Multivariate analysis of conditional relationships

Rosenberg (1968)

Elaborating, Explaining and specification

Steffen Lauritzen & Nanny Wermuth & many others

Graphical models

Graphical models in DIGRAM

Morris Rosenberg (1968): The Logic of Survey Analysis

“When a research investigator discovers a relationship between two variable, the first question he implicitly asks is: Is it real?”

“Knowing that sociological variables are block-booked, he is concerned to know whether there is an inherent link between the variables or whether it is based on an accidental connection with some associated variable. In short, he must guard against what are called spurious relationships.”

“Strictly speaking, there is no such thing as a spurious relationship; there are only spurious interpretations”

Elaboration

"The most important systematic way of examining the relationship between two variables is to introduce a third factor, called a test factor, into the analysis. This is what is meant by the process of *elaboration*."

"Typically one begins with a relationship between an independent variable and a dependent variable. One then seeks to explain this relationship by introducing an explanatory variable by introducing a test factor. The method used is to stratify on the test factor and to examine the contingent associations".

"If, when the influence of the test factor is held constant, one finds that the relationship disappears, then it may be concluded that the relationship is due to the extraneous factor".

Murder cases in Florida 1973-79

Association between the race of the murderer and death sentyencies in 4764 murder cases in Florida 1973-1979.

Murderer	Other sentence	Death sentence
Black	2448	59
	97.6 %	2.4 %
White	2185	72
	96.8 %	3.2 %

$\chi^2 = 3.1$, df = 1 p = 0.08

$\gamma = 0.16$ p = 0.08

Harder sentences to white murderers.
The association is not significant

Elaboration by the race of the victim

	Black victim		White victim	
Murderer	**Other sentence**	**Death sentence**	**Other sentence**	**Death sentence**
Sort	**2209**	**11**	**239**	**48**
	99.5 %	**0.5 %**	**83.3 %**	**16.7 %**
Hvid	**111**	**0**	**2074**	**72**
	100 %	**0 %**	**96.6 %**	**3.4 %**

$\chi^2 = 0.6$, df = 1 p = 0.59 $\chi^2 = 96.5$, df = 1 p = 0.000

$\gamma = -1.00$ p = 0.59 $\gamma = -0.71$ p = 0.000

Simpsons paradox!

Conditional relationships: Specification and description

Conditional relationships may

1) Purify or reduce contamination in the original relationship,

2) specify conditions facilitating relationships,

3) specify conditions inhibiting or blurring relationships,

4) stipulate necessary conditions,

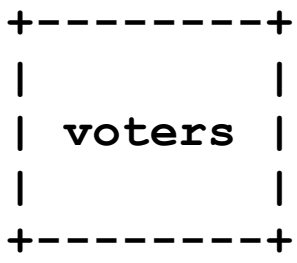
5) clarify the nature of the independent and dependent variables,

6) shed new light on the test factor categories,

7) make descriptive statements more exact.

The Voters example

The VOTERS project include five variables from a panel study of presidential selections in 1956 and 1960.




```
A:    VOTE60   -    2 nominal categories
B:   PARTY60   -    3 ordinal categories
C:    VOTE56   -    2 nominal categories
D:   PARTY56   -    3 ordinal categories
E: RELIGION    -    2 nominal categories

          CAUSAL/RECURSIVE STRUCTURE
             A <- B <- C <- D <- E


-------------- COMMENTS ------------
DATA FROM DUNCAN(1981): TWO FACES OF PANEL
ANALYSIS.
LEINHARDT, S. (ED): SOCIOLOGICAL METHODOLOGY
1981
  - PRESIDENTIAL VOTE, PARTY IDENTIFICATION,
    RELIGION AND YEAR: 1956-1960 PANEL
-------------- COMMENTS -----------------
```

A VOTE60	B PARTY60	C VOTE56
1 Democrat	1 Democrat	1 Democrat
2 Republic	2 Independ	2 Republic
	3 Republic	

D PARTY56	E RELIGION
1 Democrat	1 Catholic
2 Independ	2 Non-Cath
3 Republic	

Catholics voted for Kennedy in 1960

E RELIGION	A:--VOTE60 Democ	Repub	TOTAL	
Catho	165	37	202	
row%	81.7	18.3	100.0	
Non-C	224	360	584	
row%	38.4	61.6	100.0	X^2 = 112.7
TOTAL	389	397	786	df = 1
row%	49.5	50.5	100.0	p = 0.000
				Gam = 0.76
				p = 0.000

The methodological problem: How to elaborate the relationship Vote60 and Religion

	VOTE60	PARTY60	VOTE56	PARTY56	RELIGION
VOTE60		0.8439	0.8848	0.7788	0.7551
p		0.0000	0.0000	0.0000	0.0000
PARTY60	0.8439		0.8907	0.9292	0.4889
p	0.0000		0.0000	0.0000	0.0000
VOTE56	0.8848	0.8907		0.9022	0.2252
p	0.0000	0.0000		0.0000	0.0064
PARTY56	0.7788	0.9292	0.9022		0.2809
p	0.0000	0.0000	0.0000		0.0000
RELIGION	0.7551	0.4889	0.2252	0.2809	
p	0.0000	0.0000	0.0064	0.0000	

Marginal correlations among variable are not helpful, but a graphical model describing conditional or partial correlations may be.

Graphical models

Define models by assumptions of conditional independence and dependence instead of marginal associations.

Provide unique possibilities for tests of model fit and description of elaborated associations between variables.

Define interaction graphs or Markov graphs by elimination of edges and arrows in graphs where nodes represent variables.

An edge or arrow in in a Markov graph refers to an association that cannot be explained by elaboration and therefor need to be specified.

Block recursive models are defined by two sets of assumptions.

First, a recursive structure defined by causal, temporal or design based considerations.

Second, a list of pairwise conditional independencies give all concurrent or prior variables.

To define a graphical model we therefore have to impose initial assumptions of conditional dependence or test conditional independence of pairs of variables.

Defining graphical model in DIGRAM

Vote60 (A) $\Leftarrow$ Party60 (B) $\Leftarrow$ Vote56 (C) $\Leftarrow$ Party56 (D) $\Leftarrow$ Religion (E)

Since DIGRAM assumes that the recursive structure is included as part of the definitions of variables, the only thing we need to do is to impose assumptions of pairwise conditional dependence and independence

Several commands are available to do this.

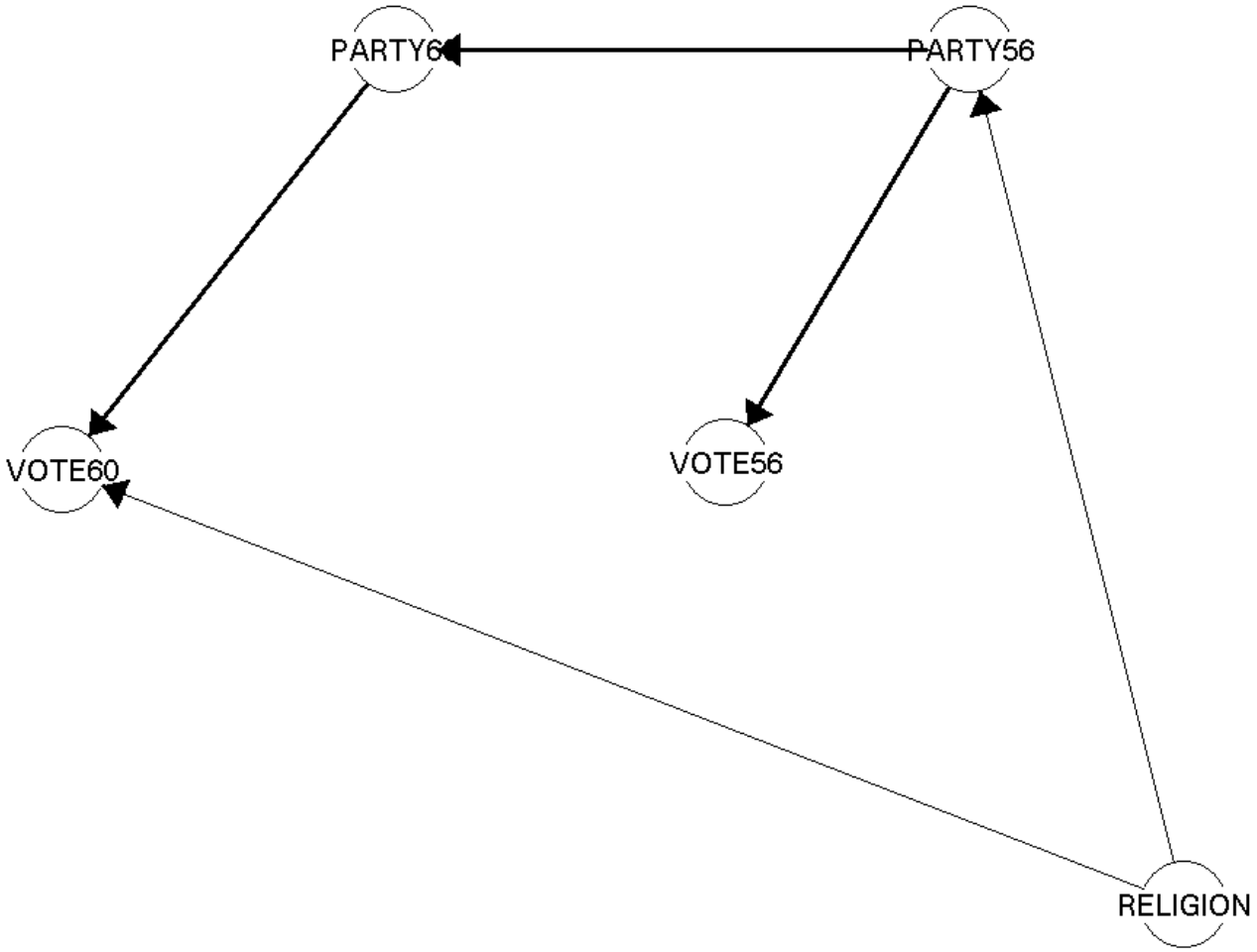
DIGRAM commands for definition of graphical models

New 1 Defines a model of independent variables

Fix AB BC CD assumes that A&B, B&C and C&D *are* conditionally dependent

ADD AE DE assumes that Vote60 and Party56 may be conditionally dependent on religion

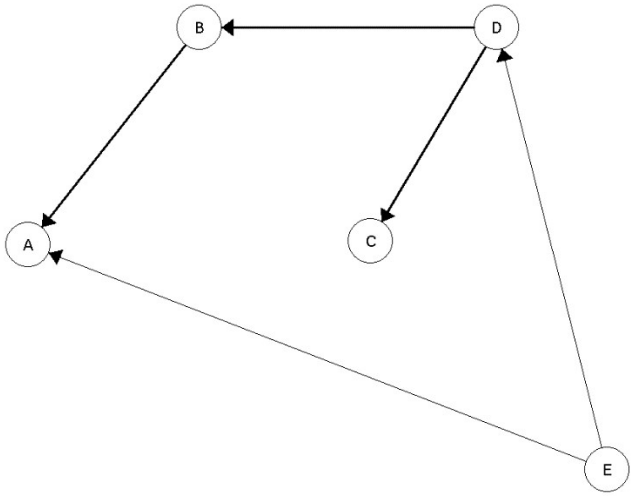
The result is a simple chain graph model defined by a directed acyclic graph (a DAG) that some would interpret as a causal model



The Vote60 ⇐ Religion is the focal relationship of interest

The Party56 ⇐ Religion association is a secondary issue

The Statistical model



Missing edges refer to conditional independence given all other current or prior variables, e.g.

$$A \perp C \mid BDE$$

$$A \perp D \mid BCE$$

$$B \perp C \mid DE$$

$$B \perp E \mid CD$$

$$C \perp E \mid DE$$

$$P(A,B,C,D,E) = P(A|B,C,D,E)\, P(B|C,D,E)\, P(C|D,E)\, P(D|E)P(E)$$

Let us assume that the model is correct and that we intend to test that Vote 60 (A) and religion (E) are conditionally independent.

The definition of the model claims that we have to elaborate the relationship in an 5-dimensional contingency table.

$$A \perp E \mid BCD$$

The global Markov properties tell us that we can simplify this relationship because we only need to elaborate with B because the table is collapsible over C and D

$$A \perp E \mid BCD \Rightarrow A \perp E \mid B$$

Graphical models may reduce challenging high-dimensional problems to a simple test of conditional independence in contingency tables with few variables.

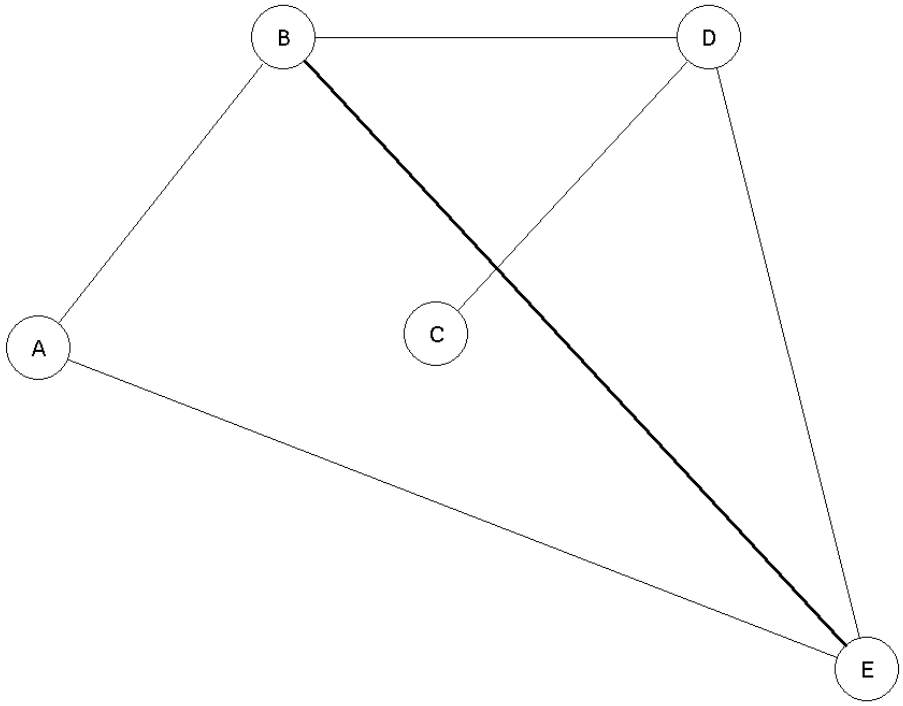
How to identify the global Markov properties

Define an undirected graphical model by moralizing the graph

Find all direct paths between D and I in the moral graph

It follows from the theorem of global Markov properties than D and I are conditional independent given a subset of nodes (variables) that may obstruct all paths between D and I

The moral graph



B & D has to be "married" because A is a "child" of both variables

To get from E to A in the moral graph we have to go through B. Cutting the path at this point therefore separates D from I from which it follows that A ⊥ E | B and that ABCDE model is collapsible with respect to the AE interaction.

Elaboration and specification

```
-------------------------------------------------------------------------------
                          p-values                  p-values (1-sided)
Hypothesis       X²  df asymp exact 99% conf.int. Gamma asymp exact 99% conf.int. nsim
-------------------------------------------------------------------------------
 1:A&E|B       75.7   3 0.000 0.000 0.000 - 0.007  0.73 0.000 0.000 0.000 - 0.007 1000 xx ++
-------------------------------------------------------------------------------
```

```
-----------------------------------------------------------
** Local testresults for strata defined by  PARTY60 (B) **
                         p-values      p-values (1-sided)
 B:   PARTY60    X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:Democrat   27.03    1 0.0000 0.0000   0.75 0.0000 0.0000
 2:Independ   12.23    1 0.0005 0.0030   0.57 0.0002 0.0020
 3:Republic   36.41    1 0.0000 0.0000   0.85 0.0007 0.0000
-----------------------------------------------------------
```

Very strong association for all Party60 categories.

The differences between the gamma coefficients for the separate parties are not significant.

However, can we trust the model?

Adding edges will change the global Markov properties.

Test of all the global Markov properties induced by the graph rejects the majority of the assumptions. We need better ways to define a graphical model.

Hypothesis	X^2	df	p-values asymp	p-values exact	Gamma	p-values (1-sided) asymp	p-values (1-sided) exact	95% confidence interval	nsim	n	
1:A&C|BE	84.1	6	0.000	0.000	0.78	0.000	0.000	[0.58 - 0.99]	1000	786	xx ++
2:A&D|BE	30.3	12	0.003	0.009	0.33	0.002	0.001	[0.11 - 0.56]	1000	786	x ++
3:B&C|D	46.9	6	0.000	0.000	0.53	0.000	0.000	[0.30 - 0.76]	1000	786	xx ++
4:B&E|D	67.0	6	0.000	0.000	0.57	0.000	0.000	[0.38 - 0.76]	1000	786	xx ++
5:C&E|D	6.9	3	0.074	0.078	-0.03	0.382	0.382	[-0.24 - 0.18]	1000	786	

Benjamini Hochberg rejects if p < 0.040 for FDR = 0.05
and p < 0.007 for FDR = 0.01

Significance of
X^2 xx : FDR = 0.01 x : FDR = 0.05
Gamma ++/-- : FDR = 0.01 +/- : FDR = 0.05

DIGRAM has procedures for model search that we will illustrate in the following DEMO on the a little more challenging EJH5 example.