delta e$		Low R^2 (Δ = 1000) $y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$								
δ	R^2	α	t-ratio	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
10	0.9993	182.01	0.90	41.90	1.22	3.85	2.77	-39.44	-1.15	0.1525
400	0.4190	182.01	0.90	1.50	1.24	3.85	2.77	0.96	1.13	0.1443
800	0.2106	182.01	0.90	0.98	1.01	3.85	2.77	1.48	3.46	0.2124
1200	0.0719	182.01	0.90	0.81	0.88	3.85	2.77	1.65	5.80	0.3264
1600	0.0380	182.01	0.90	0.72	0.80	3.85	2.77	1.74	8.13	0.4486

Table 6

Monte Carlo Multicollinearity Results

$$y = 10 + x + 2w + 2z + \Delta\epsilon$$

$$z = x + w + 10e$$
$$R^2 = 0.9995$$

$$y = \alpha + \beta x + \gamma w + \kappa z + \Delta\epsilon$$

Δ	β	t-ratio	γ	t-ratio	κ	t-ratio	R^2
5	1.20	7.04	2.22	12.95	1.79	10.47	0.9999
10	1.41	4.12	2.43	7.10	1.59	4.63	0.9995
20	1.81	2.66	2.87	4.18	1.17	1.71	0.9981
30	2.23	2.17	3.30	3.21	0.76	0.74	0.9957
40	2.64	1.93	3.73	2.72	0.34	0.25	0.9924
50	3.05	1.78	4.16	2.43	-0.07	-0.04	0.9881
100	5.09	1.49	6.33	1.85	-2.14	-0.63	0.9547

regression coefficient estimates) become apparent only when one of the regressor vectors lies closer to its fellows than does the dependent vector in relation to the full set of regressors. This thesis has been investigated in a Monte Carlo study involving an OLS regression model with three independent variables, in which two of the predictors are orthogonal to one another while the third is constructed as the sum of these plus an uncorrelated component. Taking t-ratios of 2 as a benchmark, the Monte Carlo results are clear in showing that, no matter how close the third variable may lie to the plane defined by the two orthogonal variables, multicollinearity is "harmful" only when the R^2 for that relationship is stronger than the R^2 for the model overall. Thus, a useful procedure for testing for possible ill-effects of multicollinearity in a linear regression model is to regress each of the independent variables on its fellows and then compare the resulting R^2s with the R^2 for the model overall. If the model R^2 is higher than any of these auxiliary R^2s, then, whatever problems the model might have, it can be concluded that "harmful" multicollinearity is not one of them.

It is important to note that this conclusion is empirically based, and does not, at least at this point, have a rigorous mathematical basis. While one can almost certainly say that the rule provides a sufficient condition (using a benchmark of a t-ratio of 2) for multicollinearity not to be harmful, it does not appear to be necessary. However, since the Monte Carlo results presented are pretty unequivocal, it seems likely (at least to me) that somewhere in the mathematics connecting the matrix (y, X)'(y, X) to its "sub-matrix" X'X lurks a theorem that can lead to a fully rigorous definition of harmful multicollinearity.

REFERENCES

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and The Sources of Collinearity*, Wiley Interscience.

Farrar, D. E. and Glauber, R. R. (1967), "Multicollinearity in Regression Analysis: The Problem Revisited," *The Review of Economics & Statistics*, Vol. 49, No. 1, February 1967, pp. 92 - 107.