العنوان:	Frisch / Koutsoyiannis Statistical Confluence Analysis : A further Consideration
المصدر:	المجلة العلمية للاقتصاد والتجارة
الناشر:	جامعة عين شمس - كلية التجارة
المؤلف الرئيسي:	Ali, Mostafa Ahmed
محكمة:	نعم
التاريخ الميلادي:	1983
الصفحات:	53 - 69
رقم MD:	661580
نوع المحتوى:	بحوث ومقالات
قواعد المعلومات:	EcoLink
مواضيع:	التحليل الإحصائي، الإحصاء الرياضي
رابط:	http://search.mandumah.com/Record/661580

Frisch / Koutsoyiannis

Statistical Confluence Analysis : A Further Consideration

Dr. MOSTAFA AHMED ALI

Faculty of Commerce-Ain Shams University

(I) Introduction

The purpose of this paper is to provide a further consideration to Frisch/ Koutsoyiannis statistical confluence analysis. In 1934, R. Frisch published a book the title of which was ," Statistical Confluence Analysis by Means of Complete Regression Systems." In 1973, A. Koutsoyiannis published her book," Theory of Econometrics " in which she has applied the approach utilizing data for a certain country. The author of this paper believes he has a further consideration on the subject. This further consideration is a suggested statistical test. In what follows Frisch's confluence analysis and Koutsoyiannis application of it are reviewed. Afterwards we introduce our statistical test, which we believe to be a decisive one.

(II) Frisch/Koutsoyiannis Analysis

According to R. Frisch, whenever one includes in one and the same regression equation a set of " variates" that contain two, or more, subsests which are already-taken by themselves-highly intercorrelated, there exists a great danger of obtaining "nonsensical" results. If errors of observation are present , the regression coefficients would appear in the form of an error of observation divided by another error of observation . In such a case , one would get " fictitious determinateness created by random errors." When several "variates " are included in the analysis, we may here encounter a whole hierarchy where some of the variates may form a set where a regression equation has a meaning, and others forming sets where such equations have no meaning. The study of this hierarchy is what Frisch calls " confluence analysis."

When a new "variate"is tentatively added to a previously considered set, there are three fundamental possibilities to be considered. The variate may be " useful, superfluous or detrimental ". According to A.Koutsoyiannis, the approach starts by regressing the dependent variable on each of the independent variables, one at a time separately. On the basis of a priori and statistical criteria, the elementary regression which gives the most plaussible results is choosen. Next, additional variables are gradually inserted and their effects on the following three aspects are examined;

a) The individual coefficient,
b) The overall R^2 , and ,
c) The standard errors.

The new variable is classified as ;

1. useful, if it improves the coefficient of determination without affecting the individual coefficients as to be unacceptable (on a priori consideration). In such a case, according to Frisch," we conclude that the variate added is decidedly relevant. There is no doubt that it must be considered as useful," and therefore it is retained as an explanatory variable in the relationship.

2. superfluous, if it does not improve the coefficient of determination and does not affect the individual coefficients to any considerable extent. In this case, the variable is excluded from the explanatory variables.

3. detrimental , if it affects considerably the signs or values of the coefficients. To the extent this renders the coefficients unacceptable, one may conclude that this is a warrning that intercorrelation between the independent variables is a serious problem. In Koutsoyiannis's view, this does not mean that the detrimental variable must be rejected, because this would imply ignoring information valuable to the attempts of approaching as best one can the "true" specification of the relationship.

Koutsoyiannis applied the " experimental technique," utilizing data for a certain country in order to estimate the demand function for clothing. Basic data are given in table (I). We are going to use this data when applying our approach. Koutsoyionnis has calculated simple correlations between every two independent variables. These are given as follows (definitions of variables are given in table (I)).

r_{YL} =	0.993	$r_{P_cP_o}$ =	0.991
r_{YP_c} =	0.980	r^2_{CY} =	0.995
r_{YP_o} =	0.987	r^2_{CL} =	0.967
r_{LP_c} =	0.964	$r^2_{CP_c}$ =	0.951
r_{LP_o} =	0.973	$r^2_{CP_o}$ =	0.977

The first thing to notice is the surperisingly high degree of intercorrelation between the independent variables (Y , P_c , L , P_o). Secondly, it is obvious that the correlation between the dependent variable (C) and any of the independent variables is nearly the same. Koutsoyiannis has choosen the simple regression (C/Y) as the first step, then introduced the remaining explanatory variables into the function. Finally, she arrived at the following relationship;

$$C = -12.76 + \underset{(0.01)}{0.104}\,Y - \underset{(0.07)}{0.188}\,P_c + \underset{(0.12)}{0.319}\,P_o \qquad (1)$$

$$R^2 = 0.997$$

exluding (L) after finding out that it is a superfluous variable. The foregoing equation, Koutsoyiannis accepted as the best fit.

(III) A Further Statistical Consideration

In our view the foregoing equation is still subject to questioning. We believe that the significance of the coefficients should be accepted with qualification. Due to the high intercorrelation between the explanatory variables, we believe that the coefficients are not uninfluenced by the combined effect of these variables. In what follows we are going to examine this closely and introduce a test for such a tendency. This test may be considered as an empirical criterion which can tell us to what extent the equation reflects the net relation between the dependent and the independent variables. In order to discuss this, we are going firstly to investigate the equation that would have emerged if no intercorrelation had been present among the explanatory variables. In such a case the inverse of the matrix of products/ cross products (variables measured as deviations from means) , is simply given as follows ;

$$\begin{pmatrix} (n\sigma_1^2)^{-1} & 0 & 0 & 0 \cdots\cdots & 0 \\ 0 & (n\sigma_2^2)^{-1} & 0 & 0 \cdots\cdots & 0 \\ 0 & 0 & (n\sigma_3^2)^{-1} & 0 \cdots\cdots & 0 \\ 0 & 0 & 0 & (n\sigma_4^2)^{-1} \cdots\cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \cdots\cdots & (n\sigma_m^2)^{-1} \end{pmatrix}$$

where, the first independent variable takes 1, the second independent variable takes 2, ... and so on. (σ_i^2) is the variance of variable (i).

While, (n) is the number of observations. Thus, the partial coefficient will be estimated for variable (i) simply as follows ;

$$(\sigma_i^2)^{-1} Cov_{yi}$$

where , (y) is the dependent variable in the equation. It is clear, in this case, that adding one extra independent variable would not influence either the size or the sign of the regression coefficient of any of the previously included independent variables. However, the previously computed standard errors would be changed since including more variables would decrease the number of degrees of freedom and would also, in most cases, increase the explained variation. This explained variation will be given as ;

$$n \sum_{i=1}^{m} (Cov_{yi})^2 (\sigma_i^2)^{-1}$$

But , since the total variation is given as ;

$$(n\sigma_y^2)$$

the sum of squares of the residuals (Σe^2) divided by the degrees of freedom $[n-(m+1)]$, would be determined as ;

$$\frac{n}{n-(m+1)} \left[\sigma_y^2 - \sum_{i=1}^{m} (Cov_{yi})^2 (\sigma_i^2)^{-1} \right]$$

Therefore, the standard error of the coefficient of variable (i) is given as follows ;

$$\left([n/n-(m+1)] \left[\sigma_y^2 - \sum_{i=1}^{m} (Cov_{yi}^2/\sigma_i^2)(n\sigma_i)^{-1} \right] \right)^{\frac{1}{2}}$$

This last relationship proves the fact pointed out before; that is, the standard error of any variable's coefficient would change when changing the number of independent variables ($i = 1, \ldots, m$), in spite of the fact that the intercorrelations between these independent variables are nill and therefore the coefficients themselves are stable.

Let us re-examine the above findings and their implications. As an extreme case, if there is no intercorrelation what so ever between any two independent variables, the regression coefficients would be stable but their standard errors would certainly change, as the number of independent variables increases in the equation.

The implication of the stability of the regression coefficients is clear. If we firstly regress (y) on (m-1) variables, afterwards we regress (y) on (m) variables ; the size of the coefficient of the extra explanatory variable in the last equation can simply be calculated by using the extra explained variation due to this extra variable. Suppose that the explained variation when considering (m-1) variables is (E.V.1), while it is (E.V.2) when considering (m) variables. Then the size of the coefficient of regression of the extra independent variable would be determined as ;

$$\left[(E.V.2)-(E.V.1)\right]\left[\sum y_i\right]^{-1}$$

where (i) is the extra independent variable included .

On other words, suppose we have (m) independent variables where there is no intercorrelation what so ever between them. We may estimate an equation including all these variables, where (y) is the dependent variable, (measured, similar to the (m) variables, as deviations from mean). Let us denote this equation as (Eq.all). Now , we may estimate (m) equations, each includes (m-1) independent variables. Let us denote the amount of explained variation obtained from the equation that excludes variable (1) as : (E.V.m-1) ; the amount of explained variation obtained from the equation that excludes variable (2) as : (E.V.m-2) ... and so on. Let us also denote the amount of explained variation obtained from (Eq.all) as : (E.V.all) . Then the size of the regression coefficient of each independent variable appears in equation (Eq.all) is given respectively as ;

$$[(E.V.all) - (E.V.m-1)][\Sigma y1]^{-1}$$

$$[(E.V.all) - (E.V.m-2)][\Sigma y2]^{-1}$$

... and so on. Where, ($\Sigma y1$) for example is the sum of product of the observations of variables (y) and (1) each expressed as deviations from its mean. In such a case the sum of the extra explained variation due to including each of the variables (i = 1,2... m) , gradually, one at a time would be equal to the explained variation given by the equation in which all the (m) variables appear as explanatory variables. This is true only if there is no intercorrelation what so ever as between the independent variables.

Now, let us turn to Koutsoyiannis's equation (1) and subject it to more examination. Equation (1), obviously does not satisfy the above condition since the intercorrelations between the independent variables are nearly unity. However, the previous arguement would enable us to develop a criterion according to which one may accept or reject an equation such that given by Koutsoyiannis based on Frisch confluence analysis.

Suppose we have an equation where (y) is the dependent variable, while there are x_i ($i = 1, 2 \ldots m$) independent variables subject to intercorrelation between them. Now, it is clear that the amount of explained variation due to variable (x_m), ($E.V.x_m$), is given as ;

$$(E.V.x_m) = \left[R^2_{(y \ldots x_m)} - R^2_{(y \ldots x_{m-1})} \right] (\mathrm{Var}\ y)(n)$$

where,

$R^2_{(y \ldots x_m)}$: The coefficient of multiple determination for the equation including ($x_1 \ldots x_m$) explanatory variables.

$R^2_{(y \ldots x_{m-1})}$: The coefficient of multiple determination for the whole equation that would have been obtained had variable (x_m) been elminated.

n : The number of observations.

Therefore, the regression coefficient of (x_m) that corresponds with the previous reasoning, is given as ;

$$c_m = \left[R^2_{(y \ldots x_m)} - R^2_{(y \ldots x_{m-1})}\right](\text{Var } y)/(\text{Cov } yx_m)$$

Similarly, we can estimate : c_1 , c_2 , c_3 .... c_{m-1} in addition to (c_m). The explained variation in terms of these estimates is given as follows,

$$n \sum_{i=1}^{m} (c_i)(\text{Cov } yx_i)$$

This amount excludes, of cource, the combined explanatory power due to the presence of intercorrelation between the explanatory variables.

The variance of estimate in accordance with the estimates (c's) is ,

$$\Sigma e_c^2 / n-(m+1) = \left(\Sigma y^2 - n \sum_{i=1}^{m} c_i \text{ Cov } yx_i\right) / n-(m+1)$$

Using the appropriate element on the principle diagonal of $(x'x)^{-1}$, (λ_{mm}) , the standard error of (c_m) for example is ,

$$\left[(\Sigma e_c^2 / n-(m+1))(\lambda_{mm})\right]^{\frac{1}{2}}$$

Thus one can calculate the (t) value which if turns out to be significant, the corresponding variable is accepted as a significant variable after allowing for the effect of intercorrelation between : x_1 ... x_m .

It must be noted that this approach is entirely different from the traditional one which tests the significance of the additional variable after allowing for the effect of other variables. The traditional approach tests the significance of the extra explained variation due to the variable (x_m) as compared with unexplained variation which remained after taking into consideration other variables : $x_1, \ldots, x_{m-1}$.

while the present approach tests the significance of the explained variation due to the variable (x_m) as compared with unexplained variation remained after taking into consideration other variables : $x_1, \ldots, x_{m-1}$ and after allowing for the effect of intercorrelation between other variables : $x_1, \ldots, x_{n-i}$. Obviously, according to the present approach this unexplained variation is larger than that given according to the traditional approach. Both are equal only if there is no any intercorrelation between any two independent variables. Clearly the present approach is there one both more strict and more decisive. Further, an overall test may be introduced , in accordance with the above reasoning, as follows. The coefficient of determination calculated on the basis of our approach is given as ;

$$n \sum_{i=1}^{m} c_i \, Cov_{yx_i} / \Sigma y^2 = R_*^2$$

The symbol (*) denotes the concept of eleminating the combined effect of variables ; $x_1, \ldots, x_m$. Again, this (R_*^2) will be the same as the traditional coeffecient of determination (R^2) if the intercorrelation between any two independent variables is nill, because in such a case there would be no any combined effect which requires elemination. Now for a given significance level and relevant degrees of freedom, the critical (F) value is obtained from the table and therefore one can determine the corresponding coefficient of determination, $R^2_{\alpha, \nu_1, \nu_2}$. For example, if we have (5) explanatory variables, (25) observations, and the level of significance is set at the 0.05, then the critical (F) value is 2.74 while $R^2_{0.05, 5, 19}$ = 0.419 approximately. If $R^2_* > R^2_{0.05, 5, 19}$, that is if $R^2_* > 0.419$, we conclude that the overall relationship (after excluding the combined explanatory power) is significant, and <u>vice versa</u>. In order to examine Koutsoyiannis's relationship (1) on the basis of the above approach, we have estimated the equations given in (Table) (II), each includes only two explanatory variables. For each equation the explained variation (E.V.) is given. The explained variation as determined by equation (1) is (166.159). By utilizing this information with that given in Tables(II) & (III) , we therefore estimate the c's which are given in Table IV. It is clear that these estimates can be considered negligible as compared with the estimated coefficients given in equation (1). The variance of each (c) is certainly very large since the explained variation in terms of these (c's) is very small . These estimates are given in Table (V). As regards the overall test the corresponding (R_*^2) is given as;

$$R_*^2 = (3.824) / 166.527 = 0.0230$$

For (3) independent variables, (10) observations and at the 5% level of significance the critical (F) value is 4.76 , while the corresponding (R^2) is determined as follows ;

$$4.76 = \frac{R^2 / (3)}{(1-R^2)/(10-4)}$$

that is ,

$$R^2 = 0.704$$

Since (R_*^2) $<$ (R^2) we conclude that the overall relationship (1) is not in fact significant, after excluding the combined explanatory power. This finding together with the foregoing one suggest the conclusion that the significance of equation (1) is not unquestionable. This is being so, since the significance of each variable in equation (1) is determined according to the extra explained variation due to that variable as compared with the remained variation after taking into consideration the effect of other variables which contains the combined explanatory power. If one, as we did, excludes this combined effect the result would be quite different .

Statistical Appendix

Year	Exp.on clothing (£.m) C	Dis. Income (£.m) Y	Liq. Assets (£.m) L	Price Index for clothing 1963 = 100 P_c	General Price Index 1963 = 100 P_o
1959	8.4	82.9	17.1	92	94
1960	9.6	88.0	21.3	93	96
1961	10.4	99.9	25.1	96	97
1962	11.4	105.3	29.0	94	97
1963	12.2	117.7	34.0	100	100
1964	14.2	131.0	40.0	101	101
1965	15.8	148.2	44.0	105	104
1966	17.9	161.8	49.0	112	109
1967	19.3	174.2	51.0	112	111
1968	20.8	184.7	53.0	112	111

Source : Ref.(4) P.232.

Table (II)

Coefficients are being approximated

$$C = -58.04 - \underset{(0.226)}{0.166}\, P_c + \underset{(0.287)}{0.871}\, P_o \qquad R^2 = 0.979 \quad E.V. = 163.025$$

$$C = -8.37 + \underset{(0.018)}{0.102}\, Y + \underset{(0.104)}{0.090}\, P_o \qquad R^2 = 0.996 \quad E.V. = 165.819$$

$$C = 1.40 + \underset{(0.015)}{0.126}\, Y - \underset{(0.067)}{0.036}\, P_c \qquad R^2 = 0.996 \quad E.V. = 165.775$$

Table (III)

Variables measured as deviations from means.

$\Sigma Cy = 1406.617$

$\Sigma CP_c = 306.801$

$\Sigma CP_o = 245.4$

$\Sigma C^2 = 166.527$

Table (IV)

The estimated (c's) , obtained by utilizing information given in tables (II) and (III) together with (E.V.) of equation (1) [= 166.159].

Variable (i) :	Y	P_c	P_o
c_i	0.0022	0.0011	0.0016

Table (V)

Variances of (c_i's)

Variable (i) :	Y	P_c	P_o
Variance of c_i	0.092	2.789	7.124

(E.V.) in terms of (c_i's) = (0.0022)(1406.617) + (0.0011)(306.801) + (0.0016)(245.4) = 3.824

This is also equivalent to the amount;

3 [166.159] - [163.025 + 165.819 + 165.775]

REFERENCES

1. Badi, H.B;" On Seemingly Unrelated Regressions with Error Components." Econometrica, 1980.

2. Fisher, F.M.;" Tests of Equality Between Sets of Coefficients in Two Linear Regressions." Econometrica, 1970.

3. Frisch, R;" Statistical Confluence Analysis By Means of Complete Regression Systems." Oslo, 1934.

4. Koutsoyiannis, A;" Theory of Econometrics." Macmillan, 1973.

5. Searle, S.R.;" Linear Models." John Wiley, 1971.

6. Theil, H.; " Specification Errors and the Estimation of Economic Relationships." Review of The International Statistical Institute, 1957.