1982

Regression Analysis Using Information Theory

Klaus Krippendorff
University of Pennsylvania, klaus.krippendorff@asc.upenn.edu

Regression Analysis Using Information Theory

REGRESSION ANALYSIS USING INFORMATION THEORY

by

KLAUS KRIPPENDORFF

UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA

ABSTRACT

A common task in analyzing complex systems is to explain the variation in one variable in terms of several other variables. Examples are the development of inventories of causes or effects and the analysis of experiments or predictions. When the data for such an analysis are quantitative (interval or ratio scale data) methods of this kind are known as multiple regression analysis. The paper is concerned with the analoguous analysis of qualitative data using structural models instead of regression equations and information theory to account for the extent of the explanations offered.

The paper outlines five confirmatory regression models: algebraical compensation, algebraically unique, ordinal, additive, and unique. The first two are found defective, the last three are proposed and developed. The paper also explores the possibilities of using the proposed regression models in an exploratory mode, searching primarily for a smaller set containing the most powerful explanatory variables and constraints.

The work reported here further develops previously published contributions by Ashby (e.g. 1969), Klir (e.g. 1976) and collaborators, Broekstra (1979) and by the author (e.g. Krippendorff, 1981).

INTRODUCTION

In regression analysis we seek to explain typically one variable in terms of several other variables. With the causal notions of experiments in mind, the variables to be explained are often called the dependent variables and the variables that the experimenter manipulates in order to effect the former are called the independent variables. When regression is used for evaluating predictions, one speaks of criterion variables that are explained in terms of predictor variables. Regression may also be used to develop an inventory of consequences of a certain phenomena. In system theory regression might be used to determine what links the variable(s) in one system, e.g. an environment, to the variables in another system, e.g. an organism or an organization, the dependencies within either system being either ignored or at best given a contributory status. Thus regression analysis is concerned with an analytical problem that occurs in a wide range of empirical settings marked by

(a) a clear distinction between two classes of variables, those to be explained and those used to explain,
(b) the need to know the dependencies between variables of different classes (with dependencies within variables in one class merely complicating the analysis).

In the econometric literature the most typical regression model is given by:

$$z_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir} + e_i$$

It shows the relationship between r predictor variables x, an error term e, and one criterion variable z, for the i-th observation of a larger sample of n observations. $\beta_o \ldots \beta_r$ are the regression parameters for these predictor variables which are uniformly applied to all observations in the sample. The most outstanding properties of this model are (i) that the effects of the predictors are additive and (ii) that both criterion and predictor variables must have an interval or ratio metric else the conception of parameters as coefficient would be untenable. For these two reasons the model is called linear. While non-linear models are not unknown they are difficult to evaluate and hence rarely used.

One of the statistical problems in this kind of regression analysis is to find that set of parameters $\beta_o \ldots \beta_r$ for which the discrepancy, the error term e, is minimum for the larger population from which the finite sample of observations is drawn. Another problem is to find that set of predictor variables which offers an optimal account of the criterion variable in the sense of minimizing e relative to the number of predictor variables employed.

We are here concerned with the regression analysis of qualitative data. Such data have the form of i=1,2,...,n ordered many-tuples of qualities (properties, attributes, names, states, or classes of things, people or systems):

$$\langle a,b,c,\ldots,z \rangle_i$$

What quantitative and qualitative data share is that the orginal observations define a probability distribution in a many-dimensional space. But they differ in the conception of this space as either continuous or discrete respectively and in the way the distribution in one variable can be explained or accounted for in terms of the distributions in the other variables. The most obvious obstacle is that the algebraical operations of addition and multiplication which traditional regression equations require are not definable in qualitative variates.

A qualitative analogue of the quantitative regression parameters may be found in the probability distributions themselves. Given a value of say a, we find the probability of z in the table of observed cooccurrences of a and z. We can use this table of cooccurrences as operators whose domain is the probabilities of the values of the predictor variables, and whose range is the probabilities of the values in the criterion variables. In the rather obvious equality:

$$p_z = \left(\frac{p_{abc..}}{p_{abc..|z}}\right)\left(p_{z|abc..}\right)$$

$P_{abc...}$ is the probability of observing the value <abc...> in the predictor variables, $p_{abc..|z}$ is the probability of that predictor's value given z and $p_{z|abc..}$ is the probability of the criterion's value z given the predictor value. If the observations indicate that predictors and criterion are a mere product of chance cooccurrences than for all cooccurrences, the left expression is unity and the right expression equals p_z. And if that distribution defines a mapping of ABC... to Z then the right expression is unity and the left one equals p_z. Thus the left expression relates the vector of predictor values <abc...> to the values of the criterion z whereas the right expression is an error term.

Making a jump and taking $-\sum p_{ab..z} \log_2(\quad)$ of these terms:

$$-\sum p_{abc..z} \log_2 p_z = H(Z)$$

$$-\sum p_{abc..z} \log_2 \frac{p_{abc..}}{p_{abc..|z}} = T(Z:ABC..)$$

$$-\sum p_{abc..z} \log_2 p_{z|abc..} = H_{ABC..}(Z)$$

yields the information theoretical accounting equation for regression:

$$H(Z) = T(Z:ABC..) + H_{ABC..}(Z)$$

Here H(Z) is the total entropy in the criterion variable Z. T(Z:ABC..) is the amount of information transmitted from the predictor variables ABC.. to the criterion variable, all predictor variables taken as one vector, or the amount of entropy in Z explainable from ABC.. . $H_{ABC..}(Z)$ is the amount of noise or the amount of entropy in Z not predictable from ABC.. . These three quantities are non-negative. If $H_{ABC..}(Z)=0$ then Z is perfectly explainable from the variables in the subscript. If T(Z:ABC...)=0 then the criterion Z is not explainable at all. This is the basic unanalyzed accounting equation for qualitative data and it is quite easily seen that there are no linearity assumptions implied and that the equation is applicable to quantitative data in continuous spaces as well, without requiring the idealizations typical for dealing with continuous spaces.

CONFIRMATORY APPROACHES TO REGRESSION

In a confirmatory approach one states a structural regression model and tests its fit against available data. In this context it means analyzing the quantity T(Z:ABC..) into various components including a structural error. We describe and reject two approaches to this analysis and thereby set the stage for the three approaches we do propose.

Algebraical Compensations

This approach is characterized by the following accounting equation:

$$H(Z) = \sum_I T(Z:I) + \sum_I \sum_J Q(Z:I:J) + \ldots + Q(Z:A:B:..) + H_{ABC..}(Z)$$

The first sum is the sum of all binary information terms. Since the predictor variables may be correlated, the second sum compensates for each of the algebraical differences between T(Z:IJ) and T(Z:I)+T(Z:J). The third sum compensates for each of the differences between T(Z:IJK) and T(Z:I)+T(Z:J)+T(Z:K)+Q(Z:I:K)+Q(Z:J:K)+Q(Z:I:J) and so forth. The argument against this approach is that Q-terms do not measure interaction. Unlike entropy and information quantities, Q-measures may be positive or negative and compensate for the algebraical accounting mistakes made by using one less than the full set of variables (Krippendorff, 1980).

Algebraically Unique Contributions

This approach attempts to account for the possibility that a quantity T(Z:A) is spurious in the sense that the apparent relation between A and Z may be caused by a third predictor variable not included in the measure. A's unique contribution is considered to be the difference between what all variables explain and what all variables except A explain:

$$T(Z:ABC...) - T(Z:BC...) = T_{BC..}(Z:A)$$

Carrying this argument into pairs, triples, etc. of predictor variables, and correcting for what each algebraically includes yields the following accounting equation:

$$H(Z) = \sum_I T_{\bar{I}}(Z:I) + \sum_I \sum_J Q_{\bar{I}\bar{J}}(Z:I:J) + \ldots + Q(Z:A:B:..) + H_{ABC..}$$

where $\bar{I}$ denotes the set of predictor variables with I removed, $\bar{I}\bar{J}$ all except I and J, etc. Here again uninterpretable Q-terms are the algebraical consequences of this notion. The equation is unquestionable for the unique effects of single variables but the error $T(Z:ABC...)-\sum T(Z:I)$ has no clear structural interpretation which does not make even this simple regression notion an attractive one.

For good reasons did we prefix the preceeding two approaches with algebraical for it is the algebraical computation of structural errors that cause difficulties when the underlying structure of a regression model involves loops or circular dependencies. Loops are invariably involved when dependencies among two or more predictor variables are admitted in addition to dependencies between predictor variables and the criterion. The ideas going into the above approaches are important but the algebraical method of evaluation offers no way of realizing them as intended. In the following we are proposing three approaches that evaluate the dependencies involved iteratively (Krippendorff, 1981).

Ordinal Contributions

Regression models of this kind consider the orders of complexity in the contribution the predictor variables make toward explaining the criterion variable. Besides the case in which no contribution is made (zeroth order), the contributions may be mediated through separate channels, one for each variable (first order), one for each pair of variables (second order), one for each triple of variables (third order) etc. With T_i denoting the amount of information contributed on the i-th order of complexity, we define:

$T_o = 0$

$T_1 = T(Z:ABC..) - T(ZA:ZB:ZC:..:ABC..)$

$T_2 = T(ZA:ZB:ZC:..:ABC..) - T(ZAB:ZAC:ZBC:..:ABC..)$

$T_3 = T(ZAB:ZAC:ZBC:..:ABC..) - \quad ...$

$T_r = \quad ... \quad - T(ZABC..)$

and obtain the accounting equation for the entropy in Z involving these r predictor variables by:

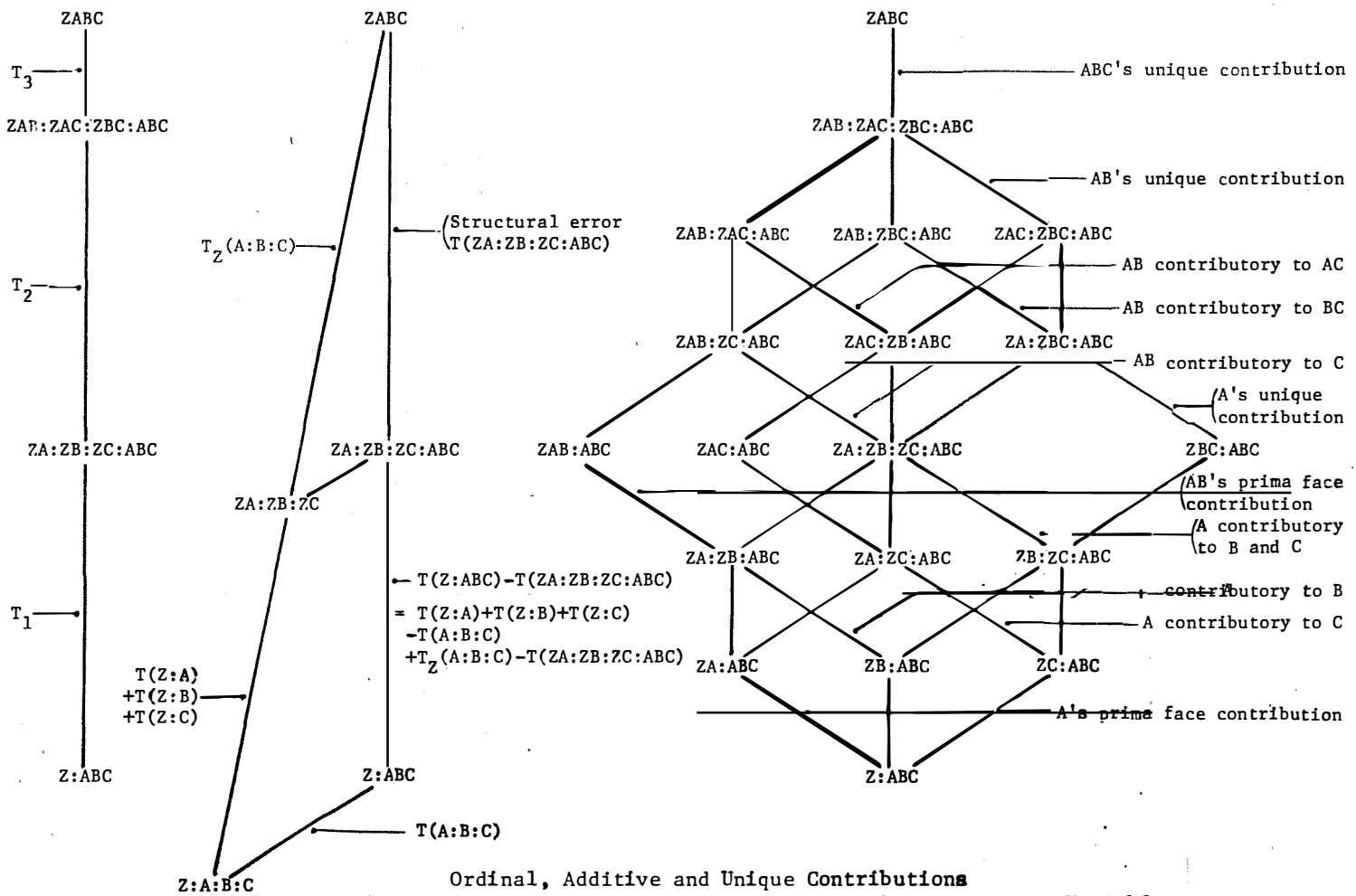
$$H(Z) = T_1 + T_2 + T_3 + ... + T_r + H_{abc..}(Z)$$

The contributions thereby summarized can be seen on the left side of Figure 1. It depicts the regression models as a scale. These models can be seen as a special case of all possible regression models which are depicted on the right side.

By examining the quantities in this equation, the researcher can quickly see where the predominant complexities lie and how large the error would be if he were to adopt a structural regression model that ignores complexities of an order higher than contained in that model.

Additive Contributions

Regression models of this kind pitch a structural model of independent predictor variables against the data. In so far as "independence" implies that effects are mediated through single channels (are of the first order) the quantity T(ZA:ZB:ZC:...:ABC..) assesses the error due to orders of effects higher than one. But there is another error, independent of the former which is due to the lack of additivity among predictor variables. This is indicated by two error quantities. T(A:B:C) responds to the presence of interdependencies within the predictor variables without correcting for the colinearities. T_ZA:B:C)-T(ZA:ZB:ZC:ABC) is a measure of the latter. Both distract from the ability



Ordinal, Additive and Unique Contributions
in the Relevant Lattices of Regression Models for Three Predictor Variables

Figure 1

to add up the individual effects of the predictor variables on the criterion, but they distract in opposite ways. Thus the accounting equation becomes:

$$H(Z) = \sum_I T(Z:I)$$
$$+ T_Z(A:B:..) - T(ZA:ZB:...:AB..) \Big\} \text{non-additivity}$$
$$- T(A:B:C:..)$$
$$+ T(ZA:ZB:...:AB..) \quad \text{structural loss}$$
$$+ H_{ABC..}(Z) \quad \text{noise}$$

It should be noted that the error due to the non-additivity of predictor variables behaves somewhat similar to a Q-measure. In fact for the two-variable case $T_Z(A:B)-T(A:B)=Q(Z:A:B)$ which has the previously stated flaws. In the current case this is tolerable because additivity is an algebraical property from which both errors, $T_Z(A:B:C:..)-T(ZA:ZB:ZC:..:ABC..)$ and $T(A:B:C:..)$ distract. For this reason it is not sufficient that their difference is zero. For predictor variables to be accepted as additive, each of these two errors must be statistically insignificant as well. The configuration of relevant models is depicted in Figure 1's center.

Unique Contributions

Regression models of this kind separate from the total amount explainable by all variables, the entropies that are uniquely attributable to the effects of any one, two, three, etc. predictor variables on the criterion and partitions the remainder into two quantities respectively assessing the contribution of an order higher than the chosen effect (the structural error) and of an order lower than the chosen effect. This involves picking out of lattice of possible regression models between ZAB... and Z:ABC.. two minimally different models that differ only in the presence/absence of a particular set of predictor variables. Table 1 exemplifies such models for the effects in one to six variables.

The unique contribution of any combination of one or more variables involves the comparison of models with loops and can hence not be evaluated algebraically. Accordingly, the algebraical expression of the contribution of one variable A, $T_{BCD..}(Z:A)$ becomes misleading as it measures not only A's unique contribution but the structural error of this contribution as well. Our accounting equation for any one unique contribution consists of the following four quantities for which Table 1 gives some examples:

$$H(Z) = T(Z:ABC..) - T(\text{unique effects removed})$$
$$+ T(\text{unique effects removed}) - T(\text{unique effects present})$$
$$+ T(\text{unique effects present})$$
$$+ H_{ABC..}(Z)$$

The first line quantitatively expresses the contribution made by all lower order effects. Relative to this quantity the unique contribution is outstanding. The third line expresses the contributions made by all higher order effects which must be removed to assess the unique effect. This is the structural error. The fourth and last line quantifies the unexplainable variation or noise. Thus:

$$H(Y) = (\text{other contrib.}) + (\text{unique contrib.}) + (\text{struct. error}) + (\text{noise})$$

The quantities of higher-order effects and of lower-order effects also include the quantities of unique effects of larger and of smaller numbers of variables respectively which can be extracted from these quantities as well. To show the options of this quantitative breakdown offers, we make use of the lattice of all possible regression models of contributions by predictor variables on the criterion etc. This lattice resembles the lattice of all possible models (Figure 7 in Krippendorff, 1981). It can be obtained by applying the algorithm described by Krippendorff (1982) on the predictor variables and by modifying the resulting models as follows: (a) expand each component of the resulting models to cover the criterion variable as well, and (b) to each such model add a component containing the predictor variables as a single vector. This lattice is exemplified in Figure 1 using three predictor variables. The quantitative differences between these models along any one path through such a lattice from the top to the bottom add up to T(Z:ABC), the contributions made by all predictor variables A,B, and C.

unique effects	number of predictor variables 2	3	4	5	6
all present	ZAB	ZABC	ZABCD	ZABCDE	ZABCDEF
ABC present		ZABC	ZABC:ZABD:ZACD:ZBCD:ABCD	ZABC:ZABDE:ZACDE:ZBCDE:ABCDE	ZABC:ZABDEF:ZACDEF:ZBCDEF:ABCDEF
ABC removed		ZAB:ZAC:ZBC:ABC	ZABD:ZACD:ZBCD:ABCD	ZABDE:ZACDE:ZBCDE:ABCDE	ZABDEF:ZACDEF:ZBCDEF:ABCDEF
AB present	ZAB	ZAB:ZAC:ZBC:ABC	ZAB:ZACD:ZBCD:ABCD	ZAB:ZACDE:ZBCDE:ABCDE	ZAB:ZACDEF:ZBCDEF:ABCDEF
AB removed	ZA:ZB:AB	ZAC:ZBC:ABC	ZACD;ZBCD:ABCD	ZACDE:ZBCDE:ABCDE	ZACDEF;ZBCDEF:ABCDEF
A present	ZA:ZB:AB	ZA:ZBC:ABC	ZA:ZBCD:ABCD	ZA:ZBCDE:ABCDE	ZA:ZBCDEF:ABCDEF
A removed	ZA:AB	ZBC:ABC	ZBCD:ABCD	ZBCDE:ABCDE	ZBCDEF:ABCDEF
all removed	Z:AB	Z:ABC	Z:ABCD	Z:ABCDE	Z:ABCDEF

Structural Models of Unique Effects on a Criterion in Different Numbers of Predictor Variables

Table 1

EXPLORATORY APPROACHES TO REGRESSION

An exploratory approach proceeds stepwise. It starts by searching for the best single predictor out of the set of possible predictor variables. It then looks for the best pair of predictor variables, etc., until either all variables are exhausted or the additional contributions become insignificant and do not add sufficiently to the predictability of the criterion to warrent inclusion.

Cumulative Contributions

The simplest stepwise procedure is reflected in the accounting equation for cumulative contributions. This equation, and its terms can be evaluated algebraically:

$$H(Z) = \underbrace{T(Z:A) + T_A(Z:B)}_{T(Z:AB)} + T_{AB}(Z:C) + .. + H_{ABC..}(Z)$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{T(Z:ABC)}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{T(Z:ABC..)}$$

Each additional variable, say L, subtracts $T_{ABC..}(Z:L)$ from the unexplained entropy and adds it to the explained entropy. By adding variables in the order of the magnitude of $T_{AB..}(Z:L)$, each of the resulting sets of predictor variables is the one with the largest explanatory power.

Ordinal Contributions

With reference to particular models of regression, the addition of one predictor variable to the equation may have not one but a variety of different effects. For example, in the model of ordinal contributions, the addition of the variable L may reduce the noise by $T_{AB..}(Z:L)$ but it will also affect $T_1, T_2, \ldots T_r$ and add T_{r+1} to the accounting equation.

$$H(Z) = T_1 + T_2 + T_3 + \ldots + T_r + H_{ABC..}(Z)$$

$$H(Z) = T'_1 + T'_2 + T'_3 + \ldots + T_r + T_{r+1} + H_{ABC..}(Z)$$

which introduces the following quantitative changes:

$$T'_1 - T_1 + T'_2 - T_2 + .. + T'_r - T_r + T'_{r+1} = T_{ABC..}(Z:L)$$

The researcher wishing to explore his data with this model will have to weigh the ordinality of the preferred explanation thus optimizing not the quantity $T_{AB...}(Z:L)$ but, say the proportion:

$$\sum_{i=1}^{v} (T'_i - T_i) \;/\; T_{ABC..}(Z:L)$$

where v is the largest preferred ordinality of the explanation. The choice of the weight obviously influences the outcome.

Additive Contributions

In this model the new variable L will add T(Z:L) to T(Z:I) and reduce the structural error, both of which is what one wants. But it may also distract from the addivity among the predictor variables. The procedure we have been following is to minimize the difference:

$$T(A:B:C:\ldots) - [T_Z(A:B:C:\ldots) - T(ZA:ZB:ZC:\ldots:ABC..)]$$

without either quantity becomming statistically significant. With this aim in mind, it is possible to proceed by adding variables as in the preceeding approaches, however, this may not lead to optimal models of additive contributions. A less efficient but more satisfactory procedure is to start by evaluating the difference for all pairs of variables. Then among those for which both terms of the difference are not statistically significant select the pairs that makes the largest contribution to the prediction of Z. Repeat the procedure for three, four etc. predictor variables until either the dependencies among predictor variables become statistically significant or computational limits are reached.

Unique Contributions

The exploratory use of the model of unique contributions is implied in Figure 1. Proceeding from the top down the right lattice, the figure shows how each succeeding structural regression model excludes one effect on the criterion variable Z. Following a path that is guided by the smallest structural error means stepwise removing unique effects, contributing effects and prima face effects until the set of contributing predictor variables has shrunk to the set for which the structural error has reached acceptable limits. The resulting set of predictor variables is the set for which further reduction becomes no longer defendable. While this procedure encounters severe computational limits not present in the cumulative and the ordinal approaches, within these limits this is one of the most powerful procedures for the univariate multiple regression analysis of qualitative data.

MULTIVARIATE MULTIPLE REGRESSION

The regression models so far considered explain one variable in terms of several other variables and are properly termed univariate multiple regression models. We now consider briefly what is involved in the generalization of the regression idea to several sets of variables which are considered as explanations of each other. The problem such a regression analysis tackles is to compute a simplification of the interdependencies between several subsystems, each of which is characterized by a different set of variables. Instead of analyzing the quantity T(Z:ABC..), as in the univariate approach, we now designate subsystems S_1, S_2,... of mutually exclusive sets of variables from a larger system $S_1S_2\ldots$ containing all these variables and we seek a regression model that accounts for the interdependencies between these systems. This involves analyzing the quantity $T(S_1:S_2:\ldots)$. The algorithm for generating all possible multivariate multiple regression models proceeds as follows:

Let a model have two kinds of components K_1, K_2,... and S_1, S_2,... The components S_i represent a partition of the variables in the whole system into mutually exclusive sets each of which designates a different subsystem of potentially explanatory

variables. The components K_i also consist of variables of the system but are not so constrained.

Given a model $m = K_1:\ldots:K_i:K_j:..S_1:\ldots:S_r:..$

→ All $i = 1,2,\ldots$ components K_i of m

Let K_i have w variables $1', 1'', \ldots, 1^w$

(1) replace K_i by the string $(K_i\text{-}1'):(K_i\text{-}1''):\ldots:(K_i\text{-}1^w)$

(2) remove any $(K_i\text{-}1)$ of the string that is equal to or contained in any other K_i or S_r of m

(3) the remaining $(K_i\text{-}1)$s, K_is and S_rs constutute one of m's possible simplifications.

For a partition of eight variables into S_1=ABC, S_2=LM and S_3=XY variables, the lattice starts with K^o=ABCLMXY on the top and terminates with the model $S_1^o:S_2:S_3$=ABC:LM:XY at the bottom. The intermediate models contain components representing dependencies between the three sets of variables, e.g. ABLMXY, BLY, and between any two sets of variables, e.g. ABL, BCLM, LMXY, MY, AX, but no dependencies within either set, e.g. AB or X are excluded from such models. Proceeding from the top to the bottom of this lattice, each removal or decomposition of a component may cause structural losses. The path with the smallest losses is the path to the simplest multivariate multiple regression model of mutually explaining variables of different, here three sets.

We note that the lattice of possible univariate multiple regression models is that special case of the above in which S_1=Z and S_2=ABC... . The procedure for testing such models is known since Klir's (1976) work. The lattice of possible regression models and the information theoretical measures for evaluating these models are new though rooted in previously published work (e.g. Broekstra, 1979; Krippendorff, 1981). Although computational limits are serious, the approach to regression analysis of qualitative data here developed is straight foreward and simple, quite unlike comparable econometric models of regression of quantitative data.

REFERENCES

Ashby, W. Ross
1969 Two tables of Identities Governing Information Flows Within Large Systems. American Society for Cybernetics Communications 1, 2:3-8.

Broeksta, Gerret
1979 Probalitistic Constraint Analysis of Structure Identification: An Overview and Some Social Science Applications. Pp. 305-334 in B. Zeigler et al. (Ed.) Methodology of Systems Modelling and Simulation Amsterdam: North Holland.

Klir, George J.
1976 Identification of Generative Structures in Empirical Data. International Journal of General Systems 3, 2: 89-104.

Krippendorff, Klaus
1980 Q: An Interpretation of the Information Theoretical Q-Measure. Proceedings, Fifth International Meeting on Cybernetics, and Systems Research. Vienna.

Krippendorff, Klaus
1981 An Algorithm for Identifying Structural Models of Multi-variate Data. International Journal of General Systems 7: 63-79.

Krippendorff, Klaus
1982 A Proposal for an Algorithm for Generating Loopless or Recursive Models of Multi-Variate Data. Proceedings, 26th Annual Conference of the Society for General Systems Research. Washington, DC.

AUTHOR Klaus Krippendorff
The Annenberg School of Communications
University of Pennsylvania, C5
Philadelphia PA 19104