Caputo, Heinicke, Pigeot:

A graphical chain model derived from a model selection strategy for the sociologists graduates study

Sonderforschungsbereich 386, Paper 73 (1997)

A graphical chain model derived from a model selection strategy for the sociologists graduates study

A. Caputo [1] A. Heinicke, I. Pigeot
Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany

Abstract

This paper objects to the arising problems due to fitting graphical chain models to multidimensional data sets. This multivariate statistical tool is used to cope with complex research questions concerning not only direct, but also indirect associations between the variables of interest. Due to this high complexity sensible strategies for fitting such models are required. Here, a data–driven selection strategy is discussed. Its application is illustrated for an empirical data example in detail.

1 Introduction

In empirical studies, typically a large number of different variables is collected for each individual. Often the researcher has a certain idea concerning the underlying association structure which implies direct as well as indirect influences on the main responses. Such influences cannot be captured by conventional regression models. Here, graphical chain models are the adequate tool since they allow for intermediate variables additional to the pure explanatories and pure responses. Such models can be regarded as a generalization of path models introduced by Wright (1921, 1923, 1934) which only treat continuous variables and univariate responses. The inclusion of discrete and continuous variables at the same time as well as the allowance for mixed multivariate responses is the result of joint work by Lauritzen and Wermuth (1989, 1990) who also started the discussion of the properties and handling of this sophisticated statistical device. Graphical

[1] Financial support by the SFB 386, Deutsche Forschungsgesellschaft is gratefully acknowledged.

chain models focus on the reduction of complexity by the use of the concept of conditional independence (Dawid, 1979).

It should be mentioned that the advantage of graphical models is twofold. Additional to the enhanced understanding of the structure among the variables due to the use of intermediates, each model has a corresponding graph of which conditional independencies can directly be read off. In such graphs variables are represented by nodes and associations by edges. Thus, conditional independence can be identified by a missing edge between the corresponding pair of variables (Lauritzen, 1996). Occurring edges should, however, be carefully interpreted since they represent no further specified associations as for instance direct influences or interactions (Cox and Wermuth, 1993). Each missing edge has a clearly defined meaning, namely a particular conditional independence. In this respect graphical models can be looked at as being binary, i.e. only allowing for an edge or a missing edge. Since they focus primarily on conditional independencies, i.e. on missing edges, the decision to exclude an edge based on a statistical analysis should be cautiously made.

Coming back to our starting point, that is to the researcher who formulates his/her research question concerning the assumed associations. By doing so he/she defines all possible associations which can occur based on subject–matter considerations. This postulated chain is then to be checked based on the collected data using appropriate statistical methods. It should, however, be noticed that such a statistical analysis cannot be regarded as a proof of causality even if the postulated association structure is confirmed by the data, since other structures are likely to be supported as well.

The main task of such a statistical analysis relates to the check of the postulated independencies which implies the selection of an adequate model. Therefore, an appropriate selection strategy is called for. Here, different competing philosophies can be found in the literature such as for instance backward and forward selections. The choice among these approaches is not data–driven, but depends on the analyst where different strategies typically do not lead to the same results. Even for conventional regression models it is not easy to decide on the selection strategy. For graphical chain model it is still harder. An extensive discussion on this topic can be found in Cox and Wermuth (1996).

In this paper, we investigate a particular selection strategy which is data–driven and which can be heuristically motivated. Here, we draw attention to the problems being related to such a strategy regarding a possible loss of the theoretic properties of the underlying graphical chain model. For this purpose, we first introduce the basic notions of such models. In Section 3 we present a concrete data set, the so–called graduates study which merely serves as an illustrative example for the selection strategy discussed in detail in Section 4. We abstain from giving any substantial interpretations and the interested reader is referred to Pigeot et al. (1997). Open questions are addressed in Section 5.

2 Graphical chain models

2.1 Preliminaries

Graphical chain models are those probability models for multivariate random observations whose independence structure is captured by a graph. This graph, a mathematical object, is also called conditional independence graph. Neither the idea of graphs is new nor that of conditional independence, but these concepts joined together by an appropriate statistical distribution is a rather new field of research. This underlying distributional assumption, is characterized by the facts that local and global Markovian properties are equivalent and that they can be read off the corresponding graph (Lauritzen and Wermuth, 1989; Lauritzen, 1996). The distribution typically considered is the Conditional Gaussian distribution (CG–distribution), where the continuous variables are multivariate normal given the discrete. CG-distributions belong to the exponential family. They are not closed under marginalization, but are closed under conditioning.

For convenience, let us briefly introduce some basic notations. Suppose we have p discrete variables and q continuous ones, Δ and Γ symbolize the sets, respectively. The corresponding random variables are given as $X_V = (I, Y)$ and a typical observation as $x_V = (i, y)$. Here, i is a p-tupel containing the values of the discrete variables, and y is a real vector of length q. We write $\mathcal{I}$ for the set of all possible i. Similarly X_d is used to denote the projections of a point x_V onto the coordinates $d \in V$. There are several possibilities to parameterize a

CG–distribution. Here, we follow the canonical approach given in Lauritzen and Wermuth (1989):

$$\begin{aligned} f(x) &= f(i,y) \\ &= \exp\{g(x_\Delta) + h(x_\Delta)^T x_\Gamma - \frac{1}{2} x_\Gamma^T K(x_\Delta) x_\Gamma\} \\ &= \exp\{g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y\}, \end{aligned} \tag{1}$$

where g is a real-valued function of i, h is a q–dimensional vector-valued function of i, K is a $q \times q$–matrix–valued function of i and v^T is the transpose of the vector v. The canonical parameters are $(g(i), K(i), h(i))$. Furthermore, CG–distributions have an expansion with interaction parameters (Lauritzen and Wermuth, 1989):

$$g(i) = \sum_{A \subseteq \Delta} \lambda_A(i), \quad h(i) = \sum_{A \subseteq \Delta} \eta_A(i), \quad K(i) = \sum_{A \subseteq \Delta} \psi_A(i),$$

where $A \in V$ and λ, η and ψ denote interactions in the following way:

- $\lambda_\emptyset$: log normalizing constant
- $\lambda_d, d \neq \emptyset$: pure discrete interactions among variables in d
- $\lambda_d, |d| = 1$: main effects of the discrete variables
- $\eta_\emptyset$: main effects of the continuous variables
- $\eta_d, d \neq \emptyset$: mixed linear interactions between a continuous variable and one of d
- ψ_d: mixed quadratic interaction matrices
- $\psi_\emptyset(i)$: pure quadratic interactions, the elements do not depend on i
- $\psi_d(i), d \neq \emptyset$: mixed quadratic interactions between variables in d and pairs of continuous variables.

A graph $\mathcal{G} = (V, E)$ now consists of a set of vertices V representing the variables and a set of edges E representing the associations between pairs of variables. E is a set of ordered pairs (A, B), A, $B \in V$. A chain graph is based on a partition of V into disjoint subsets: $V = V_1 \cup V_2 \cup \ldots \cup V_T$. The subsets are called chain components. Edges within chain components are undirected and edges between chain components are arrows pointing from components with lower index numbers to those with higher index.

We assume that there exists a joint distribution of the set of variables corresponding to V with a strictly positive density function f_V. The density function factorizes into a product of several conditional densities and one marginal density according to $\mathcal{G}$ as

$$f_V = f_{V_T|V_{T-1}\cdots V_1} \cdot f_{V_{T-1}|V_{T-2}\cdots V_1} \cdots f_{V_2|V_1} \cdot f_{V_1}. \tag{2}$$

In general the conditional densities are assumed to be CG–regressions. A CG–regression is a conditional CG–distribution. To clarify this relationship, let us consider two random vectors $X_V = (X_\Delta, X_\Gamma)^T$ and $X_{V^\star} = (X_{\Delta^\star}, X_{\Gamma^\star})^T$, the former having realizations $(i, y)^T$ and the latter $(j, z)^T$. The distribution of X_V can be looked at as being the conditional distribution of the joint distribution of the random vector $(X_V, X_{V^\star})^T$ (Lauritzen and Wermuth, 1989). Further, to each CG-regression exists a joint CG-distribution of which the CG-regression is the conditional distribution. The CG-regression does not determine the joint CG-distribution uniquely unless the parameters of the marginal distribution of $X_{V^\star}$ are known.

Let us finally mention that the factorization (2) reflects the joint density of the variables due to the underlying dependence chain. It is possible to specify a distributional assumption for the joint density by distributional assumptions of the single factors. The joint distribution of X_V is called a recursive multivariate CG-regression, if all conditional distributions of that factorization are CG-regressions (Lauritzen and Wermuth, 1989).

2.2 How to fit a graphical model to data?

As already mentioned in a first step the researcher postulates a dependence chain based on subject–matter considerations. Due to this chain the likelihood function factorizes in a product of conditional likelihood functions (cf. equation (2)). Because of its construction the parameters in each of these conditional distributions vary freely and independently of those in other conditional distributions, i.e. we have a cut in the sense of Barndorff–Nielsen (1978, p.50). Therefore, the likelihood function can be maximized by separate maximization of each factor.

The formulation of CG-distributions with interaction terms displays, that quite a large number of parameters has to be estimated for which a large number of observations is required. To reduce the number of parameters to be estimated it is not untypical to set a few parameters to zero a priori, if subjet–matter considerations recommends to do so. Since interaction terms of third and higher order can often hardly be interpreted they are usually considered as being zero. This reduces enormously the number of parameters to be estimated. Although such a proceeding seems to be plausible its implications on interaction parameters of the joint distribution as for instance on the parameters of a CG–distribution (cf. equation 1) are still unknown.

Unfortunately so far no algorithm for fitting CG–regressions as a whole has been developed and research on that topic proved to be a major task. In the pure cases, that is all variables are either continuous or discrete, the fitting does not constitute a real problem since multivariate linear regression analysis or multivariate logistic regression techniques can be applied for each single conditional density separately. A difficulty, however, arises in the interesting case where mixed responses occur. A possible solution to this problem has been suggested by Cox and Wermuth (1996), where in their data–driven strategy each conditional density of the factorization is described by a system of multiple univariate regressions. The kind of regression used depends on the measurement scale of the involved univariate response. A model defined in terms of univariate regressions has the advantage that each parameter in the system has a precise meaning as regression coefficient, if we deal with exclusively main effect regressions. If interactions terms are taken into the equation, it cannot be distinguished between the regression coefficient of the main effects and the one of the interaction terms by looking at the corresponding edge. Another problem related to the above strategy of fitting multiple univariate regressions results from neglecting the multivariate structure of the data. This may lead to a loss of efficiency, if multivariate mixed response with more than one discrete variable are modeled. Nevertheless, the strategy is tractable, close to the data, and it is justified from a heuristical point of view, but the estimation procedure does not ensure the validity of the equivalence of the Markovian properties for the whole graph.

Another nice property of fitting univariate regressions concerns the interpretation of missing edges. Consider for example the linear model, which relates the observed values by a linear equation. It is easily shown that the problem linked to the task of variable selection is identical to that of determining which edges joining the response to the covariates should be included in the graph (e.g. Whittaker, 1995, p.323, proof of proposition 10.5.1). In the conditional normal linear framework, the hypothesis that the i–th regression coefficient is zero is equivalent to the hypothesis, that Y and X_i are conditionally independent given the remaining variables of the model. This means that the graph corresponding to the regression results is easily established.

For illustration of the problems related to fitting a graphical model to empirical data, we now consider an example of a study dealing with a sociological research question.

3 An example: the graduates study

The graduates study has been conducted at the University of Munich in 1995. In total 465 questionnaires were sent to sociologists having graduated at the University of Munich between 1983–1994, 102 were not deliverable and 89 remained unanswered leading to a non–response rate of 24.52 %. Restricting the analysis to complete cases leaves us with 182 observations.

In this section we briefly describe the data set and the variables of interest. For details see Pigeot et al. (1997). We establish a dependence chain of the variables as described Section 2. That is, we divide the set of variables into purely explanatory variables, several groups of intermediates on different footing and pure responses. The partition is based on subject matter considerations only and thus it is independent of the data.

3.1 The variables

We take interest in later professional success of sociologists by which we mean satisfaction coming from work (*satisfaction*), adequacy (*adequacy*) of the kind of job they have and how much they earn (*earnings*). We think that the type of work contract (*contract*), coded as two dummies, i.e. if someone is permanently/temporarily employed or has an own business, will have an impact on professional success. The duration (*duration*) of the current job and, if the sociologists managed to settle down in the work field they wanted to (*match*=1, 0 otherwise) once they have left university is considered to be influential. Further, the sociologists' attitude (*rank*) toward the importance of work and family life as well as having at least one child (*child*=1, 0 otherwise) at the end of university is taken into account. We expect that having high values of the variable assessment (*assessment*), which measures the skill of some not–university–related abilities like being able to work in a team etc., and the amount of experience (*experience*) the sociologists gained before they have taken their current job will effect the status of success. Being dedicated (*dedication*=1, 0 otherwise) having picked up sociology out of interest (*interest*=1, 0 otherwise) and having a particular professional aim (*aim*=1, 0 otherwise) in mind relate to the way a student studies. The depth of computer–knowledge (*computer*) seems to be important and the amount of how much the university contributes to the development of non–university–related, but nevertheless indispensable skills, is captured by the variable utility (*utility*=1, if university is considered to be helpful, 0 otherwise). If a sociologist lived together with a partner (having one *partner*=1, 0 otherwise) at the end of university captures the mutual influence between private life and university demands. Also of interest is how much time the sociologists needed to get their diploma (*uni*) and the mark (*mark*) they achieved in their final exams. The field (*branch*) in which the just graduated wanted to find a job is assumed to have an impact. This variable has four categories coded as three dummies. Finally, the age (*age*) of the students at the beginning of their studies and their sex (*sex*=1: female) are supposed to effect the career of sociologists.

3.2 The chain

The 21 variables are ordered into the dependence chain given in Figure 1 due to chronological and subjet–matter considerations. This dependence chain defines all possible associations among the variables and all further statistical analysis is based on it.

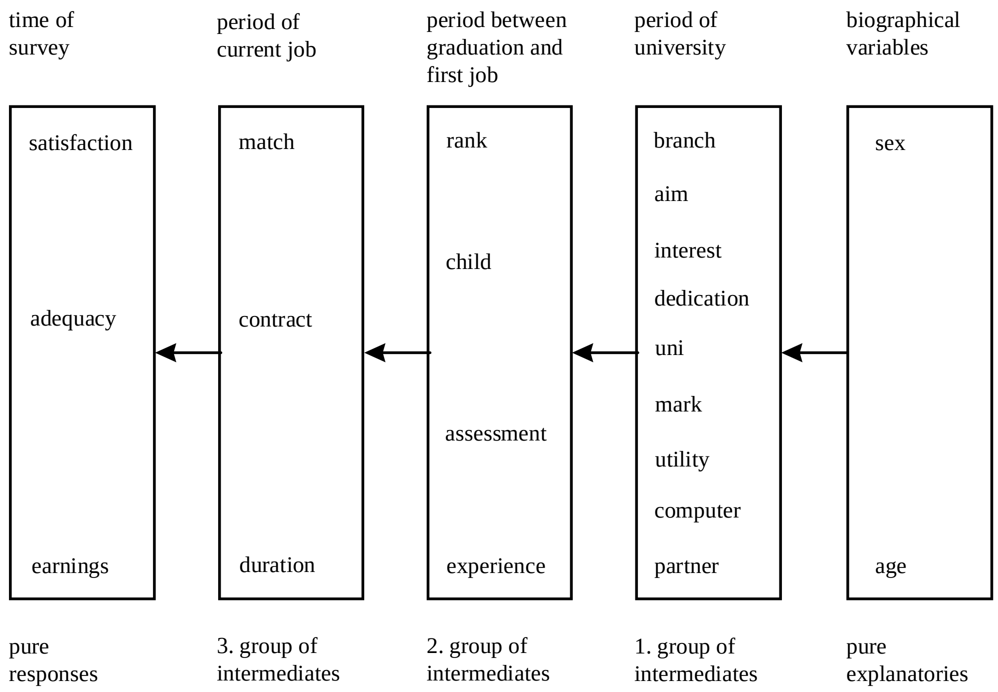

Figure 1: Association structure due to chronological and subject matter considerations

The biographical variables *age* and *sex* are placed to the very right in the same box for they are the pure explanatory ones. The first group of intermediates consists of variables covering the period the sociologists spent at university, the second the period between they graduated and their first job. The third and fourth group include variables relating to their current and former job. The pure responses, placed to the very left, can be influenced by all other variables of the chain. In general a particular variable of a particular box can be influenced by all other variables to its right as well as by all other variables of its box. In this respect it is considered to be a response variable. At the same time this variable is also viewed as potentially explanatory for all other variables to its left and its own box. In this respect it is considered to be an explanatory variable. The boxes containing the intermediates are flanked by the two extremes, to their left the box of pure responses, to their right the box of purely explanatory variables.

4 Selection strategy

This section centers on the description of one possible approach for fitting a graphical chain model to data, which is mainly based on a data–driven selection strategy (Cox and Wermuth, 1996). The analysis of the graduates study is used for illustration throughout this section.

A good variable selection strategy reduces the complexity of a model as much as possible without oversimplification of the structure among the variables. Virtually, we do not rely on automatic variable selection strategies for they force an arbitrary choice between models that fit almost just as good.

For each conditional density of the factorization we fit several multiple univariate regressions, i.e. we model a multivariate response by a system of univariate regressions, which describes the structure among the involved variables as a whole. Each univariate response is regressed on all variables to its right as well as the remaining variables of its own block.

To make this clear, say *earnings* serves as response, then all other variables are explanatory including *satisfaction* and *adequacy* as well. The three regressions as a whole describe the structure among the variables. For the third group of intermediates there are again three univariate regressions to be considered and so on. We now explain the selection strategy for one univariate regression. A flow chart of the procedure is shown in Figure 2.

The procedure for each univariate regression is split into five phases.

First Phase

The first phase deals with a screening suggested by Cox and Wermuth (1994) in order to find evidence that some interaction terms and nonlinear terms have to be considered in certain multiple univariate regressions. The idea of the screening is similar to that of normal probability plots. For interactions, i.e. cross–product terms, we examine the t–statistics from trivariate regressions such as regressing Y on X_i, X_j and $X_i \times X_j$. Due to a sufficiently large number of cases the t–statistics, i.e. the estimates divided by their standard error, approximately follow a standard normal distribution, if interactions are absent. A $|t|$–value of 4 corresponds to the 99%–quantile of the Gaussian distribution,

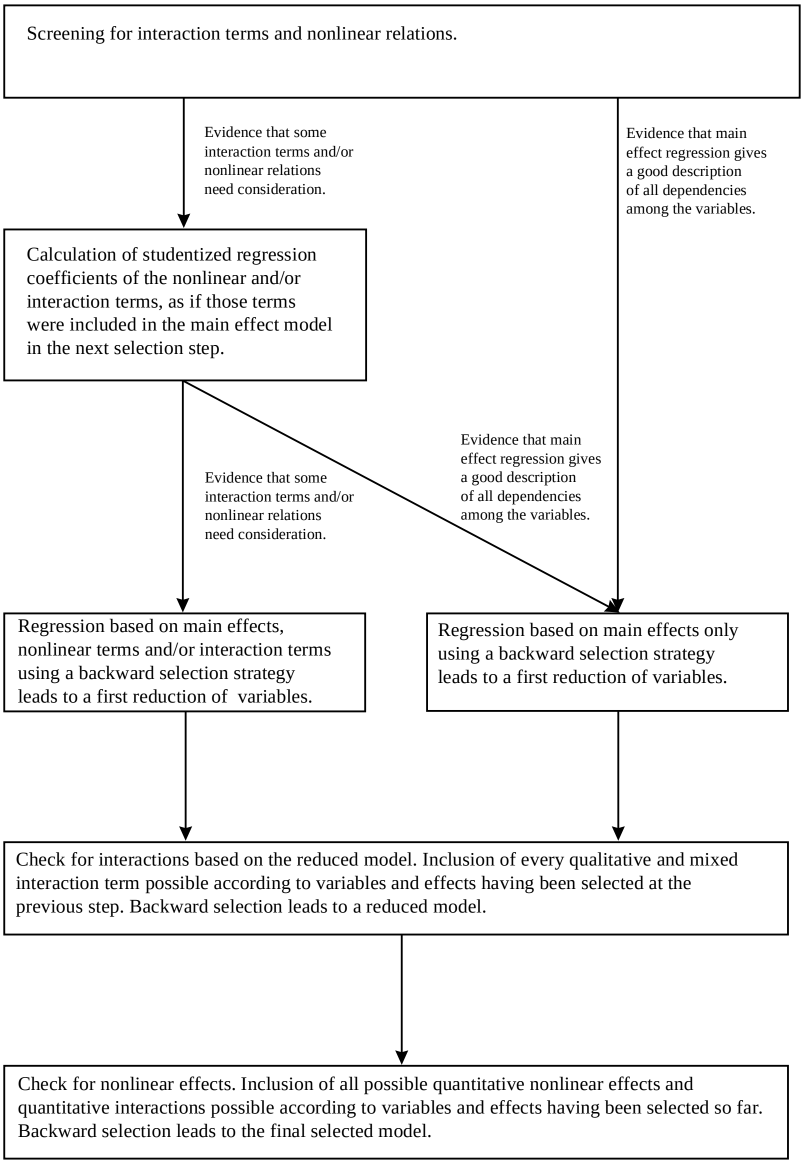
Screening for interaction terms and nonlinear relations.
Evidence that some interaction terms and/or nonlinear relations need consideration.
Evidence that main effect regression gives a good description of all dependencies among the variables.
Calculation of studentized regression coefficients of the nonlinear and/or interaction terms, as if those terms were included in the main effect model in the next selection step.
Evidence that some interaction terms and/or nonlinear relations need consideration.
Evidence that main effect regression gives a good description of all dependencies among the variables.
Regression based on main effects, nonlinear terms and/or interaction terms using a backward selection strategy leads to a first reduction of variables.
Regression based on main effects only using a backward selection strategy leads to a first reduction of variables.
Check for interactions based on the reduced model. Inclusion of every qualitative and mixed interaction term possible according to variables and effects having been selected at the previous step. Backward selection leads to a reduced model.
Check for nonlinear effects. Inclusion of all possible quantitative nonlinear effects and quantitative interactions possible according to variables and effects having been selected so far. Backward selection leads to the final selected model.

Figure 2: Flow chart of the selection strategy applied to each multiple univariate regression

the value 2 to the 97.5%–quantile. For model inclusion we apply the stricter criterion to nonlinearities and cross–product terms to avoid the inclusion of interaction terms and nonlinearities that are only artifacts. The softer one serves as inclusion criterion for main effects.

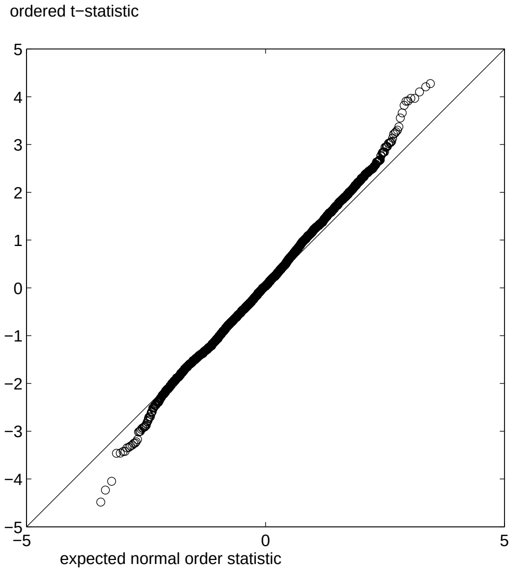

Figure 3: Plot for cross–product terms, t from trivariate regressions such as regressing Y on X, V and $X \times V$

We plot the ordered t-statistics against the expected values of the standard normal distribution. If the assumption is fulfilled that interactions are absent, the points spread along the diagonal, whereas strong divergencies imply that the assumption is violated. We proceed likewise in checking for nonlinearities.

The results of the screening are shown in Figures 3 and 4. The screening of the cross–product terms gives evidence that a few of them should be considered in the involved univariate regression. It has to be mentioned that the plots include t–statistics that do not correspond to a relevant univariate model in the sense of the dependence chain. Thus, the situation is in fact better than the plots imply. The screening related to nonlinear terms shows that all t–statistics are absolutely below 4, but the points do not spread along the diagonal nicely.

Second Phase

The second phase depends on the result of the screenings. If some terms need consideration, we calculate studentized regression coefficients of these terms, as if they were included in the main effect model in the next selection step.

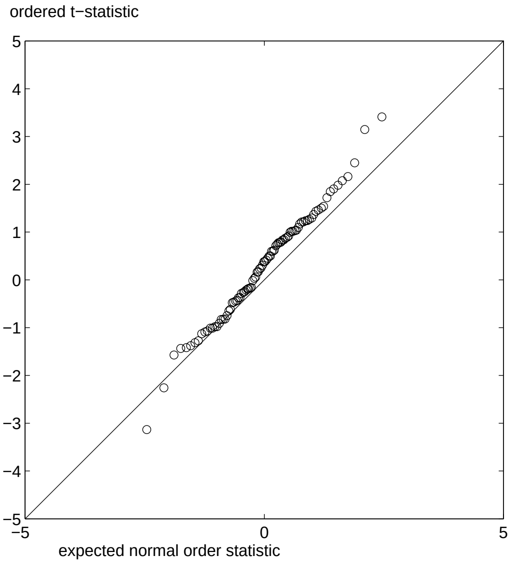

Figure 4: Plot for squared terms, bivariate regressions such as Y on X and X^2

If the corresponding value is absolutely larger than 4, the term is taken into the regression equation, if none is absolutely larger than 4 we go on to phase three.

Third Phase

If the screening does not show a striking nonlinearity and/or interaction we start with a main effect regression. If the second phase leads to the inclusion of interactions and/or nonlinearities we start with the expanded model. In other words we restrict parameters related to not explicitly included interactions and nonlinearities to zero. Now, a backward selection strategy is used and in each step we exclude the variable which has the smallest corresponding absolute t–value. To show this, we discuss the selection procedure of the multiple univariate linear regression of the continuous variable *adequacy* on all remaining variables. The screening does not give evidence that certain nonlinearities or cross–product terms need consideration. Table 1 shows the stepwise exclusion of variables. In the first step the variable *dedication* is excluded. Again, the parameters of the remaining variables are estimated and now the variable *utility* is taken out of the equation, since it has the smallest $|t|$–value of 0.01. We go on until Step 6. In Step 6 the smallest absolute t–value belongs to a dummy variable of the polytomous variable *contract*. In this case we exclude the variable, i.e. all dummies related to it, if the increase of r^2 is remarkable, i.e. if the value of the F_{change}–statistic (Rao, 1995, p.49) is "significant" to the 0.05 level with

$$F_{change} = \frac{RSS_{X_{red}} - RSS_{X_{full}} \setminus (K-p)}{RSS_{X_{full}} \setminus (T-K)},$$

where RSS denotes the residual sum of squares, X_{full} the full model, X_{red} the nested model, K the number of all regressors, p the number of regressors taken out of the equation and T the number of cases. Under the null hypothesis, that the nested model is valid the statistic F_{change} follows a $F_{K-p,T-K}$ distribution.

In this example, the increase of the r^2-value is not significant, which is also reported in Table 2. Therefore, the variable *contract* is excluded in the next selection step. After 16 steps we have our first reduced model since all remaining variables have a larger absolute t–value than 2 and the polytomous variable contributes as a whole to a significant increase in r^2.

A similar criterion is applied, if we deal with polytomous variables in multiple univariate logistic regressions. We use the term logistic regression models for models with a mixture of continuous and categorical explanatory variables. Here, the selection is based on likelihood–ratio tests. Let l_{red} denote the maximized loglikelihood of the nested model and l_{full} the loglikelihood of the full model. Under the null hypothesis that the nested model is valid, the corresponding likelihood–ratio statistic

$$\lambda = -2\{l_{red} - l_{full}\},$$

follows a χ^2_d distribution, where d is the difference of the numbers of parameters in both models.

Fourth Phase

This reduced model is the starting point of a second check for interactions and nonlinearities. First all qualitative and mixed interaction terms, where the latter means an interaction between a continuous and a discrete variable, are taken into the equation. Again a backward selection has to be carried out until all interaction terms have a larger absolute t–value of 4 and all main effects

selection– step no.	1	2	3	4	5	6	7
explanatory variable	t–values						
match	2.49	2.52	2.53	2.56	2.57	2.58	3.31
assessment	2.40	2.42	2.43	2.45	2.47	2.47	2.61
rank	2.04	2.07	2.08	2.14	2.15	2.15	2.49
aim	2.73	2.74	2.75	2.79	2.81	2.86	2.97
satisfaction2	0.38	–1.32	0.39	0.39	0.40	0.39	0.66
satisfaction3	2.33	0.38	2.35	2.36	2.39	2.39	2.51
satisfaction4	2.31	2.34	2.33	2.35	2.39	2.39	2.43
branch0	–2.65	–2.68	–2.70	–2.71	–2.74	–2.74	–3.31
branch2	–1.64	–1.65	–1.67	–1.67	–1.68	–1.67	–1.93
branch3	–1.32	–1.32	–1.33	–1.33	–1.34	–1.34	–1.21
age	0.75	0.76	0.77	0.77	0.77	0.77	0.64
child	0.86	0.86	0.87	0.87	0.86	0.85	0.85
mark	–0.47	–0.49	–0.49	–0.49	–0.51	–0.53	–0.81
computer	0.99	1.00	1.03	1.03	1.02	1.04	0.99
partner	0.70	0.70	0.70	0.71	0.73	0.75	0.76
earnings	–0.67	–0.67	–0.67	–0.70	–0.68	–0.70	–0.77
duration	1.09	1.09	1.10	1.11	1.12	1.12	0.74
experience	0.43	0.44	0.45	0.45	0.44	0.43	0.37
contract2	1.95	1.95	1.96	1.97	1.98	2.01	–
contract3	0.20	0.20	0.21	0.20	0.22	0.20	–
uni	–0.22	–0.22	–0.22	–0.22	–0.21	–	–
interest	0.17	0.18	0.18	0.18	–	–	–
sex	0.05	0.05	0.05	–	–	–	–
utility	–0.01	–0.01	–	–	–	–	–
dedication	–0.001	–	–	–	–	–	–
r^2	0.40	0.40	0.40	0.40	0.39	0.39	0.37
r^2_{adj}	0.31	0.30	0.31	0.31	0.32	0.32	0.31

Table 1: Changes of the t–statistics, r^2 and r^2_{adj} during the backward selection of the regression of *adequacy* on all remaining variables

selections								
step no.	8	9	10	11	12	14	15	16
explanatory variable	t–values							
match	3.30	3.36	3.39	3.46	3.75	3.69	3.74	3.68
assessment	2.62	2.73	2.72	2.88	2.94	2.92	3.03	3.00
rank	2.53	2.50	2.55	2.48	2.46	2.57	2.66	2.59
aim	2.97	2.94	2.92	3.09	3.21	3.32	3.28	3.25
satisfaction2	0.67	0.74	0.73	0.73	0.75	0.64	0.64	0.63
satisfaction3	2.56	2.68	2.64	2.72	2.71	2.60	2.70	2.59
satisfaction4	2.44	2.61	2.57	2.59	2.66	2.59	2.68	2.58
branch0	–3.38	0.83	–3.75	–3.73	–4.22	–4.98	–4.91	–4.87
branch2	–1.97	–3.55	–2.04	–2.12	–2.43	–2.60	–2.58	–2.48
branch3	–1.26	–2.01	–1.31	–1.42	–1.56	–1.80	–1.74	–1.71
age	0.69	0.83	0.83	0.92	0.97	0.85	1.38	–
child	0.87	0.78	0.77	0.96	0.93	0.89	–	–
mark	–0.83	–0.79	–0.76	–0.87	–0.85	–	–	–
computer	0.92	0.79	0.80	0.75	–	–	–	–
partner	0.74	0.79	0.76	–	–	–	–	–
earnings	–0.68	–0.50	–	–	–	–	–	–
duration	0.64	–	–	–	–	–	–	–
experience	–	–	–	–	–	–	–	–
contract3	–	–	–	–	–	–	–	–
contract3	–	–	–	–	–	–	–	–
uni	–	–	–	–	–	–	–	–
interest	–	–	–	–	–	–	–	–
sex	–	–	–	–	–	–	–	–
utility	–	–	–	–	–	–	–	–
dedication	–	–	–	–	–	–	–	–
r^2	0.38	0.38	0.38	0.37	0.37	0.37	0.37	0.36
r^2_{adj}	0.31	0.32	0.32	0.32	0.32	0.32	0.33	0.32

Table 1: continued

explanatory variable	F_{change}						
	6	8	9	10	11	12	13
$contract^{**}$	2.10	–	–	–	–	–	–
$branch^{*}$	–	–	–	–	–	–	8.20
$satisfaction^{*}$	–	4.82	5.18	5.18	5.01	5.58	5.15

Table 2: F_{change}–statistics of the polytomous variables *contract, branch*, and *satisfaction* during the selection of the regression *adequacy* on remaining variables. $F^{**}_{2,\infty} \approx 3.00$, $F^{*}_{3,\infty} \approx 2.60$ (95%–quantile)

larger than 2. In this example no qualitative or mixed interaction remain in the equation.

Fifth Phase

Again the reduced model developed in the foregoing phase serves as the starting model for the next phase. Nonlinear terms and quantitative interaction terms are included and a backward selection strategy is carried out leading to the final model. The translation in a representation as a graph is straightforward. Between each variable pair having a nonvanishing regression coefficient an edge is drawn. In our example, no nonlinearities are selected.

Figure 5 and 6 illustrates how the three univariate regressions of the pure responses (see Table 3) are summarized in a partial graph. The former figure shows two graphs having the same conditional independence statements. We favor, due to subject–matter reflections, the graph on the left hand side.

The combined chain model is composed of separate models for each block of responses given their covariates. We refer to Pigeot et al. (1997) for further details about extracting relevant paths and substantive results.

5 Final remarks

Summarizing our results, we address open questions and the advantages and disadvantages related to the used selection strategy. One advantage certainly is that in each univariate regression the number of parameters is comparably small,

response	*explanatories*						
adequacy	*match*	*assessment*	*rank*	*aim*	*branch0*	*branch2*	*branch3*
$\hat{\beta}$	2.83	0.31	0.19	3.17	−4.77	−2.50	−2.35
$\hat{\sigma}_{\hat{\beta}}$	0.77	0.10	0.08	0.98	0.98	1.01	3.28
	satisfact2	*satisfact3*	*satisfact4*	const			
$\hat{\beta}$	1.16	4.24	4.57	−9.69			
$\hat{\sigma}_{\hat{\beta}}$	1.84	1.64	1.77	3.28			
earnings	*duration*	*experience*	*partner*	*sex*	*branch0*	*branch2*	*branch3*
$\hat{\beta}$	1.77	1.85	2.85	−3.56	5.98	0.04	1.57
$\hat{\sigma}_{\hat{\beta}}$	0.24	0.26	1.32	1.27	1.90	1.80	2.34
	contract2	*contract3*	*adequacy*	*interact2*	*interact3*	*const*	
$\hat{\beta}$	−0.93	10.62	0.40	−0.13	−1.68	20.05	
$\hat{\sigma}_{\hat{\beta}}$	3.63	3.63	0.16	0.28	0.30	2.76	
satisfact	*adequacy*	*age*	*sex*	*const1*	*const2*	*const3*	
$\hat{\beta}$	−0.12	0.10	0.92	−5.01	−3.31	−0.12	
$\hat{\sigma}_{\hat{\beta}}$	0.03	0.04	0.31	1.11	1.05	1.02	

Table 3: Estimated regression coefficients $\hat{\beta}$ and according standard errors $\hat{\sigma}_{\hat{\beta}}$ of the selected models

and thus the selection strategy used here can be applied to data sets of moderate size. Gained experience in the handling of univariate regression models and a well developed theory makes it easy to fit these models. Additionally diagnostics concerning the fit of these models are available. For CG–regressions no such diagnostics are known.

Unquestionable, the large number of univariate regressions needed to fit the model makes evident, that our approach is meant to be exploratory so far as the final model results from repeated significant testing without any control of the multiplicity effect. At the moment no multiple test procedure is at hand to adjust the significance level in a sensible way. Thus, the obtained results cannot be regarded as significant in a strong sense.

As already mentioned one further possible disadvantage of our selection strategy is that it may not be efficient, since information is neglected. Additionally, it is still unknown to what extent the equivalence of the Markovian properties are impaired.

Despite all inconsistencies, in our opinion the above selection procedure places a sensible tool and until now no serious alternative does exist.

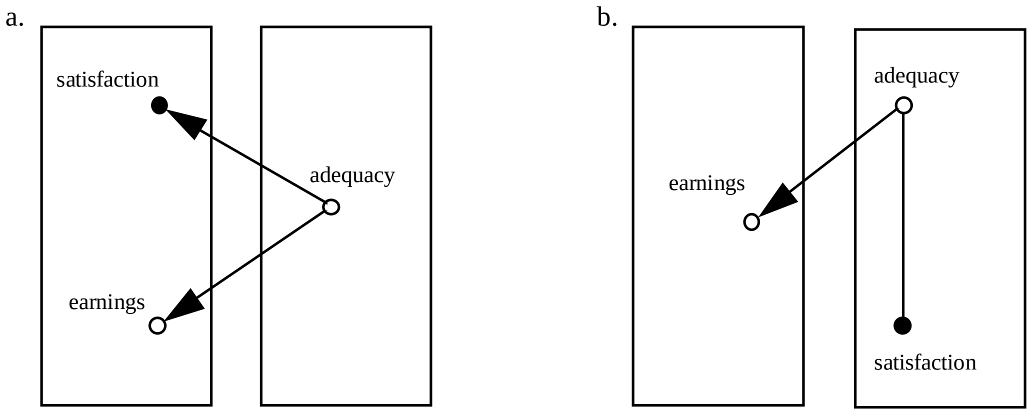

Figure 5: Two equivalent patterns of independence among the three pure response variables

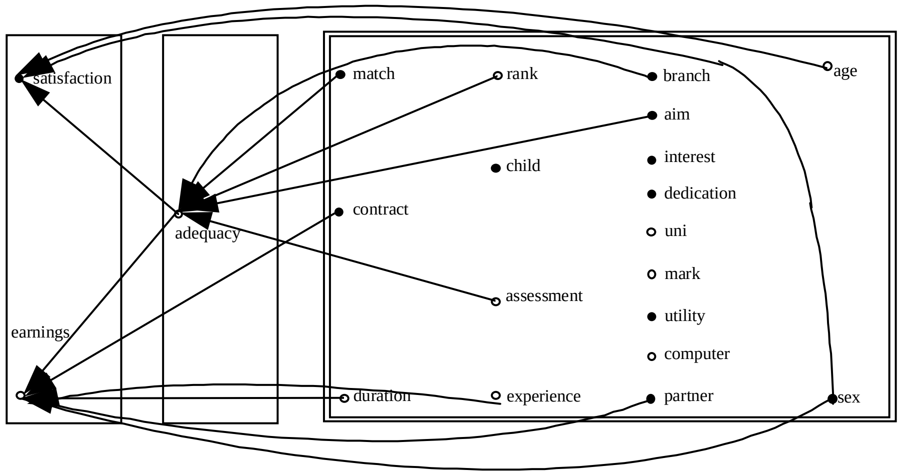

Figure 6: Graphical chain model of the three pure responses. The double box means that so far no statements about the associations among the included variables have been made

References

Barndorff–Nielsen, O. E. (1978). *Information and exponential families in statistical theory.* John Wiley and Sons, New York.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (+ discussion). *Statistical Science* **8**, 204–283.

Cox, D. R. and Wermuth, N. (1994). Tests of linearity, multivariate normality and the adequacy of linear scores. *Applied Statistics* **43**, 347–355.

Cox, D. R. and Wermuth, N. (1996). *Multivariate dependencies – models, analysis and interpretation.* Chapman and Hall, London.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society* **B 41**, 1–31.

Lauritzen, S. L. (1996). *Graphical models.* Oxford University Press, Oxford.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics* **17**, 31–57.

Pigeot, I., Heinicke, A., Caputo, A. and Brüderl, J. (1997). The professional career of sociologists: a graphical chain model reflecting early influences and associations. *SFB386 – Discussion Paper* **74**, University of Munich.

Rao, C. R. (1995). *Linear models. Least squares and alternatives.* Springer–Verlag, New York.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence, graphs and graphical chain models. *Journal of the Royal Statistical Society* **B 52**, 21–50.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics* John Wiley and Sons, Chichester.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.

Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics* **8**, 239–255.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics* **5**, 161–215.