IS COLLINEARITY A PROBLEM?

Joseph P. Newhouse

March 1971

P-4588

IS COLLINEARITY A PROBLEM?

Joseph P. Newhouse*

The Rand Corporation, Santa Monica, California

The purpose of this note is to report on some Monte Carlo experiments which make it appear as though collinearity may not be as important a problem as a casual reader of applied econometric work might infer from the frequency with which collinearity is cited to explain poor results. Specifically we used ordinary least squares to estimate the coefficients of the following equation:

(1) $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + u$, where u is normal with mean zero and variance equal to σ^2. Let b_i be the least squares estimate of β_i. The variable x_2 was constructed to be:

(2) $x_2 = x_1 + ae$, where a is a constant and e is a standardized, normally distributed variable. The constant a was chosen to control the amount of collinearity between x_1 and x_2; in the experiments run, a was chosen so that the r between x_1 and x_2 was approximately 0.50, 0.90, and 0.99. The coefficients β_o, β_1, and β_2 were set equal to 1.0. The variable x_1 was taken from a uniform distribution on the 0, 1 interval. The error term u was drawn from a standardized normal distribution and was scaled so that $E(R^2)$ was between 0.92 and 0.95.** The variables x_1

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The Rand Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The Rand Corporation as a courtesy to members of its staff.

**$E(R^2)$ was to have been set equal to 0.90. A computational error prevented this, and the error was deemed not worth correcting.

and x_2 were held fixed across repeated trials. These values of a and the scaling factor on u were chosen as approximations to the range of values which are frequently encountered in economic time series data.

Given the model specified in (1), the power of the usual t-test is a direct function of the value of the noncentrality parameter. The value of this parameter when testing $H_0: \beta_i = 0$ against $H_a: \beta_i \neq 0$ is $n\left(\frac{\beta_i d_i}{\sigma}\right)^2 \frac{1}{r^{ii}}$, where n is the number of observations, d_i is the standard deviation of x_i in the sample, and r^{ii} is the i^{th} diagonal element of the inverse of the correlation matrix among the explanatory variables. Given β, the non-centrality parameter is clearly higher, the higher the number of observations, the greater the variance in x_i, the smaller σ^2, and the lower the intercorrelation among the explanatory variables. We have fixed β_i and d_i, and, given a pattern of inter-correlation among the explanatory variables, chosen σ to yield a desired value for $E(R^2)$. Finally, we have let n and $1/r^{ii}$ vary to see how the power of the test varies with these two factors. We have approached the problem in a numerical rather than an analytical fashion because of the difficulty of solution in the general case and because these results are more easily interpreted by those doing applied research. Since the analytical solution depends upon r^{ii}, for a many variable case power depends upon specification of the entire correlation matrix. The resulting problem is, in general, in several dimensions and not easily tabled. This is not true for the canonical case, of course, and an analytical solution in the canonical case can be found in Kendall and Stuart.*

---

*Kendall, M. G. and Alan Stuart, <u>The Advanced Theory of Statistics</u>, New York: Hafner, 1946, Ch. 24.

Tables for the canonical case have also been constructed (see the references in Kendall and Stuart). But we, of course, do not have the canonical case.

It is well known that collinearity does not bias the least squares estimate of β, but only increases its standard error, and hence reduces the power of the standard test. Even when the amount of collinearity was extreme, however, least squares yielded a reasonably powerful test of the hypothesis that neither β_1 nor β_2 equalled zero in the cases we examined. Each trial produced a pair of t-statistics (one on b_1, one on b_2, a t-statistic for b_0 was not calculated); in Table 1 we show the distribution of the minimum of this pair.*

In all but one case the power of the test seems quite high. Experiments 1-3 have but 25 observations per trial, and naturally the tests were least powerful here. In Experiment 1, with relatively low intercorrelation, both variables are significant at the 1 percent level 999 times out of 1000; in the remaining case, the minimum t-statistic is significant at the 2 percent level. In Experiment 2 the amount of collinearity has substantially increased; however, both coefficients are significant at the 1 percent level 83.8 percent of the time, and one coefficient is significant at 1 percent at the other at 5 percent or better 97.3 percent of the time. A wrong sign appeared only once

*Except in Experiments 3 and 6, the larger t-statistic was always significant at 1 percent. In Experiment 3 the larger t-statistic was significant at 1 percent only around 10 percent of the time, and was significant at the 5 percent level or higher slightly under 50 percent of the time. It was significant at the 10 percent level or higher around 60 percent of the time. In Experiment 6 the larger t-statistic was significant at the 1 percent level 98 percent of the time and was always significant at the 5 percent level.

Table 1

DISTRIBUTION OF MINIMUM t-STATISTIC

(figures in percent)

Exp. No.	Number of Observations Per Trial	Number of Trials	Correlation (r) Between x_1 and x_2	$E(R^2)$	Wrong Sign Significant at (2 Tail Test) 2%	5%	10%	50%	insig.	Correct Sign Significant at (2 Tail Test) insig.	50%	10%	5%	2%	1%	Total
1	25	1000	0.4973	0.933	-	-	-	-	-	-	-	-	-	0.1	99.9	100
2	25	1000	0.8969	0.946	-	-	-	-	0.1	-	1.0	1.6	5.6	7.9	83.8	100
3	25	1000	0.9908	0.949	0.1	-	0.7	←21.4→		←77.8→		-	-	-	-	100
4	100	1000	0.5013	0.927	-	-	-	-	-	-	-	-	-	-	100 (lowest t=8.26)	100
5	100	1000	0.9009	0.945	-	-	-	-	-	-	-	-	-	-	100 (lowest t=4.94)	100
6	100	1000	0.9901	0.947	-	-	-	0.1	0.6	3.5	25.9	15.6	25.4	16.0	12.9	100
7	500	200	0.5281	0.928	-	-	-	-	-	-	-	-	-	-	100 (lowest t=22.6)	100
8	500	200	0.8998	0.945	-	-	-	-	-	-	-	-	-	-	100 (lowest t=17.0)	100
9	500	200	0.9899	0.947	-	-	-	-	-	-	-	-	-	-	100 (lowest t=4.19)	100

in 1000 trials. Experiment 3 is the case where the power of the test is quite low; there is not a single trial where both coefficients are significant at even the 10 percent level. Further, as was implied by the footnote above, both are insignificant at the 10 percent level around 40 percent of the time.

In Experiments 4-6 we have increased the number of observations to 100 per trial. Except in the extreme case (Experiment 6), both coefficients are always significant at the 1 percent level. (We also show the lowest t-statistic in 1000 trials for either coefficient for Experiments 4 and 5.) In Experiment 6, both coefficients are significant at the 5 percent level or higher more than half the time, and significant at the 10 percent level or higher 70 percent of the time. In less than one percent of the trials did a wrong sign appear. With 500 observations per trial (Experiments 7-9) collinearity was not a problem over the range of values observed. Hence, we conclude that collinearity impairs hypothesis testing severely only in the most extreme cases.

One of the reasons given for concern about collinearity is that it may impair prediction. If one is willing to assume that the structure (the relationships among the explanatory variables) will remain unchanged, it is well known that collinearity does not impair the prediction of y. However, it is feared that if the structure changes, predictions using the parameter estimates from a collinear data set will not be very good. To obtain some idea of what happens to predictive ability when the structure changes, we selected 100 coefficients at random from Experiment 4, 100 from Experiment 5, and 100 from Experiment 6. Using Equation (1) we generated 100 new values of x and y for each structure represented by these three experiments and 100 values of x and y for a structure in which x_1 and x_2 were independent of each other. (In this latter case the error term was scaled so as to make $E(R^2) \sim 0.9$.) We then used each of the 100 estimated pairs of coefficients we had selected to predict one value of y when the explanatory variables were independent. For each structure we compared the predicted y with the actual y using R^2, or the proportion of variation in y which the predicted values of y explained, as a measure of goodness.

Table 2 shows the results of this experiment. When the structure was unchanged, prediction was unimpaired. But even when the structure was changed, so that x_1 and x_2 were independent, only in Experiment C did prediction markedly worsen. Even here the falloff in predictive ability was not as large as might have been supposed. In Experiments A and B, R^2 declined very little when the structure was changed. Hence, we again conclude that only in the most extreme case is predictive ability impaired when the underlying pattern of collinearity is altered.

Table 2

PREDICTION EXPERIMENT - 100 TRIALS

		R^2 Between Actual and Predicted Y
Experiment A:	$r_{x_1x_2} = 0.50$	0.8846
	x_1x_2 independent	0.8635
Experiment B:	$r_{x_1x_2} = 0.90$	0.9063
	x_1x_2 independent	0.8582
Experiment C:	$r_{x_1x_2}$ · 0.99	0.8803
	x_1x_2 independent	0.4790

Newhouse

IS COLLINEARITY A PROBLEM?

P- 4588