Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 03-2017

Bernd Hayo

On Standard-Error-Decreasing Complementarity: Why Collinearity is Not the Whole Story

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

On Standard-Error-Decreasing Complementarity: Why Collinearity is Not the Whole Story

Bernd Hayo

University of Marburg

This version: 09 January 2017

Corresponding author:

Bernd Hayo
School of Business and Economics (FB 02)
Philipp-Universität Marburg
Universitätsstr. 24
D-35037 Marburg
Germany
Tel.: +49-(0)6421-28-23091
Fax: +49-(0)6421-28-23088
Email: hayo@wiwi.uni-marburg.de

Special thanks to Olaf Korn, Florian Neumeier, Duncan Roth, and participants of research seminars at the Universities of Dallas at Texas and Marburg for helpful comments. The usual disclaimer applies.

On Standard-Error-Decreasing Complementarity: Why Collinearity is not the Whole Story

Abstract

There is a widespread belief among economists that adding additional variables to a regression model causes higher standard errors. This note shows that, in general, this belief is unfounded and that the impact of adding variables on coefficients' standard errors is unclear. The concept of standard-error-decreasing complementarity is introduced, which works against the collinearity-induced increase in standard errors. How standard-error-decreasing complementarity works is illustrated with the help of a nontechnical heuristic, and, using an example based on artificial data, it is shown that the outcome of popular econometric approaches can be potentially misleading.

Keywords: Standard-error-decreasing complementarity, multivariate regression model, standard error, econometric methodology, multicollinearity, collinearity

JEL: C1, B4

1. Introduction

Economists often believe that including additional variables in multivariate linear regression models is problematic and should be avoided. This belief is based on the effects of multicollinearity. For instance, Baltagi (2002, p. 80) notes that in the presence of multicollinearity, ordinary least square (OLS) 'estimates are unreliable as reflected by their high variances'. Wooldridge (2013, p. 92) states that 'for estimating β_j, it is better to have less correlation between x_j and the other independent variables'. In Green (2012, p. 129), Kennedy (1992, p. 176), and Cameron and Trivedi (2005, p. 350), multicollinearity is discussed under the headings of 'Data Problems', 'Violating Assumptions', and 'Computational Difficulties', respectively. Note that we could also use alternative estimators, such as ridge regressions, to cope with collinearity. However, the price to be paid for using these models is the loss of unbiasedness. As our focus is on the estimation of, at least potentially, unbiased models, we do not discuss these modelling alternatives.

Multicollinearity or, to use the shorter term collinearity, is often defined as a 'high (but not perfect) correlation between two or more independent variables' (Wooldridge 2013, p. 91).[1] However, what I want to suggest in this note is that we should not concentrate exclusively on collinearity, as it is only one part of the story of how additional regressors affect standard errors—and not always the most important one. Put differently, the prevailing focus on collinearity, defined as a correlation between explanatory variables, is not particularly helpful and can even be highly misleading. Instead, what I believe to be of core interest for applied researchers is how adding one or more variables to a regression affects the standard errors of the coefficient estimates, i.e. the efficiency of our parameter estimates. But in general, as we will see below, information about the correlation between variables does not allow drawing conclusions about the impact of including more explanatory variable on estimation efficiency.

Thus, the main point of this note is to alert empirical researchers to the fact that the impact of collinearity as stated in many econometrics textbooks is only part of the story and that it is *a priori* unclear how adding variables will affect estimation efficiency. In fact, including additional variables may help us obtain more precisely estimated parameters. To some extent, this has already been recognised, either explicitly or implicitly, in empirical work, e.g., in randomised control trials, where providing estimates including other covariates in addition to the treatment is sometimes done so as to achieve more precision, or in the case of adding squared or higher polynomial terms to, say, a wage regression. To separate this positive influence from the negative one emphasised above, I introduce the concept of standard-error-decreasing complementarity among explanatory variables. Although this phenomenon is noted in the literature under different guises, many economists do not appear to be aware of its existence, which may lead to potentially unsatisfactory model specification choices. In this note, standard-error-decreasing complementarity is explained with an easy to understand and intuitive heuristic employing Venn diagrams and, using a practical example based on synthetic data, the potential consequences of ignoring its existence are illustrated. The note concludes by briefly outlining the implications of standard-error-decreasing complementarity for widely applied approaches to empirical research.

[1] In its perfect form, collinearity implies that OLS estimates cannot be obtained. This is the case when there are either fewer observations than parameters to be estimated or a perfect linear relationship between independent variables. Here, we are not concerned with perfect collinearity, which, in practice, is typically relevant only in models containing many dummy variables.

2. Explaining Standard-Error-Decreasing Complementarity

To illustrate the impact of adding additional explanatory variables on standard errors, we start from a multivariate regression model with k variables, which we estimate using OLS for observations i running from 1 to n.

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + e_i, \quad (1)$$

where y = dependent variable, x_j = independent variables (j = 1…k), e = residual, and $\hat{\alpha}, \hat{\beta}_j$ = estimated constant term and estimated coefficients of independent variables, respectively.

Although collinearity potentially affects the variances of all k estimated coefficients in the regression, for the sake of simplicity, we concentrate on variable j. Many textbooks (e.g., Greene 2012; Wooldridge 2013) express the variance of the estimated coefficient for variable x_j in the form of

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)} \quad (2)$$

with $\sigma^2 =$ variance of the error term, $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, and $R_j^2 =$ the R^2 obtained from regressing the j[th] independent variable on all the other independent variables.

However, for our purposes, the variance of the estimated coefficient on variable x_j can be more helpfully expressed by following Stone (1945) as

$$Var(\hat{\beta}_j) = \frac{1}{n-k}\frac{\sigma_y^2}{\sigma_j^2}\frac{(1-R^2)}{(1-R_j^2)} \quad (3)$$

where $\sigma_y^2 =$ variance of the dependent variable, $\sigma_j^2 =$ variance of the j[th] independent variable, and $R^2 =$ obtained from estimating Equation (1).

From Equation (3), we can deduce that the magnitude of $Var(\hat{\beta}_j)$ depends on six influences:

- *Number of observations*: the higher the number of observations, the lower the variance. Intuition: more observations provide more information about the estimated relationship.
- *Number of estimated parameters*: the lower the number of estimated parameters, the lower the variance. Intuition: more information is available for estimating each individual parameter.
- *Variance of y*: the smaller the variance of the dependent variable, the lower the estimated variance of the coefficient. Intuition: less movement in the dependent variable makes explanation by the independent variables easier.
- *Variance of x_j*: the larger the variance of the independent variable, the lower the coefficient's variance. Intuition: more information in the independent variable makes it easier to explain movement in the dependent variable.
- R_j^2: the weaker the relationship of the j[th] independent variable with the other independent variable, the lower the variance. Intuition: different variation in the independent variables makes it easier to discover their individual influence. Thus, a large R_j^2 causes standard errors to increase because of collinearity.

- R^2: the better the fit of the regression, the lower the variance. Intuition: the improvement in fit to the data reduces the variance of an estimated parameter. Hence, this is the driving force of standard-error-decreasing complementarity between variables.

R_j^2 and R^2 are functions of the number of explanatory variables and are weakly positively affected by the addition of another explanatory variable:

$$R^{2\prime}(k) = {dR^2(k)}/{dk} \geq 0 \tag{4}$$

$$R_j^{2\prime}(k) = {dR_j^2(k)}/{dk} \geq 0 \tag{5}$$

Thus, the variance of the estimated coefficient on variable x_j can be expressed as a function of the number of explanatory variables:

$$Var_j(k) = \frac{1}{(n-k)} \frac{\sigma_y^2}{\sigma_j^2} \frac{\left[1-R^2(k)\right]}{\left[1-R_j^2(k)\right]} \tag{6}$$

The effect on the size of the coefficient variance of the j[th] regressor due to adding a further explanatory variable can be computed as the derivative of the above expression with respect to *k*:[2]

$$\frac{dVar_j(k)}{dk} = \frac{\sigma_y^2}{\sigma_j^2} \left\{ (n-k)^{-2} \frac{\left[1-R^2(k)\right]}{\left[1-R_j^2(k)\right]} + (n-k)^{-1} \frac{-R^{2\prime}(k)\left[1-R_j^2(k)\right] + R_j^{2\prime}(k)\left[1-R^2(k)\right]}{\left[1-R_j^2(k)\right]^2} \right\} \tag{7}$$

In principle, we would also have to consider the reduction in the degrees of freedom brought about by including another variable. Except for very small sample sizes, this standard-error-increasing effect is negligible. Hence, ignoring the change in coefficient variance that is due to the reduction in the degrees of freedom leaves:

$$\frac{dVar_j(k)}{dk} = \frac{\sigma_y^2}{\sigma_j^2} (n-k)^{-1} \left\{ \frac{-R^{2\prime}(k)\left[1-R_j^2(k)\right] + R_j^{2\prime}(k)\left[1-R^2(k)\right]}{\left[1-R_j^2(k)\right]^2} \right\} \tag{8}$$

The sign of this expression is determined by the sign of the numerator of the term in parentheses, which, in general, is undetermined. This illustrates that the focus on collinearity and its common implication of a standard-error-increasing effect when variables are added could be misleading, as it depends on certain conditions.

An increase in the coefficient variance of the j[th] regressor after increasing the number of explanatory variables requires:

$$-R^{2\prime}(k)\left[1-R_j^2(k)\right] + R_j^{2\prime}(k)\left[1-R^2(k)\right] > 0 \Leftrightarrow R_j^{2\prime}(k)\left[1-R^2(k)\right] > R^{2\prime}(k)\left[1-R_j^2(k)\right] \tag{9}$$

It is apparent from Equation (9) that the relationship is not straightforward. Concentrating on the changes in the two R-squared values after adding a variable and assuming $R^2 = R_j^2$ shows that the variance increases if the increase in collinearity as measured by $R_j^{2\prime}(k)$ is greater than the increase in

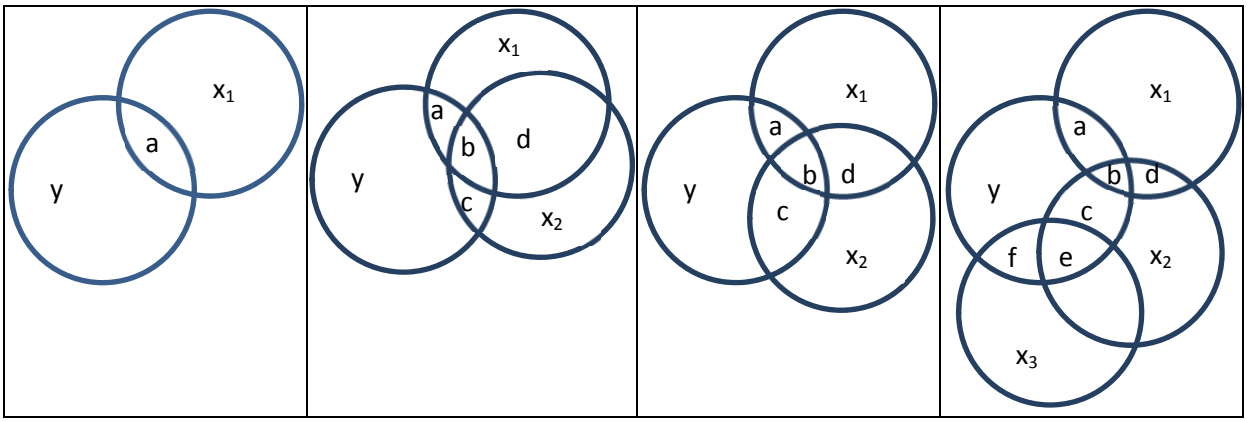
[2] Note that there are some technical issues. First, k is an integer number and, therefore, applying marginal analysis is just an approximation here. Second, in general, the derivative will reflect a particular sequence of explanatory variables.

the fit of the equation ($R^{2\prime}(k)$). But, of course, why should $R^2 = R_j^2$? This suggests that, in general, we do not know whether adding additional regressors to our model increases or decreases the efficiency of estimated parameters.

In the extant literature, the possibility of standard-error-decreasing complementarity is noted, using different names and a different perspective, by focussing on the relationship between partial R^2s and multivariate R^2. Kendall and Stuart (1973) discuss a situation where the sum of two partial R^2s from bivariate regressions of the form $y_i = \hat{\alpha}_1 + \hat{\beta}_1 x_{1i} + e_{1i}$ and $y_i = \hat{\alpha}_2 + \hat{\beta}_2 x_{2i} + e_{2i}$ is lower than the R^2 from the regression $y_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$, i.e. $R_1^2 + R_2^2 < R^2$. In these circumstances, the authors call x_2 a 'masking variable', as it 'masks' the relationship between x_1 and y. In their textbook, Daniel and Wood (1980) note that plotting x_1 against y to find out about their relationship is problematic in a multivariate context. In a statistical software manual, Ryan et al. (1985) use the term 'suppressor' for a variable that increases the significance of another variable when added to a multivariate equation. Hamilton (1987) proposes a vector geometric approach to explain how Kendall and Stuart's (1973) result is possible, i.e. that $R_1^2 + R_2^2 < R^2$. However, judging from the arguments made in many research papers and from comments heard at numerous conferences and seminars, the possibility of standard-error-decreasing complementarity of variables is not something many empirical researchers are aware of, much less take into consideration.

Thus, the current attempt to simplify the concept as much as possible by using a simple heuristic and emphasising intuition to illustrate standard-error-decreasing complementarity in the specific case of two explanatory variables. The conclusions, however, can be generalised to the case of k variables given in Equation (1). The heuristic is based on Ballentine Venn diagrams, introduced by Kennedy (1992) to explain multivariate regression, adapted to the present purpose. Figure 1 expresses the working of collinearity in a multivariate regression with the help of Venn diagrams.

Figure 1: Illustrating the Impact of Adding Variables Using Venn Diagrams



The respective variation of variables y, x_1, and x_2 is represented by circles. An overlap of circles indicates common variation between variables. The bivariate case is illustrated in panel (A) of Figure 1. There is an overlap between the circles representing the variation in y and x_1, which suggests that the latter explains parts of the former. Area a is the information set available for estimating $\hat{\beta}_1$. We assume area a to be inversely related to the estimator's standard error. In addition, the ratio of area a to the total area of y can be thought of as giving the coefficient of determination, R^2.

Multivariate regression analysis is illustrated by panels (B) and (C) of Figure 1, which show two alternative scenarios of adding another explanatory variable x_2. The R^2 of the multivariate regression can be computed by summing up areas a, b, and c. Excluding the case of orthogonality between independent and dependent variables, adding another independent variable will always increase R^2. Areas a and c are used for estimating $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. The explanative power of area b cannot be clearly allocated to either x_1 or x_2 and thus is not employed in parameter estimation. Thus, area b in our heuristic reflects collinearity. Panel (B) illustrates the standard case of adding a variable. Area b is relatively large compared to area c, where area c reflects standard-error-decreasing complementarity, which implies that the additional explanatory power of x_2 is relatively small and the overlap with x_1 in terms of explaining y is relatively large. Thus, little information is available for estimating the two parameters, which, keeping everything else constant, causes standard errors to be high.

However, as shown in Equation (9), looking at collinearity alone does not suffice for drawing general conclusions about the direction of change in coefficient standard errors. Moreover, we do not learn anything about the magnitude of the change, i.e. even if standard errors increase, coefficients may continue to be significant if the influence of collinearity is roughly offset by the influence of standard-error-decreasing complementarity. In Figure 1, this is illustrated by comparing areas b and c in panels (B) and (C): area b is larger, and area c smaller, in the former than in the latter. Hence, by solely focussing on area b, we would conclude that adding variables unambiguously leads to higher standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ in both panels (B) and (C). However, this conclusion is generally invalid, as we have to consider the change in R^2, too. In panel (C), we see that area c, reflecting standard error-decreasing-complementarity, is larger than area b, reflecting collinearity. Thus, in this example we should find lower standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ in the multivariate compared to the bivariate case.

This illustration suggests why using correlation coefficients to find out about collinearity is generally not advisable: the (squared) correlation coefficient between x_1 and x_2 is given by (b + d). Hence, in general, from the sum of the two areas we cannot infer much about the magnitude of b on its own. Theoretical exceptions are the extreme cases of perfect and zero correlation. In practice, collinearity is low if the correlation coefficient is close to zero and high if it is close to unity. But what 'close' means in this context is far from clear. Finally, consider panel (D), which adds yet another variable to our model. Now it becomes even more obvious why using correlation coefficients to find out about collinearity cannot be recommended: they ignore the presence of other variables and their potential impact on R_j^2. Thus, the intuition is that we cannot learn very much about collinearity, a multivariate phenomenon involving the dependent and the explanatory variables, by studying correlation, a bivariate phenomenon involving only explanatory variables. Put differently, even correlated explanatory variables can help explain variation in y and, thereby, decrease standard errors of estimated coefficients.

3. An Empirical Example Using Artificial Data

Let us now illustrate the working of collinearity and standard-error-decreasing complementarity in the context of a small, artificial dataset (see Table A1 in the Appendix). Variables y, x_1, and x_2 are taken from Hamilton (1987). The data-generating process is y = -4.52 + -3.1 x_1 + 1.03 x_2, whereas the variables x_3, x_4, and x_5 are newly constructed as 'nuisance influences' for the purposes of this

illustration. For the sake of the argument, let us suppose that all independent variables can be supported by good theoretical arguments, but there is no unambiguous economic theory on which to base what variables should be included in the empirical model, making it necessary to conduct what Leamer (1978) calls a 'specification search'. Put differently, there are theoretical reasons for including some or even all the variables in the final empirical model. However, in this example, a successful 'specification search' should result in a model for y that solely depends on x_1 and x_2.

We will now use the artificial dataset to demonstrate that it is possible to arrive at a wrong model by trying too hard to avoid collinearity. Our first step is to check the correlation matrix in Table 1, as is often done by applied researchers concerned about collinearity. The conclusions typically derived from this table are as follows: variables x_2, x_3, x_4, and x_5 look like promising candidates for explaining y, whereas x_1 does not seem to play a role. There is collinearity between x_1 and x_2 and between x_4 and x_5, perhaps with respect to x_3 too, and we should be wary of including them in the same model.

Researchers sometimes try to gauge the extent of collinearity in their models by using variance inflation factors (VIF), which are computed as $1/(1-R_j^2)$ and are much better at revealing the multivariate nature of collinearity than are correlation coefficients (see Belsley et al. 1980). A rule of thumb is that a VIF > 10 is indicative of collinearity. Using the Stata 14 default option of centred variables, we obtain the VIF values shown in the VIF 1 column of Table 1. All variables except x_3 appear to suffer from collinearity and their inclusion in one model could be viewed as problematic. Belsley (1984) emphasises that centring variables may produce misleading conclusions and, thus, we also compute VIF values based on uncentred variables (the VIF 2 column). The general conclusions from the centred VIF statistics hold, except that now we are warned even more strongly about collinearity involving x_1 and x_2.

Table 1: Correlation Matrix

	y	x_1	x_2	x_3	x_4	VIF 1	VIF 2
x_1	0.003	1				18.0	625.3
x_2	0.434	-0.900	1			21.6	378.4
x_3	0.808	-0.081	0.425	1		3.0	11.2
x_4	-0.753	-0.040	-0.291	-0.609	1	19.6	59.3
x_5	-0.708	-0.066	-0.247	-0.550	0.969	17.1	47.8

Note: VIF 1 (VIF 2) is based on centred (uncentred) variables.

Finally, one can compute a conditioning index for a matrix of explanatory variables. This is the ratio of the largest to the smallest eigenvalue of the matrix. Belsley et al. (1980) state that a conditioning index over 20 is indicative of collinearity. Computing the conditioning index for the five explanatory variables in our empirical example yields a value of 21, which suggests that our coefficients for this set of regressors are likely estimated imprecisely.

Using these findings, the collinearity-wary researcher would probably commence model estimation by choosing the variables having the highest correlation coefficients (in absolute terms) with the dependent variable, while trying to avoid collinearity. Given that the researcher does not know the data-generating process, these considerations suggest explaining y by x_2, x_3, and x_4. Table 2, column

(1) contains the estimation results. We apply a standard significance level of 5% in our tests.[3] Evaluating model (1) suggests that, with two significant variables and a good fit, it seems satisfactory. Moreover, diagnostic tests indicate no problems. We may want to make sure that choosing x_2 and x_4, and not x_1 and x_5, was not a mistake. Hence, we estimate model (2) to test the robustness of our model with respect to selecting specific variables out of a set of correlated variables. However, our previous results are confirmed as x_1 and x_5 behave almost exactly like x_2 and x_4, respectively, except that the fit in model (1) is slightly better.

Moreover, given their large correlation coefficient, it could be worthwhile to ensure that x_1 and x_2 are not relevant individually, which is analysed in columns (3) and (4), respectively. Neither x_1 nor x_2 are significant individually and we conclude that these variables are not explaining anything. In contrast, we expect that there is a lot of collinearity between individually significant variables x_4 and x_5, contributing to higher standard errors. To confirm this, we estimate model (5), now including both x_4 and x_5. The results seem to validate our previous analyses emphasising collinearity, as now neither x_4 nor x_5 is significant, due to more than three times higher standard errors.[4]

If this modelling approach seems a little too arbitrary, we can employ an automatic model selection algorithm. To make sure that collinearity does not interfere with obtaining precisely estimated parameters, we now use a stepwise procedure based on including one variable after the other (forward selection, p-value: 0.2, begin with empty model).

Applying the default option in Stata 14 (stepwise, pe(.2)), we obtain the output shown in Table 2, column (6). Model (6) is basically the same as model (1), except that, due to removing insignificant x_2, it boasts a superior adjusted R^2.

Rather than being satisfied with our modelling effort, let us remember that there could also be standard-error-decreasing complementarity in our dataset, which cannot be detected by studying correlations, conditioning indices, or VIFs. A practical way of finding out about standard-error-decreasing complementarity is to run a model including all relevant variables.

Model (7) of Table 2 contains the relevant estimation output and it completely contradicts our previous conclusions. Now only x_1 and x_2 emerge as being highly significant individually, two variables we previously considered to be of no relevance. How could that happen? Looking at the R^2 of this regression shows the working of standard-error-decreasing complementarity, as it has become much bigger compared to the ones computed for the earlier models. As a consequence, standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ fell dramatically and we can conclude that there is standard-error-decreasing complementarity between x_1 and x_2. Using a variance ratio test, we can reject equality of standard error estimated for $\hat{\beta}_1$ in model (2) when compared to the one estimated in model (8), which contains x_1 and x_2 as regressors.

[3] Note that given the small number of observations in our sample data, in principle, it would make a great deal of sense to follow Leamer's (1978) arguments and use a higher significance level. However, here we are only interested in illustrating a specific point and the chosen significance level works well for achieving that.

[4] Of course, we could have removed x_2 in our manual specification search, too.

Table 2: Explaining y Using Different Models (15 observations, estimator: OLS)

	(1) Specification based on correlation matrix	(2) Alternative specification	(3) Individual influence of x_1	(4) Individual influence of x_2	(5) Results for x_4 and x_5 are dominated by collinearity	(6) Best specification based on correlation matrix	(7) General model	(8) Showing standard error decreasing complementarity
x_1		0.04	0.04				3.06**	3.10**
		(0.227)	(0.416)				(0.019)	(0.012)
x_2	0.04			0.196	0.04		1.02**	1.03**
	(0.072)			(0.113)	(0.076)		(0.006)	(0.004)
x_3	0.02*	0.02**			0.02*	0.02**	0.0002	
	(0.007)	(0.007)			(0.007)	(0.006)	(0.0002)	
x_4	-0.01*				-0.01	-0.01*	0.0003	
	(0.004)				(0.015)	(0.004)	(0.0003)	
x_5		-0.01*			-0.001		-0.0005	
		(0.004)			(0.014)		(0.0003)	
constant	11.48**	11.48**	11.99	10.63	11.47**	11.74**	-4.34**	-4.51**
	(0.596)	(0.835)	(1.267)	(0.811)	(0.640)	(0.399)	(0.099)	(0.061)
F-test	F(3,11)=12.15**	F(3,11)=11.20**	F(1,13)=0.00	F(1,13)= 3.02	F(3,11)=11.71**	F(2,12)=19.09**	F(5,9)= 22376**	F(2,12)= 39220**
R^2	0.77	0.75	0.00	0.19	0.77	0.76	0.99	0.99
Adj. R^2	0.71	0.69	-0.08	0.13	0.68	0.72	0.99	0.99
Heterosc. test	F(6,8)=0.32	F(6,8)=0.42	F(2,12)=2.77	F(2,12)=0.37	F(8,6)=0.17	F(4,10)=0.63	n.a.	F(4,10)=1.08
	[p-value: 0.91]	[p-value: 0.85]	[p-value: 0.10]	[p-value: 0.70]	[p-value: 0.99]	[p-value: 0.65]		[p-value: 0.42]
RESET test	F(2,9)=0.35	F(2,9)=0.15	F(2,11)=1.38	F(2,11)=1.13	F(2,8)=0.38	F(2,10)=0.39	F(2,7)=4.61	F(2,10)=0.37
	[p-value: 0.72]	[p-value: 0.86]	[p-value: 0.29]	[p-value: 0.36]	[p-value: 0.70]	[p-value: 0.69]	[p-value: 0.06]	[p-value: 0.70]
Normality test	$Chi^2(2)$=4.70	$Chi^2(2)$=4.47	$Chi^2(2)$=2.42	$Chi^2(2)$=2.52	$Chi^2(2)$=4.92	$Chi^2(2)$=3.50	$Chi^2(2)$=3.68	$Chi^2(2)$=3.31
	[p-value: 0.10]	[p-value: 0.11]	[p-value: 0.30]	[p-value: 0.28]	[p-value: 0.09]	[p-value: 0.17]	[p-value: 0.16]	[p-value: 0.19]

Note: * (**) indicates statistical significance at a 5% (1%) level. Standard errors are given in brackets. Heteroscedasticity test is based on White (1980); RESET is based Ramsey (1969) but using squares and cubes; normality test is based on Doornik and Hansen (1994).

Similarly, we can reject equality of standard error estimated for $\hat{\beta}_2$ in model (3) when compared to the one estimated in model (8).[5] This result is an example of the situation shown in Panel C of Figure 1, where area c is much larger than area b. Thus, in spite of a correlation coefficient of -0.9, the gain in explanatory power of adding x_2 to x_1 more than offsets the collinearity between the two variables.

But surely x_4 and x_5, suffering from collinearity as shown above, are significant when tested jointly? The F-test is F(2,9) = 3.07, which is not significant at a 5% level, and nor is either of the two variables significant when included individually. In fact, if we test x_3, x_4, and x_5 jointly, we obtain an insignificant test statistics (F(3,9) = 2.71), which means that even x_3, so significant in the previous models, could be removed.

As an alternative to our manual model reduction procedure, we consider an automatic general-to-specific modelling algorithm (the default Autometrics option offered by OxMetrics 7) and obtain the output in model (8) of Table 2. In terms of the variables chosen from the given set, model (8) is orthogonal to model (6) but, statistically, it is clearly the superior model and, in fact, captures the data-generating process very well. Finally, note the decrease in standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ in model (8) compared to the model containing more variables (model (7)), which reflects the standard-error-increasing effect of collinearity.

Thus, collinear variables in a regression could be measuring important and separate determinants of the dependent variable and excluding them from the regression may be like throwing the baby out with the bathwater. Including a collinear variable often increases standard errors of the other variables in the model and they may even lose their significance. However, standard-error-decreasing complementarity works against that effect. If standard-error-decreasing complementarity is big enough, the additional fit achieved by considering an additional variable may overcompensate the increase in standard errors due to collinearity. Hence, modelling an empirical relationship without considering all relevant variables at the same time may lead to incorrect specifications, as shown in Table 2.

4. Standard-error-decreasing complementarity and omitted variable bias

Note that the effect of standard-error-decreasing complementarity must be kept conceptually separate from an omitted variable bias, as the former affects the variance of the coefficient, whereas the latter affects the coefficient itself. Figure 1 can be used to illustrate estimation biases, too. Assume we estimate $y_i = \hat{\alpha}_1 + \hat{\beta}_1 x_{1i} + e_{1i}$, i.e. we erroneously omit x_2. A bias in $\hat{\beta}_1$ arises because we now use area b in addition to area a in our estimation, even though area b cannot be solely attributed to x_1.

Looking at the coefficients in Table 2 and comparing the results for the correct specification in model (8) with the individually estimated coefficients in models (3) and (4), reveals that the point estimates are different. Using a t-test for differences in means we find that the two point estimates are also

[5] Using a two-sample variance-comparison test for testing the estimated standard error for $\hat{\beta}_1$ in model (3) against the one in model (8), we reject the null hypothesis of equal size: $F_{14,14}$ = 1200, p = 0.000. The same is true when testing the estimated standard error for $\hat{\beta}_2$ in model (4) against the one in model (8): $F_{14,14}$ = 798, p = 0.000.

statistically significant at all reasonable levels of significance.[6] Thus, the positive impact of lower standard errors on the size of the t-values is further enhanced by the existence of an omitted variable bias in the previously estimated models, which can be seen by the increase in the coefficient estimates for x_1 and x_2. Given that we know the data-generating process, we can conclude that the bivariate regressions involving x_1 and x_2 individually lead to omitted variable biases.

However, at the start of this section, it was claimed that standard-error-decreasing complementarity and omitted variable bias should be kept apart conceptually. Let us consider an empirical example that illustrates this proposition. Here, we are using the extreme case of two regressors, x_6 and x_7, which are orthogonal to each other (i.e., correlation coefficient = 0) and, by definition, do not experience any collinearity. Table 3 presents three models explaining y by employing x_6 and x_7 individually as well as jointly.

Table 3: Explaining y Using Orthogonal Regressors (15 observations, estimator: OLS)

	(9) Individual influence of x_6	(10) Individual influence of x_7	(11) Including both orthogonal variables
x_6	0.894**		0.894**
	(0.088)		(0.083)
x_7		-0.08	-0.08
		(0.150)	(0.048)
constant	10.21**	12.24**	10.46**
	(0.191)	(0.499)	(0.229)
F-test	F(1,13)= 102.2**	F(1,13)=0.29	F(2,12)=59.71**
R^2	0.89	0.02	0.91
Adj. R^2	0.88	-0.06	0.89
Heteroscedasticity test	F(2,11)=3.38	F(2,12)=1.09	F(5,9)=1.17
	[p-value: 0.07]	[p-value: 0.37]	[p-value: 0.38]
RESET test	F(2,11)=0.37	F(2,11)=0.02	F(2,10)=3.94
	[p-value: 0.55]	[p-value: 0.98]	[p-value: 0.06]
Normality test	$Chi^2(2)$=0.46	$Chi^2(2)$=1.23	$Chi^2(2)$=1.03
	[p-value: 0.80]	[p-value: 0.54]	[p-value: 0.60]

Note: See notes to Table 2.

Model (9) shows that x_6 has a significant impact on y, whereas model (10) finds no significant effect of x_7. The estimates in model (11) demonstrate the orthogonality of the two regressors, as there is no change in the coefficients when compared to the bivariate regressions. The standard errors have declined in both cases; however, However, the change is only significant from zero in the case of x_7 (even though the coefficient estimate has not become significant).[7] This illustrates the important

[6] Using a two-sample t-test with unequal variances to test the coefficient for x_1 in model (3) against the one in model (8), we reject the null hypothesis of equal size: t_{14} = -150, p = 0.000, and this is also the case when testing the coefficient for x_2 in model (4) against the one in model (8): t_{14} = -28.6, p = 0.000.

[7] Using a two-sample variance-comparison test for testing the estimated standard error for $\hat{\beta}_6$ in model (9) against the one in model (11), we cannot reject the null hypothesis of equal size: $F_{14,14}$ = 1.12, p = 0.83.

condition for standard-error-decreasing complementarity, namely, a sufficiently large increase in the fit of the regression. Adding x_6 to a model already containing x_7 increases model fit only marginally, which, given the small sample size, does not suffice to overcome the detrimental effect of the decrease in the degrees of freedom on the standard error estimate.

Thus, in principle, when entering an additional variable, there can be five outcomes on the variables already in the regression model: (i) coefficients *and* standard errors change, (ii) only coefficients change, (iii) only standard errors change, (iv) none of them changes (the additional variable is orthogonal to the other regressors and standard-error-decreasing complementarity is exactly offset by the reduction in the degrees of freedom), and (v) no estimates at all in the case of perfect multicollinearity.

This suggests that the effect of adding another variable to the regression on the variables already in the regression is, in practice, generally unpredictable. Put differently, trying to anticipate the effects of adding a variable *a priori* may invalidate statistical inference if it leads to a selective consideration of variables. There is nothing new about this conclusion; in fact, it is the basis of arguments for the superiority of general-to-specific modelling compared to a specific-to-general approach (for a lucid exposition, see Hendry 2007). However, to the best of my knowledge, the influence of standard-error-decreasing complementarity is not explicitly considered in this strand of literature.

5. Conclusion

Many empirical researchers in the field of economics believe that adding more variables to a regression model leads to unambiguously higher standard errors because of multicollinearity. This note argues that, in general, this belief is unfounded. Instead, it is shown that, although adding variables creates standard-error-increasing collinearity, equation fit may improve too, which has a negative impact on standard errors. Here, the latter effect is called standard-error-decreasing complementarity between variables. Using an intuitive heuristic and an artificial dataset, it is illustrated that standard errors of estimated coefficients may even decline after adding another variable. This suggests that, in general, we do not know whether adding additional regressors to our model increases or decreases the efficiency of estimated parameters.

An important question for empirical researchers involves the likelihood of encountering standard-error-decreasing complementarity when using real-world data. Here, we can only speculate, as we do not know the actual data generating processes (assuming they exist in the first place, of course). My own experience, as well as that of many of my colleagues, suggests that the extreme form of standard-error-decreasing complementarity revealed in our artificial dataset above is very rare, perhaps nonexistent. However, lesser forms do occur, perhaps not frequently, but regularly. For example, there are instances where two or more variables are significant when included as a group but not when considered individually. This suggests that standard-error-decreasing complementarity is not just an interesting theoretical curiosity but may actually affect everyday empirical research.

However, when testing the estimated standard error for $\hat{\beta}_7$ in model (10) against the one in model (11), we can reject the null of equally-sized standard errors: $F_{14,14} = 9.77$, $p = 0.000$.

The methodological consequences of acknowledging the potential relevance of standard-error-decreasing complementarity are likely wide ranging, as they could pose significant problems for a number of common research approaches.[8] As discussed above, one cannot expect that individually insignificant bivariate relationships will remain so in a multivariate setting—for two reasons: one is the well-known omitted variable bias and the other one is standard-error-decreasing complementarity. Thus, even if we are not overly concerned about omitted variable biases in a particular research project, e.g., because of orthogonality between the regressors, we should less readily embrace a number of widely employed empirical methodologies. I discuss three of the most important of these below.

First, many applied studies enter variables in a groupwise form, for example, initially a group of economic variables is tested for significance, then a group of political variables, and so forth. The extreme approach here would be adding single variables to some sort of base model without ever seriously considering all the variables together in one model. If there is standard-error-decreasing complementarity between individual or groups of independent variables, it will likely never be found and inference and interpretation will be flawed. Also note Gelbach's (2016) discussion of the pitfalls of adding covariates sequentially to test the robustness of the main variables of interest, particularly the sensitivity of results to the specific sequence of additions.

Second, there has been a renewed focus on the 'robustness' of empirical estimation results with respect to changes in the specification of the regression equation. The first consistent robustness approach is the 'extreme bounds analysis' developed by Leamer (1983), but the issue has been taken up recently by various authors (e.g., Plümper and Neumayer 2012). However, taking seriously models separating variables connected by standard-error-decreasing complementarity could lead to problematic inferences. Going back to our example above, should we really discard x_1 and x_2 as nonrobust only because they become insignificant as soon as they are not both included in a regression specification?

Third, it has become very popular to focus on the identification of causal effects in empirical studies, which is often combined with a movement away from estimating complicated econometric models. As Angrist and Pischke (2010) argue, attention has shifted to showing the relationships of interest in the context of small models. Thus, whether using data derived from natural or laboratory experiments, researchers tend to proceed in a *ceteris paribus* fashion, emphasising how one variable causes another variable. However, inasmuch as capturing standard-error-decreasing complementarity is necessary for obtaining significant coefficients, this implies that potentially important relationships will not be discovered, as the interconnection between several causal variables is not systematically considered. This argument is slightly related to one of the points made by Leamer (2010) in his comment on Angrist and Pischke (2010). He emphasises the potential role of correlation in finite samples even for variables that should be orthogonal in principle.

To conclude, empirical researchers using regression analyses should be aware that omitted variables may not only cause biases in the estimators, but could also lead to high coefficient standard errors. They should also understand that the effects of adding variables to a model are complex and that standard errors may either increase or decrease. Looking at correlation coefficients, conditioning

[8] An overview of econometric methodology with linkages to the philosophy of science is given by Dharmapala and McAleer (1996).

indices, or VIF values is of little help here and may even be misleading. Thus, awareness of standard-error-decreasing complementarity needs to become more widespread and its potential implications should be reflected in the methodology underlying the specification of empirical models. Without having considered the full model, i.e. the inclusion of all potentially relevant variables as argued by economic theory, one cannot rule out that the final model is statistically inferior due to ignoring the possibility of standard-error-decreasing complementarity. In my view, this constitutes another reason why a general-to-specific approach to econometrics is generally superior to specific-to-general approaches (see also Gilbert 1989).

Going beyond what I demonstrated in this note, let me conclude with a conjecture about what this implies for practitioners of applied econometrics. An important guiding principle should be that specification of the model is not based on an empirical preselection process of the variables to be included. Adding a variable to the model or subtracting one from it may change the estimated effects of the remaining variables due to omitted variable biases and/or estimation efficiency because of changes in standard errors. If smaller, more efficiently estimated models are desired as a basis for the interpretation of results (for arguments in favour of that, see, e.g., Hendry (2007), Keuzenkamp and McAleer (1995), and Hayo (1998)), then these should be derived via a consistent reduction of the general model that includes all theoretically relevant variables. Researchers should report the general model and the corresponding reduction test that supports working with a simplified model. Finally, one should not rely on collinearity diagnostics when trying to gauge the impact of adding variables on coefficient standard errors.

References

Angrist, J. D. and J.-S. Pischke (2010), The credibility revolution in empirical economics, *Journal of Economic Perspectives* 24, 3–30.

Baltagi, B. H. (2002), *Econometrics*, 3rd ed. Heidelberg: Springer.

Belsley, D. A. (1984), Demeaning conditioning diagnostics through centering, *American Statistician* 38, 73–93.

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

Cameron, A. C. and P. K. Trivedi (2005), *Microeconometrics*. Cambridge: Cambridge University Press.

Daniel, C. and F. S. Wood (1980), *Fitting Equations to Data*, 2nd ed. New York: John Wiley.

Dharmapala, D. and M. McAleer (1996), Econometric methodology and the philosophy of science, *Journal of Statistical Planning and Inference* 49, 9–37.

Doornik, J. A. and H. Hansen (1994), A practical test for univariate and multivariate normality, *Discussion Paper*, Nuffield College, Oxford.

Gelbach, J. B. (2016), When do covariates matter? And which ones, and how much?, *Journal of Labor Economics* 34, 509–543.

Gilbert, C. L. (1986), Professor Hendry's econometric methodology, *Oxford Bulletin of Economics and Statistics* 48, 283–307.

Gilbert, C. L. (1989), LSE and the British approach to time series econometrics, Oxford Economic Papers 41, 108–128.

Greene, W. H. (2012), *Econometric Analysis*, 7th ed. Upper Saddle River (USA): Pearson.

Hamilton, D. (1987), Sometimes $R^2 > r^2_{yx1} + r^2_{yx2}$: Correlated variables are not always redundant, *American Statistician* 41, 129–132.

Hayo, B. (1998), Simplicity in econometric modelling: Some methodological considerations, *Journal of Economic Methodology* 5, 247–261.

Hendry, D. F. (2007), *Econometrics: Alchemy or Science?* Oxford: Oxford University Press.

Kendall, M. G. and A. Stuart (1973), *The Advanced Theory of Statistics*, Vol. 2, 3rd ed. New York: Hafner.

Kennedy, P. (1992), *A Guide to Econometrics*, 3rd ed. Oxford: Blackwell.

Keuzenkamp, H. A. and M. McAleer (1995), Simplicity, scientific inference and econometric modelling, *Economic Journal* 105, 1–21.

Leamer, E. E. (1978), *Specification Searches*, New York: John Wiley.

Leamer, E. E. (1983), Let's take the con out of econometrics, *American Economic Review* 73, 31–43.

Leamer, E. E. (2010), Tantalus on the road to Asymptopia, *Journal of Economic Perspectives* 24, 31–46.

Plümper, T. and E. Neumayer (2012), Model uncertainty and robustness tests: Towards a new logic of statistical inference, *mimeo*, London School of Economics.

Ramsey, J. B. (1969), Tests for specification errors in classical linear least squares regression analysis, *Journal of the Royal Statistical Society* B 31, 350–371.

Stone, R. (1945), The analysis of market demand, *Journal of the Royal Statistical Society*, B7, 297.

White, H. (1980), A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroscedasticity, *Econometrica* 48, 817–838.

Wooldridge, J. M. (2013), *Introductory Econometrics*, 5th ed. Andover (UK): Cengage Learning.

Appendix

Table A1: Artificial Data for Illustrating the Effect of Collinearity and Standard-Error-Decreasing Complementarity

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
12.37	2.23	9.66	33	36	36	2	1
12.66	2.57	8.94	45	50	38	3	1
12.00	3.87	4.40	24	40	43	2	4
11.93	3.10	6.64	14	23	20	2	3
11.06	3.39	4.91	10	60	50	1	3
13.03	2.83	8.52	54	14	23	3	5
13.13	3.02	8.04	87	12	25	3	4
11.44	2.14	9.05	16	56	61	1	2
12.86	3.04	7.71	54	43	25	3	2
10.84	3.26	5.11	15	125	130	1	4
11.20	3.39	5.05	23	49	33	1	1
11.56	2.35	8.51	54	30	28	2	5
10.83	2.76	6.59	12	130	130	1	5
12.63	3.90	4.90	45	23	23	3	3
12.46	3.16	6.96	56	35	30	2	2