Fickel, Norman

Working Paper

Sequential regression: a neodescriptive approach to multicollinearity

Diskussionspapier, No. 33/2000

Provided in Cooperation with:
Friedrich-Alexander-University Erlangen-Nuremberg, Chair of Statistics and Econometrics

Suggested Citation: Fickel, Norman (2000) : Sequential regression: a neodescriptive approach to multicollinearity, Diskussionspapier, No. 33/2000, Friedrich-Alexander-Universität Erlangen-Nürnburg, Lehrstuhl für Statistik und Ökonometrie, Nürnberg

This Version is available at:
https://hdl.handle.net/10419/29605

Sequential Regression:
A Neodescriptive Approach to Multicollinearity[1]

Norman Fickel

Friedrich-Alexander-University Erlangen-Nuremberg
Faculty of Economics and Social Sciences
Department of Economics
Lange Gasse 20
90403 Nuremberg
Germany

Abstract

Classical regression analysis uses partial coefficients to measure the influences of some variables (regressors) on another variable (regressand). However, a descriptive point of view shows that these coefficients are very bad measures of influence. Their interpretation as an average change of the regressand is only valid if the regressors are weakly correlated, and they are useless when the degree of multicollinearity is high. Despite these obvious flaws there is a lack of alternative ideas to measure influences. On that score this paper proposes two new coefficients of influence: (1) A *supplementary* coefficient measures the additional influence of a regressor when certain variables are already taken into account. (2) A *particular* coefficient, which is a mean of certain supplementary coefficients, allocates the influence of a regressor within the collective influence of all regressors. Both new coefficients can directly be interpreted as average changes of the regressand.

---

[1] Please do not cite this paper! Comments are welcome. E-Mail: Norman.Fickel@wiso.uni-erlangen.de

1 Introduction

Statistical methodology can be divided into two realms: descriptive and inductive statistics. Inductive statistics is characterized by the usage of probability theory to make stochastic significance test and confidence intervals available. To utilize these stochastic methods, some textbooks present multiple regression analysis as a tool of inductive statistics. However, in many applications, especially when social and economical data are concerned, the assumptions underlying these stochastic methods are not compelling and often they are obviously violated (for important examples cf. Hahn and Meeker 1993). Despite this fact, descriptive methods have not been developed adequately in statistical theory for the last decades since the current paradigm demands new methods to be stochastically motivated. This point of view appeared in 1944 when Haavelmo wrote in his preface of "The Probability Approach in Econometrics" (1994: iii): "For *no tool developed in the theory of statistics has any meaning* – except, perhaps, for descriptive purposes – *without being referred to some stochastic scheme*." [emphases as in the original]. This paradigm has been widely spread for several reasons in spite of its decreasing usefulness in applied statistics (cf. Nester 1996). I think the time has come to focus on new forms of descriptive analysis more intensely. To achieve this, this paper makes use of the framework of neodescriptive statistics.

1.1 Neodescriptive Statistics

Kruskal (1987: 6) defines neodescriptive statistics as "serious consideration, without shame, of what functionals on distributions are useful in analysis and understanding". This short statement is completed by the following characterization (Fickel 1999: 30–36) which reshapes the three levels of indication, determination and inference as proposed by Mosteller and Tukey (1977: 21f):

- *Indication:* The data set given is aggregated to one or more numbers indicating a specific property of this data set. This property is directly connected to the real-world application.
- *Determination:* The indication's result can be assessed by some coefficient of determination which describes how expressive the indication is. This helps the analyst to discriminate between substantial and unimportant coefficients.
- *Interpretation:* To achieve an intuitive understanding of the data set, a verbal statement of the measurement's meaning is provided. As far as a causal language is used, the dependence of the applied concept is clearly pointed out.

The notion of neodescriptivism gives space to alternative ideas of measurement, significance and usage of causal language, and so the framework of statistics is enlarged when the focus is not on stochastic models. This idea is somewhat related to the theory of fuzzy sets since the concept of determination can be seen as a special case of a fuzzy set which refers to the statement "The indication's value describes the data expressively". So a high value of determination says that the indication is highly expressive whereas a low determination renders the indication relatively inexpressive.

What is neodescriptive regression analysis like? Descriptive elements of regression analyses become neodescriptive when they are combined according to the above characterization. For example, the usage of simple regression can be summarized as follows (cf. *4.10.2*):

- *Indication:* The slope coefficient indicates the regressor's influence which is measured in terms of the scaling units of regressor and regressand.
- *Determination:* The percentage of the regressand's variance explained by the regression equation gives a coefficient of determination which helps to assess the meaning of the slope coefficient.
- *Interpretation:* The indication describes the average change of the regressand, when the regressor is varied by one scaling unit. This statement is based on the correlation in the data and does not imply any causal relationship.

When transferring these elements to multiple regression, the problem of multicollinearity can arise.

1.2 Problem of Multicollinearity

The aim of regression analysis is to measure the influences of various regressors on the regressand. To achieve this, for every regressor the ordinary least squares method provides a coefficient which can be interpreted as average change of the regressand when the regressor is varied by one unit. In *simple* regression analysis, every regressor is individually handled and so the influences of other regressors are ignored, which gives a total coefficient. In *multiple* regression analysis, the regressor is adjusted for the other regressors' influences and so its partial coefficient describes the influence when the other regressors are "held constant". However, this condition of holding other regressors constant decisively depends on the degree of multicollinearity in the data set, that is the extent to which the regressors are correlated among one another. Three levels of multicollinearity can be distinguished (cf. Mosteller and Tukey 1977: 270ff; Gunst 1983; Morrow-Howell 1994):

1. *Weak Multicollinearity:* The partial coefficient differs only slightly from the total coefficient and so multiple regression does not give more information than separated simple regressions do. Essentially, the interpretation of both coefficients is the same.
2. *Medium Multicollinearity:* Partial and total coefficient have noticeably different values. By assuming the stochastic standard model of regression, it might be said that the partial coefficient is "true" and the total coefficient is biased. Yet the descriptive meaning of the total coefficient is always clear whereas in this case the partial coefficient often has a dubious interpretation due to the existing correlation.
3. *Strong Multicollinearity:* When multicollinearity is perfect, partial coefficients cannot be determined. Otherwise the condition of "holding the other regressors constant" is meaningless. In addition, small changes in the data can sharply effect the value of a partial coefficient.

Literature knows two strategies to handle multicollinearity: A first is to determine partial coefficients for an altered set of regressors which are less correlated. For example, the forward selection algorithm only chooses a subset of the initial regressors, cluster analysis provides

cluster representatives, and possibly the regressors have interpretable transformations which are less correlated. A second strategy is to keep the initial regressors and modify the least square method to get varied partial coefficients. These modification methods can be categorized as follows:

a) *Shrinkage Type:* Multicollinearity often causes unplausibly large partial coefficients. So shrinkage methods, such as ridge regression, systematically give smaller coefficients. A certain parameter allows to select the degree of smallness. A resulting coefficient can be interpreted as *shrinked* average change of the regressand. (Cf. Hoerl and Kennard 1970; Kadiyala and Oberhelman 1994; Breiman 1995; Tibshirani 1996.)

b) *Suppression Type:* Principal component analysis might give interpretable results for some data sets. In this case the influences can be adjusted for "irrelevant" parts not belonging to the principal components chosen. A small number of principal components gives a stronger adjustment. The resulting coefficient is interpreted as an average change of the regressand when certain influences are *suppressed*. (Cf. Hawkins 1973; Hadi and Ling 1998.)

c) *Mixed Type:* The types of shrinkage and suppression can be combined, which yields coefficients modified twice. (Cf. Stone and Brooks 1990.)

Modification methods can give convincing results in certain applications. Yet they rely on assumptions on the size of "plausible" coefficients or on the existence of meaningful principal components. Statistics lacks a method applicable as standard approach for measuring influences in multicollinear data sets.

2 Measuring Influences

Modification methods try to provide results which are as close as possible to partial coefficients. Alternatively, this paper suggests measurement concepts not motivated by closeness to partial coefficients.

2.1 Supplementary Coefficient

Multicollinearity causes an interpretation problem because regressors have common influences which must be split among the various regressors. So the question arises what the additional influence of a regressor is when all common influences are removed. A first step to measure this additional or "supplementary" influence is to adjust the regressor for the common influences by replacing it by its residuals (cf. *4.1.3*). These residuals are given by a multiple regression of the actual regressor on the other regressors. When multicollinearity is high, this regression explains a large percentage of the actual regressor's variance. In this case, the percentage of unexplained variance, called tolerance of the regressor, is small.

The slope of a simple regression on this adjusted regressor coincides with the partial coefficient of multiple regression (cf. *4.2.5*). Very small tolerances lead to very small residuals and therefore this slope is very large in tendency. Obviously, one unit change in the original regressor does not come along with one unit change in its adjusted version. Thus the partial coefficient is no measure of supplementary influence.

In order to achieve an appropriate measurement, the adjusted version is synchronized with the original regressor in such a way that one unit change of the regressor corresponds with one unit change in its adjusted and synchronized form. This is done by multiplying the adjusted version by a suitably chosen number, which is the reciprocal of the regressor's tolerance (Fickel 1999: 105ff). This reciprocal is called variance inflation factor (VIF), which refers to its property in the standard stochastic model of regression analysis. In a word, one can synchronize a regressor by multiplying it by its VIF or, equivalently, by dividing it by its tolerance. Now the slope of simple regression on this adjusted and synchronized regressor measures the supplementary influence and the resulting *"supplementary coefficient"* has the following properties (cf. *4.4.7*ff):

- *Simple If Not Correlated:* The supplementary coefficient is identical with the slope of a simple regression if the regressors are uncorrelated with each other. That is why, the regressor's tolerance is 100 %. In this case, the supplementary coefficient has the same value as the total influence.
- *Zero If Perfectly Correlated:* It is zero if there is perfect multicollinearity with respect to the actual regressor, that is, the regressor can be described completely as a linear function of the other regressors, and therefore has a tolerance of 0 %. In this case, the regressor contains no additional information.
- *Natural Scaling Units:* It is given in units of the regressand per units of the regressor. Thus the size of the supplementary coefficient can be appraised directly.
- *Product of Tolerance and Partial:* It is always the product of the regressor's tolerance and partial coefficient, unless the partial coefficient does not exist because there is perfect multicollinearity. This factorization can be seen as the supplementary influence being part of the partial influence. The absolute value of the partial coefficient is never greater than the absolute value of the supplementary coefficient, and both coefficients have always the same sign.

The supplementary coefficient indicates the additional influence of its regressor. The determination of its indication can be quantified by means of the increment of the regressand's explained variance when the regressor is additionally taken into account. This increment is the difference of variance percentages: The first is the percentage of variance explained by all regressors and the second is the percentage of variance explained when the actual regressor is omitted. This difference can be easily transformed in only stochastically motivated coefficients such as t-values, partial F-values and P-values (cf. Bring 1994: 212f). So the supplementary coefficient, proposed in this paper as a new tool of data analysis, complements known measures of additional effect. To sum up, it may be said that the supplementary measurement establishes a neodescriptive tool as follows (cf. *4.10.3*):

- *Indication:* A supplementary coefficient measures the additional influence of a regressor apart from its common influence together with the remaining regressors. The scaling unit is the same as in simple regression.
- *Determination:* The expressiveness of a supplementary coefficient can be measured by the percentage variance which is additionally explained.

- *Interpretation:* A supplementary coefficient describes the average change of the regressand, when the regressor is varied by one scaling unit and its common effects with the other regressors are respected. In this case, all regressors are included.

The concept of supplementary measurement provides a new understanding of the partial coefficient when multicollinearity is severe: The influences of all regressors are not represented by their partial coefficients which show only the directions of the supplementary influences. The absolute value of a partial coefficient is misleading and can only be seen as the formal product of the supplementary coefficient and the VIF.

The supplementary coefficient can be understood as a part of the total coefficient and their difference describes the common influences of the actual regressor with the remaining regressors. This difference is the tolerance's complement multiplied by the "anti-partial" coefficient which is defined as the slope of a simple regression of the regressand on the actual regressor's residuals in this paper. These residuals are the version of the actual regressor when adjusted for all other regressors and so the variance of these residuals just defines its tolerance. So a total coefficient is a weighted arithmetic mean of a partial and anti-partial coefficient where the weights are the tolerance and the tolerance's complement, respectively (cf. *4.5.3*). The anti-partial coefficient describes the regressor's influence as far as it can be represented by a variation of the other regressors.

A partial influence is graphically demonstrated by a "partial regression plot" (also known as "added variable plot"), where the x-axis shows the regressor and the y-axis shows the regressand, and both are adjusted for the remaining regressors (cf. *4.2.6*; Chambers 1983: 268ff). Since the supplementary coefficient is also the slope of a simple regression line, it can be depicted graphically, too. To achieve this, the x-axis has to show the values of the adjusted and synchronized regressor and the y-axis the original values of the regressand. Alternatively, a diagram shows the same slope when the x-axis represents the original values of the regressor and the y-axis the adjusted (but not synchronized) values of the regressand (cf. *4.4.9*). This is a known plot used to reveal dependencies on the additional regressor (cf. Draper and Smith 1998: 68). Yet in this case the simple coefficient of determination can differ from the determination of the supplementary coefficient as described above.

2.2 Supplementary Sequence

A supplementary coefficient only describes an additional influence of just one regressor when all remaining regressors are already taken into account. Common influences cannot be seen and a supplementary influence is very small if all regressors are strongly correlated. In this case little information is gained by looking at the supplementary (and hence partial) coefficients.

Yet the analyst can get an overview over the data by choosing an ordering of the regressors. This ordering can be natural like the sequence of questions in a questionnaire or it can be arbitrary with the more interesting regressors in the first places. In any case the chosen ordering gives a sequence of supplementary coefficients in the following way (cf. *4.6.2*):

1. The first element of the sequence is the total coefficient of the first regressor. This coefficient actually coincides with the supplementary coefficient because no other regressor is taken into account.
2. The second element is the supplementary coefficient of the second regressor when only the first regressor is taken into account.
3. The third element is the supplementary coefficient of the third regressor when the first and the second regressor are taken into account.
4. ... and so on...
5. The last element is the supplementary coefficient when the first up to the last but one regressor are taken into account. In this case no regressor is omitted.

This sequence helps the analysts to understand their data set with respect to its correlation structure. High multicollinearity in tendency leads to small supplementary coefficients at the end of the sequence. This new tool of data analysis complements sequential sums of squares which consist of the percentage of variance explained by the first one, two, and so on, regressors of a sequence (cf. Rawlings et al. 1998: 196f; Draper and Smith 1998; 151f). So a *"supplementary sequence"* fulfils the characteristics of neodescriptive statistics:

- *Indication:* The supplementary sequence indicates the additional influences when step by step a regressor is taken into account. The scaling unit of each individual coefficient may be different according to the scaling unit of its regressor.
- *Determination:* The accompanying percentages of variance which is additionally explained show the expressiveness of the regressors' influences with respect to the chosen ordering.
- *Interpretation:* Each element of the sequence describes average changes of the regressand while step by step additional regressors are included. Clearly, the inclusion of a new regressor does not change the supplementary influences of the regressors included previously.

A supplementary sequence depends on the choice of an ordering. This dependence is the stronger the more correlated the regressors are. The analyst can place the more interesting regressors at the beginning of the ordering, and by varying the ordering, a more detailed insight into effects of multicollinearity can be gained. A 'scree plot' graphically shows the determination of a supplementary sequence by depicting the cumulative percentage of unexplained variance at the y-axis against the number of regressors used at the x-axis. A sharp reduction within a 'scree plot' demonstrates that the explained variance is concentrated on certain regressors with respect to the chosen ordering.

In case of a large number of regressors (say more than four) an ordering might be chosen with the help of a formal criterion which also maximizes the explained variance. The following strategies are possible:

- *Stepwise Maximizing:* The ordering is constructed step by step. The first regressor is taken in such a way that its total coefficient of determination is maximal. The second regressor has to maximize the increment in the coefficient of determination, and so on.

- *Fixed Number Maximizing:* A certain number governs this strategy as a parameter: A subset of this number of regressors is selected to maximize the coefficient of determination. The regressors within and outside the chosen subset are ordered in the stepwise manner.
- *Maximizing Concentration:* The ordering is taken in such a way that the increments of variance concentrate on the beginning of the sequence. A concentration ratio can be used to measure the degree of concentration.

The strategies of stepwise and fixed number maximizing are closely related to subset selection in multiple regression: The algorithm of forward selection gives the same ordering as stepwise maximizing when applied to all given regressors, and the algorithm of "best" subset selection just gives the subset of regressors used in fixed number maximizing (cf. Miller 1990: 43ff; Bring 1994: 212f). The adjusted coefficient of determination suggested by Bomsdorf (1993, 1994) can be used for choosing the parameter adequately.

In order to maximize concentration, various measurement concepts of concentration are available (cf. Piesch 1975). In this paper Rosenbluth's ratio is suggested, which does not depend on any parameters and can easily be represented graphically; yet it has to be modified slightly to get appropriate results. Since a maximal concentration on the beginning of the sequence (and not just on arbitrarily located regressors) is aimed at, the modified ratio respects the position of all regressors including their increments (cf. *4.7.1*). When no multicollinearity is present, the increments do not depend on the regressors' ordering; consequently, the modified ratio does not differ from Rosenbluth's original ratio. For highly concentrated orderings, this difference is often small even if the regressors are strongly correlated.

The starting point of sequential regression was to take regressors into account step by step. There are known procedures also using an ordering of regressors and giving a sequence of coefficients. One of these is the so-called "method of stepwise least squares", which adjusts in every step the regressand for the actual regressor (cf. Malinvaud 1966: 21ff). Another procedure is the "method of successive elimination", which adjusts in every step the actual regressor for all previous regressors (cf. Ezekiel and Fox 1959: 169ff). When the latter method is used, the coefficient of the last regressor is identical with its partial coefficient in multiple regression. In such a way a partial coefficient can be computed (yet there are more efficient algorithms).

2.3 Particular Coefficient

The measurement of supplementary influences provides a basis for indicating the particular influence of a regressor, which is the regressor's part of the collective influences of all regressors. Thus there is no dependence on a certain ordering. General modification methods do not give a measure of particular influence since they only alter the partial coefficient which is meaningless for highly multicollinear data.

The coefficient of a regressor within a supplementary sequence indicates its additional influence with respect to the actual ordering. By inspecting all possible orderings of the given regressors, the analyst gets a set of supplementary coefficients for each regressor. The range of this set shows the dependence of the additional influence on the present multicollinearity. In this paper, the particular influence of a regressor is defined as the arithmetic mean of all its

supplementary coefficients existing (Fickel 1999: 134ff). Other numbers but the arithmetic mean might also describe the set of supplementary coefficients, but the arithmetic mean is simple and so induces properties which are easy to understand. To sum up, the following statements can be made about the *"particular coefficient"* of a regressor (cf. *4.8.3*):

- *Simple If Not Correlated:* The particular coefficient is identical with the simple coefficient if the regressors are uncorrelated. Naturally, in this case, it also coincides with the partial and supplementary coefficient.
- *Proportional If Perfectly Correlated:* If all regressors are identical, it is the simple coefficient of one of these regressors divided by the number of all regressors. In the special case of just two regressors, it is half of the simple coefficient of one of the two regressors.
- *Natural Scaling Units:* Like the supplementary and partial coefficient it is given in units of the regressand per units of the regressor.
- *Independent of Ordering:* Unlike the supplementary coefficient (in a supplementary sequence) it does not depend on a certain ordering since all orderings are equally taken into account.

A regressor's part of the collective influences is indicated by the particular coefficient of this regressor. The determination of this indication can be measured by dividing up the multiple coefficient of determination correspondingly, as discussed by Kruskal (1987): The sequential sum of squares are computed for every ordering of the regressors and then a regressor's particular part is the average of all its percentages. As a result, the particular parts of all regressors sum up to the multiple coefficient of determination (cf. *4.8.2*). Kruskal suggested these percentages to assess the relative importance of regressors. Yet these percentages only measure the determination of a particular influence and so the particular coefficients proposed in the present paper complete the averaged percentages to a tool of neodescriptive statistics:

- *Indication:* A particular coefficient indicates the part of a regressor within the collective influence of all given regressors.
- *Determination:* A particular coefficient is the more expressive, the larger its regressor's percentage of the regressand's variance is.
- *Interpretation:* The particular coefficient of a regressor describes the average change of the regressand when the regressor is varied by one unit on condition that all remaining regressors are held constant on average.

Particular influences can be measured for every degree of multicollinearity and so give an insight into the data's structure even if standard regression analysis using partial coefficients fails.

2.4 Components of Total Coefficient

It sounds natural to say that the total influence of a regressor consists of an additional and a common part with respect to the other regressors. Can this statement be made precise by using the concepts mentioned above? How can the common part be analysed in detail? This paper introduces *"components"* as a possible approach (Fickel 1999: 142ff).

As a starting point, some ordering of the regressors is chosen again. Yet the actual regressor has to be the last in the ordering. A sequence of parts of the total coefficient of the actual regressor is constructed as follows: By taking the ordering's first regressor into consideration, the supplementary coefficient of the actual regressor can be computed. The difference of total and supplementary coefficient indicates the common influence of the two regressors at hand. By taking the ordering's first and second regressor into account, the change in the actual regressor's supplementary coefficient shows the effect of the second regressor. This procedure can be repeated until all regressors of the ordering are considered. Since the last regressor is the actual regressor itself, the last coefficient is its supplementary coefficient with respect to all other regressors. If every stepwise difference is used, the total coefficient of the actual regressor can be expressed as sum of these differences and its supplementary coefficient (cf. *4.9.2*).

Yet all differences depend on the chosen ordering. By averaging over all orderings which end with the actual regressor, components of each regressor are constructed. So the total coefficient is the sum of its supplementary coefficient and all other regressor's components. These components have the following properties (cf. *4.9.2*ff):

- *Zero If Not Correlated:* A component is zero if its regressor is not correlated with the actual regressor. If there is no multicollinearity, all components are zero.
- *Proportional If Perfectly Correlated:* If all regressors are perfectly correlated, then every component is identical to the total coefficient divided by the number of regressors (without the actual regressor).
- *Identical Scaling Units:* Every component has the same scaling unit as the total coefficient of the actual regressor.
- *Independent of Ordering:* The components are independent of a certain ordering.

This partition of the total coefficient can be made for every regressor. A tabular representation with a column for each partition and a row for each regressor gives a comprehensive analysis of effects of multicollinearity: The column sums are total coefficients and the diagonal contains the supplementary coefficients. The components are shown in the cells outside the diagonal. This representation is similar to commonality analysis (cf. Newton and Spurrell 1967: 53ff; Pedhazur 1982: 199ff), yet it contains different measures of influence.

Analogously, a regressor's total coefficient of determination can be divided among all given regressors in such a way that the parts are non-negative numbers. To achieve this, adjusted versions of the regressors are used. The part of the actual regressor itself is the product of its total coefficient of determination and its tolerance. In the special case of perfect correlation the tolerance is zero and so this part vanishes (cf. *4.9.5*).

The partition of a regressor's total influence into components for the remaining regressors can be seen as neodescriptive:

- *Indication:* A regressor's component indicates its influence entangled with the total influence of the actual regressor. The actual regressor itself is represented by its supplementary coefficient.

- *Determination:* A component is the more expressive, the larger its percentage of the total coefficient of determination is. The expressiveness of the supplementary coefficient directly depends on the tolerance of the actual regressor.
- *Interpretation:* A regressor's part describes the average change in the regressand when the actual regressor is varied by one unit via the regressor.

By partitioning a total influence into components a better understanding of the results of a simple regression analysis can be achieved when correlated variables are present.

3 Real-World Example

The new neodescriptive tools are exemplified with a data set on the Gross Domestic Product (GDP) of the 15 members of the European Union in 1997. GDP is measured in millions of European Currency Units (ECU) at market prices. The regressors of the analysis are the economically active population (EAP) in one thousand persons, the (total) population in one thousand persons and the area of the member country in square kilometre (data source: Federal Statistical Office of Germany 1999: 36ff). Obviously, these three regressors are highly correlated since the size of a member country dominates their values. The tolerances are about 1 % for EAP as well as for the population and 55 % for the area.

Table 1: Influences on Gross Domestic Product

Variable	Type of Influence			
	Total	Partial	Supplementary	Particular
Indication **Million ECU Per Unit of Variable**				
EAP	45.4	54.5	0.6	20.4
Population	20.5	-3.9	-0.04	9.0
Area	1.71	-0.07	-0.04	0.53
Determination **% of Variance of GDP**				
EAP	97	×	1	44
Population	96	×	< 0.1	43
Area	32	×	< 0.1	11
Sum	×	×	×	**98**

Table 1 shows different types of influence: All total coefficients are positive and so in simple regression their variables are all positively correlated with GDP. Their coefficient of determination is high for EAP (97 %) and for the population (96 %), and relatively low for the area (32 %). Using all three regressors together yields a multiple coefficient of determination of

98 %. Although technically easily computed the partial coefficients are hard to interpret. In this case, the partial coefficient of EAP with 45.5 Million ECU per 1,000 persons seems rather high when compared with the corresponding total coefficient. Analogously, the negative value of the population cannot be seen as its adequate part of the collective influence on GDP. The supplementary influences make the effect of multicollinearity clear: Every regressor has a very small additional influence according to its coefficient, and the corresponding variance percentages are not more than 1 %. A partition of the collective influence is given by the particular coefficient: EAP has the value 20.4 Million ECU per 1,000 persons, which is about twice the coefficient of the population. Their variance percentage differs only slightly with 44 % versus 43 %, and so both regressors are equally expressive. The area has only a percentage of 11 % which is just about one forth of each of the other values and about one eighth of the other values summed up.

Table 2: Components of Total Influence on Gross Domestic Product

Row Variable	Column Variable			**Total**
	EAP	Population	Area	
	Indication			
	Million ECU Per Unit of Row Variable			
EAP	*0.6*	37.1	7.8	**45.4**
Population	16.9	*-0.04*	3.7	**20.5**
Area	0.84	0.91	*-0.04*	**1.71**
	Determination			
	% of Variance of GDP			
EAP	*1*	79	17	**97**
Population	76	*1*	19	**96**
Area	7	8	*18*	**32**

A closer look on the structure of the total influences is provided by table 2: The total coefficient 45.5 of EAP can be split up into its supplementary part 0.6 (see table 1) and a common part of 44.9 (= 37.1 + 7.8), which is attributable to the population with a variance percentage of 79 % and to the area with 17 %. Hence the area's contribution is relatively small within the total influence of EAP on GDP. The structure of the total influence of the area is different, because its supplementary coefficient is small and negative. The positive value of the total coefficient can be attributed to EAP and the population in roughly equal parts (0.84 and 0.91). Yet the variance percentage (32 %) belongs to more than one half to the area itself. Only 7 % respectively 8 % can be attributed to the other regressors. So the area is separated from EAP and population, which is understandable since both latter variables directly refer to the population structure of a member country.

To conclude, it may be said that the new statistics give a plausible breakdown of the collective influences of EAP, the population and the area on GDP, and so provide a deeper insight into the data. However, one can assess the quality of a coefficient in detail by using scatter plots of GDP against the (adjusted and synchronized) variable and so outliers can be detected. Since no stochastic model is tested, no checking of the according assumptions is necessary to interpret the results descriptively.

4 Mathematics

4.1 Influence and Adjustment

Let *n* be a natural number.

4.1.1 Notation. A mapping $E : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined by

$$E(x,y) := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{for all } x, y \in \mathbb{R}^n$$

(if the nominator is zero, then $E(x,y) := 0$).

4.1.2 Remark. If $\bar{x} = \bar{y} = 0$ and $x'x > 0$ then $E(x,y) = (x'y)/(x'x)$.

4.1.3 Notation. A mapping $B : \mathbb{R}^n \times \mathsf{P}(\mathbb{R}^n) \to \mathbb{R}^n$ is defined by

$$B(x,M) := x - X(X'X)^{-1}X'x \quad \text{for all } x \in \mathbb{R}^n, M \subseteq \mathbb{R}^n$$

where the columns of *X* are a basis of the vector space $\operatorname{span}(\{1_n\} \cup M)$.

4.1.4 Remark. Let be $x \in \mathbb{R}^n$ and $M \subseteq \mathbb{R}^n$. Then

a) $B(x,\varnothing) = x - \bar{x}$,

b) $B(y,\{x\}) = (y - \bar{y}) - E(x,y) \cdot (x - \bar{x})$,

c) *x* and *M* are not correlated if $B(x,M) = x - \bar{x}$,

d) *x* and *M* are perfectly correlated if $B(x,M) = 0$.

4.1.5 Notation. For $x, y \in \mathbb{R}^n, M \subseteq \mathbb{R}^n$ let be $e := B(x,M)$ and $f := B(y,M)$.

4.1.6 Lemma. For all $x \in \mathbb{R}^n, M \subseteq \mathbb{R}^n$ it holds

a) $e = 0$ if $x \in M$,

b) $B(e,M) = e$.

Proof (cf. Greene 1993: 178). With $H := E_n - X(X'X)^{-1}X'$ one has $e = Hy$. Straightforward matrix calculations show $H^2 = H$ and $H' = H$. For a): For $x \in M$ there is a $a \in \mathbb{R}^{1+q}$ such that $x = Xa$ and so $B(x,M) = HXa = (X - X(X'X)^{-1}(X'X))a = (X - X)a = 0$. For b): $B(e,M) = He = H(Hx) = H^2x = Hx = e$.

4.1.7 Lemma. Let $x, y \in \mathbb{R}^n$ and $M \subseteq \mathbb{R}^n$. Then

a) $E(e,y) = E(e,f)$,

b) $E(e,x) = 1$ if $e \neq 0$,

c) $E(x,e) \cdot E(e,y) = E(x,f)$.

Proof. Without restricting the proof, $\bar{x} = \bar{y} = 0$ is assumed. For a): By using H as in the proof of lemma *4.1.6* one gets $e'y = (Hx)'y = x'Hy = x'H^2y = x'H'Hy = (Hx)'(Hy) = ef$ and so property a). For b): The special case $x = y$ gives $E(e,x) = E(e,e) = 1$. For c): Now one has

$$E(x,e) \cdot E(e,y) = \frac{x'e}{x'x}\frac{e'y}{e'e} = \frac{e'x}{x'x}\frac{e'y}{e'e} = \frac{e'e}{x'x}\frac{e'f}{e'e} = \frac{e'f}{x'x} = E(x,f).$$

4.1.8 Theorem. For $x, y \in \mathbb{R}^n$ and $M \subseteq \mathbb{R}^n$ one has

$$B(y, M \cup \{x\}) = B(f, \{e\}).$$

Proof. Let a matrix X be such that its columns are a basis of the vector space $\operatorname{span}(\{1_n\} \cup M)$ and define $W := (X, e)$. Somewhat tedious matrix calculations using the linear independence of X's columns and e show

$$W(W'W)^{-1}W' = X(X'X)^{-1}X' + ee'$$

and so by using a)

$$\begin{aligned} B(y, M \cup \{x\}) &= y - W(W'W)^{-1}W'y = y - X(X'X)^{-1}X'y - ee'y = f - ee'y \\ &= f - (e'y) \cdot e = f - (e'f) \cdot e = B(f, \{e\}). \end{aligned}$$

4.2 Partial Coefficient

Let be $k = 1, \ldots, q$.

4.2.1 Definition. A *regression task* (with q regressors) in $\mathbb{R}^n$ is a tupel $(x_1, \ldots, x_q, y)$ such that $x_1, \ldots, x_q, y \in \mathbb{R}^n$. It is called *regular*, if the columns of $X := (1_n, x_1, \ldots, x_q)$ are linearly independent and $\sum_{i=1}^{n} (y_i - \bar{y})^2 > 0$.

4.2.2 Notation. Let $(x_1, \ldots, x_q, y)$ be a regular regression task. Then

$$p := (p_0, p_1, \ldots, p_q)' := (X'X)^{-1}X'y.$$

4.2.3 Remark. The number p_k is the partial coefficient of x_k.

4.2.4 Notation. $M := \{x_1, \ldots, x_q\}$, $M_k := M \setminus \{x_k\}$, $e_k := B(x_k, M_k)$ and $f_k := B(y, M_k)$.

4.2.5 Lemma (cf. Ezekiel and Fox 1959: 170ff; Ryan 1997: 168ff). Let $(x_1, \ldots, x_q, y)$ be a regular regression task. Then $p_k = E(e_k, y)$.

Proof. One has

$$y = \sum_{l=0}^{q} p_l x_l + f = \bar{y} + \sum_{l=1}^{q} p_l x_l + f$$

and so by using lemma *4.1.7*

$$E(e_k, y) = \sum_{l=1}^{q} E(e_k, p_l x_l) + E(e_k, f) = \sum_{l=1}^{q} p_l E(e_k, x_l) + E(e_k, f)$$

$$= p_k E(e_k, x_k) + 0 = p_k \cdot 1 = p_k .$$

4.2.6 Corollary (Frisch and Waugh 1933: 391ff; cf. Greene 1993: 180). $p_k = E(e_k, f_k)$.

Proof. This follows from lemma *4.2.5* together with lemma *4.1.7*.

4.2.7 Remark. Lemma *4.2.5* completes the definition of p_k for non-regular regression tasks.

4.3 Coefficient of Determination

Let $(x_1, \ldots, x_q, y)$ be a regular regression task in $\mathbb{R}^n$.

4.3.1 Notation. $\hat{y} := X(X'X)^{-1}X'y$.

4.3.2 Notation. $R := \dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$.

4.3.3 Remark. R is the multiple coefficient of determination (also known as R square coefficient) of the regression task.

4.3.4 Lemma. $R = 1 - E(y, f)$.

Proof. Without restricting the proof $\bar{y} = 0$ can be assumed. Then

$$R = (\hat{y}'\hat{y})/(y'y) = (y - f)'(y - f)/(y'y) = (y'y - 2y'f + f'f)/(y'y)$$

$$= 1 - 2E(y, f) + (f'f)/(y'y) .$$

By using lemma *4.1.7* one gets $E(y, f) = (y'f)/(y'y) = (f'f)/(y'y) \cdot (y'f)/(f'f) = (f'f)/(y'y) \cdot E(f, y) = (f'f)/(y'y) \cdot 1 = (f'f)/(y'y)$, which completes the proof.

4.3.5 Notation. A mapping $R : \mathsf{P}(\mathbb{R}^n) \times \mathbb{R}^n \to [0, 1]$ is defined by

$$R(M, x) := 1 - E(x, e) \quad \text{for all } x \in \mathbb{R}^n, M \subseteq \mathbb{R}^n .$$

4.3.6 Remark. $R(x, y) := R(\{x\}, y) = E(x, y) \cdot E(y, x)$ for all $x, y \in \mathbb{R}$.

4.4 Supplementary Coefficient

Let $(x_1, \ldots, x_q, y)$ be a regression task and $k = 1, \ldots, q$.

4.4.1 Notation. A mapping $B_s : \mathbb{R}^n \times \mathsf{P}(\mathbb{R}^n) \to \mathbb{R}^n$ is defined by

$$B_s(x, M) := E(x, e)^{-1} \cdot e \quad \text{for all } x \in \mathbb{R}^n, M \subseteq \mathbb{R}^n$$

(if $E(x, e) = 0$ then $B_s(x, M) := 0$).

4.4.2 Lemma. For all $x \in \mathbb{R}^n, M \subseteq \mathbb{R}^n$ it holds $E(x, B_s(x, M)) = 1$ if $B_s(x, M) \neq 0$.

Proof. $E(x, B_s(x, M)) = E(x, E(x, e)^{-1} \cdot e) = E(x, e)^{-1} E(x, e) = 1$.

4.4.3 Definition. The *supplementary coefficient* of x_k is

$$s_k := E(B_s(x_k, M_k), y).$$

4.4.4 Notation. $T_k := E(x_k, e_k)$.

4.4.5 Lemma. T_k is the tolerance of x_k, that is $T_k = 1 - R(M_k, x_k)$.

Proof. Lemma *4.3.4* gives $R(M_k, x_k) = 1 - E(x_k, e_k)$ and so $T_k = E(x_k, e_k) = 1 - R(M_k, x_k)$.

4.4.6 Remark. $T_k = 1$ if x_k and M_k are not correlated, and $T_k = 0$ if x_k and M_k are perfectly correlated.

4.4.7 Theorem (Fickel 1999: 221). $s_k = T_k \cdot p_k$.

Proof. $s_k = E(E(x_k, e_k)^{-1} \cdot e_k, y) = E(x_k, e_k) \cdot E(e_k, y) = T_k \cdot p_k$.

4.4.8 Corollary. $s_k = E(x_k, y)$ if x_k and M_k are not correlated, and $s_k = 0$ if x_k and M_k are perfectly correlated.

Proof. This follows from remark *4.4.6*.

4.4.9 Theorem. $s_k = E(x_k, f_k)$.

Proof. From theorem *4.4.7* together with lemma *4.1.7* follows

$$s_k = E(x_k, e_k) \cdot E(e_k, y) = E(x_k, e_k) \cdot E(e_k, f_k) = E(x_k, f_k).$$

4.4.10 Definition. The *supplementary increment* of x_k is $\Delta R_k := R(M, y) - R(M_k, y)$.

4.4.11 Lemma. $\Delta R_k = R(e_k, y)$.

Proof. By using lemma *4.1.7* two times one gets

$$\begin{aligned} R(M, y) &= 1 - E(y, f) = 1 - E(y, B(y, M)) = 1 - E(y, B(y, M_k \cup \{x_k\})) \\ &= 1 - E(y, B(B(y, M_k), e_k)) = 1 - E(y, B(f_k, e_k)) = 1 - E(y, f_k - E(e_k, f_k) \cdot e_k) \\ &= 1 - E(y, f_k) + E(e_k, f_k) \cdot E(y, e_k) = (1 - E(y, f_k)) + E(e_k, y) \cdot E(y, e_k) \\ &= R(M_k, y) + R(e_k, y) \end{aligned}$$

and so the proposition follows by definition *4.4.10*.

4.5 Anti-Partial Coefficient

Let $(x_1, \ldots, x_q, y)$ be a regression task and $k = 1, \ldots, q$.

4.5.1 Notation. $\hat{x}_k := x_k - e_k$.

4.5.2 Definition. The *anti-partial coefficient* of x_k is $\hat{p}_k := E(\hat{x}_k, y)$.

4.5.3 Theorem. $E(x_k, y) = (1 - T_k)\hat{p}_k + T_k p_k$.

Proof. Without restricting the proof let be $\bar{x}_k = \bar{y} = 0$. Then

$$E(x_k,y) \;=\; E(\hat{x}_k+e_k,y) \;=\; \frac{\hat{x}_k'\hat{x}_k+e_k'y}{x_k'x_k} \;=\; \frac{\hat{x}_k'\hat{x}_k}{x_k'x_k}E(\hat{x}_k,y)+\frac{e_k'y}{x_k'x_k}E(e_k,y)$$

$$= R(x_k,M_k)E(\hat{x}_k,y)+(1-R(x_k,M_k))E(e_k,y) \;=\; (1-T_k)\hat{p}_k+T_k p_k .$$

4.5.4 Remark.

a) $E(x_k,y)=(1-T_k)\hat{p}_k+s_k$,

b) $R(x_k,y)=(1-T_k)R(x_k,y)+T_kR(x_k,y)$,

c) $(1-T_k)R(x_k,y)\geq 0$, $T_kR(x_k,y)\geq 0$.

4.6 Supplementary Sequence

Let *n* and *q* be natural numbers.

4.6.1 Notation. Let S_q be the set of permutations of $\{1,\ldots,q\}$, that is all bijective mappings from $\{1,\ldots,q\}$ onto $\{1,\ldots,q\}$.

4.6.2 Notation. Let $(x_1,\ldots,x_q,y)$ be a regression task in $\mathbb{R}^n$. Then for $\sigma\in S_q$

a) $s^\sigma_{\sigma(k)} := E(B_s(x_{\sigma(k)};x_{\sigma(1)},\ldots,x_{\sigma(k-1)}),y)$ for all k = 1, ..., q,

b) $\Delta R^\sigma_{\sigma(k)} := R(B(x_{\sigma(k)};x_{\sigma(1)},\ldots,x_{\sigma(k-1)}),y)$ for all k = 1, ..., q

(where $\{x_{\sigma(1)},\ldots,x_{\sigma(k-1)}\}:=\emptyset$ if k = 1).

4.6.3 Lemma. Let $(x_1,\ldots,x_q,y)$ be a regression task and $\sigma\in S_q$. Then

$$\sum_{k=1}^{q}\Delta R^\sigma_{\sigma(k)} = R .$$

Proof. This follows by using lemma *4.4.11* iteratively and $R(B(x_{\sigma(1)};\emptyset),y) = R(x_{\sigma(1)},y)$.

4.6.4 Notation. $1_{a,b} := \begin{cases}1, & \text{if } a\neq b\\ 0, & \text{if } a=b\end{cases}$ for all $a,b\in\mathbb{R}$.

4.6.5 Lemma. For every k = 1, ..., q it holds

a) $s^\sigma_k = E(x_k,y)$ and $\Delta R^\sigma_k = R(x_k,y)$, if x_k and M_k are not correlated,

b) $s^\sigma_k = E(x_k,y)\cdot 1_{\sigma(k),1}$ and $\Delta R^\sigma_k = R(x_k,y)\cdot 1_{\sigma(k),1}$, if x_k and M_k are perfectly correlated.

Proof. This follows from lemma *4.4.8*.

4.7 Concentration Ratio of Rosenbluth

Let $(x_1,\ldots,x_q,y)$ be a regression task in $\mathbb{R}^n$ and $\sigma\in S_q$.

4.7.1 Notation. $K^\sigma_R := \dfrac{1}{(\frac{2}{R}\sum_{k=1}^{q}k\Delta R^\sigma_k)-1}$ (if $R=0$ then $K^\sigma_R := 1$).

4.7.2 Lemma. $\frac{1}{2q-1} \le K_R^\sigma \le 1.$

Proof. Using lemma *4.6.3* gives

$$R = \sum_{k=1}^{q} \Delta R_k^\sigma \le \sum_{k=1}^{q} k\Delta R_k^\sigma \le \sum_{k=1}^{q} q\Delta R_k^\sigma = qR,$$

and so, if $R > 0$, by transforming $x \mapsto 2x/R - 1$

$$1 \le \frac{2}{R}\sum_{k=1}^{q} k\Delta R_k^\sigma - 1 \le 2q-1.$$

Hence the proof is completed by taking reciprocals.

4.8 Particular Coefficient

Let $(x_1,\ldots,x_q,y)$ be a regression task and $k = 1, ..., q$.

4.8.1 Notation. Define

a) $\bar{s}_k := \frac{1}{q!}\sum_{\sigma\in S_q} s_k^\sigma$,

b) $\Delta\bar{R}_k := \frac{1}{q!}\sum_{\sigma\in S_q} \Delta R_k^\sigma$.

4.8.2 Lemma. $\sum_{k=1}^{q} \Delta\bar{R}_k = R.$

Proof. This follows from lemma *4.6.3* when averaged for every ordering $\sigma \in S_q$.

4.8.3 Lemma.

a) $\bar{s}_k = E(x_k, y)$ and $\Delta\bar{R}_k = R(x_k, y)$, if x_k and M_k are not correlated.

b) $\bar{s}_k = E(x_k, y)/q$ and $\Delta\bar{R}_k = R(x_k, y)/q$, if x_k and M_k are perfectly correlated.

Proof. By using lemma *4.6.5* one gets:

For a): $\bar{s}_k = \frac{1}{q!}\sum_{\sigma\in S_q} E(x_k, y) = \frac{1}{q!} q!\, E(x_k, y) = E(x_k, y).$

For b): $\bar{s}_k = \frac{1}{q!}\sum_{\sigma\in S_q} E(x_k, y)\cdot 1_{\sigma(k),1} = \frac{1}{q!}(q-1)!\, E(x_k, y) = E(x_k, y)/q.$

The properties of $\Delta\bar{R}_k$ can be followed analogously.

4.9 Components of Total Coefficient

Let $(x_1,\ldots,x_q,y)$ be a regression task and $k = 1, ..., q$.

4.9.1 Notation. Let $\sigma \in S_q$ with $\sigma(q) = k$. Then

$$a_{k,l}^{\sigma} \coloneqq E(B_s(x_k; x_{\sigma(1)}, \ldots, x_{\sigma(l)}), y) \quad \text{for all } l = 1, ..., q$$

and $a_{k,0}^{\sigma} \coloneqq E(B_s(x_k, \varnothing), y)$. Further

$$\Delta a_{k,l}^{\sigma} \coloneqq a_{k,l-1}^{\sigma} - a_{k,l}^{\sigma} \quad \text{for all } l = 1, ..., q.$$

4.9.2 Lemma. Let $\sigma \in S_q$ with $\sigma(q) = k$. Then $E(x_k, y) = \sum_{l=1}^{q} \Delta a_{k,l}^{\sigma}$ and $\Delta a_{k,q}^{\sigma} = s_k$.

Proof. One has $a_{k,0}^{\sigma} = E(\bar{x}_k, y) = E(x_k, y)$, $a_{k,q-1}^{\sigma} = s_k$, $a_{k,q}^{\sigma} = E(0, y) = 0$ and so

$$E(x_k, y) = E(x_k, y) - 0 = a_{k,0}^{\sigma} - a_{k,q}^{\sigma} = \sum_{l=1}^{q} \Delta a_{k,l}^{\sigma}$$

and $\Delta a_{k,q}^{\sigma} = a_{k,q-1}^{\sigma} - a_{k,q}^{\sigma} = s_k - 0 = s_k$.

4.9.3 Notation. Let $\sigma \in S_q$ with $\sigma(q) = k$. Then

$$\Delta R_{k,l}^{\sigma} \coloneqq R(x_k, \tilde{x}_l^{\sigma}) \cdot R(x_k, y) \quad \text{for all } l = 1, ..., q$$

where $\tilde{x}_l^{\sigma} \coloneqq B_s(x_{\sigma(l)}; x_{\sigma(1)}, \ldots, x_{\sigma(l-1)})$.

4.9.4 Lemma. For all $l = 1, ..., q-1$ and $\sigma \in S_q$ with $\sigma(q) = k$ it holds

a) $\Delta a_{k,l}^{\sigma} = 0$ and $\Delta R_{k,l}^{\sigma} = 0$, if x_k and M_k are not correlated,

b) $\Delta a_{k,l}^{\sigma} = E(x_k, y) \cdot 1_{1,l}$ and $\Delta R_{k,l}^{\sigma} = R(x_k, y) \cdot 1_{1,l}$, if x_k and M_k are perfectly correlated.

Proof. For a): Then $\Delta a_{k,l}^{\sigma} = E(x_k, y) - E(x_k, y) = 0$ and $\Delta R_{k,l}^{\sigma} = 0 \cdot E(x_k, y) = 0$.

For b): Then for $l = 1$

$$\Delta a_{k,l}^{\sigma} = E(B_s(x_k, \varnothing), y) - E(B_s(x_k, x_{\sigma(1)}), y) = E(x_k, y) - 0 = E(x_k, y),$$

$$\Delta R_{k,l}^{\sigma} = R(x_k, x_{\sigma(1)}) R(x_k, y) = R(x_k, x_k) R(x_k, y) = R(x_k, y),$$

and for $l > 1$

$$\Delta a_{k,l}^{\sigma} = E(0, y) - E(0, y) = 0,$$

$$\Delta R_{k,l}^{\sigma} = R(x_k, 0) R(x_k, y) = 0 \cdot R(x_k, y) = 0.$$

4.9.5 Lemma. Let $\sigma \in S_q$ with $\sigma(q) = k$. Then

a) $\Delta R_{k,l}^{\sigma} \geq 0$ for all $l = 1, ..., q$,

b) $R(x_k, y) = \sum_{l=1}^{q} \Delta R_{k,l}^{\sigma}$,

c) $\Delta R_{k,q}^{\sigma} = T_k \cdot R(x_k, y)$.

Proof. For a): This holds since $\Delta R_{k,l}^{\sigma}$ is a product of two non-negative factors.

For b): Applying lemma *4.6.3* and using lemma *4.3.4* gives

$$\sum_{l=1}^{q-1} R(\tilde{x}_l^{\sigma}, x_k) = R(x_{\sigma(1)}, \ldots, x_{\sigma(q-1)}; x_k) = 1 - E(x_k, \tilde{x}_q^{\sigma})$$

and hence by using lemma *4.1.7*

$$\sum_{l=1}^{q} R(\tilde{x}_l^{\sigma}, x_k) = 1 - E(x_k, \tilde{x}_q^{\sigma}) + R(x_k, \tilde{x}_q^{\sigma}) = 1 - E(x_k, \tilde{x}_q^{\sigma}) + E(x_k, \tilde{x}_q^{\sigma}) \cdot 1 = 1.$$

Multiplying both sides by $R(x_k, y)$ proofs b).

For c): Since $\tilde{x}_q^{\sigma} = B_s(x_k, M_k)$ it follows with notation *4.4.4* and lemma *4.1.7* that $T_k = E(x_k, e_k) = R(x_k, e_k) = R(x_k, \tilde{x}_q^{\sigma})$ and hence c).

4.9.6 Notation. By using $S_{q,k} := \{\sigma \in S_q \mid \sigma(q) = k\}$ one defines

a) $\Delta\bar{a}_{k,l} := \frac{1}{(q-1)!} \sum_{\sigma \in S_{q,k}} \Delta a_{k,r}^{\sigma}$ for all $l = 1, \ldots, q$,

b) $\Delta\bar{R}_{k,l} := \frac{1}{(q-1)!} \sum_{\sigma \in S_{q,k}} \Delta\bar{R}_{k,r}^{\sigma}$ for all $l = 1, \ldots, q$

(where $r := \sigma^{-1}(l)$ varies in the sums).

4.9.7 Lemma. For all $l = 1, \ldots, q$ it holds

a) $\Delta\bar{a}_{k,l} = E(x_k, y) \cdot 1_{k,l}$ and $\Delta\bar{R}_{k,l} = R(x_k, y) \cdot 1_{k,l}$, if x_k and M_k are not correlated,

b) $\Delta\bar{a}_{k,l} = E(x_k, y)/(q-1) \cdot 1_{k,l}$ and $\Delta\bar{R}_{k,l} = R(x_k, y)/(q-1) \cdot 1_{k,l}$, if x_k and M_k are perfectly correlated.

Proof. This directly follows with the help of lemma *4.9.4*.

4.9.8 Theorem (Fickel 1999: 242ff).

a) $E(x_k, y) = \sum_{l=1}^{q-1} \Delta\bar{a}_{k,l} + s_k$,

b) $\Delta\bar{R}_{k,l} \geq 0$ for all $l = 1, \ldots, q$,

c) $R(x_k, y) = \sum_{l=1}^{q} \Delta\bar{R}_{k,l} + T_k \cdot R(x_k, y)$.

Proof. These are direct consequences of lemma *4.9.2* and lemma *4.9.5*.

4.10 Neodescriptive Statistics

4.10.1 Definition. A *neodescriptive statistic* on a set M is a pair (I, D) such that $I : M \to \mathbb{R}$ and $D : M \to [0, 1]$ are mappings.

4.10.2 Remark. The pair (E, R) is the neodescriptive statistic on $\mathbb{R}^n \times \mathbb{R}^n$ measuring total influence (cf. section 1.1).

4.10.3 Remark. Let n and q be natural numbers. For every $k = 1, \ldots, q$ each of the following pairs is a neodescriptive statistic on the set of all regression tasks with q regressors in $\mathbb{R}^n$:

a) $(s_k, \Delta R_k)$ measures the supplementary influence of regressor x_k (cf. section 2.1),

b) $(s_k, T_k R(x_k, y))$ measures the supplementary part of the total influence of regressor x_k (cf. section 2.1),

c) $(s_k^\sigma, \Delta R_k^\sigma)$ measures the supplementary influence of regressor x_k within the sequence $\sigma \in S_q$ (cf. section 2.2),

d) $(\bar{s}_k, \Delta \overline{R}_k)$ measures the particular influence of regressor x_k (cf. section 2.3),

e) $(\Delta \bar{a}_{k,l}, \Delta \overline{R}_{k,l})$ measures the component of regressor x_l in the total influence of regressor x_k for every $l = 1, ..., q$ (cf. section 2.4).

5 References

Bomsdorf, Eckart (1993): "A new adjusted coefficient of determination in multiple linear regression (German)". *Allgemeines Statistisches Archiv* 77: 233-239.

Bomsdorf, Eckart (1994): "Some comparative remarks on the behavior of adjusted coefficients of determination (German)". *Allgemeines Statistisches Archiv* 78: 197-206.

Breiman, Leo (1995): "Better subset regression using the nonnegative garrote". *Technometrics* 37: 373-384.

Bring, Johan (1994): "How to standardize regression coefficients". *The American Statistician* 48: 209-213.

Chambers, John M., [et al.] (1983): *Graphical Methods for Data Analysis*. Duxbury, Boston (Massachusetts).

Draper, Norman R.; Smith, Harry (1998): *Applied Regression Analysis*. 3rd edition. Wiley, New York.

Ezekiel, Mordecai; Fox, Karl A. (1959): *Methods of Correlation and Regression Analysis: Linear and Curvilinear*. 3rd edition. Wiley, New York.

Federal Statistical Office of Germany (1999): *Statistical Yearbook 1999 for Foreign Countries (German).* Metzler-Poeschel, Stuttgart.

Fickel, Norman (1999): *Sequential Regression: A Neodescriptive Solution of the Multicollinearity Problem Using Stepwisely Adjusted and Synchronized Variables (German).* Unpublished Habilitationsschrift. Friedrich-Alexander-University, Nuremberg.

Frisch, Ragnar; Waugh, Frederick V. (1933): „Partial time regressions as compared with individual trends". *Econometrica* 1: 387-401.

Greene, William H. (1993): *Econometric Analysis.* 2nd edition. Macmillan, New York.

Gunst, Richard F. (1983): „Regression analysis with multicollinear predictor variables: definition, detection, and effects". *Communications in Statistics: Theory and Methods* 12: 2217-2260.

Haavelmo, Trygve (1944): *The Probability Approach in Econometrics.* Supplement to Econometrica, Volume 12. University of Chicago, Chicago (Illinois).

Hadi, Ali S.; Ling, Robert F. (1998): "Some cautionary notes on the use of principal components regression". *The American Statistician* 52: 15-19.

Hahn, Gerald J.; Meeker, William Q. (1993): "Assumptions for statistical inference". *The American Statistican* 47: 1-11.

Hawkins, Douglas M. (1973): “On the investigation of alternative regressions by principal component analysis”. *Applied Statistics* 22: 275-286.

Hoerl, Arthur E.; Kennard, Robert W. (1970): „Ridge regression: biased estimation for non-orthogonal problems“. *Technometrics* 12: 55-67.

Kadiyala, K. R.; Oberhelman, Dennis (1994): “A comparison of Stein-like procedures for estimating linear regression models with multicollinear data”. *Communications in Statistics, Theory and Methods* 23: 1447-1469.

Kruskal, William (1987): “Relative importance by averaging over orderings”. *The American Statistician* 41: 6-10.

Malinvaud, Edmond. (1966): *Statistical Methods of Econometrics*. North-Holland, Amsterdam.

Miller, Alan J. (1990): *Subset Selection in Regression*. Monographs on Statistics and Applied Probability 40. Chapman and Hall, London.

Morrow-Howell, Nancy (1994): „The M word: multicollinearity in multiple regression“. *Social Work Research* 18: 247-251.

Mosteller, Frederick; Tukey, John W. (1977): *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading (Massachusetts).

Nester, Marks R. (1996): “An applied statistician’s creed”. *Applied Statistics* 45: 401-410.

Newton, R. G.; Spurrell, D. J. (1967): „A development of multiple regression for the analysis of routine data“. *Applied Statistics* 16: 51-64.

Pedhazur, Elazar J. (1982): *Multiple Regression in Behavioral Research: Explanation and Prediction.* 2nd edition. Holt, Rinehart and Winston, New York.

Piesch, Walter (1975): *Statistische Konzentrationsmaße: Formale Eigenschaften und verteilungstheoretische Zusammenhänge*. Mohr, Tübingen.

Rawlings, John O.; Pantula, Sastry G.; Dickey, David A. (1998): *Applied Regression Analysis: A Research Tool*. 2nd edition. Springer, New York.

Ryan, Thomas P. (1997): *Modern Regression Methods*. Wiley, New York.

Stone, M.; Brooks, R. J. (1990): “Continuum regression: crossvalidated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression”. *Journal of the Royal Statistical Society, Series B* 52: 237-269.

Tibshirani, Robert (1996): “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society, Series B* 58: 267-288.