In: Population Health Research, 181-203. K. Dean (ed).
London: Sage

Association Structures with Few Variables: Characteristics and Examples

Nanny Wermuth

Introduction

The usefulness of any graphical representation depends on the ease with which its implications can be deduced and on whether it has an unambiguous interpretation or not.

Graphical chain representations were suggested (Lauritzen and Wermuth, 1989) to represent complex association structures among variables which may be qualitative or quantitative. The word 'association' is used broadly to include both symmetric associations for variables treated on an equal footing and directed relations concerned with the dependence of a response on explanatory variables, sometimes called influences. Symmetric associations occur not only when there are no response variables at all, but also when some variables are joint responses or joint influences, or when they are joint intermediate variables in the sense of being responses to one set of variables and influences on another.

Figure 9.1 shows as an example one possible graphical chain representation for six variables. Variable A is a direct response to variables B, X and C and an indirect response to D, Y; variables B and X are joint intermediate variables; and C, Y and D are regarded only as influences on X, B, as well as on A via X, B.

The purpose of this chapter is to illustrate some of the essential features of graphical chain representations, and to relate them to more traditional formulations of models as well as to familiar tasks in analysing data. To this end we define chain graphs and mention some related distributional assumptions. We discuss differences between using a chain graph to characterize a statistical model or a substantive research hypothesis. We present reasons for analysing the associations among influences in addition to the type of dependence of responses on the explanatory variables. We explain why interaction effects known from analyses of variance models or, more generally, for regression models are not reflected in graphical

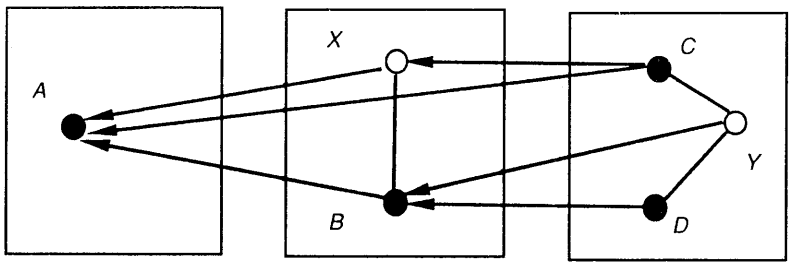

Figure 9.1 *Example of a graphical chain representation. A discrete variable* A *depends directly on the continuous variable* X *and on two discrete variables* B, C *and only indirectly on variables* Y *and* D. *The variables* X *and* B *are intermediate in the sense of being influences on variable* A *and joint responses to variables* C, D *and* Y, *while* C, D *and* Y *are regarded only as influences, that is explanatory variables*

chain representations. We illustrate the different types of analyses and chain graphs that can correspond to the familiar research question of whether a further explanatory variable will improve the prediction of a response. Finally, we discuss some relations of graphical chain models to linear structural relation models.

In this chapter the emphasis is on examples with few variables, but the theory has been developed for many variables. One of the main advantages of graphical chain representations is that problems with many variables which appear complex at first sight might be split up into a sequence of analyses, most of which involve far fewer variables.

Definition of Chain Graphs and Distributional Assumptions

The statistical models for the association structures that we consider consist of distributional assumptions and an independence structure, that is a set of independencies represented by a chain graph.

Chain Graphs

The chain graphs consist of points for variables, and of at most one line or one arrow as the connection of a variable pair, and can be arranged to form a chain of boxes. The chain structure has to be supplied from subject-matter knowledge about responses and potential influences. A chain graph drawn with boxes is viewed as a substantive research hypothesis (Wermuth and Lauritzen, 1990) about direct and indirect relations among variables and not only as a

statistical model. In a graph of a statistical model, that is one drawn without boxes, a connected variable pair just means an unrestricted association, while in a graph of a substantive research hypothesis it stands for a non-vanishing association.

Points represent two types of property of observational units: variables with a nominal scale, called categorical or qualitative (drawn as dots); and variables for which numerical measurements are obtained, called quantitative (drawn as circles). There are two types of associations: the directed associations (drawn as arrows) for variable pairs, where one variable is regarded as a response and the other as an explanatory or influencing variable; and the symmetric associations (drawn as lines without arrowheads), where no direction of dependence has been specified.

If instead the graph just represents a statistical model, that is it is drawn without boxes, then such a model is defined for sets of discrete (dots) and continuous (circles) random variables in terms of specific distributional assumptions and a set of conditional independence restrictions. The graph depicts the independencies, since the set of missing direct connections for variable pairs corresponds to a specific conditional independence structure.

The convention adopted for the chain models of Lauritzen and Wermuth (1989) to ensure an unambiguous interpretation of each pairwise relation is that the conditioning variables of each pair are the remaining 'concurrent variables'. In graphs with dashed lines and arrows of multivariate regression chains (Wermuth and Cox, 1992a) – not discussed here – a different convention is used. The set of concurrent variables is obtained for a given pair (U,V) by ignoring all variables, to which U *and* V are potential influences, that is it is found by deleting from the picture all those boxes to which arrows from both U and V could point. In Figure 9.1, for example, the concurrent variables to (A, Y) are all six variables; to (X, Y) are all variables except A; and to (C, D) are C, D and Y.

Thus the missing link between C and D means conditional independence of C and D given Y ($C \perp D \mid Y$); the missing arrow between B and C says that (B, C) is conditionally independent given X, D, Y ($B \perp C \mid (X, D, Y)$); the arrow from Y to B means a dependence of B on Y given X, C, D; the line between X and B represents a symmetric association between (X, B) given C, Y, D; and the single response A is conditionally independent of the indirect influences D, Y given the directly related explanatory variables B, X and C, that is $A \perp (D, Y) \mid (B, X, C)$.

This last independence statement is derived from the pairwise independencies and is one application of a result for general chain graphs by which one can read off the graph *all* implied

independencies (Frydenberg, 1990). Though this result is most important for understanding and interpreting complex structures, it is less needed for the structures with few variables considered in this chapter.

Distributional Assumptions

The joint density fv in a graphical chain model can be expressed in terms of densities for the different sets of the concurrent variables, for instance for Figure 9.1 as

$$fv = f_{a|bc} f_{b|c} f_c \tag{9.1}$$

where $a=\{A\}$, $b=\{X,B\}$, $c=\{C,Y,D\}$ are called the elements of the dependence chain $C = (a,b,c)$, and with three sets of concurrent variables $a \cup b \cup c$, $b \cup c$ and c. In principle a large number of different distributions corresponding to different special assumptions about the factors determining fv can belong to a chain graph; however, algorithms for estimating associations and for testing independencies are at present not available for many.

In the examples discussed here we assume conditional Gaussian (CG) distribution and regressions (Lauritzen and Wermuth, 1989). Special cases are as follows. For a single continuous response a CG regression can be a linear regression, an analysis of variance, or an analysis of covariance; for a single discrete response it is a linear or a quadratic logistic regression. A CG distribution takes the continuous variable to have a joint normal distribution conditional in each cell defined by the level combinations of the discrete variables; it leads to log-linear models if there are no continuous variables and to a joint normal distribution if there are no discrete variables.

Other distributional assumptions are possible. Some results of how such different assumptions will affect estimation and test results are available (Cox and Wermuth, 1992a).

Substantive Research Hypotheses versus Statistical Models

It can be helpful to distinguish between a statistical model represented by a chain graph drawn without boxes and a substantive research hypothesis represented by a chain graph drawn with boxes. A substantive research hypothesis depends strongly on subject-matter knowledge. Such knowledge typically involves not only indirect relations (modelled in connection with chain graphs via missing direct links which mean independencies) but also the relative strength and the type of associations among variables. In fact, the aim of much substantive research is to establish evidence for relations which are rather strong as compared to weak and

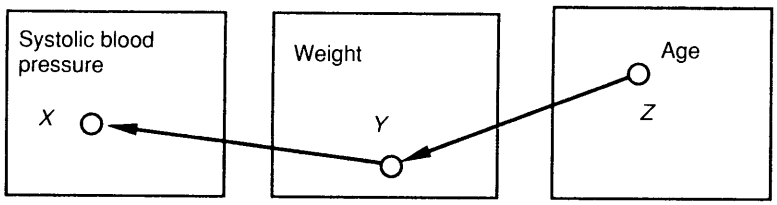

Figure 9.2 *Example for a simple substantive hypothesis: there is a non-vanishing dependence of systolic blood pressure* X *on weight* Y *and of weight* Y *on age* Z; X *is conditionally independent of* Z *given* Y ($X \perp Z \mid Y$), *that is systolic blood pressure* X *depends only indirectly on age* Z *since it is independent of age given information on weight* Y.

negligible ones. This is reflected in the graphical representation, that is connections in chain graphs drawn with boxes mean non-vanishing associations.

In contrast a statistical model for associations can be defined and studied without connection to any specific substantive issues. In it a relation between a variable pair is either restricted by an independence statement and gives a missing direct link in the chain graph, or regarded as unrestricted, that is permitted to vary freely within the limits specified by distributional assumptions.

For instance, for patients with hypertension, strong positive linear dependencies of systolic blood pressure X on both degree of overweight Y and age Z are expected, and a plausible hypothesis is that the dependence on age becomes rather unimportant given the information on degree of overweight. This substantive research hypothesis is expressed with the graph in Figure 9.2. It says that there is a non-negligible correlation between overweight and age (ρ_{yz}), and between systolic blood pressure and overweight (ρ_{xy}), but that knowing the age of a patient does not improve prediction of the degree of hypertension provided the information on degree of overweight is available ($\rho_{xz.y} = 0$). This research hypothesis implies in particular that the simple correlation ρ_{xz} of blood pressure and age is non-zero but is less strong than the smaller of ρ_{yz} and ρ_{xy}, since $\rho_{xz.y} = 0$ implies $\rho_{xz} = \rho_{yz}\rho_{xy}$ and correlations are smaller than one.

In contrast to the research hypothesis, the statistical model underlying Figure 9.2, which could have been specified as a trivariate normal distribution with $\rho_{xz.y} = 0$, does not imply a non-zero marginal association ρ_{xz}. In fact, it is consistent with either or both of ρ_{yz} and ρ_{xy} being zero, in which case ρ_{xz} would also be zero. All that can be derived from the statistical model is that $\rho_{xz} = 0$ is

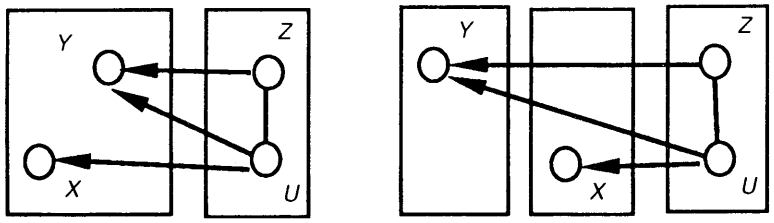

Figure 9.3 *Two distinct research hypotheses which correspond to the same statistical model; the missing edge* (X,Z) *means* $X \perp Z \mid (Y,U)$ *in the left graph and* $X \perp Z \mid U$ *in the right graph; the missing edge between* X *and* Y *means* $X \perp Y\ (Z,U)$, *in both graphs. A compact description of the set of independencies is* $X \perp (Y,Z) \mid U$ *in both graphs*

not implied, that is marginal independence of X and Z is not implied. This is a much weaker implication than the one derived from the research hypothesis.

It may also happen that several research hypotheses are compatible with the same statistical model. One example is given in Figure 9.3. In the left graph Z and U are potential joint influences on the joint responses X and Y, while in the right graph X is regarded as a potential influence on Y as well. Consequently, the meaning of pairwise relations can differ: for instance, the arrow from U to X is a dependence given Y and Z in the left graph, while it is a dependence given Z in the right graph.

To understand the meaning of a research hypothesis, it is crucial to know the dependence chain, since it assigns a specific meaning to each pairwise relation. This is not the case for the corresponding conditional independence structure since it depends only on the underlying chain graph (Frydenberg, 1990). A chain graph used to characterize a statistical model can be obtained from a graph characterizing a research hypothesis by deleting the boxes, that is by ignoring the specific ordering of the variables given in terms of a dependence chain.

The data in Table 9.1 for 98 healthy male adults (Hodapp et al., 1988) show that the research hypothesis of Figure 9.2 is also compatible with observations for persons not suffering from hypertension. The observed partial correlation $r_{xz.y}$ is almost zero and the marginal correlations are all positive, though the strength of the correlations for this collective of healthy persons is smaller than expected for a collective of hypertensive patients.

No statistical test could have rejected the hypothesis $\rho_{xz} = 0$, since the observed correlation is rather small ($r_{xz} = 0.139$). How-

Table 9.1 *Risk factors for cardiovascular diseases: observed marginal correlations (lower half), observed partial correlations (upper half) and further data summaries, n=98*

Variable	*X* Systolic blood pressure	*Y* Weight	*Z* Age
X Systolic blood pressure	1	0.348	−0.007
Y Weight relative to height	0.371	1	0.369
Z Age	0.139	0.390	1
Mean	128.31	0.42	32.74
Standard deviation	13.47	0.04	11.67

ever, it would be unwise to use such a test and its result in the present context: it would mean to ignore the available subject-matter knowledge, in particular the implication of the research hypothesis in Figure 9.2.

Reasons for Analysing Relations among Explanatory Variables

If observations become available from a particular study, there will be expectations on the part of the investigator regarding strength, direction or lack of associations not only for the response variables but also for the potential explanatory variables. This alone is an important reason to investigate relations among explanatory variables.

If unexpected findings are encountered there may be systematic errors in the data or there may be selection effects. For instance, in a study of effects of different pre-operative sedative treatments an unexpected strong association between vigilance, a strategy to cope with anxiety and stress, and gender of patients was observed. The reason for this turned out to be a selection strategy: the anaesthetist had allocated to the control group, that is to no pre-operative treatment at all, only those patients who appeared to be least excited. As a consequence, patients in the control group had characteristics quite distinct from patients in treated groups.

Another important reason for analysing relations among explanatory variables is to investigate whether moderation in the confounding sense (Breslow and Day, 1980; Wermuth, 1992a) can be a feature of the investigated relations. This means that an association which coincides in several subgroups is qualitatively different overall, that is without a split into subgroups. In the literature on

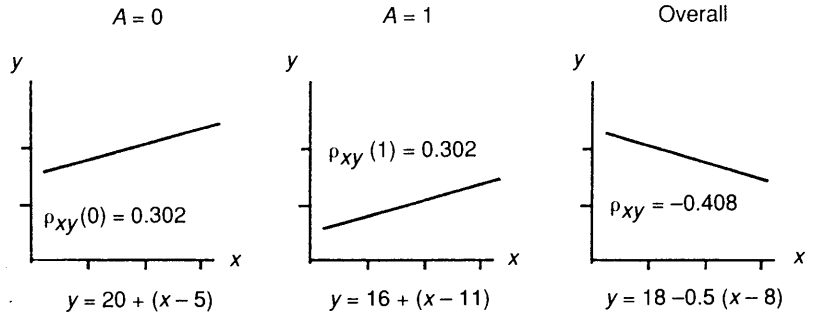

Figure 9.4 *Example for moderation in the confounding sense with:* $\mu_y(0) = 20$, $\mu_x(0) = 5$, $\sigma_{yy}(0) = 11$, $\sigma_{xx}(0) = 1$, $\sigma_{xy}(0) = 1$; $\mu_y(1) = 16$, $\mu_x(1) = 11$, $\sigma_{yy}(1) = 11$, $\sigma_{xx}(1) = 1$, $\sigma_{xy}(1) = 1$; *and* $Pr(A = 0) = Pr(A = 1) = 0.5$

contingency tables, such situations have been called the Yule-Simpson paradox (Simpson, 1951; Wermuth, 1989). A particularly striking version of it is shown for parallel linear regressions in Figure 9.4.

Moderation in the confounding sense is further illustrated with fictitious data for a contingency table in Table 9.2. A reversal in the direction of dependence after marginalizing over one of the explanatory variables cannot occur with independent explanatory variables; it is more likely the stronger the explanatory variables are associated. It is most important to be aware of moderation in the confounding sense if results from two studies are to be compared; confounding can be the explanation for results which appear contradictory at first sight.

Chain Graphs and Interactions

A missing link in a chain graph means a conditional independence, and a direct connection between a variable pair means a particular – not further specified – conditional association. For purposes of interpretation it is important to understand how these concepts relate to more traditional definitions of interactions.

Fisher (for example, 1956: Chapter 42) had introduced interaction in a two-way analysis of variance context to mean a different dependence of the response on one explanatory variable for different levels of the second variable, that is as a two-factor interaction in a model for dependence of a quantitative response X on two qualitative explanatory variables A and B. X had a normal distribution given A and B, possibly differing in means but not in

Table 9.2 *Examples with strongly consistent results within sites in terms of relative risks, which have value 1.5 at each site, and appear reversed overall in case (a) but are replicated in case (b)*

(a) Yule-Simpson paradox (moderation in the confounding sense)

	Clinic X: treatment		Clinic O: treatment		Overall: treatment	
Outcome	A	B	A	B	A	B
1	600	40	300	2000	900	2040
	(30%)	(20%)	(75%)	(50%)	(38%)	(49%)
2	1400	160	100	2000	1500	2160
Sum	2000	200	400	4000	2400	4200
	$\frac{\pi 1\|AX}{\pi 1\|BX} = 1.5$		$\frac{\pi 1\|AO}{\pi 1\|BO} = 1.5$		$\frac{\pi 1\|A}{\pi 1\|B} = 0.78$	

(b) Relative risk is collapsible since explanatory variables are independent

	Clinic X: treatment		Clinic O: treatment		Overall: treatment	
Outcome	A	B	A	B	A	B
1	60	400	300	2000	360	2400
	(30%)	(20%)	(75%)	(50%)	(60%)	(40%)
2	140	1600	100	2000	240	3600
Sum	200	2000	400	4000	600	6000
	$\frac{\pi 1\|AX}{\pi 1\|BX} = 1.5$		$\frac{\pi 1\|AO}{\pi 1\|BO} = 1.5$		$\frac{\pi 1\|A}{\pi 1\|B} = 1.5$	

variance. Conditional independence of X of A given B can in this model be expressed as $g_{x|ij}^{X|AB} = g_{x|j}^{X|B}$, and implies that there is no main effect of A and no two-factor interaction of A,B on X.

These notions have been extended to other dependence models. For instance, for a linear regression of X on a quantitative influence Y and a qualitative influence A, a two-factor interaction of A, Y on X means changing slopes of the linear regressions of X on Y at the different levels of A; that is, the lack of a two-factor interaction implies parallel regressions. Conditional independence of the response X of A requires in addition that the main effect is missing. More precisely, $X \perp A \mid Y$ implies that the parallel regression lines coincide, that is have equal intercepts of all levels of A.

The same interpretation applies to other CG regressions, that is logistic regressions with qualitative explanatory variables $g_{i|jk}^{A|BC}$ or with mixed explanatory variables $g_{i|jx}^{A|BX}$, and also to corresponding probit regressions.

Bartlett (1935) had given a definition of a three-factor interaction in a three-dimensional contingency table: changing odds ratios of two variables for different levels of the third variable. Though this definition for log-linear models appears to be quite similar to Fisher's definition, there is the important distinction in that it concerns interaction in a joint distribution, that is in a model for symmetric associations, and not the more commonly considered interaction in models for dependencies (Cox, 1984).

Conditional independence of A of B given C in a long-linear model can be expressed as $g_{ijk}^{ABC} = g_{ik}^{AC} g_{jk}^{BC} / g_{k}^{C}$ and it implies in particular that there is no log-linear two-factor interaction of A, B and no three-factor interaction. Since $g_{i|jk}^{A|BC} = g_{ijk}^{ABC} / g_{jk}^{BC} = g_{i|k}^{A|C}$ there is a correspondence between missing one- and two-factor interactions (B, BC) in a regression of A on B, C and missing two- and three-factor interactions (AB, ABC) in a joint distribution: they are equivalent formulations for conditional independence of (A,B) given C.

Bartlett's notion of interaction in a joint distribution and its relation to interactions in corresponding regression models has been extended to other than log-linear models with CG distributions and corresponding CG regressions. In any CG distribution a variable pair is conditionally independent given all of the remaining variables if and only if the two-factor interaction and all higher-order interactions of this pair vanish. Furthermore, vanishing of two- and higher-order interactions in the joint distribution implies the vanishing of a main effect and higher-order interactions in a corresponding CG regression which has one variable of the pair as a univariate response. This gives the precise meaning of a missing line and of a missing arrow in a chain graph in terms of interactions.

Similarly, an arrow in a chain graph means the presence of a main effect *or* of a two-factor *or* of a higher-order interaction in a regression, and a line means the presence of a two-factor *or* of a higher-order interaction in a joint distribution. This explains why a graphical chain representation completely describes independencies but only incompletely specifies the type of associations which are present. To give an example we take the symmetric association structure for three symptoms of EPH gestosis (Wermuth and Koller, 1976), an illness occurring during pregnancy, and symptoms after LSD intake (Lienert, 1970). The symptoms for the gestosis data are A yedema, B proteinuria and C hypertension, and for the LSD data they are distortions in: A thinking, B consciousness and C affective behaviour (Table 9.3).

In both cases the graphical chain representation is a complete graph with lines connecting all three symptom pairs. However, the

Table 9.3 *Counts for combinations of three symptoms of EPH gestosis and after LSD intake*

Data set	Symptom	Levels and counts							
	A	1	0	1	0	1	0	1	0
	B	1	1	0	0	1	1	0	0
	C	1	1	1	1	0	0	0	0
LSD intake[1]		21	4	2	11	5	16	13	1
EPH gestosis		14	9	36	45	26	44	609	2342

[1] One observation has been added to each cell.

association structures are quite different in the two situations. There is no log-linear three-factor interaction for the gestosis data but a strong log-linear three-factor interaction for the LSD data. All two-way margins show strong associations for the EPH data, but rather weak associations for the LSD data. These differences are not captured in the graph. The graphs just reflect the conditioning sets for each substantial dependence or association and the conditional independencies if there are any. Thus they show, for large sets of variables, to which sequences of smaller problems the analysis may be simplified.

Prediction of a Response: Is It Improved by an Additional Explanatory Variable?

In many research situations in which past research has established the dependence of a response variable X on a single explanatory variable $Y\,(h_{x|y}^{X|Y} \neq h_x^X)$, a natural next question is whether prediction of the response might be further improved by another variable Z. If this is not the case we have $X \perp Z \mid Y$ or $h_{x|yz}^{X|YZ} = h_{x|y}^{X|Y}$, and hence, the type of dependence of X on Y is not moderated at all by Z. If, however, the response is dependent on the additional explanatory variables, it becomes necessary to describe the type of dependence; in particular, it can be an issue whether there is moderation in the interactional or in the confounding sense (Wermuth, 1992a). In the case of moderation in the confounding sense the direction of a dependence can appear reversed after a second variable is included in the regression. In the case of moderation in the interactional sense the dependence of the response on an explanatory variable differs with different levels of the other explanatory variable.

Quite different types of analyses are needed in such situations depending on whether the involved variables are qualitative or quantitative. We give three examples in Figure 9.5. In the first

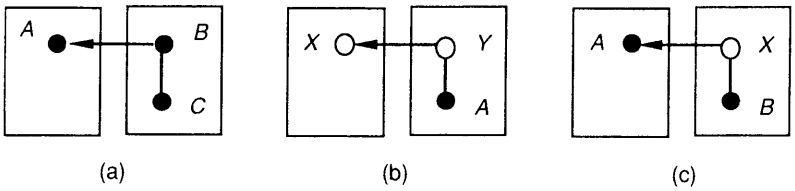

Figure 9.5 *Three regression chain graphs corresponding to the same type of research hypothesis: of two associated explanatory variables, only one is needed to predict the response. (a) Logistic regression chain with two qualitative explanatory variables; (b) linear regression chain with mixed explanatory variables; (c) logistic regression chain with mixed explanatory variables*

example (Figure 9.5a) observations on $n = 25{,}777$ women (National Institutes of Health, 1972) are available for three qualitative variables A, B, C defined as follows:

- A perinatal death ($i = 1$ yes; $i = 2$ no)
- B survival state of last prior child ($j = 1$ living; $j = 2$ child death; $j = 3$ foetal death; $j = 4$ neonatal death; $j = 5$ unknown)
- C skin colour of woman ($k = 1$ light; $k = 2$ dark).

The counts and the observed risks (in percentage rates) for perinatal mortality A at each of the level combinations of the potential influences B, C are given in Table 9.4. The research hypothesis is that prediction of the risk of perinatal mortality is not improved by the information on the skin colour of the mother provided a good indicator for the medical and socioeconomic situation of the mother is available. The test results in Table 9.5 and the mortality rates estimated under this hypothesis (Table 9.4) show how well the hypothesis is compatible with the observations. As a final summary the estimated relative risks for perinatal mortality are displayed in Table 9.4. The increase in risk for perinatal mortality as compared with the best condition (the last child prior to the present is alive) is substantial. The relative risk of 4.2, for instance, says that the risk for perinatal mortality is four times higher under the worst condition (survival status of last child unknown) as compared with the best condition. This risk increase due to poor socioeconomic conditions of the mother is higher than most risk increases owing to medical factors reported in National Institutes of Health (1972).

In the second example (Figure 9.5b) there are observations for n=40 patients prior to an operation on the jaw (Krohne et al., 1989). None of the patients have been treated with sedative drugs,

Table 9.4 *Counts and other data summaries for perinatal deaths (A), survival status of last prior child (B), and skin colour of woman (C)*

Levels					Estimates under $\pi_{ijk} = \pi_{ij}$		
A	*B*	*C*	Observed count	Observed % rate	Count	% rate	Relative risk
1	1	1	270	0.28	297.5	0.32	
2	1	1	9,148		9,120.5		
1	2	1	3	0.27	3.4	0.31	1
2	2	1	108		107.6		
1	3	1	134	0.74	132.8	0.73	2.3
2	3	1	1,678		1,679.2		
1	4	1	17	0.90	19.3	1.02	3.2
2	4	1	173		170.7		
1	5	1	56	1.26	59.3	1.33	4.2
2	5	1	389		385.7		
1	1	2	371	0.34	343.5	0.32	
2	1	2	10,502		10,529.5		
1	2	2	5	0.34	4.6	0.31	1
2	2	2	144		144.4		
1	3	2	154	0.72	155.2	0.73	2.3
2	3	2	1,963		1,961.8		
1	4	2	37	1.08	34.7	1.02	3.2
2	4	2	305		307.3		
1	5	2	46	1.44	42.7	1.33	4.2
2	5	2	274		277.3		

Source: National Institutes of Health, 1972: 187

Table 9.5 *Tests for conditional independencies, Table 9.4 data*

Pair	Concurrent variables	Values of chi-square statistic	Degrees of freedom	Corresponding fractile or *p*-value
(A,B)	ABC	279.12	8	<0.001
(A,C)	ABC	6.03	5	0.302
(B,C)	BC	68.81	4	<0.001

and – owing to a selection strategy of the anaesthetist – all have in common that they are non-vigilant, that is they do not use vigilant behaviour as a strategy in copying with stress. The variables are as follows:

X level of free fatty acids measured in the blood just before the operation

Table 9.6 *Effects of pre-operative anxiety Y on free fatty acids X*

	Overall		$A=1$		$A=2$	
	X	Y	X	Y	X	Y
Mean	390.8	41.7	434.8	44.0	346.7	39.5
Standard deviation	156.6	11.2	165.9	12.3	136.8	9.8
Correlation	0.38		0.54		0.06	
Number of patients	40		20		20	

Y level of anxiety measured with a state anxiety questionnaire on the morning of the day of the operation

A coping strategy 'cognitive avoidance' (1 = not employed; 2 = employed; categories were obtained by median dichotomizing the corresponding questionnaire scores U).

The research hypotheses are that either the coping strategy does not modify the dependence of free fatty acids on anxiety (as displayed in Figure 9.5b) or, if it does, a stronger dependence of physiological reaction on anxiety is expected if the patients do not use a strategy to cope with anxiety than if they do. When such a change in association is expected, any analysis based only on simple and partial correlations of the variables X, Y and U would not be suitable.

Table 9.6 gives the basic data summaries and Figure 9.6 shows the scatter plots between X and Y for $A=1$ and $A=2$, to ascertain that the observed changes in association between X and Y are not due to outliers or other irregularities of the data. Thus the data lead to a rejection of the hypothesis that information on the coping strategy is not needed to predict the level of free fatty acids from the level of anxiety. Instead the data support the described expectations regarding changes in associations, that is there is the expected interaction effect of Y and A on X.

In the third example (Figure 9.5c) the response variable is again qualitative, it is known to depend on a quantitative explanatory variable, and a potential qualitative moderator variable is considered. The observed variables for $n=149$ patients (Schmitt, 1990) are as follows:

A success of treatment (0 = no; 1 = yes) obtained from dichotomizing a more detailed score

X stage of chronic pain, a constructed indicator with possible values from 4 to 12

B gender.

In this context the hypothesis in Figure 9.5c says that the depen-

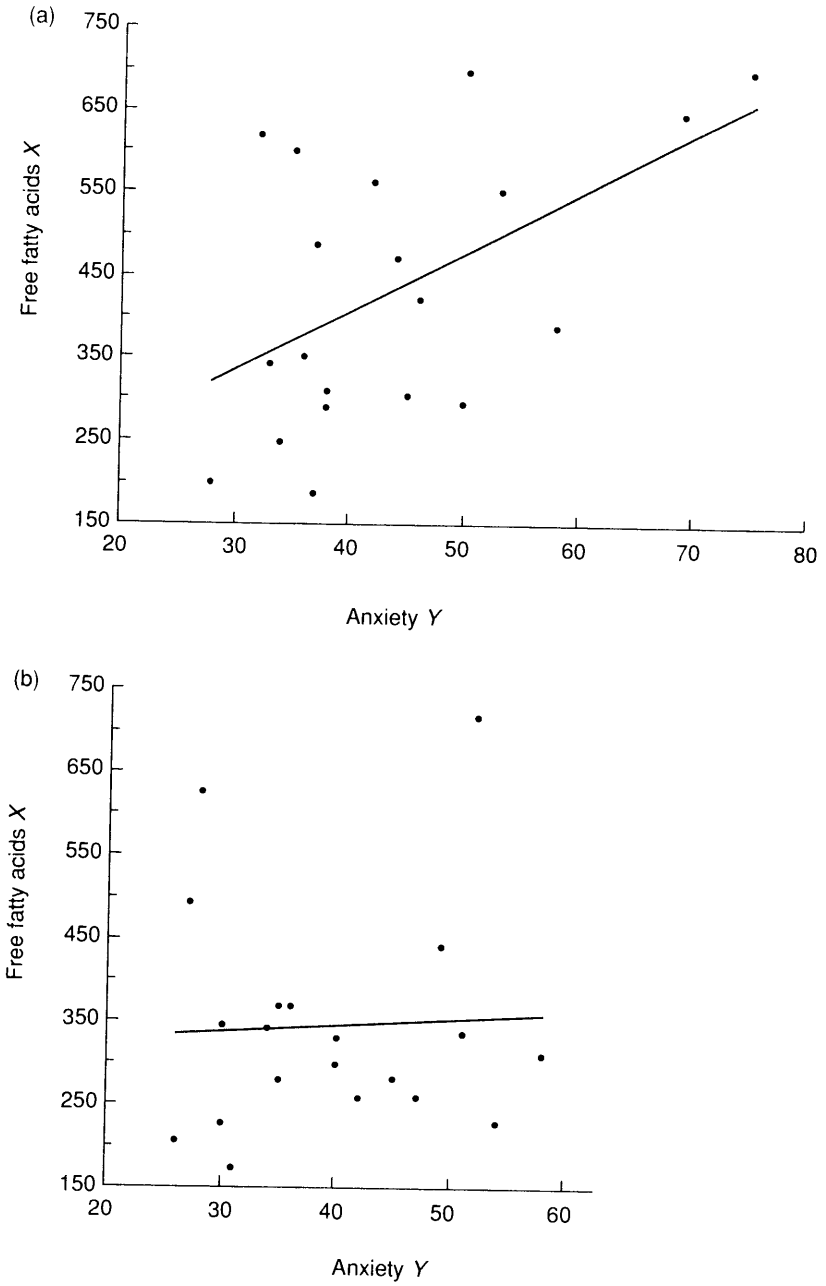

Figure 9.6 *Dependence of free fatty acids* X *on anxiety* Y *if the coping strategy of cognitive avoidance is (a) not employed* (A=*1) and (b) employed* (A = *2)*

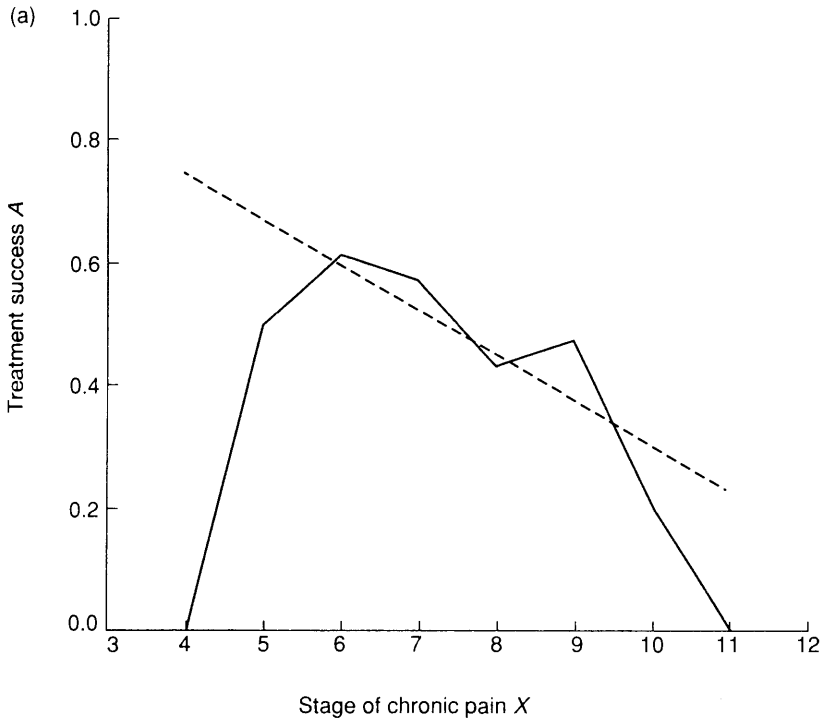

Figure 9.7 *Observed frequencies for success of treatment* A *as it depends on stage of chronic pain* X *and probabilities estimated by (a) linear logistic regression and (b) quadratic logistic regression. Solid line is smoothed observed values, broken lines are predictions*

dence of success of treatment on the stage of chronic pain is the same for males and females.

These data for a logistic regression provide one of the many examples in which an automatic search procedure leads to misleading results because the distributional assumptions in the research procedure amount to an overspecification; that is, if in this case the stepwise logistic regression of BMDP is used, which is a model search based only on global test statistics in linear logistic regression. More precisely, if one starts by assuming a linear logistic regression for X on Y it appears as if gender has no moderating effect on the dependence of success of treatment on stage of chronic pain. The computed goodness-of-fit statistics do not point to a bad fit of the models. This poor fit is, however, easily discovered from plotting fitted against observed probabilities of success as shown in Figure 9.7a.

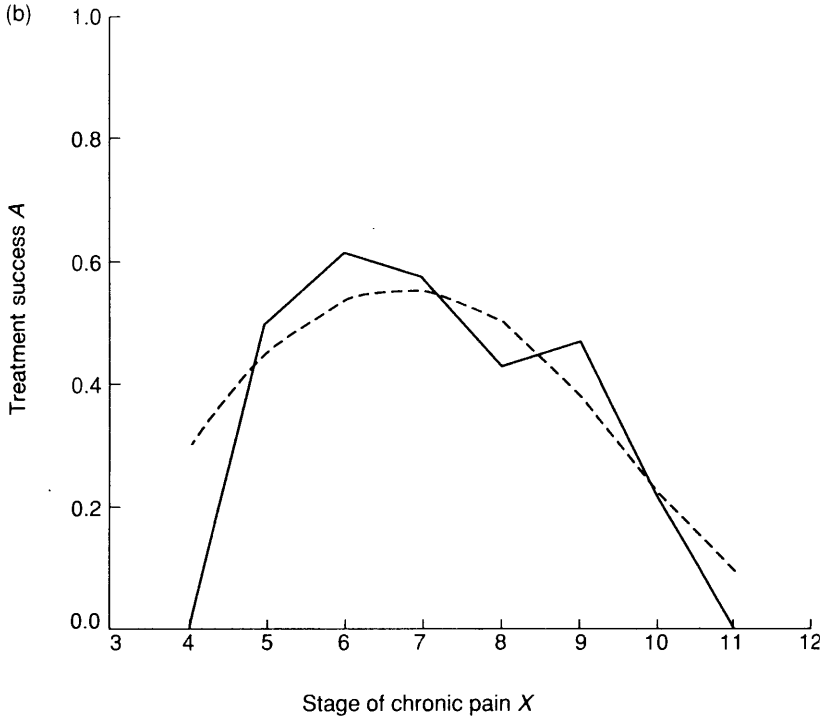

Figure 9.7 *continued*

If instead of assuming linear logistic regression we permit a quadratic dependence, that is we use the assumption of a non-homogeneous CG regression chain, then not only do we obtain a good fit overall as shown in Figure 9.7b, but gender emerges as an important moderator in the interactional sense. Observed and estimated success rates are displayed in Table 9.7 and in Figure 9.8. In this case the estimation results point to an unexpected interaction effect. It turns out that the patients with low success rates of treatment in spite of low scores for the stage of chronic pain are female headache patients. Further observations will be needed to judge the relevance of the result.

Relations to Other Models

Linear regressions and probit regressions are also assumptions used for models of linear structural relations (Jöreskog, 1977; Muthén, 1984). However, graphical representations of the latter cannot in

Table 9.7 *Scores for stage of chronic pain (X), counts, and estimated probabilities of success of treatment when leaving the clinic (A) for n = 149 patients given gender (B) and stage (X)*

Stage of chronic pain	Total count		Number of successes		Probabilities estimated by: observed relative frequencies		quadratic logistic regression	
	Females	Males	Females	Males	Females	Males	Females	Males
x	$n_{.1x}$	$n_{.2x}$	n_{11x}	n_{12x}	$n_{1\mid 1x}$	$n_{1\mid 2x}$		
4	1	–	0	–	0.00	–	0.06	–
5	2	–	1	–	0.50	–	0.21	–
6	5	8	1	7	0.20	0.88	0.39	0.80
7	12	9	7	5	0.58	0.56	0.50	0.63
8	36	15	16	6	0.44	0.40	0.50	0.47
9	20	14	10	6	0.50	0.43	0.39	0.34
10	14	10	2	3	0.14	0.30	0.20	0.27
11	1	2	0	0	0.00	0.00	0.06	0.24

general be interpreted as chain graphs. One exception is when they correspond to systems of univariate recursive equations (see Wold, 1954; Wermuth and Lauritzen, 1983). Such graphs have been called univariate recursive (Wermuth and Lauritzen, 1990) or directed acyclic graphs (Pearl, 1988). Another exception is if they correspond to a multivariate regression or to a block-regression model (Wermuth, 1992b).

Models defined for univariate recursive systems within either framework, that is as a structural relation model or as a CG regression chain model, can be identical, similar or rather different. They are identical if all responses are continuous variables. They are rather similar if they only differ in probit versus logistic regression, since linear logistic regressions are virtually indistinguishable from linear probit regressions (Cox, 1966; Cox and Snell, 1989). They differ substantially if a quadratic logistic regression appears in the CG regression chain model but only a linear probit regression in the corresponding structural relation model.

There exist conditional independence structures which can be tested directly within the framework of graphical chain models but not within the framework of linear structural relations, and vice versa. A simple example of the former is displayed in Figure 9.9 and Table 9.8. An example for the latter is $X \perp U \mid Z$ *and* $Y \perp Z \mid U$, which can be interpreted as a hypothesis in a multivariate regression of X and Y on U and Z (see Cox and Wermuth, 1993).

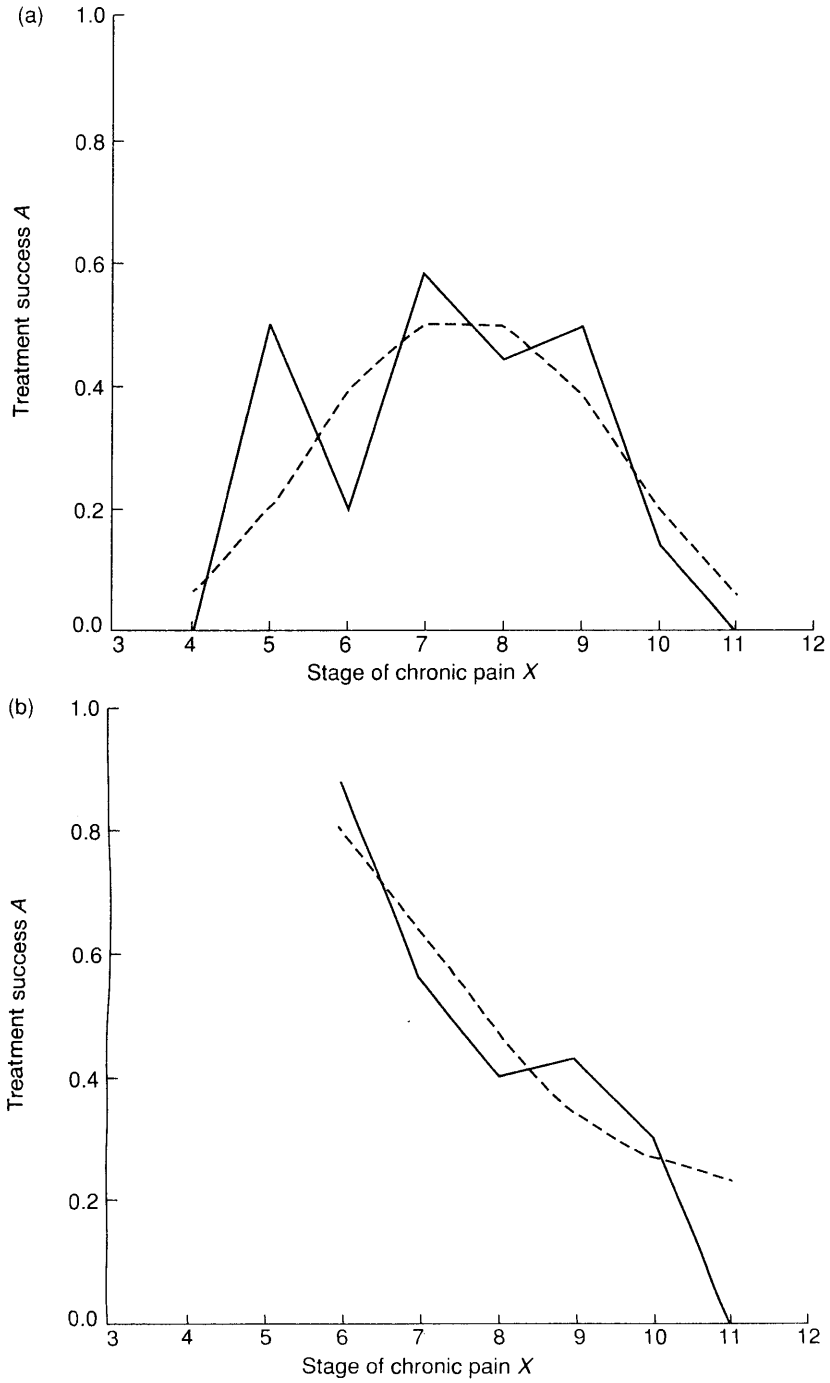

Figure 9.8 *Observed frequencies for success of treatment* A *and probabilities estimated by quadratic logistic regression given stage of chronic pain* X *and gender* B *for (a) females and (b) males. Solid line is smoothed observed values, broken lines are predictions*

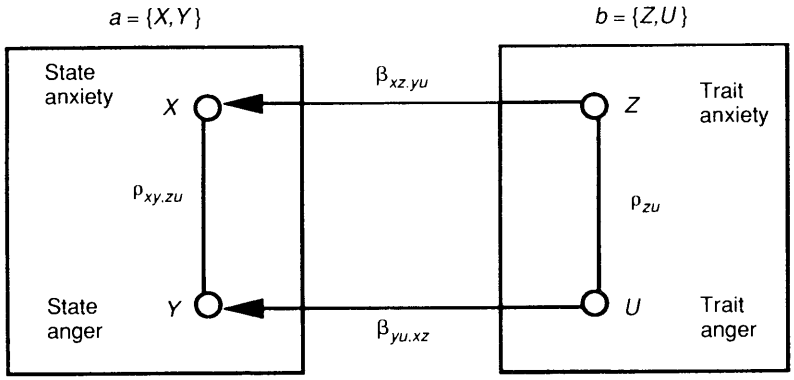

Figure 9.9 *A chain graph with a nondecomposable hypothesis:* $X \perp U \mid (Y,Z)$ *and* $Y \perp Z \mid (X,U)$

The main disadvantages of linear structural equation models are:

1 The interpretation of each model, that is the meaning of parameters and of missing direct connections, has to be derived from scratch in most situations, since no general results are available to deduce them.
2 The meaning of equation parameters in linear structural relations is not tied to the notion of independence. It may, in particular, occur that by imposing one more zero restriction in a model, one suddenly has a situation in which some parameters are unidentifiable.

Table 9.8 *Observed marginal correlations (lower half) and observed partial correlations given all remaining variables (upper half), and further data summaries, n=684*

Variable	*X* State anxiety	*Y* State anger	*Z* Trait anxiety	*U* Trait anger
X State anxiety	1	0.45	0.47	− 0.04
Y State anger	0.61	1	0.03	0.32
Z Trait anxiety	0.62	0.47	1	0.32
U Trait anger	0.39	0.50	0.49	1
Mean	18.87	15.23	21.20	23.42
Standard deviation	6.10	6.70	5.68	6.57

Source: C.D. Spielberger, personal communication of data on anger, anxiety (1983).

3 A discrete variable enters as a response variable only by assuming an underlying normal variate which has been partitioned to give the categorized variable. This excludes nominal scaled variables and models for symmetric associations with three-factor interactions, that is explanations for data like those reported by Lienert (Table 9.3).

These disadvantages are not shared by graphical chain models. Their main disadvantages are of a different kind:

1 No programmed algorithms are widely available yet which permit the computation of estimates for each model in this class; such a development is likely to build on work by Frydenberg and Edwards (1989), Cox and Wermuth (1990; 1991) and Jensen et al. (1991).
2 The statistical theory for models with latent variables which is needed in many applications is not yet well developed.
3 More examples of good analyses with many variables would be helpful.

Though a considerable amount of new work on different special aspects of models for multivariate dependencies and associations has been published, for instance Cox and Wermuth (1992a; 1992b; 1992c; 1993) and Wermuth and Cox (1992a; 1992b; 1992c) much more needs to be done.

References

Bartlett, M.S. (1935) 'Contingency table interactions', *Supplement, Journal of th Royal Statistical Society*, 2: 248–52.

Breslow, N. and Day, N. (1980) *The Analysis of Case-Control Studies: Statistica Analysis in Cancer Research*. Lyon: International Agency for Research on Cancer

Cox, D.R. (1966) 'Some procedures connected with the logistic qualitative respons curve', in F.N. David (ed.), *Research Papers in Statistics: Essays in Honour c J. Neyman's 70th Birthday*. London: Wiley. pp. 55–71.

Cox, D.R. (1984) 'Interaction', *International Statistical Review*, 52: 1–31.

Cox, D.R. and Snell, E.J. (1989) *Analysis of Binary Data*, 2nd edn. Londor Chapman and Hall.

Cox, D.R. and Wermuth, N. (1990) 'An approximation to maximum-likelihoc estimates in reduced models', *Biometrika*, 77: 747–61.

Cox, D.R. and Wermuth, N. (1991) 'A simple approximation for bivariate ar trivariate normal integrals', *International Statistical Review*, 59: 263–9.

Cox, D.R. and Wermuth, N. (1992a) 'Response models for mixed binary ar quantitative variables', *Biometrika*, 79: 441–61.

Cox, D.R. and Wermuth, N. (1992b) 'On the calculation of derived variables in tl analysis of multivariate responses', *Journal of Multivariate Analysis*, 42: 162–7(

Cox, D.R. and Wermuth, N. (1992c), 'A comment on the coefficient of determ nation for binary responses', *American Statistician*, 46: 1–4.

Cox, D.R. and Wermuth, N. (1993) 'Linear dependencies represented by chain-graphs.' To appear with discussion in *Statistical Science*.

Fisher, R.A. (1956) *Statistische Methoden für die Wissenschaft*, 12th edn. Edinburgh: Oliver and Boyd.

Frydenberg, M. (1990) 'The chain graph Markov property', *Scandinavian Journal of Statistics*, 17: 333–54.

Frydenberg, M. and Edwards, D. (1989) 'A modified iterative proportional scaling algorithm for estimation in regular exponential families', *Computational Statistics and Data Analysis*, 8: 143–53.

Hodapp, V., Neuser, K.W. and Weyer, G. (1988) 'Job stress, emotion, and work environment: toward a causal model', *Personality and Individual Differences*, 9: 851–9.

Jensen, S.T., Johansen, S. and Lauritzen, S.L. (1991) 'Globally convergent algorithms for maximizing a likelihood function', *Biometrika*, 78: 867–78.

Jöreskog, K.G. (1977) 'Structural equation models in the social sciences: specification, estimation and testing', in P.R. Krishnaiah (ed.), *Applications of Statistics*. Amsterdam: North-Holland. pp. 267–87.

Krohne, H.W., Kleemann, P.P., Hardt, J. and Theisen, A. (1989) 'Beziehungen zwischen Bewältigungsstrategien und präoperativen Streßreaktionen', *Zeitschrift für Klinische Psychologie*, 18: 350–64.

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

Lienert, G.A. (1970) 'Konfigurationsfrequenzanalyse einiger Lysergsäure-diäthylamid-Wirkungen', *Arzneimittel-Forschung*, 20 (7): 912–13.

Muthén, B. (1984) 'A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators', *Psychometrika*, 49: 115–32.

National Institutes of Health (1972) *The Women and their Pregnancies*, DHEW publication (NIH) 73–379. Washington, DC: US Government Printing Office.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Schmitt, N. (1990) 'Stadieneinteilung chronischer Schmerzen'. Medical dissertation, University of Mainz.

Simpson, E.H. (1951) 'The interpretation of interaction in contingency tables', *Journal of the Royal Statistical Society, Series B*, 13: 238–41.

Spielberger, C.D. (1983) *Manual for the State-Trait Anxiety Inventory*. Palo Alto: Consulting Psychologists Press.

Wermuth, N. (1989) 'Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable', *Journal of the Royal Statistical Society, Series B*, 49: 353–64.

Wermuth, N. (1992a) 'On moderating effects in the interactional and in the confounding sense, a reply', *Methodika*, 6: 5–7.

Wermuth, N. (1992b) 'On block-recursive regression equations (with discussion)', *Brazilian Journal of Probability and Statistics*, 6: 1–56.

Wermuth, N. and Cox, D.R. (1992a) 'Derived variables calculated from similar responses: some characteristics and examples', *Computational Statistics and Data Analysis*.

Wermuth, N. and Cox, D.R. (1992b) 'On the relations between interactions obtained with alternative codings of discrete variables', *Methodika*, 6: 76–85.

Wermuth, N. and Cox, D.R. (1992c) 'Graphical models for dependencies and

associations' in Y. Dodge and J. Whittaker (eds), *Computational Statistics, vol. 1*. Heidelberg: Physica, pp. 235–49.

Wermuth, N. and Koller, S. (1976) 'Systematik multivariater Korrelationsmuster angewandt auf die Symptomkorrelation von Krankheiten', in S. Koller and J. Berger (eds), *Klinisch-Statistische Forschung*. Stuttgart: Schattauer. pp. 111–20.

Wermuth, N. and Lauritzen, S.L. (1983) 'Graphical and recursive models for contingency tables', *Biometrika*, 70: 537–52.

Wermuth, N. and Lauritzen, S.L. (1990) 'On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion)', *Journal of the Royal Statistical Society, Series B*, 52: 21–72.

Wold, H.O. (1954) 'Causality and econometrics', *Econometrica*, 22: 162–77.