A C T A   U N I V E R S I T A T I S   L O D Z I E N S I S

FOLIA OECONOMICA 54, 1986

Iwona Konarzewska*, Władysław Milo**

SOME NOTES ON APPLICABILITY OF VARIANCE-DECOMPOSITION-PROPORTIONS METHOD

1. Introduction

Let us consider the equation of linear regression of the form

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \xi, \tag{1}$$

where $\mathbf{X} = [X_1, X_2, \dots, X_k]$ is a random vector of explanatory variables, Y is an explained variable and ξ is a disturbance term. We assume about these random variables the following: $\mathbf{X}$ is normally distributed with the expected value $\mathcal{E}(\mathbf{X}) = \mu = [\mu_1, \mu_2, \dots, \mu_k]$ and with the variance-covariance matrix $\mathcal{D}(\mathbf{X}) = \mathbf{\Sigma}$; ξ is normally distributed with the expected value $\mathcal{E}(\xi) = 0$ and the variance var $(\xi) = \sigma^2$; ξ is independent from $\mathbf{X}$. Consequently, the dependent variable Y is normally distributed with the expected value $\mathcal{E}(Y) = \mu'\beta_x + \beta_0$ and variance $var(Y) = \beta_x' \mathbf{\Sigma} \beta_x + \sigma^2$, where $\beta_x = [\beta_1, \beta_2, \dots, \beta_k]$, $\mathcal{E}$, $\mathcal{D}$ are the operators of expected value and variance-covariance.

Parameters $\beta_0, \beta_1, \dots, \beta_k$ of the model (1) can be estimated when we have a matrix of sample observations on explanatory variables and a vector of sample observations on explained variable. Under the assumption that the matrix of observations on explanatory variables is a fixed, not-random matrix we can obtain the following model

$$\mathcal{M}_0 = (\mathcal{R}^{n\times(k+1)}, \mathcal{S}, \mathbf{Y} = \mathbf{z}\beta + \Xi, k_0 \leqslant k + 1, n_0 = n, \tag{2}$$

*Senior Assistant, Institute of Econometrics and Statistics, University of Łódź.

**Lecturer, Institute of Econometrics and Statistics, University of Łódź.

$\mathscr{P}_{\mathbf{Y}} = \mathscr{N}_{\mathbf{Y}}(\mathbf{x}\beta, \sigma^2 I)),$

where $\mathscr{R}^{nx(k+1)}$ - a set of real n x k + 1 matrices, $\mathcal{S}$ - a probability space with the complete measure $\mathscr{P}$, $\mathbf{x} = [\mathbf{1} \vdots \mathbf{x}]$, $\mathbf{1}$ - the column vector of units, $\mathbf{x} \in \mathscr{R}^{nxk}$ - a matrix of observations on explanatory variables, $\boldsymbol{\beta} = [\beta_0\ \boldsymbol{\beta}_x]$, $\beta \in \mathscr{R}^{(k+1)x1}$ - vector of parameters, $\mathbf{Y}$, Ξ - random n x 1 vectors, $k_0 = \text{rank}(\mathbf{x})$, $n_0 = \text{rank}(\mathscr{D}(\mathbf{Y}))$.

The model (2) can be treated as a sample realization of the model (1).

Now we carry out the process of standardization:the model (1) with respect to theoretical means and covariances, the model (2) with respect to sample means and covariances. We obtain the following forms of standardized versions of considered relations:

$$Y^* = \beta^*_1 X^*_1 + \ldots + \beta^*_k X_k + \xi^*, \tag{1a}$$

where $\mathscr{P}_{Y^*, X^*_i} = \mathscr{N}_{Y^*, X^*_i}(0, 1)$, $i = \overline{1, k}$, $X^*_i = \dfrac{X_i - \mu_i}{\sigma_i}$, $Y^* = \dfrac{\mathbf{Y} - \mu'\boldsymbol{\beta}_x}{\sqrt{\text{var}(Y)}}$,

$$\sigma_i = \sqrt{\text{var}(X_i)},\ \mathscr{P}_{\xi^*} = \mathscr{N}_{\xi^*}\left(0, \frac{\sigma_\xi^2}{\sqrt{\text{var}(Y)}}\right),\ \beta^*_i = \frac{\sigma_i}{\sqrt{\text{var}(Y)}}\beta_i,$$

$\mathscr{D}(\mathbf{x}^*) = \Sigma^*$, Σ^* - a matrix of simple correlation coefficients between variables X_i, $i = \overline{1, k}$;

$$0\mathscr{N}\mathscr{M}_0 = \left(\mathscr{R}^{nxk}, \mathcal{S}, \mathbf{Y}^* = \mathbf{x}^*\boldsymbol{\beta}^*_x + \Xi^*, k_0 \leqslant k, n_0 < n,\right.$$
$$\left.\mathscr{P}_{Y^*} = 0\mathscr{N}_{Y^*}\left(\mathbf{x}^*\boldsymbol{\beta}^*_x, \frac{\sigma^2}{d_y}\mathbf{M}\right)\right), \tag{2a}$$

where $\mathbf{x}^* = \mathbf{M}\mathbf{x}(D^*)^{-1}$, $\mathbf{Y}^* = \frac{1}{d_y}\mathbf{M}\mathbf{Y}$, $\mathbf{M} = I_n - \frac{1}{n}\mathbb{1}\cdot\mathbb{1}'$,

$\mathbf{D} = [d_{ij}]$ $i, j = \overline{1, k} = \frac{1}{n}\mathbf{x}'\mathbf{M}\mathbf{x}$, $\mathbf{D}^* = \text{diag}\left(d_{11}^{1/2}, \ldots, d_{kk}^{1/2}\right)$,

$d_y = \sqrt{\frac{1}{n}\mathbf{y}'\mathbf{M}\mathbf{y}}$, $\mathbf{y}$ - a sample realization of the random vector $\mathbf{Y}$,

$\frac{1}{n} x^{*\prime}x^{*}$ - a matrix of sample simple correlation coefficients between variables X_i, $i = \overline{1, k}$.

The standardized versions (1a) or (2a) are usually the base to study the problem of extreme dependencies among explanatory variables called multicollinearity problem. In general the main effects of standardization of the equation (1) is reparametrization and the fact that the variance-covariance matrix Σ becomes the matrix of simple correlation coefficients. For the model (2) the effect of standardization is much deeper - we observe a change of parameter vector, matrix $\frac{1}{n} x^{*\prime} x^{*}$ becomes simple sample correlation matrix and additionally the normal distribution of the vector Y changes to singular normal of the vector Y^{*} because of idempotency of the matrix M.

2. Multicollinearity - diagnostical measures

Standarization reduces all model variables to the same scale. Further on, we analyze problems of interdependencies among model variables on the basis of standardized versions (1a) and (2a). In the case of (1a), we can speak about stochastical multicollinearity when rank $(\Sigma^{*}) < k$. In the case of (2a), we can speak about numerical multicollinearity when rank $(x^{*\prime}x^{*}) < k$. We have to mention that if we are dealing with stochastical multicollinearity in the (1a) then in its sample realization model (2a) there will appear numerical multicollinearity with probability one. On the other hand, if one is dealing with numerical multicollinearity in the case of (2a) there is no certainty whether it is caused by statistical dependency among variables X_i, $i = \overline{1, k}$ or whether it is a property of the individual sample matrix of statistical data (in the sense that expanding the matrix x^{*} by a new row of observations on explanatory variables we can get rid off multicollinearity problem). The problem of non-full rank of the $x^{*\prime}x^{*}$ matrix is related to exact linear dependency between columns of the matrix x^{*}. The parameters of this linear relationship are the elements of the eigen vector of $x^{*\prime}x^{*}$ matrix connected with the eigen value equal to zero. It can be easily shown by using the matrix $x^{*\prime}$ x^{*} spectral decomposition.

Let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ be the diagonal matrix with the eigen values of $x^{*\prime}x^*$ on the main diagonal and $V = [v_1, \ldots, v_k] \in \mathcal{R}^{k \times k}$ be the matrix of normalized eigen vectors corresponding to $\lambda_1, \ldots, \lambda_k$. It holds

$$\Lambda = V'x^{*\prime}x^*V \tag{3}$$

and

$$\lambda_j = (x^*v_j)'(x^*v_j) = \sum_{i=1}^{n}\left(\sum_{r=1}^{k} x^*_{ir} v_{rj}\right)^2, \quad j = \overline{1, k}.$$

The singularity of $x^{*\prime}x^*$ matrix means that at least one of eigen values λ_j $j = \overline{1, k}$, say λ_s, is equal to zero. Therefore

$$\lambda_s = 0 \Longrightarrow \sum_{i=1}^{n}\left(\sum_{r=1}^{k} x^*_{ir} v_{rs}\right)^2 = 0, \tag{4}$$

which means that the elements v_{rs} $r = \overline{1, k}$ are the parameters of linear relationship between columns of x^*.

The exact multicollinearity is a rare phenomenon in econometrical models. We are dealing mostly with near-multicollinearity problems. We can apply various measures of the strength of near-multicollinearity. These measures can be divided into two groups: numerical and statistical. Numerical measures are based on condition number of $x^{*\prime}x^*$ and x^* matrix. As a condition number we use (see B e l s l e y, K u h, W e l s c h 1980)

$$\varkappa(x^*) = \sqrt{\varkappa(x^{*\prime}x^*)} = \sqrt{\frac{\lambda_{max}(x^{*\prime}x^*)}{\lambda_{min}(x^{*\prime}x^*)}}, \tag{5}$$

where $\lambda_{max}(x^{*\prime}x^*)$ and $\lambda_{min}(x^{*\prime}x^*)$ are respectively the maximal and minimal eigen value of $x^{*\prime}x^*$.

Among statistical measures we distinguish:

- simple correlation coefficients r_{ij} (the elements of $\frac{1}{n}x^{*\prime}x^*$ matrix),

- partial correlation coefficients R_{ij}

$$R_{ij} = \frac{-r^{ij}}{\sqrt{r^{ii}}\sqrt{r^{jj}}}, \tag{6}$$

where r^{ij} is (i, j) element of $\frac{1}{n}(x^{*\prime}x^{*})^{-1}$,

- sample multiple correlation coefficients R_i between X_i $i = \overline{1, k}$ and other explanatory variables,
- variance inflation factors (VIF)

$$VIF_i = r^{ii} = \frac{1}{1 - R_i^2}, \tag{7}$$

- measure based on additional contributions of variables (see Theil 1971)

$$\delta = R^2 - \sum_{j=1}^{k} \left(R^2 - \tilde{R}_j^2\right), \tag{8}$$

where R is multiple correlation coefficient between Y and $\mathfrak{X} = \left\{X_1, \ldots, X_k\right\}$ and R_j is multiple correlation coefficient between Y and $\mathfrak{X} \setminus \left\{X_j\right\}$.

The coefficients r_{ij} and R_{ij} give us very small amount of information in the case when more than two of explanatory variables (columns of x^*) participate in the relationship. Additionally when at least one eigen value of $x^{*\prime}x^*$ tends to zero all R_{ij} are effected in the sense that $\forall\, i,j \lim_{\lambda_s \to 0} R_{ij} = \pm 1$, which can be easily shown applying spectral decomposition (see Konarzewska, Milo 1982). Greater amount of information can be obtained from the analysis of R_i and VIF_i $i = \overline{1, k}$ about the strength of dependencies and variables included. We have to notice that $R = \max_i R_i$ is a bounded measure in the range (0, 1) but the limit for each VIF_i when $\lambda_s \to 0$ and $v_{is} \neq 0$ is $+\infty$. The measure δ developed by Theil (1971) is bounded in the range $(-k + 1, 0)$ and is zero in the case of orthogonality of $x^{*\prime}x^*$ and $-k + 1$ in the case of extreme multicollinearity. Both two groups of multicollinearity measures are useful in determining the number and strength of relationships among explanatory variables. However, the analysis of eigen values and eigen vectors of $x^{*\prime}x^*$ matrix gives us the possibility to go deeper into the nature and consequences of the observed dependencies.

3. Variance-Decomposition-Proportions Method

The method allows us to find
- the number of relationships among columns of x^*,
- which variables (columns of x^*) are involved in an individual relationship,
- which of the model parameters can be estimated by least squares method without great imprecision.

The method was firstly described and analyzed by Belsley, Kuh, Welsch (1980). The main idea of this method, called variance-decomposition proportions method, is as follows.

Let us rewrite the estimated sample variance of the estimator

$$b_i^* = (x^{*\prime}x^*)_i^{-1}x^{*\prime}Y^*,$$

where $(x^{*\prime}x^*)_i^{-1} = \left[r^{i1} \dots r^{ik}\right]$, in the form

$$s^2(b_i^*) = \hat{\sigma}^2 r^{ii} = \hat{\sigma}^2 \sum_{l=1}^{k} \frac{v_{il}^2}{\lambda_l}, \tag{9}$$

where $\hat{\sigma}^2$ is an estimated variance of disturbance term.

We construct a matrix π as the matrix with the elements π_{li}, where:

$$\pi_{li} = \frac{v_{il}^2/\lambda_l}{\sum_{r=1}^{k} v_{ir}^2/\lambda_r}. \tag{10}$$

It can be noticed that $\forall\, i \sum_{l=1}^{k} \pi_{li} = 1$. For matrices x^* with mutually orthogonal columns it holds $\pi = I_k$. For an example we can consider the matrix π of the form

η	$var(b_1^*)$	$var(b_2^*)$	$var(b_3^*)$
η_1	1	0	0
η_2	0	0.01	0.1
η_3	0	0.99	0.9

where η_i $i = \overline{1, 3}$ are condition indexes defined as follows:

$$\eta_r = \frac{\lambda_{max}(x^{*\prime}x^*)}{\lambda_r(x^{*\prime}x^*)} \quad r = \overline{1, k}. \tag{11}$$

The maximal condition index is, by definition, equal to the matrix x^* condition number $\mathcal{K}(x^*)$.

If η_3 is "great" (what means that η_3 is greater than 15) we can say that among the explanatory variables exists one dependency. (B e l s l e y, K u h, W e l s c h 1980) state that as the multiple correlation coefficients which characterize the dependency increase along the progression <.9,.9, .99, .999, .9999, ... the condition indexes increase along the progression 3, 10, 30, 100, 300, ... This dependency is between X_2 and X_3 and effects variances of estimators b_2^* and b_3^*.Generally steps of the variance-decomposition proportions method can be summarized as follows:

1. Standardization of matrix x.
2. Computing condition number and condition indexes $\mathcal{K}(x^*)$, η_r $r = \overline{1, k}$.
3. Choice of limit values η^* and π^*(f.i. $\eta^* = 15$, $\pi^* = 0.5$).
4. Computing π matrix.
5. Examining elements of rows of π corresponding to $\eta_r > \eta^*$ if there exist at least two of them for which $\pi_{rl} > \pi^*$.

Existence of one relationship causes no problems for diagnosis. Problems arise when two or more relationships coexist. The first kind of problems is with dominating dependencies (one of "great" condition indexes is much higher than others "great" condition indexes). The second kind of problems is with competing dependencies (two or more "great" condition indexes with similar condition indexes). To solve the arising problems one has to build auxiliary regressions to check which variables are involved in which relationship.

In our opinion variance-decomposition proportions method in spite of the above mentioned problems is a supreme one in comparison with the methods based on the analysis of multiple correlation coefficients because it gives us deeper insight into the nature of dependencies and possibility of quantification of near-multicollinearity consequences on estimation precision. We should not forget about the second factor of sample estimator's variance

$= \hat{\delta}^2$. Although the estimated variance can be extended by the component connected with high condition index, at the same time it may be shrunken towards zero by near-zero value of $\hat{\delta}^2$(in the case of very high value of R^2).

Now, we will show two practical examples of variance-decomposition proportions method to following equations:

I $FSQFP_t = f_1\ (FSQFP_{t-1}, WUQF_t, PYPR_t, U75)$, t = 1963-1977,

II $XQMH_t = f_2\ (KQMW_t, NUQM_t, T)$, t = 1961-1979,

where: FSQFP - means monthly wage in fuel and power industries, WUQF - productivity of work in fuel and power industries, PYPR - index of living costs of mean employee's family, XQMH - ln of production in metalurgic, chemical and mineral industries (m.c.m.), KQMW - ln of productive capital stock in m.c.m. industries, WUQM - number of employees in m.c.m. industries, ln, T - time trend, U75 - dummy variable - U75 = 1 in 1975, U75 = 0 in other years of sample period.

We obtained the following results:

I.

$$x^{*\prime}x^{*} = \begin{bmatrix} 1.0000 & & & \\ 0.9201 & 1.0000 & & \\ 0.9843 & 0.9404 & 1.0000 & \\ 0.3029 & 0.3098 & 0.2984 & 1.0000 \end{bmatrix}$$

η	b_1^*	b_2^*	b_3^*	b_4^*
1.0000	0.0032	0.0114	0.0024	0.0186
5.8290	0.0893	0.8399	0.0210	0.0020
14.8568	0.9067	0.1465	0.9759	0.0036
1.8734	0.0008	0.0022	0.0007	0.9758

We can observe one not very strong dependency between $FSQFP_{t-1}$ and $PYPR_t$.Only var(b_1^*) and var(b_3^*) are effected by this relationship.

II.

$$x^{*\prime}x^{*} = \begin{bmatrix} 1.0000 & & \\ 0.9281 & 1.0000 & \\ 0.9982 & 0.9488 & 1.000 \end{bmatrix}$$

η	b_1^*	b_2^*	b_3^*
1.0000	0.0002	0.0050	0.0001
5.8576	0.0042	0.3546	0.0014
66.5111	0.9957	0.6404	0.9985

Similarly, we can observe one strong dependency between $KQMW_t$ and T - NUQM variable is also interrelated with these two but in a bit weaker way. All estimates of parameters are effected by near-multicollinearity.

References

B e l s l e y D. A., K u h E., W e l s c h R.E., 1980, Regression Diagnostics, Wiley.

K o n a r z e w s k a I., M i l o W., 1982, Diagnostyka współliniowości między zmiennymi objaśniającymi modelu ekonometrycznego - omówienie metod, typescript, R. III.9.4.3.

T h e i l H., 1971, Principles of Econometrics, Wiley.

Iwona Konarzewska, Władysław Milo

KILKA UWAG NA TEMAT MOŻLIWOŚCI STOSOWANIA METODY UDZIAŁÓW W ZDEKOMPONOWANEJ WARIANCJI

W artykule rozważono metodę diagnozy związków między zmiennymi objaśniającymi liniowego modelu regresji. Podstawą tej metody jest analiza numeryczna macierzy obserwacji na tych zmiennych. Metoda jest skonstruowana przy zastosowaniu regresji według wartości osobliwych. Obliczane są proporcje udziału każdego składnika tej wariancji w całej sumie. Metoda ta, nazywana metodą udziałów w zdekomponowanej wariancji, wprowadzona przez B e l s l e y, K u h, W e l s c h (1980) pozwala obok możliwości wykrycia ilości związków także na specyfikację zmiennych związanych relacjami. Dzięki temu możliwa jest diagnoza, które współczynniki w modelu mogą być oszacowane względnie precyzyjnie.

Autorzy przedstawili wyniki zastosowania tej metody na przykładach zbudowanych na bazie banku danych modelu W-3 gospodarki narodowej Polski oraz własną opinię na temat możliwości wykorzystania tej metody.