MIT LIBRARIES

3 9080 00441 5326

LIBRARY
OF THE
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

Digitized by the Internet Archive
in 2011 with funding from
Boston Library Consortium Member Libraries

<http://www.archive.org/details/effectofsimplesp00fish>

THE EFFECT OF SIMPLE SPECIFICATION ERROR ON THE
COEFFICIENTS OF "UNAFFECTED" VARIABLES

By

Franklin M. Fisher

Working Paper No. 194

January 1977

THE EFFECT OF SIMPLE SPECIFICATION ERROR ON THE
COEFFICIENTS OF "UNAFFECTED" VARIABLES

Franklin M. Fisher

Massachusetts Institute of Technology

A basic assumption of least-squares regression theory¹ is that the regressors are uncorrelated with the disturbance. When this assumption fails, so does the consistency of least squares. Such a failure can typically be regarded as a specification error and many such errors can be cast into this form. Failure may involve, for example, the omission of regressor-correlated variables from the equation or measurement error in the regressors² or simply the treating of endogenous variables as exogenous through the neglect of simultaneous feedback effects.

¹ Although the results of this paper are stated in terms of least squares, they obviously apply to any estimator which can be put in a least-squares-like form, hence to two-stage least squares, instrumental variables, and the like. It does not seem worthwhile making such obvious generality explicit in the notation or the text.

² This is the case treated in M. D. Levi, "Errors in the Variables Bias in the Presence of Correctly Measured Variables," Econometrica, 41, September, 1973, pp. 985-86. Levi considers the case of only one variable measured with error (which is one way of regarding the problem treated in the present paper) and obtains the result as to the direct effect given below, showing also that the signs of the indirect effects depend on the cofactors of the cross-product matrix of regressors.

0731574

It sometimes happens that the crucial assumption can be maintained with respect to most of the regressors but fails as regards one of them. Thus, a single variable may be measured with error or a single variable involved in a simultaneous feedback with all other variables safely exogenous. In such a circumstance, the effect of the error on the estimated coefficient of the affected variable would often be easy to analyze, at least as to sign, were that variable the only one in the regression. In essence, one simply examines the sources of the problem to determine the sign of the correlation between the disturbance and the variable in question. The effect of the error on the estimated coefficient will be in the same direction.

Where the affected variable is not the only regressor, however, the situation is not quite so clear. If the remaining, "unaffected" variables are not all orthogonal to the affected one, will the sign of the direct effect remain as in the simple case? Moreover, it is an elementary error to suppose that the effects of the specification error are confined to the coefficient of the directly affected variable. In general, there will be indirect effects as well, and various conjectures arise as to them.

Thus, for example, one might suppose that (after adjusting for the units in which the variables are measured) the direct effect will be larger than the indirect effects. One might also suppose that a particular indirect effect will be larger, the larger is the correlation between the unaffected variable in question and the directly affected variable.

In the present paper, I examine these and related questions. I show that while the sign of the direct effect does indeed remain the same as

in the simple case, the conjectures as to the indirect effects are wrong unless there is only one unaffected regressor (apart from the constant term) or unless all unaffected regressors are mutually orthogonal. In particular, it is possible to produce an example in which the ratio of every indirect effect to the direct effect can be made indefinitely large. Thus, not only is it false to suppose that the effects of simple specification error are confined to the coefficient of the directly affected variable, it is wrong to suppose that such coefficient suffers from the principal effect.

Fortunately, the signs of the indirect effects can be estimated from the data, as can the ratios of the indirect effects to the direct effect, so that the analyst need not throw up his hands. Indeed, most of the results are obtained by showing that the ratios in question are given by the negative of the vector of regression coefficients when the affected variable is regressed on the unaffected ones.

I now proceed to the formal analysis. In the usual notation, the equation to be estimated is:

$$(1) \quad Y = X\beta + \epsilon,$$

where Y is a T -vector of observations on the dependent variable, X is a $T \times k$ matrix of observations on k regressors, β is a k -vector of parameters to be estimated, and ϵ is a T -vector of disturbances. I assume that $Q \equiv \text{Plim } (X'X/T)$ exists and is non-singular and that:

$$(2) \quad \text{Plim } \left(\frac{X'\epsilon}{T} \right) = \lambda e_k,$$

where λ is a non-zero scalar and e_k is the k -vector with last component unity and remaining components zero. Thus the crucial assumption of least squares is violated only for X_k , the last regressor. The sign of the correlation between X_k and ε is that of λ .

In view of (2), least squares will not be consistent, and, by Theil's specification error theorem, the inconsistency, denoted by δ , will be given by:¹

$$(3) \quad \delta \equiv \text{Plim } (X'X)^{-1}X'Y - \beta = \text{Plim } (X'X)^{-1}X'\varepsilon = \lambda Q^{-1}e_k.$$

I have already implicitly denoted the j th column of X by X_j . Now denote by $X(j)$, the matrix formed from X by deleting the j th column.

Define:

$$(4) \quad \gamma \equiv \text{Plim } \{X(k)'X(k)\}^{-1}X(k)'X_k$$

so that γ is asymptotically the vector of regression coefficients when the directly affected variable, X_k , is regressed on the remaining variables.

I now prove:

Theorem: (a) $\lambda \delta_k > 0$; (b) $\delta_j/\delta_k = -\gamma_j$, $j \neq k$.

Proof: (a) From (3), $\delta_k = \lambda(Q^{-1})_{kk}$ and $(Q^{-1})_{kk} > 0$, since Q is positive definite.

(b) From (3), $\delta_j = (Q^{-1})_{kj}$. Hence $\delta_j/\delta_k = Q^{kj}/Q^{kk}$, where Q^{ij} denotes the cofactor of Q_{ij} . Without loss of generality suppose that

¹ I work in terms of inconsistency rather than bias principally because the results as to indirect effects involve ratios. Some of the results would hold for bias under appropriate assumptions.

$j = 1$. Then

$$\begin{aligned}
 (5) \quad \frac{Q^{k1}}{Q^{kk}} &= \frac{(-1)^{1+k} \text{Det Plim } \{X(k)'X(1)/T\}}{(-1)^{2k} \text{Det Plim } \{X(k)'X(k)/T\}} \\
 &= (-1)^{1+k} \text{Plim } \{\text{Det } (X(k)'X(k))^{-1} X(k)'X(1)\} \\
 &= (-1)^{1+k} \begin{vmatrix} 0 & \dots & 0 & \gamma_1 \\ \hline & & I & \vdots \\ & & & \gamma_{k-1} \end{vmatrix},
 \end{aligned}$$

where the last equality follows from the fact that the first $k-1$ columns of $X(1)$ are identical to the last $k-1$ columns of $X(k)$. Expanding by cofactors of the first row yields

$$(6) \quad \delta_1 / \delta_k = (-1)^{1+k} (-1)^{1+(k-1)} \gamma_1 = -\gamma_1,$$

and the theorem is proved.

Thus the direct effect is the same as in the simple case, while the ratios of the indirect effects to the direct effect are given by the negative of the vector of regression coefficients when the affected variable is regressed on the remaining variables. We now go on to use these results in considering the conjectures raised above.

I begin by considering the relative sizes of the indirect and direct effects. Clearly such relative size depends in the first instance on the units in which the different variables are measured. This is evident either

directly from the fact that δ_j/δ_k is not unit free or from the fact that γ_j is not. Hence, by choosing units in which the variance of X_k is made very small relative to the variance of any X_j , $j \neq k$, we can make the ratios of indirect to direct effects as large as we like (unless X_k is orthogonal to all the other X_j). This does not say very much.

If the conjecture that direct effects are likely to be larger than indirect effects has any content, therefore, it must involve the case in which units are chosen to take out such trivial scale effects. Accordingly, I now assume that all variables are measured from their means and divided by their standard deviations; hence, all variables now have mean zero and variance unity and the constant term is suppressed. I shall call these "standard units". Q now becomes a matrix of correlation coefficients, with

$$(7) \quad Q_{ij} = r_{ij} \quad ,$$

the correlation between variables i and j .

It is now easy to prove.

Corollary: In standard units, if either $k = 2$ or $r_{ij} = 0$ for all $i, j, k \neq i \neq j \neq k$, then $\left| \delta_j/\delta_k \right| < 1$.

Proof: In either of the two cases, γ_j will be the same as the regression coefficient in the simple regression of X_k on X_j . With the units chosen as described, however, this is r_{kj} .

Unfortunately, no similar result holds in more general cases. To

see this, it suffices to examine a three-variable case with $r_{12} \neq 0$.

Thus, choose $r_{12} = \mu - 1$, with $\mu > 0$. Choose $r_{13} = r_{23} = r > 0$ and

$r = h \sqrt{\mu/2}$, where $0 < h < 1$. Positive definiteness of Q requires

$\mu < 2$ and $\text{Det } Q > 0$. Examining the latter condition, we see that:

$$\begin{aligned} (8) \quad \text{Det } Q &= 1 + 2(\mu - 1)r^2 - 2r^2 - (\mu - 1)^2 \\ &= (2r^2 - \mu)(\mu - 2) = \mu(h^2 - 1)(\mu - 2) > 0 \end{aligned}$$

in view of the restrictions placed on h and μ .

From the theorem, however, we need only examine γ_1 and γ_2 . In this case,

$$(9) \quad -\gamma_1 = -\gamma_2 = \frac{r(\mu-1)-r}{1-(\mu-1)^2} = \frac{r(\mu-2)}{2\mu-\mu^2} = -r/\mu = -\frac{h}{\sqrt{2\mu}}$$

The absolute value of which tends to infinity as μ tends to zero. Hence every indirect effect can be indefinitely large relative to the direct effect.

It is clear from the corollary and the above example that colinearity among the unaffected variables can produce relatively big indirect effects. However, it is wrong to conclude that such colinearity has a monotonic effect. Consider the general three variable case (still in standard units). Then:

$$(10) \quad \delta_1/\delta_3 = -\gamma_1 = \frac{r_{12}r_{23} - r_{13}^2}{1 - r_{12}^2}.$$

If r_{13} and r_{23} have the same sign, for example, an increase in r_{12} from zero can change the sign of γ_1 , and therefore, of δ_1 as well (since $\delta_3 > 0$, by the theorem).

Similarly, all the indirect effects come about because the unaffected variables are correlated with the directly affected variable. One might therefore think that the unaffected variable most correlated with the directly affected one will be the unaffected variable whose coefficient has the biggest problem. This is clearly false as a general proposition, since there is no reason that the ratios of the γ_j which are multiple regression coefficients must lie on the same side of unity as the ratios of the r_{kj} , the correlation or simple regression coefficients. Indeed, it would be odd if this conjecture were true, since the grounds for believing it are similar to, but weaker than those for believing that the direct effect must be larger than the indirect ones.

Moreover, it is not true that increasing the correlation between the directly affected variable and a particular unaffected variable necessarily worsens the inconsistency in the estimate of the latter variable's coefficient. Thus, in the three-variable case in (10), an increase in r_{13} , other things equal, can produce a sign reversal in γ_1 and hence δ_1 .

What is true about an increase in such a correlation is that it has an unambiguously negative effect on the inconsistency involved (which may itself be positive or negative). To see this (still in standard units), differentiate (4) with respect to r_{kk-1} , obtaining:

$$(11) \quad \partial \gamma / \partial r_{kk-1} = \text{Plim} \{ (X(k)'X(k))^{-1} e_{k-1} \}$$

where $\bar{e}_{k-1}$ is the $(k-1)$ -vector with last component unity and other components zero. Comparing this with (3), it is evident that the proof of the theorem shows:

$$(12) \quad \partial \gamma_{k-1} / \partial r_{kk-1} > 0 ; \quad \frac{\partial \gamma_1 / \partial r_{kk-1}}{\partial \gamma_{k-1} / \partial r_{kk-1}} = -\bar{\gamma}_j \quad j \neq k-1$$

where $\bar{\gamma}$ is the probability limit of the vector of regression coefficients obtained when X_{k-1} is regressed on the first $k-2$ variables. (Note that this result applies to any regression and has nothing directly to do with specification error.)

Now,

$$(13) \quad \partial \delta_k / \partial r_{kk-1} = \partial \left(\frac{Q^{kk}}{\text{Det } Q} \right) / \partial r_{kk-1} = - \left(\frac{Q^{kk}}{(\text{Det } Q)^2} \right) (2 Q^{kk-1})$$

$$= - 2 \delta_k \delta_{k-1} = 2 \delta_k^2 \gamma_{k-1} ,$$

Thus:

$$(14) \quad \partial \delta_{k-1} / \partial r_{kk-1} = \partial (-\gamma_{k-1} \delta_k) / \partial r_{kk-1} = - 2 \gamma_{k-1}^2 \delta_k^2 - \delta_k \partial \gamma_{k-1} / \partial r_{kk-1} < 0 ,$$

from (12) and (13).

Thus, the larger is r_{kk-1} , other things equal, the smaller (algebraically) will be the inconsistency in the estimate of β_{k-1} . We have already observed that this may or may not correspond to an increase in the absolute value of that inconsistency. Does this ambiguity disappear as r_{kk-1}^2 approaches unity so that increased correlation is deleterious for coefficients of variables already very highly correlated with the directly affected one?

Like the conjecture that indirect effects are small relative to direct effects, this turns out to be true only for special cases. It is obviously true for $k = 2$, or for X_{k-1} orthogonal to $X_1, \dots, X_{k-2}$. Moreover, when either $k = 3$ or the variables, $1 \dots k-2$ are mutually orthogonal, the parallel relation to (10) shows that as r_{kk-1}^2 approaches unity, the sign of δ_{k-1} becomes that of $-r_{kk-1}$ so that, from (14), the absolute value of δ_{k-1} is also growing (and, indeed, will be maximized at either 1 or -1). For more general cases, this is not true, essentially for the same sort of reasons as lie behind the counterexample to the proposition that indirect effects are large relative to direct ones. Similarly, it is not true that an increase in r_{kk-1} has a bigger effect on δ_{k-1} than on δ_j , $j \neq k-1$.

The failure of so many simple conjectures is not a matter for depression, however. It is clear from the theorem that the signs of the effects of simple specification error and, indeed, their relative magnitudes can be readily determined by regressing the directly affected variable on the remaining ones. What is necessary is to refrain from behaving as though such a multiple regression is equivalent to simple ones.

T-J5 E19 w no.191
Krugman, Paul./Contractionary effects
730156 D*BKS 00031715

3 9080 000 795 440

T-J5 E19 w no.192
Diamond, Peter/Welfare analysis of imp
730159 D*BKS 00031714

3 9080 000 795 424

T-J5 E19 w no.193
Weitzman, Mart/Optimal revenue functio
730161 D*BKS 00031713

3 9080 000 795 408

HB31.M415 no.194
Fisher, Frankl/The effect of simple sp
731574 D*BKS 00037733

3 9080 000 867 090

HB31.M415 no.195
Diamond, Peter/Tax incidence in a two
731576 D*BKS 00037734

3 9080 000 867 116

MIT LIBRARIES DUPL

3 9080 004 415 300

MIT LIBRARIES

3 9080 004 415 318

HB31.M415 no.198
Dixit, Avinash/Quality and quantity co
731582 D*BKS 00037737

3 9080 000 867 173

HB31.M415 no.199
Friedlaender, /Capital taxation in a d
731584 D*BKS 00037738

3 9080 000 867 199

MIT LIBRARIES

3 9080 004 415 326