Informační Bulletin

České Statistické Společnosti English issue, 1993, vol. 4.

"Dressing up" with correlation

Stanislav Komenda

Not only statisticians know that interrelations existing between the events and variables offer an opportunity to consider the known state of a variable as a source of information when the state of another variable is to forecast. Moreover, the statisticians know how to bring this opportunity to life – i. e. to give a rule, a formula, an algorithm of computation to serve as an information channel. The essential part of the respective methods is to search in the statistical textbooks within the chapter headed "Regression analysis".

Statistical methods take into account that the statement one empirical variable makes on the other one can never be perfect, in any real situation, leaving some space for ambiguity or uncertainty. Information transmission always has the character of a noisy channel.

The information flow and the transmission are not the only use of interrelations existing between the variables. Applications are countless – some among them with extremely important consequences. As an example anthropometrical standardization in frame of preparatory phases in the mass production where the products have to consider body dimensions of the respective consumer – is to introduce. The garment (ready-made clothes) industry and the shoe producing industry as well are situations of this kind.

Individual made-to-order production of a cloth by a tailor or of a shoe by a shoemaker also took doubtless into consideration the interrelations existing between body dimensions – as recognized within the hundreds of years old experience. But, it is important that both the tailor and the shoemaker as well had the opportunity to measure the respective body dimensions of a consumer and moreover to control the fit again in a certain phase of the product manufacture and to make corrections and changes needed.

On the contrary, starting position of the mass ready-made production, as represented by the within-the-wars enterprises of Rolný and Nehera in the town of Prostějov or Baťa in Zlín, was different. Ready-made products are oriented towards an anonymous client. Nothing is known in advance on the customer who is to put on just this cloth or shoes in the future. That is the reason why in the preparation of manufacturing the reasoning must be included concerning the "sizes" to be produced so that the producer would "hit" just the space of the size figures actually existing in the population of potential future clients. And that is just this part of our play when correlation is entering on the scene – the task of a sufficient fit for the anonymous customers being unsolvable without it.

To manufacture a product of this kind we consider, 10 – 20 (in case of clothes, coats, trousers, skirts and overalls) or 5 – 7 (in case of shoes) body dimensions are taken into account. By the experience, an acceptable fit of a product is attainable only within certain tolerance interval specific for each dimension. These tolerances equal about 6 cm for body height, 4 – 6 cm for chest circumference or waist circumference and so on. If the Nature (genetics, anatomy, physical anthropology) composed our bodies using the principle of free combinations of these tolerance intervals in the multidimensional space, thousands of type figures would results in the case of five dimensions only. If it were the case, no mass ready-made manufacture would be possible.

The possibility of the ready-made manufacturing is due to existing correlation. To be more precise – due to the non-zero correlation. I met

Mr. Thomas Baťa twice – at the occasion of awarding him the *Pro merito* medal at Palacký University (1990) and one year later at the occasion of his *Honorary Doctor Degree* awarded by the Prague School of Economy – but I missed the chance to persuade him that the Baťa family should include the symbol of correlation coefficient into its heraldic shield.

It is just due to existing (and high enough) correlation among the body dimensions that the set of type figures able to cover – under the given tolerances of the body dimensions considered – the clients' population contains significantly lower amount of elements. The majority of combinations do not occur in the actual populations of clients at all or occur only with the frequency negligibly low. The computations considering fit tolerances of body dimensions and their variation within the respective populations show that the mass ready–made manufacture would suffice with a few tens of the size types, the number of them in case of shoes being substantially lower. Obviously, all computations of this kind have to consider age and sex of the potential clients, too.

As demonstrated by numerous anthropometrical studies analyzing human body length dimensions (length of the upper and lower extremities and their segments, trunk length etc) correlate highly positively with the stature (body height) while the width, depth and circumference dimensions have a high positive correlation with the chest circumference and waist circumference, respectively. As the regression equations where stature and chest circumference play the role of regressors (independent variables) and the other body dimensions necessary to take into account in the garment manufacture play the role of regressands (dependent variables) prove, the type figures are sufficient to define as the suitable combinations of the values chosen within the scales of these two basic dimensions. The reason is that residual variances of these non-basic dimensions important from the viewpoint of garment manufacturing (like the arm length, thigh and calf circumference, for example) become low enough so that the 2- or 4-residual sigma interval does not cross over the tolerance interval guaranteeing good fit. The same remains valid for the shoe production, too – with the difference that the number of body dimensions important in the manufacture under consideration is lower – in comparison with that needed in garment manufacturing.

Let me introduce a remark. When discussing on residual variance we consider its value in a point within the multidimensional space of regressors (multidimensional probability distribution taken as a model). But the problem introduced needs to consider residual variance (of a respective non-basic body dimension) not in a point but in an interval of the regressors' values. This last variance is to expect somewhat higher – with the possibility to compute or at least to estimate by how much higher.

Several years ago I met with an unusual system of garment manufacturing.

There was a tailor in Warsaw, Poland, Mr. A. Elert (tailoring for "better" people – they told me) who published a paper in the Polish garment industry journal explaining there that each dimension of human body can be expressed as a certain multiple of the radius of the hypothetic sphere with the same mass and density as that of human body (this average density being approximately 1.1 $\mathrm{g.cm}^{-1}$). His system of garment manufacturing was derived from this fundamental concept – only the values of stature and body mass (weight) of a person were necessary to be known for this manufacturing.

I was asked to comment on the paper. After some period of hesitation when this project seemed to be very odd to me – I conceived that the information richness of the body weight as a regressor is to verify. And, surprisingly, body weight proved a high reliability in its ability to forecast depth, width and circumference body dimensions! Residual variances of the circumferences (of the neck, thigh, calf etc) decreased, within the subjects with the same category of body weight, to a small fraction of the original, unconditioned variances. Length dimensions proved to be predicted from the weight with a very low efficiency, on the other side. But, the prediction supported by both stature and body weight reached the efficiency comparable with that of the three regressors – stature, chest circumference and waist (hip) circumference. Thus the conclusion was – manufacturing garment by means of the known value of body weight seems to be, from the informational viewpoint, quite reasonable.

As far as the Elert's canons concerns (according to which each dimension would be a constant multiple of the radius of the sphere derived from the body weight), this concept is to be considered idealized and naive, due to its non-statistical character. The reason is that anthropometrical data prove easily existing variability (higher or lower, but always a non-zero one) in case of any body dimension, in a subpopulation of subjects with the identical body weight (specified by the respective age and sex category). As said in advance, circumferences vary in such a subpopulation relatively only little, while length dimensions vary relatively more. The concept of these canons is applicable only within the "central" part of the space of body dimensions – with certain approximation.

By the way, there are some constants definable on human body. E. g. the index computed as the fraction with the *caput femoris* circumference in the numerator and the diameter of *caput femoris* in the denominator defines the irrational number π – due to the very spheric shape od *caput femoris*.

To support my considerations in a little more quantitative way, some body dimensions interesting from the viewpoint of garment manufacturing are introduced (the symbol W being used for them as regressands), together with their correlations with the regressors M (body weight) and X (stature).

Besides it, residual variances of these regressands are given, in case of various regressors' systems: (M), (X, M), (X, Y, Z), where Y is chest circumference and Z is the waist circumference (for men) or hip circumference (for women). We limit ourselves in the computation for the adult population – the solution having some specific features in case of children and youth.

Obviously, correlation of non-basic body dimensions with the stature X and body weight M has a complementary character – each dimension being "loaded" by the information almost exclusively either from M or from X, a balanced loading occurring only exceptionally. That is the reason why information efficiency of both systems (X, M) and (X, Y, Z) seems to be comparable, for almost all body dimensions W under consideration.

MEN

Body dimension W	r_{WM}	r_{WX}	s_{WW}	$s_{W.M}$	$s_{W.XM}$	$s_{W.XYZ}$
Neck circumference	0.70	0.20	4.76	2.41	2.31	2.09
Thigh circumference	0.79	0.29	11.65	6.72	6.59	7.56
Lower arm circumference	0.78	0.13	7.91	3.08	2.56	2.23
Waist height	0.38	0.91	26.87	23.05	4.75	3.50
Knee height	0.36	0.76	7.50	6.54	3.13	2.87
Height of the 7th neck vertebra	0.50	0.97	36.04	27.21	2.21	0.34
Gluteal furrow height	0.33	0.84	18.09	16.06	5.30	4.69
Sitting height	0.38	0.80	21.26	18.26	7.50	6.76
Frontal shoulder width	0.51	0.38	4.93	3.67	3.53	2.94
Frontal hip width	0.68	0.45	4.64	2.51	2.39	1.80
Profile chest width	0.77	0.18	4.69	1.91	1.72	1.19

WOMEN

Body dimension W	r_{WM}	r_{WX}	s_{WW}	$s_{W.M}$	$s_{W.XM}$	$s_{W.XYZ}$
Neck circumference	0.70	0.11	5.45	2.81	2.78	2.81
Thigh circumference	0.82	0.15	29.10	9.60	9.50	7.79
Lower arm circumference	0.86	-0.13	11.57	2.99	2.26	2.43
Waist height	0.34	0.88	22.52	19.89	4.72	4.65
Knee height	0.21	0.71	7.73	7.39	3.82	3.81
Height of the 7th neck vertebra	0.31	0.97	31.98	28.82	1.99	1.98
Gluteal furrow height	0.15	0.86	16.93	16.55	4.29	4.21
Sitting height	0.25	0.76	21.55	20.19	9.03	8.85
Frontal shoulder width	0.52	0.41	2.92	2.12	1.88	1.87
Frontal hip width	0.80	0.22	7.75	2.75	2.75	1.71
Profile chest width	0.82	0.04	8.70	2.75	2.54	1.66

Table. Correlation coefficients, variances and residual variances of the anthropometrical systems to compare. All variances are in cm^2.

How Sports Game Outcomes Depend on Intermediate Game Scores

Jiří Anděl

If you find the name Frederic Mosteller among authors of a paper you can be sure that it is something interesting for reading. This is confirmed by the paper *Predicting professional sports game outcomes from intermediate game scores*, written by H. Cooper, K. M. DeNeve, F. Mosteller, and published in the journal *Chance*, Vol. 5, 1992, No. 3 – 4, pp. 18 – 22.

It is our own experience that fans leave stadium in many cases already before the end of the game if they believe that the result will not change. (Some people claim that many fans in the last time do not come to stadium at all.) On the other hand it is not so much known in Europe that many people in the U.S.A. come to see basketball just before the end of the game, because one says that the last quarter of an hour of the game (or even the last two minutes) is the most dramatical and the score is changing. Is it true? Statistics can help to answer this question.

There are data about 200 basketball matches, 100 baseball matches, 100 ice–hockey matches and 100 football matches. From the practical point of view it is random sample from games played in 1991/1992. Define "at the beginning" and "before the end" in the following way:

game	at the beginning	before the end
basketball	after 1/4 of the game	after 3/4 of the game
baseball	after 3 runs	after 7 runs
ice-hockey	after 1st period	after 2nd period
football	after 1/4 of the game	after 3/4 of the game

In the basketball, ice-hockey and football the team who was loosing before the end succeeded to win the game in about 20cases. In baseball it was only in 6 % cases.

The dependence between the beginning and the final result is more complicated. The team which was loosing at the beginning succeeded to win in 30 % in basketball, in 19 % in baseball, in 31 % in ice-hockey and even in 45 % in football.

It is mainly the home team who is able to change the unfavorable score before the end of the game, at least in basketball. The home team did it in 33 % cases, the visiting team only in 10big. There is a small difference in ice–hockey but in football the chances to change the unfavorable score are the same for the home team as for the visiting team. There is no information about baseball because the number of games where the team loosing before the end finally won, was too small.

Which model is the right one
or
 Handpick Your coefficient of determination

Karel Zvára

1. Introduction

Let we have the common normal linear model $\boldsymbol{y} \sim \mathsf{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ where parameter $\boldsymbol{\mu}$ is estimated by least squares method as $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$. The residual sum of squares is equal to the square of vector $\boldsymbol{y}-\hat{\boldsymbol{y}}$ length, which is equal to $RSS = \|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Describing the original variability of the dependent variable by $\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\|^2 = \sum_{i=1}^n (y_i - \bar{y})^2$, residual sum of squares expresses the part of this variability of dependent variable, which cannot be explained by the supposed dependency. The coefficient determination R^2 is the fraction of the sum of squares of deviations of dependent variable from its mean that is attributable to the regression:

$$R^2 = 1 - \frac{\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\|^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Unlike the residual sum of squares, the coefficient determination is a dimensionless variable. We remind that our interpretation of coefficient of determination is reasonable only for regression models with intercept. We will suppose in our paper that this assumption is fulfilled.

Let we remind the some modifications of the model, which does not influence the mentioned characteristics. The residual sum of squares depends only on the distance of the vectors $\boldsymbol{y}$ a $\hat{\boldsymbol{y}}$. It stay the same after the adding of some fixed vector to both vectors or when instead of matrix $\boldsymbol{X}$ we use the matrix $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{D}$, where $\boldsymbol{D}$ is a nonsingular matrix. If follows from the fact that the vector $\hat{\boldsymbol{y}}$ is a projection of the vector $\boldsymbol{y}$ on a linear space of the columns of matrix $\boldsymbol{X}$, which is the same as the linear space of columns of the matrix $\boldsymbol{X}^*$. Especially, the residual sum of squares does not depend on scale of regressors, which values are in the columns of the matrix $\boldsymbol{X}$. The adjustment of the scale of the dependent variable influences the residual sum of squares, but in the fraction of coefficient of determination the square of the transformation constant is shortened, therefore this operation does not influence value of this coefficient.

2. Examples

In their paper Radek and Partyková (1984) deal with dependency of the yield of potatoes on the mean month temperature and mean month rainfalls. We do not wont do comment their work, we only will use their data. When we tried to explain the variability of yields by the rainfalls on all of 12 months preceding the gathering, we found that only two months rainfalls (September's and October's) collects nearly all information.

From our point of view it is interesting that predictions of yields by

$$\hat{v} = 21{,}0102 + 0{,}0595 s_9 - 0{,}0585 s_{10}$$

or

$$\hat{v} = 21{,}0102 - 0{,}0585(s_{10} - s_9) + 0{,}0010 s_9$$

give the same residual sums of squares $RSS = 4.3536$ and the same coefficient of determination $R^2 = 57.8\%$. It is a good illustration of independence these two statistics to a choice of the base of the linear space. It is not too interesting for us, that last the regression coefficient's standard error is equal to 0.0358 so that this regression coefficient is statistically not significant.

Let us try to predict the October's rainfalls from known September's ones. The Least squares method gives the estimate

$$\widehat{s_{10}} = 59{,}0833 - 0{,}3915 s_9$$

with the slope's standard error 0.2057, with the residual sum of squares $RSS = 529{,}52$ and with coefficient of determination $R^2 = 17{,}6\%$. We can explain so small part of variability that we cannot reject on 5% level the hypothesis that the slope is equal to zero. Let us try to explain by means of September's rainfalls value of $s_{10} - s_9$ (we saw that the prediction of yield was a function of this difference). Instead of relation $s_{10} = \alpha + \beta s_9$ we go into the relation $s_{10} - s_9 = \alpha + (\beta - 1)s_9$. Geometrically, in the relation $\boldsymbol{y} = \mathbf{1}\alpha + \boldsymbol{x}\beta + \boldsymbol{e}$ we subtracted fixed vector of October's temperatures from both sides of the relation. Therefore both of vectors $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$ are changed by this constant vector so that the residual sum of squares was no changed. The same is true for the residual variance. The length of the regression vector $\boldsymbol{x}$ was not changed, therefore the standard error of the slope is the same, too. Indeed, the cited data gave

$$\widehat{y_{10} - y_9} = 59.0833 - 1.3915 s_9$$

with the standard error 0.2057, the residual sum of squares $RSS = 529.52$, but with the coefficient of determination $R^2 = 72.9\%$. The slope is significantly non-null for every meaningful level. The coefficient of determination was dramatically changed. It is true in any circumstances?

Let we try to deal with another data. A group of boys was watched for a few years. We know their stature in ten and in twenty year. The prediction of twelve year stature is given by

$$\widehat{v}_{12} = 43.0441 + 0.7684 v_{10}$$

with slope's standard error 0.2729, the residual sum of squares $RSS = 143.2542$ and with the coefficient of determination $R^2 = 53.1\%$. Searching for the prediction of the increment of stature for two years we get

$$\widehat{v_{12} - v_{10}} = 43.0441 - 0.2316 v_{10}$$

with the standard slope's error 0.2729, with the residual sum of squares $RSS = 143.2542$, but with the coefficient of determination $R^2 = 9.3\%$.

Let us try to explain the suggested problem.

3. The coefficient of multiple correlation

We know, that the coefficient of determination is the square of coefficient multiple correlation. For simplicity, let we have only the simple linear regression. Then, the coefficient of determination is equal to square of the sample correlation coefficient r^2_{yx}. Using the estimates of the variance and the covariance we can write for fixed c

$$\begin{aligned} r^2_{y-cx,x} &= \frac{\mathsf{cov}^2(y-cx,x)}{\mathsf{var}(y-cx)\,\mathsf{var}(x)} = \frac{[\mathsf{cov}(x,y)-c\,\mathsf{var}(x)]^2}{[\mathsf{var}(y)-2c\,\mathsf{cov}(x,y)+c^2\,\mathsf{var}(x)]\,\mathsf{var}(x)} \\ &= \frac{\mathsf{cov}^2(x,y)}{\mathsf{var}(x)\,\mathsf{var}(y)}\,\frac{\mathsf{var}(y)\,[\mathsf{cov}(x,y)-c\,\mathsf{var}(x)]^2}{\mathsf{cov}^2(x,y)\,[\mathsf{var}(y)-2c\,\mathsf{cov}(x,y)+c^2\,\mathsf{var}(x)]} \\ &= r^2_{yx}\,\frac{\mathsf{var}(y)\,[\mathsf{cov}(x,y)-c\,\mathsf{var}(x)]^2}{\mathsf{cov}^2(x,y)\,[\mathsf{var}(y)-2c\,\mathsf{cov}(x,y)+c^2\,\mathsf{var}(x)]}\,. \end{aligned}$$

It follows that in the model for $\boldsymbol{y} - c\,\boldsymbol{x}$ coefficient of determination $r^2_{y-cx,x}$ is greater than the coefficient of determination r^2_{yx} in the model for $\boldsymbol{y}$, if and only if the fraction in the last equation is greater than 1. It is true if and only if

$$\begin{aligned} &\mathsf{var}(y)\,[\mathsf{cov}(x,y)-c\,\mathsf{var}(x)]^2 > \\ &\qquad\qquad \mathsf{cov}^2(x,y)\,[\mathsf{var}(y)-2c\,\mathsf{cov}(x,y)+c^2\,\mathsf{var}(x)] \end{aligned}$$

$$\begin{aligned} &-2c\,\mathsf{var}(x)\,\mathsf{cov}(x,y)\,\mathsf{var}(y)+c^2\mathsf{var}^2(x)\,\mathsf{var}(y) > \\ &\qquad\qquad -2c\,\mathsf{cov}^3(x,y)+c^2\,\mathsf{var}(x)\,\mathsf{cov}^2(x,y) \end{aligned}$$

$$\begin{aligned} &\mathsf{var}(x)\,\mathsf{var}(y)[-2c\,\mathsf{cov}(x,y)\ +c^2\,\mathsf{var}(x)] > \\ &\qquad\qquad \mathsf{cov}^2(x,y)[-2c\,\mathsf{cov}(x,y)+c^2\,\mathsf{var}(x)]. \end{aligned}$$

Because of $\mathsf{var}(x)\,\mathsf{var}(y) \geq \mathsf{cov}^2(x,y)$, it is true that

$$c\,[2\,\mathsf{cov}(x,y)-c\,\mathsf{var}(x)] < 0.$$

The estimate of the slope is equal to $b_{y|x} = \mathsf{cov}(x,y)/\mathsf{var}(x)$, therefore the last inequality can be rewritten as $c[b_{y|x} - c/2] < 0$. Especially, for $c = 1$ in our examples we get the inequality $r^2_{y-x,x} > r^2_{yx}$ if and only if $b_{y|x} < 0,5$.

The lapidary exhibition of the difference of two values of the coefficient of determination may be the next one. For the simple linear regression the coefficient of determination is the square of the correlation coefficient. The last coefficient is significantly non-null if and only if the same is true for the slope of simple regression. The hypothesis that the slope in the model for $\boldsymbol{y} - c\,\boldsymbol{x}$ is equal to zero is equivalent to the hypothesis that the slope is equal to the constant c in the model for y. Because of the standard error of slope is the same in the both of models, the value of test statistics (correlation coefficient, coefficient of determination) depends on the relation among the estimate $b_{y|x}$ of slope and the value c. The coefficient of determination in the model for $\boldsymbol{y}$ have to be greater than the same coefficient in the model for $\boldsymbol{y} - c\,\boldsymbol{x}$, if and only if the estimate $b_{y|x}$ is nearer to 0 than to c.

Reference

[1] . K. Shah (1991) *Relationship between the coefficients of determination of algebraically related models.* The American Statistician **45**, 300–301.

[2] . Radek, E. Partyková (1984) *The dependency of the potatoes yield on the weather* (in Czech). Rostlinná výroba **30**, 729–738.

Author's address: KPMS MFF UK, Sokolovská 83, 186 00 Praha 8-Karlín.

Additional concepts of the coefficient of determination

Josef Kozák

1. Introduction

RNDr. K. Zvára, CSc., showed in the two issues of the Information Bulletin of the Czech Statistical Society (No. 1 in February and No. 2 in May 1993) that the well known measure of the "quality" of the regression model – the coefficient of determination - represents a solution to interesting methodical considerations. The goal of these remarks is the continuation of these kind of considerations.

2. Basic results

Let us consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}_n, \quad E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}_n, \tag{1}$$

where Y denotes the n-element vector of the known observations of the explained variable, $\mathbf{X}$ means the non-stochastic $n \times K$, matrix of explanatory variable , $1 \leq K < n$, with full rank, $\boldsymbol{\beta}$ denotes the unknown K-element vector of parameters and $\boldsymbol{\epsilon}$ means the n-dimensional normally distributed random vector with the above mentioned properties where $\sigma^2 > 0$ denotes an unknown scalar. Let us assume that the elements of the vector $\mathbf{Y}$ as well as of the columns of the matrix $\mathbf{X}$ have zero averages, i.e. using the notation $\mathbf{1}_n$ for the n-element vector of unities, the relations

$$\mathbf{Y}'\mathbf{1}_n = 0 \qquad \text{a} \qquad \mathbf{X}'\mathbf{1}_n = \mathbf{0}_K \,. \tag{2}$$

hold. Regarding the vector $\boldsymbol{\beta}$ and the scalar σ^2, only the usual least-squares estimators will be considered

$$\mathbf{b}(\mathbf{Y},\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,, \tag{3}$$

$$s^2 = (\mathbf{Y} - \mathbf{Y}(\mathbf{X}))'(\mathbf{Y} - \mathbf{Y}(\mathbf{X}))/(n-K)\,, \tag{4}$$

where

$$\mathbf{Y}(\mathbf{X}) = \mathbf{X}\mathbf{b}(\mathbf{Y},\mathbf{X}) \tag{5}$$

denotes the estimate of the vector of the deterministic component $(\mathbf{X}\boldsymbol{\beta})$. Introducing the notation

$$\mathbf{U}(\mathbf{X}) = \mathbf{Y} - \mathbf{Y}(\mathbf{X}) \tag{6}$$

for the estimator of the vector $\boldsymbol{\epsilon}$ of random disturbances (the so called vector of residuals), it is not difficult to prove identities

$$\mathbf{Y} = \mathbf{Y}(\mathbf{X}) + \mathbf{U}(\mathbf{X})\,, \tag{7}$$

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{Y}(\mathbf{X}))'\mathbf{Y}(\mathbf{X}) + (\mathbf{U}(\mathbf{X}))'\mathbf{U}\ (\mathbf{X})\,. \tag{8}$$

The last relation represents the solution to the construction of the coefficient of determination

$$D(\mathbf{X}) = \frac{(\mathbf{Y}(\mathbf{X}))'\mathbf{Y}(\mathbf{X})}{\mathbf{Y}'\mathbf{Y}} = 1 - \frac{(\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X})}{\mathbf{Y}'\mathbf{Y}}\,; \tag{9}$$

because $n^{-1}\mathbf{Y}'\mathbf{Y}$, resp. $n^{-1}(\mathbf{Y}(\mathbf{X}))'\mathbf{Y}(\mathbf{X})$ can be interpreted as the variance of the empirical values, resp. the variance of the estimates of the theoretical values, the above mentioned measure indicates the ratio of the variance of the empirical values, which can be explained by the variance of the estimates of the theoretical values.

3. Change of space

As given for instance in [2], p.386, the initial identity (7) can be substituted with the given non-zero K-element vector $\mathbf{c}$ by the identity

$$\mathbf{Y} - \mathbf{X}\mathbf{c} = (\mathbf{Y}(\mathbf{X}) - \mathbf{X}\mathbf{c}) + \mathbf{U}(\mathbf{X})\,, \tag{10}$$

i.e. on the both sides of the identity (7) the vector $(\mathbf{X}\mathbf{c})$ can be substracted, especially due to the fact that instead of the analysis of the vector $\mathbf{Y}$, the analysis of the "shifted" vector $(\mathbf{Y} - \mathbf{X}\mathbf{c})$ seems to be more rational (at least from the point of view of material interpretation). Simple examples of this kind of transformations for $K = 1$ are given in [2] and i [1] ; unfortunately, a more general explanation of this idea is not easy in spite of the expectation.

Respecting (7) and (3) the relation (10) can be used for the derivation of the identity

$$(\mathbf{Y}-\mathbf{X}\mathbf{c})'(\mathbf{Y}-\mathbf{X}\mathbf{c}) = (\mathbf{Y}(\mathbf{X})-\mathbf{X}\mathbf{c})'(\mathbf{Y}(\mathbf{X})-\mathbf{X}\mathbf{c})+(\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X}) \tag{11}$$

"competing" the identity (8) and offering another variant of the coefficient of determination

$$\begin{aligned} D(\mathbf{X};\mathbf{c}) &= \\ &= \frac{(\mathbf{Y}(\mathbf{X}) - \mathbf{X}\mathbf{c})'(\mathbf{Y}(\mathbf{X}) - \mathbf{X}\mathbf{c})}{(\mathbf{Y} - \mathbf{X}\mathbf{c})'(\mathbf{Y} - \mathbf{X}\mathbf{c})} = 1 - \frac{(\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X})}{(\mathbf{Y} - \mathbf{X}\mathbf{c})'(\mathbf{Y} - \mathbf{X}\mathbf{c})}\,, \end{aligned} \tag{12}$$

which can be distinguished from the measure (9) by the name "shifted coefficient of determination" (with a certain degree of hesitation); this measure gives the ratio, which explains the variance of the estimates of the "shifted" theoretical values $n^{-1}(\mathbf{Y}(\mathbf{X}) - \mathbf{X}\mathbf{c})'(\mathbf{Y}(\mathbf{X}) - \mathbf{X}\mathbf{c})$ as compared with the variance of "shifted" empirical values $n^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{c}))'(\mathbf{Y}-\mathbf{X}\mathbf{c}))$.

As stated above, there does not exist any reason from the material point of view for the comparison of both defined coefficients. Nevertheless, if we are interested in such comparison, it is not difficult to get the result

$$\begin{aligned} D(\mathbf{X};\mathbf{c}) - D(\mathbf{X}) &= \\ &= \frac{(\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X})}{(\mathbf{Y}'\mathbf{Y})(\mathbf{Y}-\mathbf{X}\mathbf{c}))'(\mathbf{Y}-\mathbf{X}\mathbf{c}))}\mathbf{c}'\mathbf{X}'\mathbf{X}(\mathbf{c}-2\mathbf{b}(\mathbf{Y},\mathbf{X}))\,, \end{aligned} \tag{13}$$

by utilizing (8) and (3), which can be commented as follows: the mutual relation of both coefficients depends on the relation of the vectors $\mathbf{c}$ and $\mathbf{b}(\mathbf{Y},\mathbf{X})$. The relation (13) represents a generalization of the consideration for $K = 1$ from [1] .

4. Model with $K > 1$

In practical applications a regression model with two kinds of explanatory variables is, as a rule, employed, i.e. the (1) one with

$$\mathbf{X} = [\mathbf{T}|\mathbf{F}]\,, \tag{14}$$

$J + H = K,\ 2 \le K < n$, where $\mathbf{T}$, $\mathbf{F}$ denote the full rank matrices $n \times J,\ n \times H$, respectively, $1 \le J < n,\ 1 \le H < n$. Let us mention some consequences of this case.

(a) First, it seems to be useful to remember some results from [3], p.3–12: let us introduce a square matrix of order n

$$\mathbf{M} = \mathbf{I}_n - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\,, \tag{15}$$

with properties

$$\mathbf{M} = \mathbf{M}' = \mathbf{M}^2\,, \qquad \mathbf{M}\mathbf{T} = \mathbf{0}_{n\times J}\,, \tag{16}$$

and let us define a $n \times H$ matrix

$$\mathbf{F}_T = \mathbf{M}\mathbf{F} = \mathbf{F} - \mathbf{T}\mathbf{b}(\mathbf{F},\mathbf{T})\,, \qquad \mathbf{b}(\mathbf{F},\mathbf{T}) = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{F} \tag{17}$$

containing the deviations of values of variables of the second group from their estimates gained under the utilization the variables of the first kind; finally, introducing vectors

$$\mathbf{b}(\mathbf{Y},\mathbf{T}) = (\mathbf{T}'\mathbf{T}) - \mathbf{1}'_T\mathbf{Y}\,, \qquad \mathbf{b}(\mathbf{Y},\mathbf{F}_T) = (\mathbf{F}'_T\mathbf{F}_T)^{-1}\mathbf{F}'_T\mathbf{Y}\,; \tag{18}$$

it is not difficult to prove that the vector of the estimates of the theoretical values $\mathbf{Y}(\mathbf{X})$ introduced in (5) can be under the notations

$$\text{(19)} \qquad \mathbf{Y}(\mathbf{T}) = \mathbf{T}\mathbf{b}(\mathbf{Y},\mathbf{T})\,, \qquad \mathbf{Y}(\mathbf{F}_T) = \mathbf{F}_T\mathbf{b}(\mathbf{Y},\mathbf{F}_T)\,.$$

expressed as

$$\text{(20)} \qquad \mathbf{Y}(\mathbf{X}) = \mathbf{Y}(\mathbf{T}) + \mathbf{Y}(\mathbf{F}_T)\,,$$

thus, the following can be observed: working with the model with two kinds of explanatory variables in the above mentioned sense, the vector $\mathbf{Y}(\mathbf{X})$ can be understood as the sum of two mutually independent vectors, namely the vector $\mathbf{Y}(\mathbf{T})$ containing the estimates of the theoretical values depending only on the first kind of explanatory variables, and the vector $\mathbf{Y}(\mathbf{F}_T)$ containing the estimates of theoretical values derived only on the basis of the second kind of explanatory variables, the influence of the first kind of explanatory variables being excluded.

(b) Respecting (20), the identity (7) can be rewritten as $\mathbf{Y} = \mathbf{Y}(\mathbf{T}) + \mathbf{Y}(\mathbf{F}_T) + \mathbf{U}(\mathbf{X})$. Further, in the analogy to (6) $\mathbf{U}(\mathbf{T}) = \mathbf{Y} - \mathbf{Y}(\mathbf{T})$ denotes the vector of residuals associated with the model with the first kind of explanatory variables . As a consequence, there is the identity

$$\text{(21)} \qquad \mathbf{U}(\mathbf{T}) = \mathbf{Y}(\mathbf{F}_T) + \mathbf{U}(\mathbf{X})\,,$$

which - due to the fact that $(\mathbf{Y}(\mathbf{F}_T))'\mathbf{U}(\mathbf{X}) = 0$ – further leads to the identity

$$\text{(22)} \qquad (\mathbf{U}(\mathbf{T}))'\mathbf{U}(\mathbf{T}) = (\mathbf{Y}(\mathbf{F}_T))'\mathbf{Y}(\mathbf{F}_T) + (\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X})$$

representing a further "competitor" of the initial identity (8). This one offers to construct a further version of the coefficient of determination

$$\text{(23)} \qquad D(\mathbf{F}_T) = \frac{(\mathbf{Y}(\mathbf{F}_T))'\mathbf{Y}(\mathbf{F}_T)}{(\mathbf{U}(\mathbf{T}))'\mathbf{U}(\mathbf{T})} = 1 - \frac{(\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X})}{(\mathbf{U}(\mathbf{T}))'\mathbf{U}(\mathbf{T})}\,,$$

which declares, how much of the variance of the theoretical values based on the second kind of variables by exclusion the influence of the first kind of variables "explains" the amount of the variance of the residual values including only the first kind of variables ; this version will be called the "partial coefficient of determination" (again with a certain amount of hesitation).

In this connection the natural question arises, whether there exists any connection between the "general" coefficient of determination $D(\mathbf{X})$ and "individual" coefficients $D(\mathbf{T})$ and $D(\mathbf{F}_T)$, where in analogy to (9)

$$D(\mathbf{T}) = \frac{(\mathbf{Y}(\mathbf{T}))'\mathbf{Y}(\mathbf{T})}{\mathbf{Y}'\mathbf{Y}} = 1 - \frac{(\mathbf{U}(\mathbf{T}))'\mathbf{U}(\mathbf{T})}{\mathbf{Y}'\mathbf{Y}} \tag{24}$$

denotes the coefficient of determination corresponding with the model using only the first kind of explanatory variables. Because of the fact that under given circumstance the identity

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{Y}(\mathbf{T}))'\mathbf{Y}(\mathbf{T}) + (\mathbf{Y}(\mathbf{F}_T))'\mathbf{Y}(\mathbf{F}_T) + (\mathbf{U}(\mathbf{X}))'\mathbf{U}(\mathbf{X})\,,$$

holds, by utilization of the definitions (9), (23) and (24) it is not difficult to prove the identity

$$D(\mathbf{X}) = 1 - (1 - D(\mathbf{T}))(1 - D(\mathbf{F}_T)) \tag{25}$$

which generalizes the known relation between the partial correlation coefficients and the multivariate one in the case with $K > 1$. Finally, by elementary arrangement of the (25) we get the following result

$$1 - D(\mathbf{X}) = (1 - D(\mathbf{T}))(1 - D(\mathbf{F}_T))\,, \tag{26}$$

which characterizes the degree, to which the ratio of the unexplained variability of the model gets smaller.

The situation (14) takes place for instance by analyses of the time-series of economic indicators where **T** denotes the matrix of functions of the time-variable, whose inclusion into the model is, as a rule, without any kind of doubt, and **F** is a matrix of the so called factor variables (symptomatic or causal variables), which are included in the model, as a rule, always with a given degree of uncertainty. It is not clear yet in this connection, in what kind of "philosophy" is useful to treat to the problem of including the factor variables into the model; in this connection the relations (25) and (26) suffer some inspiration (in my opinion at least). Utilizing the time-series for prediction purposes, we should try to find a "good" model leading to the relative "high" coefficient of determination $D(\mathbf{X})$; as it was mentioned above, the matrix **T** is practically well known and with its connected coefficient $D(\mathbf{T})$ practically equals unity, and, therefore, we do not to bother ourselves with the selection of any kind of matrix F of factor variables. On the contrary, in the case of the econometric models, where we try to find a ”good” model of behaviour of the explained variable on the basis of the factor explanatory variables, it seems to be necessary to get rid at the influence of variables from any matrix $\mathbf{T}$, i.e. in this kind of analysis we have to orient ourselves on the matrix $\mathbf{F}$, which maximizes the resulting partial coefficient of determination $D(\mathbf{F}_T)$ and to omit the coefficients $D(T)$ and $D(\mathbf{X})$.

5. Acknowledgement

I would like to express my gratitude for help and discussion to my friends and colleagues J. Arlt and M. Coleman.

Reference

[1] Zvára, K.: *What Kind of the Model Is the Right One or Choose the Coefficient of Determination* (in Czech). Information Bulletin of the Czech Statistical Society, No 2, May 1993

[2] Spanos, A.:*Statistical Foundations of Econometric Modelling.* Cambridge University Press, Cambridge, 1986

[3] Kozák, J.:*The Summary of More Recent Results About the Prognostic Exploitation of the Linear Model of Time-Series* (in Czech). Manuscript, 1989

The main Activities of Czech Statistical Society in 1993

- 27th January 1993,*the third general meeting* was held at University of Economics, Prague. The programme was divided into three parts: lectures, discussion and the election. The new Society committee was elected with the following members:

  prof. Ing. Václav Čermák, DrSc., president
  RNDr. Jaromír Antoch, CSc., viceprezident
  Prof. Ing. Jaroslav Jílek, CSc., viceprezident
  RNDr. Gejza Dohnal, CSc., scientific secretary
  Ing. Hana Řezanková, CSc., treasurer
  Prof. RNDr. Jiří Anděl, DrSc.
  Doc. Ing. Richard Hindls, CSc.,
  RNDr. Felix Koschin, CSc.
  Prof. Ing. Jiří Likeš, DrSc.
  Ing. Josef Machek, CSc.
  Ing. Zdeněk Roth, CSc.
  Doc. Ing. Eduard Souček, CSc.
  RNDr. Karel Zvára, CSc.

- The Society committee advertised the regular scholarship which enables young statisticians (students) to participate in statistical conferences or seminars in Czech Republic. To request this scholarship, the candidate must fulfil the following conditions:
  - to be a member of Czech Statistical Society,
  - to be younger than 35,
  - to submit a contribution on desired conference or seminar.

  The commission appointed by Society committee will passed judgement on all obtained requirements.

- 17th September 1993, *the seminar "Today statistics"* was held at University of Economics Prague. The lectures was prepared by the Department of probability and mathematical statistics of Charles University, faculty of mathematics and physics and by the Department of statistics and probability of University of Economics, Prague.

  The seminar was introduced by the president of Czech Statistical Society, prof. Čermák. Lectures covered 4 areas: teaching (Prof. Jílek, Doc. Hebák), medical statistics (Mrs. Mazánková), theoretical statistics (Dr. Zvára) and networks services (Dr. Antoch).

- 3^{th} December, 1993, the state meeting in honour of centenary of Prof. Janko's birth was held at the University of Economics, Prague. Prof. Janko was the eminent expert in the field of insurance mathematics and he was one of the founders of the modern mathematical–statistical methods in our country. The meeting was organized by the University of Economics, Charles University and Czech Statistical Society.

Contents

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání a jednou v roce v anglické verzi. Předseda společnosti: Prof. Ing. V. Čermák, DrSc., VŠE, nám. W. Churchila 3, 130 00 Praha 3, E-mail: vaac@vse.cz.

Redakce: Dr. Gejza Dohnal, Jeronýmova 7, 130 00 Praha 3, E-mail: dohnal@fsik.cvut.cz.