

---

# *The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations*

CARL F. MELA\* and PRAVEEN K. KOPALLE†

Duke University, Durham, North Carolina 27708, USA and †Dartmouth College, Hanover, New Hampshire, USA 03755

---

The purpose of this paper is to ascertain how collinearity in general, and the sign of correlations in specific, affect parameter inference, variable omission bias, and their diagnostic indices in regression. It is found that collinearity can reduce parameter variance estimates and that positive and negative correlation structures have an asymmetric effect on variable omission bias. It is also shown that the effects of collinearity are moderated by the relationship between the dependent variable and the regressors, a consideration not incorporated into most commonly used collinearity diagnostics. The formulae derived enable researchers to assess the sensitivity of regression results to the underlying correlation structure in the data.

## I. INTRODUCTION

The problem of collinearity in empirical research is among the most endemic concerns raised by marketers. In fact, a recent search in EconLit revealed 154 studies discussing collinearity or multicollinearity in their abstracts. A similar full text search of *Applied Economics* (using Infotrac) yielded 220 articles since 1991. Various econometric references have indicated that collinearity increases estimates of parameter variance, yields high  $R^2$  in the face of low parameter significance, and results in parameters with incorrect signs and implausible magnitudes (Belsley *et al.*, 1980; Kmenta, 1986; Greene, 1990).

To assess whether collinearity is indeed problematic, various diagnostics are frequently employed (see Appendix 1). If passed, then the results are assumed to be free of ‘problems.’ Green *et al.* Albaum (1988), Tull and Hawkins (1990) and Lehmann *et al.* (1998) respectively suggest 0.9, 0.35 and 0.7 as a threshold of bivariate correlations for the harmful effect of collinearity. Regarding other collinearity diagnostics, Belsley *et al.* (1980) and Johnston (1984) suggest condition indices (*CI*) less than 20 are not problematic, while Hair *et al.* (1995) suggest variance inflation factors (*VIF*) less than 10 are indicative

of inconsequential collinearity. In addition, the determinant of the regressor correlation matrix is also commonly used as a diagnostic of collinearity (Johnston, 1984).

In spite of the hegemony of these issues, there exist few, if any studies, that explicitly (i.e., analytically) link the correlation matrix to (i) the problems that these correlations cause and (ii) the ability of collinearity diagnostics to detect them. These diagnostics assume that different correlation matrices which happen to yield identical diagnostics will have the exact same effect on regression. It is shown that this is not often the case; some correlation structures will be more problematic than others even if they yield the same diagnostic. In particular, it is found that positive and negative correlations of equal magnitude can have very different effects on variable omission bias and variance inflation (i.e., the inflation of the parameter variance estimates) but often yield the same collinearity diagnostic.

Second, it is quite possible for one correlation structure to affect variable omission bias severely while having little effect on variance inflation (and vice versa), yet the diagnostic techniques are the same for both problems. This suggests that it can be misleading to use a single diagnostic to assess multiple problems. Third, the effects of collinearity are moderated by the correlations between the regres-

\* Corresponding author. E-mail mela@duke.edu.

Table 1. Applied Economics articles referencing the effect of collinearity

Reference	R reported	Diagnostic used	Effects of collinearity discussed	R problem?	Action taken
Hayo (1999)	No	Not reported	Variance inflation	Not reported	None
Panaopoulou and Tsakloglou (1999)	Partial	Bivariate correlations	Variance inflation	Yes	Change model
Vanhoudt (1999)	No	Not reported	No	No	None
Paci and Pigliaru (1999)	Partial	Bivariate correlations	Variance inflation	No	None
Aiello (1999)	No	Effect of variable addition on parameter errors	Variance inflation	No	None
Lofstrom (1999)	No	Condition Index	No	Yes	Change variable
Liu and Lynk (1999)	No	Effect of variable addition on parameter errors	Variance inflation	Yes	Change model
Greene (1999)	No	Effect of variable addition on $F$ and $R^2$	Variance inflation	Yes	Disaggregate data
Gerdtham <i>et al.</i> (1999)	Yes	Condition Index, $VIF$	No	No	None
Menahem (1999)	No	Magnitude of $R^2$	No	No	None
Natke (1999)	No	Not reported	Inability to conduct J-test	Yes	Change model
Erickson <i>et al.</i> (1999)	No	Singular value decomposition of data matrix	Variance inflation	Yes	Auxilliary regression of independent variables
Doroodian <i>et al.</i> (1999)	No	None	No	Not reported	Difference variables
Dutta and Ahmed (1999)	No	Correlation matrix	No	No	None
Brunton and Alexander (1999)	No	None	Variance inflation	Yes	None
Hansen (1999)	No	Bivariate correlations	Variance inflation	Yes	Change variable
Billington (1999)	No	None	Incorrect signs	No	None

sors and the dependent variable. All the aforementioned diagnostics ignore this fact. This suggests that it is imperative to consider these correlations when imputing the effects of correlation on regression.

To exemplify these points, Table 1 presents an overview of 18 articles appearing in *Applied Economics* that raised more than a passing concern about collinearity (this summary is limited to the last calendar year in order to conserve space). The plethora of recent articles concerned with collinearity suggest the importance of the topic. One of these 18 articles presents a correlation matrix. Regarding diagnostics employed, four papers used bivariate correlations (all with different cutoffs), one used a  $VIF$ , and two used condition indices. Two others used the effect of a variable omission on the significance of remaining parameters to assess whether collinearity was problematic. Ten articles outlined the potential consequences of collinearity (e.g., variable omission bias, impact on parameter variance estimates, sign reversal etc.) or addressed how the diagnostics measure them. Given (i) the prevalence of the issue, (ii) the lack of consensus on its effect, and (iii) the variation in measures used to assess it, it is helpful to analyse the effects of collinearity in greater detail.

Accordingly, the contribution of this paper is three-fold. First, precise analytical expressions are derived regarding the effect of collinearity on variable omission bias and variance inflation. It is shown that, *ceteris paribus*, collinearity can actually *deflate* parameter variance estimates in certain instances. Second, the conditions are presented under

which the impact of negative correlations on bias and variance inflation is greater than the impact of equivalent positive correlations and vice versa. Third, how various collinearity diagnostics perform are assessed under different correlation structures.

The paper proceeds as follows. First are derived multiple regressor expressions for the parameter estimates, parameter variance estimates,  $t$ -statistics, and  $R^2$  as a function of the underlying correlation structure in the data. Then, the asymmetric effects of the *magnitude* of positive and negative, correlations are illustrated within the context of two regressors. Next, two correlation matrices published in *Applied Economics* are analysed to assess the impact of collinearity on regression in an applied, 'real' context. Finally a number of general insights are concluded.

## II. MODEL

Classical statistics (e.g., Farrar and Glauber, 1967) suggests that there are two major factors that bear upon how insights from regression are affected by collinearity. These factors and their effects are categorized in Table 2. They include whether or not the data used to estimate the model are population data and whether or not the 'true' theoretical model is known and fully specified.

In cell 1, the model is properly specified and regression estimation is predicated upon the entire population with all appropriate variables. In this instance, parameter estimates

Table 2. Implications of collinearity for regression analysis

		Model specification (affects columns of the data matrix)	
		Model known and variables available	Model unknown and/or variables unavailable
Sampling (affects rows of the data matrix)	Population level	<i>Cell 1</i> • No problem	<i>Cell 2</i> • Variable omission bias • Inclusion of irrelevant variable bias
	Sample level	<i>Cell 3</i> • Variance inflation • Variance contraction	<i>Cell 4</i> • Variable omission bias • Inclusion of irrelevant variable bias • Variance inflation • Variance contraction

are all unbiased and collinearity presents no problem to the researcher. However, in many cases, the researcher does not have population data, does not know the true model, or does not have all the variables needed to properly specify the model (cell 2). When the model is misspecified, but estimated using population data, the parameter estimates in regression are biased (Johnston, 1984). This problem often arises when it is unclear which variables among a large set of potential regressors should be included in a regression model.

When the researcher has properly specified the model, but estimated it on a sample of the data (cell 3), parameters are unbiased even in the face of collinearity. However, collinearity can cause parameter variance estimates to increase (Lehmann *et al.*, 1998; Wittink, 1988). Surprisingly, as will be shown, it can also cause parameter variance estimates to decrease.

Finally, it is possible to both estimate the regression model at the sample level and have the model specification unknown (cell 4). This may be the most common case because (i) it is typically infeasible to collect population data, thus leading to sampling issues, (ii) theory is not typically complete enough to know a model's specification with certainty, and (iii) data are often either not complete enough to include all relevant variables, or contain more variables than can possibly be included in a regression. In these conditions, both variable omission bias and variance inflation (deflation) of parameter variance estimates can be problematic.

Given the forgoing concerns, it is helpful to derive formal expressions for omitted variable bias and variance inflation as a function of the correlation matrix of the regressors. This is done in the next section.

*The effects of collinearity on regression*

Given the forgoing concerns, it is helpful to derive formal expressions for omitted variable bias and variance inflation as a function of the correlation matrix. Accordingly, a

commonly observed case in social science research is considered, i.e.,  $n$  sample observations of a dependent variable,  $\mathbf{y}$ , and independent variables,  $\mathbf{X}$ , are observed by a researcher who is trying to ascertain the relationship between  $\mathbf{y}$  and  $\mathbf{X}$ .

This study commences with the case of a  $P$  variable regression ( $p = 1, 2, \dots, P$  indexes the regressors),

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{1}$$

where  $\mathbf{y}$  is a vector of  $n$  observations of the dependent variable,  $\mathbf{X}$  is a  $n \times P$  matrix of regressors,  $\beta$  is a  $P \times 1$  vector of population parameters, and  $\mathbf{e}$  is a  $n \times 1$  vector of errors,  $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ . To facilitate exposition, without loss of generality, standardized variables are used. Writing the expression for a  $P \times 1$  variable vector of standardized parameter estimates,  $\hat{\beta}$ , in terms of the sample correlation matrices (Kmenta, 1986; Johnson and Wichern, 1988; Morrison, 1990)

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) = \begin{bmatrix} n & nr_{12} & \dots & nr_{1p} \\ nr_{21} & n & \dots & nr_{2p} \\ \dots & \dots & \dots & \dots \\ nr_{p1} & nr_{p2} & \dots & n \end{bmatrix}^{-1} \begin{bmatrix} nr_{y1} \\ nr_{y2} \\ \dots \\ nr_{yP} \end{bmatrix} \\ &= \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{y1} \\ r_{y2} \\ \dots \\ r_{yP} \end{bmatrix} \\ &= R_{xx}^{-1}R_{xy} \end{aligned} \tag{2}$$

where  $r_{ij}$  is the correlation between  $x_i$  and  $x_j$  ( $i, j = 1, \dots, P$ ),  $r_{yi}$  is the correlation between  $x_i$  and  $y$  and  $n$  is the number of observations. The result in Equation 2 suggests that correlations impact other areas of substantive interest regarding parameter estimates. For example, following Johnston (1984, p. 260), it is noted that the effect of omitting  $j$  relevant variables,  $z_j$ , in Equation 1

biases the parameter estimates for the remaining  $P-j$  variables. This bias can be expressed as  $\mathbf{R}_{xx}^{*-1}\mathbf{R}_{xz}\beta_z$ , where  $\mathbf{R}_{xx}^*$  is the correlation matrix among the  $P-j$  independent variables;  $\mathbf{R}_{xz}$  represents the correlation between the  $j$  omitted variables and the  $P-j$  regressors; and  $\beta_z$  are the population parameters for the omitted variables in the fully specified ( $P$  variable) regression. As these population parameters are unknown, they are replaced with their estimators,  $\hat{\beta}_z$ .

To calculate the estimate of the variance of  $\hat{\beta}$ , we noted that the variance-covariance matrix<sup>1</sup> of  $\hat{\beta}$  is given by

$$E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\hat{\Sigma}_{\hat{\beta}} \equiv \hat{\sigma}^2\left(\frac{1}{n}\mathbf{R}_{xx}^{-1}\right) \quad (3)$$

where  $\mathbf{R}_{xx}$  is the  $(P) \times (P)$  correlation matrix of the regressors and  $\hat{\sigma}^2$  is the estimate of  $\sigma^2$ . As  $\hat{\sigma}^2$  is the conditional variance of  $\mathbf{y}$  given  $\mathbf{X}$  (Morrison, 1990), it may be expressed as,

$$\begin{aligned} \hat{\sigma}^2 &= s^2 = \text{var}(\mathbf{y}|\mathbf{X}) = \left(\frac{n}{n-P}\right)(1 - \mathbf{R}'_{xy}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}) \\ &= \left(\frac{n}{n-P}\right) \frac{|\mathbf{R}|}{|\mathbf{R}_{xx}|} \end{aligned} \quad (4)$$

where  $\mathbf{R}$  is the  $(P+1) \times (P+1)$  correlation matrix of the dependent variable and regressors.

Using Equations 2-4, the  $t$ -statistic for the  $p$ th parameter is given by

$$t_{\hat{\beta}_p} = \frac{\hat{\beta}_p}{\sqrt{\hat{\Sigma}_{\hat{\beta}_p}^2}} \quad (5)$$

where  $\hat{\Sigma}_{\hat{\beta}_p}^2$  is the  $p$ th diagonal element of the estimate of the variance-covariance matrix for the  $p$ th parameter estimate given in Equation 2. When data are standardized, the total sum of squares (SST) is given by  $n$ . As  $s^2 = \mathbf{e}'\mathbf{e}/(n-P)$  the error sum of squares (SSE) is given by  $s^2(n-P)$ , or, using Equation 4

$$\begin{aligned} SSE &= s^2(n-P) = \frac{n}{(n-P)}(1 - \mathbf{R}'_{xy}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy})(n-P) \\ &= n \frac{|\mathbf{R}|}{|\mathbf{R}_{xx}|} \end{aligned} \quad (6)$$

Therefore,  $R^2$  can be expressed as,

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = 1 - \frac{|\mathbf{R}|}{|\mathbf{R}_{xx}|} \quad (7)$$

The  $F$ -statistic is then derived as follows,

$$F_{P-1, n-P} = \frac{n-P}{P-1} \frac{R^2}{(1-R^2)} = \frac{(|\mathbf{R}_{xx}| - |\mathbf{R}|)(n-P)}{(|\mathbf{R}|)(P-1)} \quad (8)$$

<sup>1</sup>This analysis applied to both nonstochastic and stochastic regressors. When comparing different sample draws from the same distribution, the assumption of nonstochastic regressors may be untenable (Pindyck and Rubinfeld, 1981). However, under the assumption that the stochastic regressors are uncorrelated with the error term, Johnston (1984 p. 281-84) shows that  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  remains an unbiased estimator of the population level variance-covariance matrix,  $\Sigma$ , and that  $\hat{\beta}$  remains an unbiased estimator of the population parameters,  $\beta$ .

### Two views on regression

It is important to note that there exist two views on regression, and that the foregoing analysis is concordant with one, which is denoted 'the researcher's view.'

*The researcher's view.* While the researcher may or may not control the  $\mathbf{X}$  (observations of independent variables), the true process by which the corresponding observations of the dependent variable,  $\mathbf{y}$ , are generated is unknown to the researcher. Thus, the researcher hypothesizes a linear additive model as a paramorphic representation of the process that generates the  $\mathbf{y}$ . Next, given the  $\mathbf{X}$  and  $\mathbf{y}$ , the objective is to estimate regression parameters,  $\hat{\beta}$ , that provide the best fit to this model and to determine the precision of those estimates (Johnston, 1984; Kmenta, 1986; Stewart, 1987; Morrison, 1990). In this view, the parameters including  $\sigma^2$  are determined by the data. This implies, for example, that  $\hat{\beta} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ . Moreover, if  $\mathbf{R}_{xx}$  differs across samples of data,  $\mathbf{R}_{xy}$  does not necessarily need to differ as well.

*An alternative view.* This analysis may be compared to another approach where it is assumed there exists a known linear additive relationship that is indeed the true underlying process that generates the  $\mathbf{y}$  (Johnson *et al.*, 1989; Mason and Perreault, 1991). The parameters, including  $\sigma^2$ , are viewed to be constants and are independent of the  $\mathbf{X}$ . Thus, in this view,  $\mathbf{R}_{xy} = \mathbf{R}_{xx}\beta$  and the data are a function of the parameters (note that the  $\beta$  are true parameters, not estimates,  $\hat{\beta}$ ). Thus, if  $\mathbf{R}_{xx}$  changes, so, too will  $\mathbf{R}_{xy}$ . In this case, the parameters are known and do not change with the data.

This alternative view is not adopted in the current analysis for two reasons. First the true model is not typically known. The researcher knows only  $\hat{\beta}$ , not  $\beta$ . Thus, although the alternative view is of theoretical interest, it may be of little practical value to researchers. Second, there do exist samples of data where  $\mathbf{R}_{xx}$  vary and  $\mathbf{R}_{xy}$  do not. The alternative view presupposes this is not possible. Third, the linear model is an approximation (often very good) of an unknown process that combines the regressors into a dependent variable. In the researcher's view, that approximation is acceptable if predictions are good.

It may be shown that the two approaches are mathematically equivalent when  $\hat{\beta} = \beta$  (Johnson and Wichern, 1988). Nonetheless, it should be noted that the present findings must be interpreted in light of the differences in assumptions that distinguish these two approaches. For example,

in the alternative view, it can be shown that the assumption that  $\sigma^2$  is independent of the data suggests variance estimates *can not* decrease.

*The two regressor case*

Equations 1–8 are difficult to interpret in matrix form. Therefore, the approach is illustrated with the two regressor model in order to provide intuition for the asymmetric effect of positive and negative collinearity among the regressors. This enables a clear demonstration of the effect of collinearity on regression and portray some conditions under which the various collinearity diagnostics are informative and some conditions where they are misleading.

*Effect of collinearity on variable omission bias.* The regression model is given by,

$$y = \beta_1 x_1 + \beta_2 x_2 + e \quad e \sim N(0, \sigma^2) \quad (9)$$

where  $y$  is the dependent variable,  $x_1$  and  $x_2$  are regressors, and  $\beta_1, \beta_2$ , and  $\sigma^2$  are the population level parameters. As  $x_1, x_2$ , and  $y$  are standardized, there is no intercept in Equation 9.

From Equation 2, it is seen that the estimates of  $\beta$  in Equation 9 are given by (see also Johnston, 1984 p. 81)

$$\hat{\beta} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix} = \begin{bmatrix} \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \\ \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \end{bmatrix} \quad (10)$$

where  $r_{yj}$  is the sample correlation between  $y$  and the independent variable  $x_j, j = 1, 2$  and  $r_{12}$  is the sample correlation between  $x_1$  and  $x_2$ .

Recall, the omitted variable bias is given by,  $\mathbf{R}_{xx}^{-1} \mathbf{R}_{xz} \hat{\beta}_z$ . Using this result, it may be shown that the omission of  $x_j$  biases the regression coefficient of  $x_i$  by  $r_{12} \hat{\beta}_j$  (where  $i, j = 1, 2; j \neq i$ ; and  $\hat{\beta}_j$  is the parameter estimate for variable  $j$  in the fully specified model). When  $r_{12}$  is 0, there is no variable omission bias, consistent with the findings of Johnston (1984). However, when collinearity is present, there is variable omission bias, and in Proposition 1, the nature of the bias is elaborated on.<sup>2</sup>

*Proposition 1:*

- (a) For  $r_{yi}, r_{yj} > 0$ , the bias in  $\hat{\beta}_i$ , upon omitting variable  $j$  ( $i, j = 1, 2; j \neq i$ ), is strictly increasing in  $r_{12} < 0$ . The bias is strictly increasing (decreasing) in  $r_{12} > 0$ , if  $r_{yj}/r_{yi} > (<)2r_{12}/[1 + r_{12}]^2$ .

- (b) For  $r_{yi}, r_{yj} < 0$ , the bias in  $\hat{\beta}_i$ , upon omitting variable  $j$ , is strictly decreasing in  $r_{12} < 0$ . It is strictly increasing (decreasing) in  $r_{12} > 0$ , if  $r_{yj}/r_{yi} < (>)2r_{12}/[1 + r_{12}]^2$ .
- (c) For  $r_{yi} < 0$  and  $r_{yj} > 0, i, j = 1, 2, j \neq i$ , the bias in  $\hat{\beta}_i$ , upon omitting variable  $j$  is strictly increasing in  $r_{12} > 0$ . It is strictly increasing (decreasing) in  $r_{12} < 0$ , if  $r_{yj}/r_{yi} < (>)2r_{12}/[1 + r_{12}]^2$ . The bias in  $\hat{\beta}_j$ , upon omitting variable  $i$  is strictly decreasing in  $r_{12} \forall r_{12} > 0$ . It is strictly increasing (decreasing) in  $r_{12} < 0$ , if  $r_{yi}/r_{yj} > (<)2r_{12}/[1 + r_{12}]^2$ .

*Proof.* See Appendix 2.

For example, when  $r_{yi} = r_{yj} = r_y > 0$ , omitted variable bias is negative and decreasing as the correlation ( $r_{12}$ ) with the omitted variable becomes increasingly negative. The bias is positive and increasing when  $r_{12}$  becomes increasingly positive. Omitted variable bias is zero when  $r_{12} = 0$ . Further,  $|\partial \text{bias } \beta_i / \partial r_{12}|, i = 1, 2$ , is greater for negative correlations than it is for equivalent positive correlations (see Appendix 2). Thus, there exists an asymmetry in omitted variable bias to the underlying correlation structure. For example, when  $r_y > 0$ , the slope,  $|\partial \text{bias } \beta_i / \partial r_{12}|$ , at  $r_{12} = 0.5$  is  $0.67r_y$ , whereas the slope at  $r_{12} = -0.5$  is  $4r_y$ , six times  $|\partial \text{bias } \beta_i / \partial r_{12}|$  at  $r_{12} = 0.5$ .

To more clearly illustrate this relationship, (i) how omitted variable bias varies with different possible values of  $r_{12}$  given a fixed  $r_{y1} = r_{y2} = r_y$  (following Wittink, 1988, p. 89–90)<sup>3</sup> and (ii) the degree to which diagnostics of collinearity capture the omitted variable bias, were assessed. Figure 1 depicts the relationship between  $r_{12}$  and the bias in  $\hat{\beta}_i$ , upon omitting variable  $j$ , for  $r_{y1} = r_{y2} = 0.5$ . It also portrays the *VIF*, Condition Index, and determinant. Figure 1 shows that when  $r_y > 0$ , the effect of variable omission bias is much greater when a negatively correlated variable is omitted than when a corresponding positively correlated variable is omitted. Second, this asymmetry becomes greater as  $r_y$  increases. The opposite result holds when  $r_y < 0$ .

Figure 1 reveals some disconcerting insights regarding common diagnostics of collinearity. First, as can be seen, omission bias is much worse for negative than positive correlations of equivalent magnitude. Yet the *VIF* is the same value for negative and positive correlations of equal magnitude, as are the *CI* and the determinant. Even though the bias is over three times greater for  $r_{12} = -0.5$  than for  $r_{12} = 0.5$ , each of the respective diagnostics (*VIF*, *CI*, determinant) suggest the omission bias should be the same.

<sup>2</sup> Note, in this and subsequent propositions, a change in  $r_{12}$  implies a change across samples or populations. It is impossible for  $r_{12}$  to change within a sample or population. Note that comparisons across samples or populations are common; collinearity diagnostics are designed to compare differences in inter-regressor correlations across samples or populations.

<sup>3</sup> Note that values for  $r_{12}$  must be valid (i.e., the correlation matrix among the independent variables and the dependent variable must be positive semidefinite).

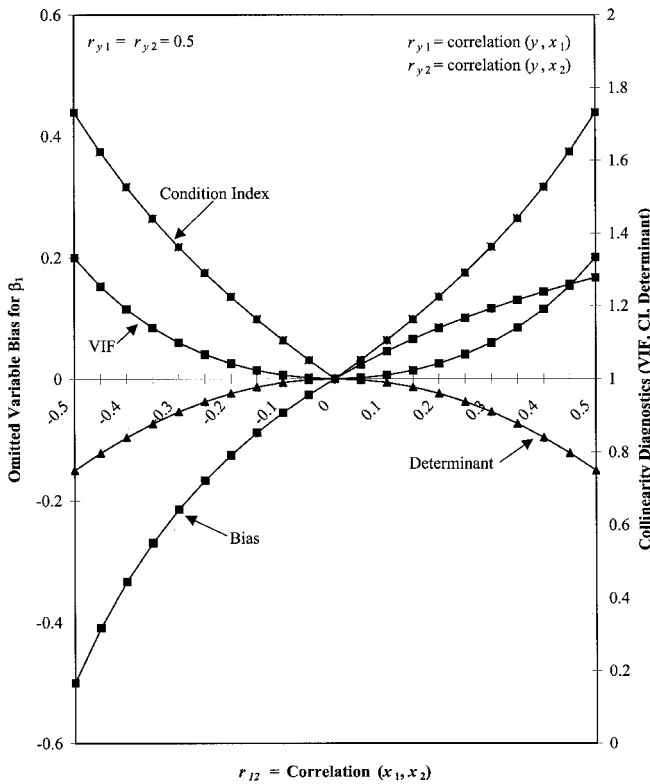


Fig. 1. Effect of collinearity on variable omission bias  
 True model:  $y = \beta_1 x_1 + \beta_2 x_2$ ; Omitted variable model:  $y = \beta_1 x_1$

Second, the direction of the omission bias reverses with the sign of the collinearity (that is, the omission bias is negative for negative correlations and positive for positive correlations). The diagnostics do not capture this. Conversely, formulas regarding the impact of collinearity (such as the ones we propose) offer a more detailed insight regarding whether collinearity is a concern.

*Effect of collinearity on parameter variance estimates.* To express the effect of collinearity on parameter variance estimates, using Equations 3–4, and 10 and simplifying gives,

$$\hat{\text{var}}(\hat{\beta}_2) = \hat{\text{var}}(\hat{\beta}_3) = \frac{(1 - r_{y1}^2 - r_{y2}^2 - r_{12}^2 + 2r_{y1}r_{y2}r_{12})}{(n - 2)((1 - r_{12}^2))^2} \quad (11)$$

The presence of  $r_{12}$  in the numerator of Equation 11 suggests that positive and negative correlations have differing effects.

*Proposition 2.* The parameter variance estimates for both  $x_i$  and  $x_j$  ( $i, j = 1, 2, j \neq i$ ) are strictly increasing (decreasing) in  $r_{12}$  if  $r_{yi}r_{yj} > (<)r_{12}[(1 - r_{12}^2 - 2D)/(1 - r_{12}^2)]$ , where  $D \in [0, 1]$  is the determinant of the correlation matrix of  $y, x_1$  and  $x_2$ .

*Proof.* See Appendix 2.

Proposition 2 suggests that the result found in most econometric texts, i.e., correlations among the regressors raise parameter variance estimates, holds only under certain conditions. Surprisingly, Proposition 2 indicates that increasing collinearity can decrease parameter variance estimates. For example, when  $r_{yi} < 0, r_{yj} > 0, D > 0.5$ , and  $r_{12} < 0$ , the parameter variance estimate strictly decreases as  $r_{12}$  becomes more negative. In this instance, increasing collinearity yields lower parameter variance estimates. Similarly,  $\partial(\hat{\text{var}}(\hat{\beta}))/\partial r_{12} > 0$  when  $r_{12} = 0$  and  $r_{y1}r_{y2} > 0$ . In this case, any small negative correlation decreases the variance estimate of  $\hat{\beta}$ . For simplicity, let  $r_{y1} = r_{y2} = r_y$ . Upon substitution into the expression in the Appendix 2 for  $\partial(\hat{\text{var}}(\hat{\beta}))/\partial r_{12}$  and simplifying, it is found that  $\hat{\text{var}}(\hat{\beta})$  is increasing (decreasing) in  $r_{12} < 0$  when  $r_y^2 > (<)r_{12}(1 + r_{12})/3r_{12} - 1$ . If  $r_{12} > 0$ , then  $\hat{\text{var}}(\hat{\beta})$  is increasing in  $r_{12}$ . Thus, for sufficiently large  $r_{y1}$  and  $r_{y2}$ , increasing negative collinearity decreases parameter variance estimates. To further illustrate the properties of parameter variance estimates outlined in Proposition 2, Fig. 2 depicts the complex relationship between  $\hat{\text{var}}(\hat{\beta})$  and  $r_{12}$ , when  $r_{y1} = r_{y2} = r_y = 0.5$ . The common collinearity diagnostics as a function of  $r_{12}$  has been overlaid.

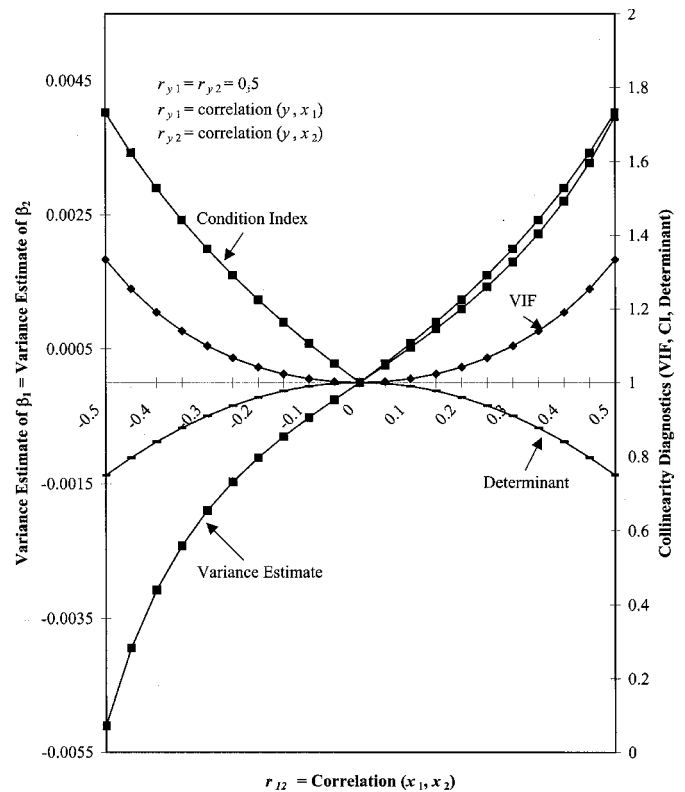


Fig. 2. Effect of collinearity on parameter variable estimates  
 Model:  $y = \beta_1 x_1 + \beta_2 x_2$

As with parameter estimates, the effect of positive and negative collinearity is asymmetric. Although not depicted in Fig. 2,  $r_y$  affects the monotonicity of the relationship. When  $r_y$  is sufficiently high, increasing  $r_{12}$  yields a greater parameter variance estimate. When  $r_y$  is sufficiently low, the relationship can become non-monotonic (inverted u-shaped).

All collinearity diagnostics suggest parameter variance estimates will inflate as the correlation becomes negative, when in fact, it can contract. Thus, the diagnostics could suggest that collinearity is a problem (inflates parameter variance) when it is not (deflates parameter variance). As with Fig. 1, the effect of correlations on diagnostics is symmetric (for positive and negative correlations of equal magnitude) even though the effect of collinearity on regression is not.

*Parameter variance simulation.* To illustrate the effect of negative and positive correlations on parameter variance estimates (i.e., make inferences regarding how well regression recovers population parameters), sample data for  $y$ ,  $x_1$ , and  $x_2$  were generated from a multivariate normal distribution. The following three population correlation matrices (where  $\rho_{12} = \{-0.5, 0, 0.5\}$ ) were used among the dependent and the independent variables.

Given the design in Table 3, the results presented in Proposition 2 and Fig. 2 lead to the following hypothesis:

*Hypothesis:* When  $r_{12}, r_{y1}$ , and  $r_{y2}$  are determined using data generated from Table 1, the variance of the parameter estimates in a negatively correlated environment

Table 3. Simulation design for population correlations

	$y$	$x_1$	$x_2$
$y$	1.0		
$x_1$	0.4	1.0	
$x_2$	0.4	$\rho_{12}$	1.0

Table 4. Simulation results

$\rho_{12}$	mean $\hat{\beta}_1$ estimated	mean $\hat{\beta}_2$ estimated	$\beta_1, \beta_2$ population parameters	$\text{var}(\hat{\beta}_2)$ estimated	$\text{var}(\hat{\beta}_2)$ estimated	var $\beta_1$ , var $\beta_2$ population parameters
0.5	0.266	0.265	0.267	0.011	0.011	0.011
0.0	0.391	0.400	0.400	0.007	0.007	0.007
-0.5	0.797	0.806	0.800	0.005	0.005	0.005

( $\rho_{12} = -0.5$ ) will be lower than those generated in a positively correlated environment ( $\rho_{12} = 0.5$ ).

To generate the data implied by Table 3, calculating population level parameters for  $\beta_1, \beta_2$ , and  $\sigma$  from the population correlation matrix was undertaken. Next,  $\rho_{12}$  was used to generate 100 sample observations of  $x_1$  and  $x_2$ . Using these values for  $\beta_1, \beta_2, x_1, x_2$ , and  $\sigma$ , 100 sample observations of  $y$  were generated.

This process was repeated 500 times yielding 500 sample data sets of 100 observations each. For each of these 500 data sets,  $y$  was regressed on  $x_1$  and  $x_2$  yielding 500 sample level estimates of  $\beta_1$  and  $\beta_2$ . Using the mean of the estimated parameters, the parameter variance estimates for  $\beta_1$  and  $\beta_2$  was calculated. Finally, this process was repeated for each of the three values of  $\rho_{12}$ .<sup>4</sup>

The results of the simulation are presented in Table 4. As predicted in the Hypothesis, the results in Table 4 demonstrate that parameter variance estimates are substantially lower in negatively (versus positively) correlated environments. By comparing the first and second columns to the third column, it may be observed that the mean parameter estimates appear unbiased (cf. Johnston, 1984). Second, by comparing the last three columns, it can be noted that the simulated parameter variance estimates are very close to those predicted by Equation 11. The findings of this simulation demonstrate that the sign of correlation among the independent variables affects parameter inference in regression and that certain correlation structures can actually reduce parameter sampling variance.

*Rescaling.* Rescaling a regressor by changing its sign does not mitigate the asymmetry outlined in this paper. For example, reversing the sign of  $x_2$  changes a correlation structure of  $r_{y1}, r_{y2}, -r_{12}$  to  $r_{y1}, -r_{y2}, r_{12}$ . As can be seen in Equations 10–13, rescaling in this fashion simply changes the sign of the parameter and its  $t$ -statistic while having no effect on omission bias,  $R^2$ , or parameter variance estimates. Thus, the size and direction of the asymmetry remain robust to rescaling for omission bias,  $R^2$ ,

<sup>4</sup>An additional 500 samples were generated for each of the three conditions by holding  $r_{12}$  fixed across each of the 500 samples. This was done by using the same data for  $X$  in each of the 500 samples. The  $X$  in each of the three conditions were chosen to reflect the population correlations  $r_{12}$ . The results were identical.

and the parameter variance estimate. However, the direction of the asymmetry changes for the  $t$ -statistic.

### III. APPLICATIONS

In this section, the two regressor results are applied using two examples from *Applied Economics*. The key purpose of this section is to demonstrate that the computation of the formulae in Section II yield a more accurate depiction of the effects of collinearity than the various collinearity indices previously enumerated. In particular, real examples are used to demonstrate that variable omission bias can be greater even when (i) variance inflation factors are lower, (ii) determinants are higher, and (iii) condition indices are lower.

Specifically, the following subset of the correlation matrices reported by Hojman (1992) and Nguyen and Cosset (1995) in *Applied Economics* are considered: Notice the negative correlations in the Hojman (1992) matrix. These suggest that variable omission bias may be worse in that instance. To see this, standardized regression coefficients were first calculated for each of the two two-regressor models in Table 5. Then, dropping the 'HB' variable from the first model and the 'FCON' variable from the second model, the standardized coefficients for the remain-

ing regressor were recalculated in each of the two models. Table 6 presents the resulting regression diagnostics and the mean omitted variable bias for each regression model (see Section II for the formula).

Table 6 vividly depicts the nonmonotonicity in the collinearity diagnostics. Recall, higher VIF, lower determinants and higher condition indices are all suppose to be indicative of *greater* problems arising from collinearity. Table 6 portrays a case where problems are *lesser* even though the VIF is higher, the determinant is lower, and the condition index is higher. Perhaps an even more misleading aspect of the diagnostics is that they suggest no collinearity problems are likely to exist in either case as the CI is well below the threshold of 30 and the VIF is substantially below 10. Yet, in the Hojman case the parameter estimate for  $U$  changes by 225% when HB is added to a model with only  $U$  as the independent variable.<sup>5</sup>

These problems arise because the diagnostics do not accommodate differences between negative and positive correlations, nor do they consider relationships with the dependent variable. The formulae presented in Section II more accurately assess the impact of collinearity on any variable of interest and enable better decisions regarding whether corrective techniques (e.g., ridge regression) need to be taken.

Table 5. *Correlation matrices*

Hojman (1992)

	$IMR$ (Birth rate) <sup>1</sup>	$U$ (Unemployment rate)	$HB$ (Ratio between health expenditures per capita and birth rate)
$IMR$ <sup>1</sup>	1		
$U$	-0.77	1	
$HB$	-0.28	-0.35	1

Nguyen and Cosset (1995)

	$FS$ (Firm sales) <sup>1</sup>	$FEMP$ (Foreign employees)	$FCON$ (Foreign countries)
$FS$ <sup>1</sup>	1		
$FEMP$	0.47	1	
$FCON$	0.58	0.40	1

Note: <sup>1</sup>Dependent variable.

Table 6. *Omission bias and collinearity diagnostics*

Matrix	Mean omission bias	Determinant	Condition index	$VIF$
Hojman (1992)	0.31	0.88	2.08	1.14
Nguyen and Cosset (1995)	0.15	0.84	2.33	1.19

<sup>5</sup> The coefficient for  $U$  changes from -0.77 to -0.99 when HB is added to the model. The coefficient for HB changes from -0.28 to -0.63 when  $U$  is added. Similarly, the coefficient for FEMP changes from 0.47 to 0.28 when FCON is added. FCON changes from 0.58 to 0.47 when FEMP is added.

## IV. CONCLUSION

*Key findings*

In this paper the effects of collinearity on omitted variable bias and parameter variance estimates are examined. It is found that, consistent with prior results, positive correlations can yield less precise estimates, can induce parameters to switch signs, and affect  $R^2$ . These findings are extended in three ways. First, it is shown that this effect is moderated asymmetrically by the signs of the inter-regressor correlations and correlations between the dependent variable and the regressors (denoted by  $\mathbf{R}_{xy}$ ). Second, some conditions are suggested under which the impact of negative correlations on parameter inference and model fit is greater than the impact of corresponding positive correlations and vice versa. Finally, analytical results are derived for the impact of collinearity and discuss the implications of the results for collinearity diagnostics in the context of omitted variable bias and variance inflation. The following summarize the findings for the case of all elements in  $\mathbf{R}_{xy} > 0$  (the other cases are presented in Appendix 2):

- Negative correlations among the independent variables have a much greater impact on variable omission bias than equivalent positive correlations. This has important ramifications for model specification in negatively correlated environments.
- A commonly stated result is that collinearity increases parameter variance estimates (e.g., Greene, 1990; Johnston, 1984; etc.). It is found that, under certain conditions, collinearity can reduce parameter variance estimates.
- Condition indices, variance inflation factors, and the determinant of the correlation matrix, as collinearity diagnostics, do not discriminate well between positive and negative correlations. Also, these diagnostics do not capture the effects of  $\mathbf{R}_{xy}$ , which have been shown to be an important moderator of collinearity's effects.
- The formulae derived can be used to perform a sensitivity analysis of key model statistics.

*Future research*

The analysis also suggests a number of extensions to other linear models. Structural models (Bagozzi, 1977), one of the most widely used linear models in business economics, may be subject to the same effects and the analysis can be extended to that methodology.

Second, the analysis pertains to the effects of collinearity on fit and inference. Assessing these effects is important because it is necessary to know whether collinearity is problematic or not before attempting to correct for it. Therefore, a useful extension of the work would be the development of collinearity diagnostics that incorporate

correlation signs and the regressor-dependent variable correlations.

Finally, regression is not the only technique subject to the problems arising from collinearity. Nonlinear models and limited dependent variable models are likely also affected by collinearity. A rigorous analysis of negative and positive correlations in these domains would also be important to further understand the reliability and validity of parameter estimates in those settings.

Given the ramifications and broad applicability of the principles outlined in the analysis, it is hoped that the findings inspire further work regarding the effect of collinearity in general, and the asymmetric impact of negative and positive correlations in particular. Fox and Monette (1992, p. 183), in the *Journal of the American Statistical Association*, note:

A reviewer of an earlier version of this article suggested that 'collinearity is not so much a problem as a state of nature – like the law of gravity – and that railing against collinearity is rather like complaining about not being able to fly by flapping your arms.' Although we have some sympathy with this view, we believe it overstates the case: The identification of specific sources of imprecision in estimation may, in certain instances, suggest how estimates can be improved, for example, by collecting additional data (abandoning arm flapping and trying an airplane) . . . the respecification of a statistical model or reorientation of the goals of a study.

It is hoped that this paper addresses 'these sources of imprecision' and leads to further research in this domain.

## ACKNOWLEDGEMENTS

The authors would like to thank Kusum Ailawadi, Neil Beckwith, Kamel Jedidi, Donald R. Lehmann, Lawrence Marsh, Bill McDonald, H. Fred Mittelstaedt, Scott Neslin, Stephen Powell and seminar participants at Dartmouth College and the University of Notre Dame for their helpful comments on earlier versions of this manuscript.

## REFERENCES

- Aiello, F. (1999) Effects of STABEX on ACP's economic growth: further evidence, *Applied Economics*, **31**(9), 1033–42.
- Bagozzi, R. P. (1977) Structural equation models in experimental research, *Journal of Marketing Research*, **14**, 209–26.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics – Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.
- Billington, N. (1999) The location of foreign direct investment: an empirical analysis, *Applied Economics*, **31**(1), 65–76.
- Brunton, R. A. and Alexander, R. W. (1999) Aggregate investment in New Zealand pre and post-restructuring *Applied Economics*, **31**(3), 287–92.

- Doroodion, K., Jung, C. and Boyd, R. (1999) The *J*-curve effect and US agricultural and industrial trade, *Applied Economics*, **31**(6), 687–95.
- Dutta, D. and Ahmed, N. (1999) An aggregate import demand function for Bangladesh: a cointegration approach, *Applied Economics*, **31**(4), 465–72.
- Ereckson, O. H., Platt, G., Whistler, C. and Ziegart, A. (1999) Factors influencing the adoption of state lotteries, *Applied Economics*, **31**(7), 875–84.
- Farrar, D. E. and Glauber, R. R. (1967) Multicollinearity in regression analysis: the problem revisited, *Review of Economics and Statistics*, **49**, 92–107.
- Fox, J. and Monette, G. (1992) Generalized collinearity diagnosis, *Journal of the American Statistical Association*, **193**(87), 178–83.
- Gerdtham, U. G., Rehnberg, C. and Tambour, M. (1999) The impact of internal markets on health care efficiency: evidence from health care reforms in Sweden, *Applied Economics*, **31**(8), 935–45.
- Green, E., Tull, D. S. and Albaum, G. (1988) *Research for Marketing Decisions*, 5th edn, Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Greene, C. A. (1999) On the impossibility of a stable and low GDP elasticity of money demand: the arithmetic of aggregation, replication and income growth, *Applied Economics*, **31**(9), 1119–27.
- Greene, W. H. (1990) *Econometric Analysis*, Macmillan Publishing Company, New York.
- Hair, J. F., Jr., Anderson, R. E., Tatham, R. L. and Black, W. C. (1995) *Multivariate Data Analysis*, 3rd ed, Macmillan Publishing Company, New York.
- Hansen, E. (1999) A pricing-to-market model with unobserved variables: explaining New Zealand's import prices, *Applied Economics*, **31**(1), 3–8.
- Hayo, B. (1999) Money-output Granger causality revisited: an empirical analysis of EU countries, *Applied Economics*, **31**(11), 1489–501.
- Hojman, D. E. (1992) Evolution of child and infant mortality in Chile: A model, *Applied Economics*, **24**(10), 1173–79.
- Johnson, E. J., Meyer, R. J. and Ghose, S. (1989) When choice model fail: compensatory models in negatively correlated environments, *Journal of Marketing Research*, **26**(3), 255–70.
- Johnson, R. A. and Wichern, D. W. (1988) *Applied Multivariate Statistical Analysis*, 2nd edn, Prentice Hall, Englewood Cliffs, New Jersey.
- Johnston, J. (1984) *Econometric Methods*, 3rd edn, McGraw-Hill Publishing Company, New York.
- Kmenta, J. (1986) *Elements of Econometrics*, 2nd edn, Macmillan Publishing Company, New York.
- Lehmann D. R., Gupta, S. and Steckel, J. (1988) *Marketing Research*, Addison-Wesley Educational Publishers, Inc., Reading, Massachusetts.
- Liu, Z. and Lynk, E. L. (1999) Evidence on market structure of the deregulated US airline industry, *Applied Economics*, **31**(9), 1083–92.
- Lofstrom, A. (1999) Can job evaluation improve women's wages? *Applied Economics*, **31**(9), 1053–60.
- Mason, C. H. and Perreault Jr., W. D. (1991) Collinearity, power and interpretation of multiple regression analysis, *Journal of Marketing Research*, **28**(3) 268–80.
- Menahem, G. (1999) A target level of risk model of respiratory pathologies and smoking behavior, *Applied Economics*, **31**(6), 709–22.
- Morrison, D. F. (1990) *Multivariate Statistical Methods*, 3rd edn, McGraw-Hill Publishing Company, New York.
- Natke, P. A. (1999) Financial repression and firm self-financing of investment: empirical evidence from Brazil, *Applied Economics*, **31**(8), 1009–19.
- Nguyen, T. and Cosset, J.-C. (1995) The measurement of the Degree of Foreign Involvement, *Applied Economics*, **27**(4), 343–51.
- Paci, R. and Pigliaru, F. (1999) Is dualism still a source of convergence in Europe? *Applied Economics*, **31**(11), 1423–36.
- Panapoulou, G. and Tsakloglou, P. (1999) Fertility and economic development: theoretical considerations and cross-country evidence, *Applied Economics*, **31**(11), 1337–51.
- Pindyck, R. S. and Rubinfeld, D. L. (1981) *Econometric Models and Economic Forecasts*, 2nd edn, McGraw-Hill Inc, New York.
- Stewart, G. W. (1987) Collinearity and least squares regression, *Statistical Science*, **32**(1), 68–100.
- Tull, D. S. and Hawkins, D. I. (1990) *Marketing Research*, 5th edn, Macmillan Publishing Company, New York.
- Vanhoudt, P. (1999) Are public and private outlays for physical and knowledge capital accumulation equally productive, *Applied Economics*, **31**(11), 1401–410.
- Wittink, D. R. (1988) *The Application of Regression Analysis*, Simon & Schuster, Needham Heights, Massachusetts.

## APPENDIX 1

### *Diagnostics for Collinearity*

Several collinearity diagnostics are commonly employed in business economics. One such diagnostic, the condition index, is the square root of the ratio of largest to smallest eigenvalues in the correlation matrix of the independent variables. Belsley *et al.* (1980) and Johnston (1984, p. 250) suggest that condition indices in excess of 20 are problematic. However, this diagnostic does not consider correlations with the dependent variable, which have already been shown to moderate the effects of collinearity. Second, in the two variable case, regardless of whether  $r_{12} = 0.5$  or  $r_{12} = -0.5$ , the eigenvalues are [1.5, 0.5]. Thus, the eigenvalues do not reflect the direction of the correlation as the eigenvalues are the same in both cases and the resulting condition index is 1.73. Yet, it has been shown that the two correlation structures have very different effects.

A second commonly used diagnostic is the regressor correlation matrix determinant. This diagnostic ranges from 1 when there is no collinearity, to 0 when there is perfect collinearity. Like the condition index, this diagnostic does not incorporate the moderating effect of correlations with the dependent variable. Second, in the two regressors model, regardless of whether  $r_{12} = 0.5$  or  $r_{12} = -0.5$  the determinant is 0.75. The determinant suggests the effects of collinearity are equivalent in each case even though we have shown the effects to be quite different.

A third collinearity diagnostic, the variance inflation factor (VIF), is also commonly used. The VIF for regressor  $x_i$  is given by  $VIF_i = 1/(1 - r_i^2)$  where  $r_i^2$  is the  $R^2$  of a regres-

sion of regressor  $x_i$  on all the remaining regressors. When a regressor is orthogonal,  $VIF = 1$ .  $VIF$ 's in excess of 10 ( $r_i^2 > 0.90$ ) are considered to be problematic (Hair *et al.*, 1995). In the foregoing example, the  $VIF$  is 1.33, regardless of whether  $r_{12} = 0.5$  or  $r_{12} = -0.5$ . Thus, the diagnostic has the same limitations as the condition index and the determinant in its ability to discriminate between positive and negative correlations. Second, like the other diagnostics, it does not consider the moderating role of the correlations between the dependent and the independent variables.

APPENDIX 2

Impact on omitted variable bias: proof of proposition 1a-1c

Bias in  $\hat{\beta}_i = bias_i = \hat{\beta}_j r_{12}$ ,  $i, j = 1, 2, j \neq i$ . Taking the partial derivative w.r.t.  $r_{12}$ , and simplifying yields,

$$\frac{\partial bias_i}{\partial r_{12}} = \frac{r_{yj}(1 + r_{12}^2) - 2r_{yi}r_{12}}{(1 - r_{12}^2)^2}$$

For a fixed magnitude of  $r_{yi}$  and  $r_{12}$ , and  $r_{yi} > 0$ ,  $\partial bias_i / \partial r_{12}$  when  $r_{12} > 0$  is always less than  $\partial bias_i / \partial r_{12}$  when  $r_{12} < 0$ . This suggests that variable omission bias is more sensitive to negative correlations when  $r_{yi} > 0$ . The opposite holds when  $r_{yi} < 0$ . Further, from the expression for  $\partial bias_i / \partial r_{12}$  it may be readily seen that

$$\text{Sign} \left[ \frac{\partial bias_i}{\partial r_{12}} \right] = \text{Sign} [r_{yj}(1 + r_{12}^2) - 2r_{yi}r_{12}].$$

Therefore,

- (a) For  $0 < r_{yi}, r_{yj} < 1$ , it is found that  $\partial bias_i / \partial r_{12} > 0$  for  $r_{12} < 0$ .

For  $r_{12} > 0, \partial bias_i / \partial r_{12} \geq 0$

$$\text{if } \frac{r_{yj}}{r_{yi}} \geq \frac{2r_{12}}{1 + r_{12}^2}.$$

- (b) For  $-1 < r_{yi}, r_{yj} < 0$ , it is found that  $\partial bias_i / \partial r_{12} < 0$  for  $r_{12} < 0$ .

For

$$r_{12} > 0, \frac{\partial bias_i}{\partial r_{12}} \geq 0 \text{ if } \frac{r_{yj}}{r_{yi}} \geq \frac{2r_{12}}{1 + r_{12}^2}.$$

- (c) Regarding correlations with mixed signs, without loss of generality let  $-1 < r_{yi} < 0$  and  $r_{yj} > 0$ ,  $i, j = 1, 2, j \neq i$ . The bias in  $\hat{\beta}_{ij} = bias_j = \hat{\beta}_i r_{12}$ . Taking the partial derivative w.r.t.  $r_{12}$ , and simplifying gives,

$$\text{Sign} \left[ \frac{\partial bias_j}{\partial r_{12}} \right] = \text{Sign} [r_{yi}(1 + r_{12}^2) - 2r_{yj}r_{12}].$$

Thus, it is found that for  $r_{12} < 0, \partial bias_i / \partial r_{12} \geq 0$  if  $r_{yj} / r_{yi} \geq 2r_{12} / (1 + r_{12}^2)$ ;  $\partial bias_j / \partial r_{12} \geq 0$  if  $r_{yi} / r_{yj} \geq 2r_{12} / (1 + r_{12}^2)$ . For  $r_{12} > 0, \partial bias_i / \partial r_{12} > 0, \partial bias_j / \partial r_{12} < 0$ .

Impact on parameter variance estimate: proof of proposition 2

Taking the partial derivative of Equation 11 with respect to  $r_{12}$  and simplifying gives,

$$\text{Sign} \left[ \frac{\partial(\text{var}(\hat{\beta}))}{\partial r_{12}} \right] = \text{Sign} [(1 - r_{12}^2)(r_{y1}r_{y2} + r_{12}) + 2r_{12}D]$$

where  $D \in [0, 1]$  is the determinant of  $\mathbf{R}_{xx}\mathbf{R}_{xy}$ . Thus,

$$\frac{\partial(\text{var}(\hat{\beta}))}{\partial r_{12}} \geq 0 \text{ if } r_{y1}r_{y2} \geq \frac{r_{12}(1 - r_{12}^2 - 2D)}{1 - r_{12}^2}$$