A guided tour through DIGRAM 2.0

Analysis of contingency tables by chain graph models

Svend Kreiner
Dept. of Biostatistics, University of Copenhagen
2012

This is the first of what is supposed to be a series of guided tours through the different types of statistical analyses supported by DIGRAM.

We will use the EJH5 project to illustrate the analysis of contingency tables by chain graph models in DIGRAM. The project contains information from a panel study of living conditions in Denmark, containing information collected from adolescents in the eighth grade of the Danish public school in 1967, information on education later in life and finally information on health, income and unemployment collected in 1992. The analysis of these data will attempt to answer the following questions:

1) Does intelligence have a direct effect on outcome variables collected 25 years after the measurement of intelligence?
2) In what way does the social background modify the effect of intelligence?

To answer these questions we will attempt to develop a chain graph model and see, what the model can tell us on the effect of intelligence.

During the guided tour of the analysis of the EJH5 project you will learn how to

a) describe data,
b) create and analyze tables,
c) define, display and analyze graphical models,
d) generate tables and hypotheses by analysis of graphs,
e) test model based hypotheses,
f) select models,
g) check models,
h) describe relationships,
i) analyze data by loglinear models.

The guided tour is intended to give you a rough idea about DIGRAM's capabilities. Technical details will be skipped and many commands will not be discussed here. The complete set of commands will (in time) be described in the user guide and technical details documented in separate papers dedicated to special topics.

Some of the procedures that we are going to show you depend on a graphical model that we have some belief in. The model is shown in Figure 1. We will, as part of the guided tour, show you how the model was assembled. The 11 variables included in the model is shown in Figure 2.

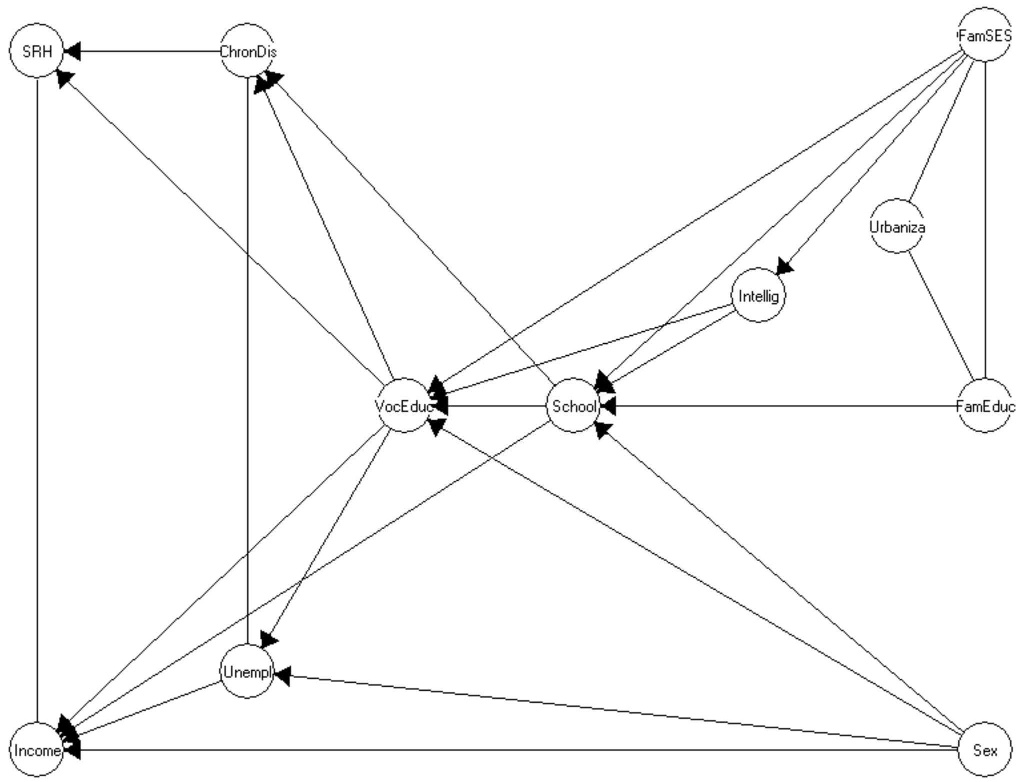


Figure 1. The EJH5 project graph

```
D:   Income  -   7 ordinal categories
C:      SRH  -   4 ordinal categories
A: ChronDis  -   2 ordinal categories
B:   Unempl  -   2 ordinal categories
F:  VocEduc  -   5 ordinal categories
G:   School  -   4 ordinal categories
I: Intellig  -   5 ordinal categories
J: Urbaniza  -   4 ordinal categories
K:   FamSES  -   5 ordinal categories
L:  FamEduc  -   6 ordinal categories
M:      Sex  -   2 ordinal categories

CAUSAL/RECURSIVE STRUCTURE
D,C <- A,B <- F <- G <- I <- J,K,L,M

+-----------------------------------------+
|  D   Income |  C      SRH |  A ChronDis |
|-------------+-------------+-------------|
|  1    1 - 2 |  1 VeryGood |  1     None |
|  2 100.000- |  2     Fair |  2       1+ |
|  3 150.000- |  3 LessFair |             |
|  4 200.000- |  4      Bad |             |
|  5 250.000- |             |             |
|  6    7 - 8 |             |             |
|  7   9 - 11 |             |             |
+-----------------------------------------+
+-----------------------------------------+
|  B   Unempl |  F  VocEduc |  G   School |
|-------------+-------------+-------------|
|  1 < 1 year |  1     LANG |  1    0 - 2 |
|  2 1+ years |  2 MELLEMLA |  2    3 - 4 |
|             |  3     KORT |  3    5 - 8 |
|             |  4 LæRLINGE |  4   9 - 12 |
|             |  5    INGEN |             |
+-----------------------------------------+
+-----------------------------------------+
|  I Intellig |  J Urbaniza |  K   FamSES |
|-------------+-------------+-------------|
|  1        - |  1 KøBENHAV |  1        I |
|  2    26-30 |  2 PROVINSB |  2       II |
|  3    31-35 |  3   MINDRE |  3      III |
|  4    36-40 |  4 LANDKOMM |  4       IV |
|  5      41+ |             |  5        V |
+-----------------------------------------+
+---------------------------+
|  L  FamEduc |  M      Sex |
|-------------+-------------|
|  1   HØJERE |  1     Male |
|  2    ANDEN |  2   Female |
|  3 LÆRLINGE |             |
|  4  KORTERE |             |
|  5 TILLÆRIN |             |
|  6    INGEN |             |
+---------------------------+
```

Figure 2. The variables included in the EJH5 project

Examining and describing data

SYS and TAB files are generated the first time DIGRAM opens a new project. Before you proceed with the analysis you should however examine data to check whether variab-les have been properly defined. Three commands are available for initial examination and description of data:

FREQUENCIES <variables>	produces marginal frequencies for the variables
DESCRIBE <variable>	generate tables showing the conditional distribution of the remaining project variables given the variable referred to by the parameter of the DESCRIBE command.
GAMMA M	produces a matrix of marginal γ coefficients measuring the correlations among the project variables
COLLAPS <variable>	Examines whether or not some of the categories of the variable given as a parameter to the COLLAPS command are collapsible in the sense that differences between conditional distributions of other variables given these variables. Collapsibility across categories will be examined for all polytomous variables if the COLLAPS command is issued without parameters.

Remember that you only have to include the first three characters of commands.

Note that DESCRIBE and COLLAPS with more than one variable and GAMMA without the M parameter result in somewhat different analyses then those shown here.

Parts of the description of the EJH5 data description are shown in Figures 3 to 6. The p-values associated with Goodman and Kruskall's γ coefficients are 1-sided. We recognize that the use of one-sided p-values is somewhat unorthodox, but trust that you as a user will be able to multiply p-values with two whenever two-sided p-values are appropriate.

```
  +-----------+
  |           |
  | D: Income |
  |           |
  +-----------+

Reference no.   1
Variable no.   14

This variable was categorized

  Values  D Count      Pct  CumPct
---------------------------------
   < 100  1     160   6.8     6.8
 100-200  2     397  17.0    23.8
 150-200  3     636  27.2    50.9
 200-250  4     600  25.6    76.6
 250-300  5     239  10.2    86.8
 300-400  6     206   8.8    95.6
    400+  7     104   4.4   100.0


TOTAL  2342

Missing:
BELOW   809
```

Figure 3. ***FREQUENCIES D:*** *Marginal frequencies are generated for all variables. This figure only shows the frequencies for the first project variable taken from column 14 of the original data matrix.*

The marginal associations among variables can be described in two different ways. The DESCRIBE command produces tables and the GAMMA command produces a matrix containing marginal gamma coefficients measuring the association among the ordinal variables of the project.

```
  +-------------------------------------+
  |                                     |
  | ****  Description of D - Income **** |
  |                                     |
  +-------------------------------------+


Income  is directly associated with

   C -      SRH

Income  is directly dependent on

   B -   Unempl
   F -  VocEduc
   G -   School
   M -      Sex

Conditional distributions given Income(D)

                       Income(D)
                    < 100 100-2 150-2 200-2 250-3 300-4 400+   Current status
-------------------------------------------------------------------------------
                        %     %     %     %     %     %     %
         Intellig(I)                                               *** cond.ind.

                    -  13.8  10.6  12.9   9.3   8.4   4.4   4.8
                26-30  14.4  13.9  15.6  13.5  10.9  10.2   1.0
                31-35  23.1  25.2  23.1  18.8  16.7  13.6  13.5
                36-40  21.9  27.2  23.3  22.8  18.4  25.2  15.4
                  41+  26.9  23.2  25.2  35.5  45.6  46.6  65.4

                Total   160   397   636   600   239   206   104

-------------------------------------------------------------------------------
***: At least one p-value less than or equal to 0.001
 **: At least one p-value less than or equal to 0.01
  *: At least one p-value less than or equal to 0.05
```

Figure 19. ***DESCRIBE D:*** *The description includes a summary of direct connections to D in the current model (Figure 1) and the conditional distributions of all project variables given Income (D). (The conditional distribution of intelligence (I) given Income is shown here). The current status refers to the relationships between variable defined by the current graphical project model.*

Marginal Gamma coefficients

	D	C	A	B	F	G	I	J	K	L	M
D: Income	.	-0.230	-0.184	-0.462	-0.388	0.267	0.198	-0.107	-0.176	-0.192	-0.605
C: SRH	-0.230	.	0.820	0.244	0.193	-0.159	-0.055	-0.026	0.109	0.088	-0.051
A:ChronDis	-0.184	0.820	.	0.263	0.202	-0.130	-0.070	-0.064	0.145	0.045	-0.095
B: Unempl	-0.462	0.244	0.263	.	0.287	-0.169	-0.153	0.028	0.103	0.069	0.179
F: VocEduc	-0.388	0.193	0.202	0.287	.	-0.660	-0.428	0.110	0.354	0.328	0.019
G: School	0.267	-0.159	-0.130	-0.169	-0.660	.	0.528	-0.163	-0.456	-0.464	0.151
I:Intellig	0.198	-0.055	-0.070	-0.153	-0.428	0.528	.	-0.123	-0.253	-0.248	0.028
J:Urbaniza	-0.107	-0.026	-0.064	0.028	0.110	-0.163	-0.123	.	0.136	0.489	0.006
K: FamSES	-0.176	0.109	0.145	0.103	0.354	-0.456	-0.253	0.136	.	0.608	0.031
L: FamEduc	-0.192	0.088	0.045	0.069	0.328	-0.464	-0.248	0.489	0.608	.	0.035
M: Sex	-0.605	-0.051	-0.095	0.179	0.019	0.151	0.028	0.006	0.031	0.035	.

Figure 4. ***GAMMA M:*** *Gamma coefficients measuring the marginal associations among ordinal variables.*

The number of categories associated with the variables is an issue, since a large number of categories will generate large sparse tables where we have problems both with the asymptotic properties and with the power of test statistics. DIGRAM has methods that take care of the problems with asymptotics, but power is still a problem. For this reason it is generally recommended to try to avoid using more categories than necessary. During the initial examination of data you may perform an analysis of category collapsibility to check whether some of the categories can be collapsed. To invoke this analysis you must use the COLLAPS command as shown in Figure 5.

```
+--------------------------------------------+
|                                            |
| Analysis of collapsibility of I-categories |
|                                            |
+--------------------------------------------+


      I:Intellig     Categories =  1-2  3  4  5

Test against      chi**2   df    p     gamma     P*
   D   Income       7.8    6 0.249   0.072  0.126
   C      SRH       4.6    3 0.204  -0.092  0.128
   A ChronDis       0.5    1 0.469  -0.067  0.235
   B   Unempl       0.1    1 0.739   0.028  0.370
   F  VocEduc       8.9    4 0.063  -0.116  0.050
   G   School       9.3    3 0.026   0.131  0.018 *  +
   J Urbaniza       2.2    3 0.528  -0.081  0.102
   K   FamSES       1.6    4 0.816   0.017  0.396
   L  FamEduc       2.5    5 0.778  -0.036  0.306
   M      Sex       3.2    1 0.073   0.133  0.036    +

               - output omitted here -

      I:Intellig     Categories =  1  2  3  4-5

Test against      chi**2   df    p     gamma     P
   D   Income      41.5    6 0.000   0.226  0.000 ** ++
   C      SRH       2.4    3 0.487  -0.056  0.156
   A ChronDis       2.6    1 0.110  -0.101  0.057
   B   Unempl       1.5    1 0.224  -0.073  0.113
   F  VocEduc      71.4    4 0.000  -0.302  0.000 ** --
   G   School     101.0    3 0.000   0.409  0.000 ** ++
   J Urbaniza      11.0    3 0.012  -0.131  0.001 *  --
   K   FamSES      57.1    4 0.000  -0.284  0.000 ** --
   L  FamEduc      41.3    5 0.000  -0.247  0.000 ** --
   M      Sex       4.3    1 0.038  -0.104  0.019 *  -
```

* *p-values for γ coefficientsa are 1-sided*

Figure 5. ***COLLAPS I:*** *Examination of collapsibility across categories of Intelligence (I). two-way tables relating Intelligence to other variables. For ordinal variables collapsibility will only be considered for adjacent categories. There is little evidence against collapsing the first two categories of I, but strong evidence against collapse of the two last categories.*

Each of the tests in Figure 5 is a test comparing the distribution of a variable in two different groups defined by adjacent Intelligence categories. The first test shows that the differences between the distributions of income in the first two Intelligence categories are not significant ($\chi^2 = 7.8$, df = 6, p = 0.249; $\gamma = 0.249$, p = 0.126).

In addition to the analysis shown in Figure 5, DIGRAM also performs an analysis of category collapsibility in two way tables with Intelligence and all the variables that are directly connected to Intelligence in the current project model. Analysis of collapsibility is illustrated in the next section of these notes.

Creating and analyzing multidimensional contingency tables 1: Tests of conditional independence

DIGRAM is first of all a program for analysis of multidimensional contingency tables. In this section we will show you the basics of creating and analyzing tables. You will learn how to

1) tabulate data and display tables,
2) create and test hypotheses of conditional independence,
3) fit loglinear models,
4) test for collapsibility across categories in multidimensional tables.

We distinguish between model-based and model-free tables. DIGRAM fits chain graph models to the complete set of project variables. Collapsibility properties and global Markov properties of these models generate marginal tables and models in which specific problems may be addressed. At any point of time during the analysis you may ask for tables generated by the model. You are, however, not restricted to looking at model-based tables. You can of course create any table you want to and ask for any analysis of this table, as long as the problems to be addressed respect the recursive structure of the data.

This section only describes model-free analyses of tables. Model based analysis will be described after the next section on definition and modification of graphical models.

The first problem we will address is whether there is marginal effect of Intelligence on Income and the degree to which Sex and the socioeconomic status of the family changes our understanding of the effect of intelligence on Income.

TABULATE DI   Creates the marginal table[1].

Figure 6 shows you what you will see when the table is ready.

```
A new table has been created.

The marginal DI model:

Variables   DI
        : D **
        : I **


  +-----------------------------+
  |                             |
  | Report on missing responses |
  |                             |
  +-----------------------------+

D   Income  Observed =   2342  Missing =    809
I Intellig  Observed =   3151  Missing =      0

Number of cases with complete responses = 2342 out of 3151

Marginal loglinear model DI   Fixed: DI
```

Figure 6. **TABULATE DI:** Report on the marginal model after tabulating

The report after tabulating includes information on missing values and on the number of persons with complete responses on the variables of the table. You can obtain such reports on all variables at any point of time by invoking the **MISS <variables>** command.

[1] Previous versions of DIGRAM permitted only tables with up to eight variables. This limitation has been relaxed, but tables are still limited with respect to the number of cells in the table. Use **SHOW L** to get information on the limitations in DIGRAM.

Whenever you create a new table, DIGRAM will derive the marginal model for the table from the chain graph model for the complete set of variables. In the marginal DI model, DIGRAM regards the DI association as a fixed interaction since an insignificant test of marginal independence does not imply that there should be no edge between D and I in the complete model.

This does not mean that you should not test the hypothesis of marginal independence. In fact, DIGRAM sets this hypothesis up for you since it is the only hypothesis of independence that can be tested in the DI-table. To test this hypothesis you just have to invoke the TEST command. The result is shown in Figure 7.

```
Table 1. The DI distribution.

     +Intellig
     | | D:--Income                                |
     I | < 100 100-2 150-2 200-2 250-3 300-4  400+ | TOTAL |
-------+-------------------------------------------+-------+
     - |    22    42    82    56    20     9     5 |   236 |
 26-30 |    23    55    99    81    26    21     1 |   306 |
 31-35 |    37   100   147   113    40    28    14 |   479 |
 36-40 |    35   108   148   137    44    52    16 |   540 |  X² = 141.9
   41+ |    43    92   160   213   109    96    68 |   781 |  df =  24
-----------------------------------------------------------+   p = 0.000
 TOTAL |   160   397   636   600   239   206   104 |  2342 | Gam =  0.20
-----------------------------------------------------------+   p = 0.000
```

Figure 7. **TEST:** Two-way tables are always printed in connection with a test of independence. Parameters may be added, controlling the output as described below

Figure 7 shows that there is a moderate, but highly significant, effect of Intelligence on Income, to see whether the effect exists for both men and women or whether the effect depends on the social background you must perform a stratified analysis where the Intelligence-Income association is elaborated with respect to Gender (M) and Socio economic status (K).

The first thing to do is to create the table,

TABULATE DIGM

No hypotheses are automatically created, since there are several hypotheses of conditional independence that can be defined for a four dimensional table. To test the hypothesis that Intelligence and Income are conditionally independent given gender and socio economic statue we first invoke the

HYPOTHESES DI

command followed by the

TEST T

command. The results can be seen in Figures 8a – 8.c.

The output from a test of conditional independence consists of several parts:

1) First the table. This output is optional. It is only included if you add a T parameter or one of the other TEST parameters to the TEST command. Figures 8a and 8b show the first three socio economic strata for respectively men and women. Note that test statistics are calculated for each strata of the table.
2)

```
Table 1. The DI|KM distribution.

Conditioning variables:

+---------------------------+
|  K   FamSES |  M      Sex |
|-------------+-------------|
|  1        I |  1     Male |
|  2       II |  2   Female |
|  3      III |             |
|  4       IV |             |
|  5        V |             |
+---------------------------+

+-----Sex
|+--FamSES
||     +Intellig
||     | | D:--Income                                 |
MK     I | < 100 100-2 150-2 200-2 250-3 300-4  400+ | TOTAL |
---------+-------------------------------------------+-------+
11     - |     0     0     0     3     1     1     0 |     5 |
     row%|   0.0   0.0   0.0  60.0  20.0  20.0   0.0 | 100.0 |
   26-30 |     0     0     2     1     0     1     0 |     4 |
     row%|   0.0   0.0  50.0  25.0   0.0  25.0   0.0 | 100.0 |
   31-35 |     2     0     1     3     0     1     1 |     8 |
     row%|  25.0   0.0  12.5  37.5   0.0  12.5  12.5 | 100.0 |
   36-40 |     0     2     0     1     0     2     0 |     5 |
     row%|   0.0  40.0   0.0  20.0   0.0  40.0   0.0 | 100.0 |
     41+ |     1     3     2     3     4     6     8 |    27 |
     row%|   3.7  11.1   7.4  11.1  14.8  22.2  29.6 | 100.0 |  X² =  33.3
-------------------------------------------------------------+  df =  24
   TOTAL |     3     5     5    11     5    11     9 |    49 |   p = 0.098
     row%|   6.1  10.2  10.2  22.4  10.2  22.4  18.4 | 100.0 | Gam =  0.28
-------------------------------------------------------------+   p = 0.016
12     - |     0     2     2     0     2     0     0 |     6 |
     row%|   0.0  33.3  33.3   0.0  33.3   0.0   0.0 | 100.0 |
   26-30 |     0     0     1     1     1     1     0 |     4 |
     row%|   0.0   0.0  25.0  25.0  25.0  25.0   0.0 | 100.0 |
   31-35 |     0     0     4     6     3     5     2 |    20 |
     row%|   0.0   0.0  20.0  30.0  15.0  25.0  10.0 | 100.0 |
   36-40 |     1     2     3     8     6     6     3 |    29 |
     row%|   3.4   6.9  10.3  27.6  20.7  20.7  10.3 | 100.0 |
     41+ |     1     1     4    11    14    18    14 |    63 |
     row%|   1.6   1.6   6.3  17.5  22.2  28.6  22.2 | 100.0 |  X² =  32.1
-------------------------------------------------------------+  df =  24
   TOTAL |     2     5    14    26    26    30    19 |   122 |   p = 0.124
     row%|   1.6   4.1  11.5  21.3  21.3  24.6  15.6 | 100.0 | Gam =  0.34
-------------------------------------------------------------+   p = 0.000
-------------------------------------------------------------+   p = 0.000
13     - |     2     5     8    13     6     3     3 |    40 |
     row%|   5.0  12.5  20.0  32.5  15.0   7.5   7.5 | 100.0 |
   26-30 |     1     4    17    19    11     6     1 |    59 |
     row%|   1.7   6.8  28.8  32.2  18.6  10.2   1.7 | 100.0 |
   31-35 |     2     6    15    26     9    13     4 |    75 |
     row%|   2.7   8.0  20.0  34.7  12.0  17.3   5.3 | 100.0 |
   36-40 |     2     5    18    22     9     9     5 |    70 |
     row%|   2.9   7.1  25.7  31.4  12.9  12.9   7.1 | 100.0 |
     41+ |     3     5    18    37    33    36    25 |   157 |
     row%|   1.9   3.2  11.5  23.6  21.0  22.9  15.9 | 100.0 |  X² =  43.0
-------------------------------------------------------------+  df =  24
   TOTAL |    10    25    76   117    68    67    38 |   401 |   p = 0.010
     row%|   2.5   6.2  19.0  29.2  17.0  16.7   9.5 | 100.0 | Gam =  0.28
-------------------------------------------------------------+   p = 0.000
```

Figure 8a. **HYPOTHESES DI and TEST T:** Use the T parameter after TEST, if you want to see the table. Figure 8.a shows the first part of the table.

```
+-----Sex
|+--FamSES
||     +Intellig
||     | | D:--Income                              |
MK     I | < 100 100-2 150-2 200-2 250-3 300-4  400+ | TOTAL |
---------+-----------------------------------------+-------+
21     - |     0     0     1     1     2     0     0 |     4 |
     row%|   0.0   0.0  25.0  25.0  50.0   0.0   0.0 | 100.0 |
   26-30 |     2     0     2     0     0     0     0 |     4 |
     row%|  50.0   0.0  50.0   0.0   0.0   0.0   0.0 | 100.0 |
   31-35 |     0     0     5     2     0     0     0 |     7 |
     row%|   0.0   0.0  71.4  28.6   0.0   0.0   0.0 | 100.0 |
   36-40 |     0     2     3     1     1     1     0 |     8 |
     row%|   0.0  25.0  37.5  12.5  12.5  12.5   0.0 | 100.0 |
     41+ |     2     5     5    10     5     4     1 |    32 |
     row%|   6.3  15.6  15.6  31.3  15.6  12.5   3.1 | 100.0 |  X² =  31.4
-----------------------------------------------------------+  df =  24
   TOTAL |     4     7    16    14     8     5     1 |    55 |   p = 0.144
     row%|   7.3  12.7  29.1  25.5  14.5   9.1   1.8 | 100.0 | Gam =  0.21
-----------------------------------------------------------+   p = 0.092
22     - |     0     0     1     2     0     0     0 |     3 |
     row%|   0.0   0.0  33.3  66.7   0.0   0.0   0.0 | 100.0 |
   26-30 |     0     2     0     2     0     1     0 |     5 |
     row%|   0.0  40.0   0.0  40.0   0.0  20.0   0.0 | 100.0 |
   31-35 |     0     3     2     3     0     0     1 |     9 |
     row%|   0.0  33.3  22.2  33.3   0.0   0.0  11.1 | 100.0 |
   36-40 |     2     9     3    11     1     1     1 |    28 |
     row%|   7.1  32.1  10.7  39.3   3.6   3.6   3.6 | 100.0 |
     41+ |     2     7    15    18     5     3     2 |    52 |
     row%|   3.8  13.5  28.8  34.6   9.6   5.8   3.8 | 100.0 |  X² =  18.0
-----------------------------------------------------------+  df =  24
   TOTAL |     4    21    21    36     6     5     4 |    97 |   p = 0.803
     row%|   4.1  21.6  21.6  37.1   6.2   5.2   4.1 | 100.0 | Gam =  0.11
-----------------------------------------------------------+   p = 0.172
23     - |     5     8     8     5     0     0     0 |    26 |
     row%|  19.2  30.8  30.8  19.2   0.0   0.0   0.0 | 100.0 |
   26-30 |     7    20    18     9     0     1     0 |    55 |
     row%|  12.7  36.4  32.7  16.4   0.0   1.8   0.0 | 100.0 |
   31-35 |    13    38    40    15     6     1     1 |   114 |
     row%|  11.4  33.3  35.1  13.2   5.3   0.9   0.9 | 100.0 |
   36-40 |     8    38    40    27     6     5     1 |   125 |
     row%|   6.4  30.4  32.0  21.6   4.8   4.0   0.8 | 100.0 |
     41+ |    14    29    44    50    12     7     1 |   157 |
     row%|   8.9  18.5  28.0  31.8   7.6   4.5   0.6 | 100.0 |  X² =  36.8
-----------------------------------------------------------+  df =  24
   TOTAL |    47   133   150   106    24    14     3 |   477 |   p = 0.046
     row%|   9.9  27.9  31.4  22.2   5.0   2.9   0.6 | 100.0 | Gam =  0.25
-----------------------------------------------------------+   p = 0.000
```

Figure 8b. **HYPOTHESES DI and TEST T:** Use the T parameter after TEST, if you want to see the table. Figure 8.b shows the last part of the table.

```
****  Summary of results  ****

----------------------------------------------------------
                        p-values   p-values (1-sided)
Hypothesis       X²  df asymp exact Gamma asymp exact
----------------------------------------------------------                [I]
 1:D&I|KM     319.5 236 0.000        0.24 0.000           xx ++
----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by   FamSES (K) **
                         p-values      p-values (1-sided)
 K:   FamSES   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:        I  64.65   48 0.0547          0.24 0.0099
 2:       II  50.11   48 0.3897          0.26 0.0003
 3:      III  79.84   48 0.0026          0.26 0.0000
 4:       IV  75.09   48 0.0075          0.23 0.0000
 5:        V  49.85   44 0.2521          0.18 0.0001
-----------------------------------------------------------
                                                                          [II]
-----------------------------------------------------------
** Local testresults for strata defined by      Sex (M) **
                         p-values      p-values (1-sided)
 M:      Sex   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    Male 192.37  120 0.0000          0.27 0.0000
 2:  Female 127.16  116 0.2254          0.21 0.0000
-----------------------------------------------------------

+-------------------------------------------------------+
|                                                       |
|   Summary of gamma coefficients in separate strata    |
|                                                       |
| Significance evaluated by asymptotic 2-sided p-values |
|                                                       |
+-------------------------------------------------------+
                                                                          [III]

  gamma    p >0.05      0.01<p<=0.05      p<0.01
--------------------------------------------------------------
Negative       0              0              0
Positive       2              3              5
```

Figure 8c. **HYPOTHESES DI and TEST T:** Summary of test results.

The result of the test of conditional independence of Income (D) and Intelligence (I) is summarized in Figure 8c. The summary consists of three parts.

I: First the global (overall) test result summarizing the results from the different strata. The default test statistics is the sum of the χ^2 tests for each of the different strata and the partial γ coefficient (a weighted mean of the γ coefficients from the different strata). Both test statistics are highly significant. The partial γ coefficient is a little stronger than the marginal γ coefficient in Figure 7.

II: Second, local test statistics calculated separately for different strata defined by one of the conditioning variables. If the conditioning set of variables consists of more than one variable, the global test statistics are partial test statistics. The χ^2 test in the first FamSES stratum (Social class I) is the sum of the χ^2 tests for men and women in this stratum while the γ coefficient is a weighted mean of the γ coefficients for men and women in Social Class I.

III: Finally, the output includes a table showing the distribution of the γ coefficients from the different strata with respect to the sign of the coefficients and the assessment of significance. All 10 γ coefficients are positive: 8 significant and two insignificant.

Parameters for tests of conditional independence.

At this point we have to talk a little about the options available for the statistical tests. You can see in Figure 8c that the tables with global test results have columns with space for exact p-values. Such p-values are of interest when you are analyzing large sparse tables with lots of zeros, since we know that the asymptotic distributions of test statistics are extremely poor approximations of the exact distributions of the test statistics. For this reason DIGRAM offers exact p-values, or rather much better unbiased Monte Carlo approximations of the exact p-values, than the asymptotic p-values.

Also, in Table 8c, p-values associated with γ coefficients are 1-sided and not 2-sided. There are good reasons for that, but it is of course op to you to decide whether you disagree with DIGRAM's default parameters. For this reason there are a number of commands that will let you change these defaults. These parameters are collected in Table 1.

Table 1. Test parameters

Command	Effect	Default
EXACT <Nsim feed>	Monte Carlo approximation by analysis nsim random tables	1000 9
SEQUENTIAL <Nsim feed Alpha>	Sequential Monte Carlo approximation by analysis nsim random tables	1000 9 0.05
REPEATED <Nsim feed Alpha risk>	Repeated Monte Carlo approximation by analysis nsim random tables	1000 9 0.05 0.001
ASYMPTOTIC	p-values approximated by the asymptotic distribution of test statistics	
ONE	1-sided assessment of significance	
TWO	2-sided assessment of significance	
GLOBAL	Global test results only	
LOCAL	Local test results	
CHI	χ^2 Test of conditional independence against a saturated alternative	
DEVIANCE	Likelihood ratio test of conditional independence against a saturated alternative	
PARTIAL	Likelihood ratio test of conditional independence against a log-linear model without higher order interaction	

Monte Carlo testing may be time-consuming. The Sequential and repeated Monte Carlo tests are included in order to save you some time. Sequential Monte Carlo tests stop, when it is absolutely sure that the test result will not be significant at the given critical level (alpha). The repeated Monte Carlo test stops when the risk that the test result will be significant is small (risk). Figure 8d repeats the results from Figure 8c, but now with Monte Carlo approximation of exact p-values of 2-sided tests based on a random sample of 1000 tables. Note that the global test results also present 99% confidence intervals of the estimates of the exact p-values. The test against the partial 2-factor alternative has been included. Note that monte Carlo approximation of the exact p-value is very time-consuming, since DIGRAM has to fit a loglinear model to each of the 1000 random tables

```
****  Summary of results  ****

NSIM = 1000 tables generated for exact p-values

--------------------------------------------------------------------------------------
                         p-values                 p-values (2-sided)
Hypothesis       X²  df asymp exact 99% conf.int. Gamma asymp exact 99% conf.int. nsim
--------------------------------------------------------------------------------------
 1:D&I|KM     319.5 236 0.000 0.000 0.000 - 0.007  0.24 0.000 0.000 0.000 - 0.007 1000
xx ++
--------------------------------------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by   FamSES (K) **
                          p-values      p-values (1-sided)
 K:   FamSES   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:       I  64.65   48 0.0547 0.0620   0.24 0.0099 0.0130
 2:      II  50.11   48 0.3897 0.3910   0.26 0.0003 0.0000
 3:     III  79.84   48 0.0026 0.0030   0.26 0.0000 0.0000
 4:      IV  75.09   48 0.0075 0.0050   0.23 0.0000 0.0000
 5:       V  49.85   44 0.2521 0.2350   0.18 0.0001 0.0010
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by      Sex (M) **
                          p-values      p-values (1-sided)
 M:      Sex   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    Male 192.37  120 0.0000 0.0000   0.27 0.0000 0.0000
 2:  Female 127.16  116 0.2254 0.2600   0.21 0.0000 0.0000
-----------------------------------------------------------


Test against two-factor association
             lr =    142.1
             df =      24
  asymptotic p  =   0.0000
       exact p  =   0.0000
```

Figure 8d. **EXACT, TWO, PARTIAL**: Figure 8c revisited. Assessment of significance with Monte Carlo estimates of 2-sided p-values. The test against the partial 2-factor alternative has been included.

We return later to take a look at a number of advanced methods for analysis of contingency tables. Before that, the next item on the agenda is the test of the model-based hypotheses

Model based hypotheses

We now turn to the definition of tables required for analysis of model based hypotheses.

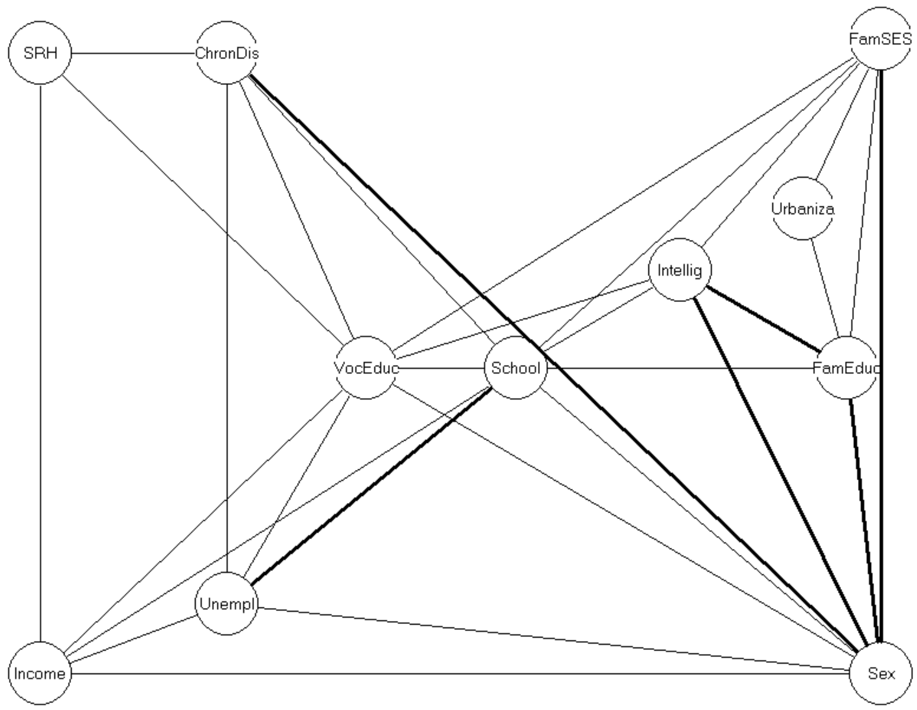
Model based hypotheses are referred to as GMP hypotheses, since they are defined by the global Markov properties of the graphical models. DIGRAM has three commands that you may use to generate such hypotheses

SEPARATE **<variable pairs>** defines GMP hypotheses in regression graphs

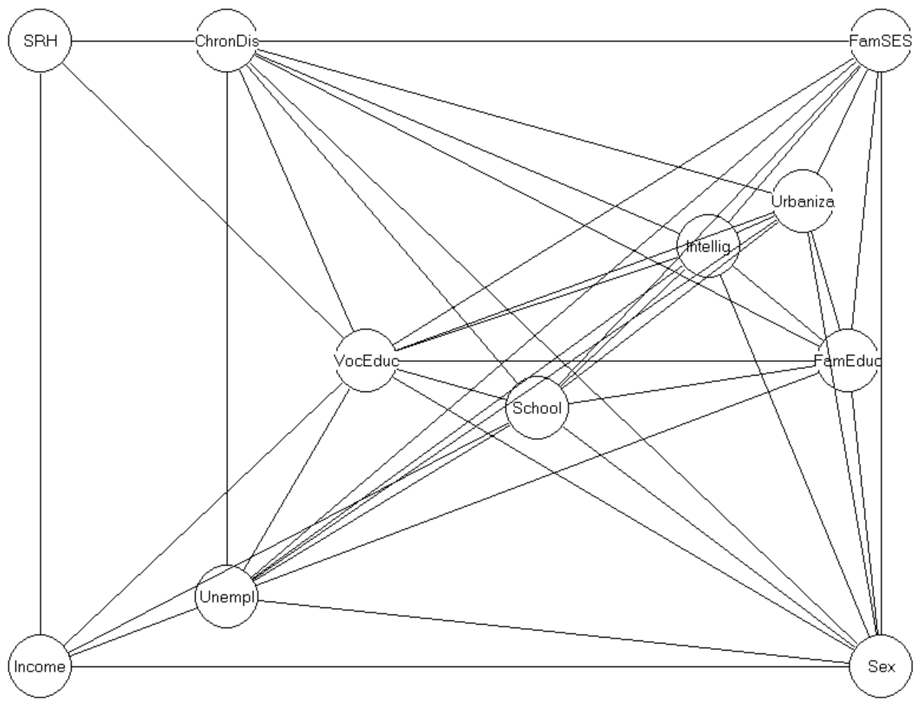
GMP **<variable pairs>** defines GMP hypotheses in moral graphs

REDUCE **<variable pairs>** defines GMP hypotheses by decomposition of Regression graphs

The regression graph and the moralized graph are shown in Figures 9a and 9b. Both graphs are undirected graphs. In the regression graph (corresponding to the regression model describing the conditional distribution of Income and SRH given all other variables) all the nodes of the explanatory variables are directly connected. The set of explanatory variables is, in other words, a clique in the regression graph. In the moralized graph, edges have only been added between the explanatory variables if they are required for moralization of the graph.



(a)



(b)

Figure 9. The moral graph (a) and the regression graph (b) associated with the graph shown in Figure 1. The bold edges in the moral graph has been added to moralize relationships

SEPARATE DI and GMP DI finds minimal sets of variables separating income (D) and Intelligence (I) from each other in respectively the regression graph and the moral graph. To find the separators DIGRAM first finds all paths between D and I and then identify the minimal sets of nodes interrupting all these paths. The moral graph has fewer edges than the regression graph. Hence, minimal separating sets in the moral graph may be smaller than the separating sets in the regression graph.

SEPARATE and GMP identifies conditioning sets of variables for tests of hypotheses of conditional independence that are true, if the variables are conditionally independent given all other current and prior variables in the model. The test statistic does not have to be the same as the test of conditional independence given all these variables. REDUCE identifies the minimal set of conditioning variables required for calculation of the test statistic in the complete model. This set will in many cases be larger than the conditioning sets derived by SEPARATE and GMP.

Figures 10-12 shows the model-based hypotheses for tests of conditional independence of Income (D) and Intelligence (I) derived under the model in Figure 1.

```
Separation hypotheses:

2 Hypotheses:

HYPOTHESIS  1:  D & I  |  C B F G M
HYPOTHESIS  2:  D & I  |  A B F G M
```

Figure 10. **SEPARATE DI**. The conditioning sets are the minimal set of separators of all paths between D and I in Figure 9a

```
  +--------------------------------------+
  |                                      |
  | Overview of smallest GMP hypotheses. |
  |                                      |
  +--------------------------------------+

1 Hypothesis:

HYPOTHESIS  1:  D & I  |  F G M
```

Figure 11. **GMP DI**. The conditioning set is the minimal set of separators of all paths between D and I in Figure 9b

```
Hypothesis implied by reducibility:

1 Hypothesis:

HYPOTHESIS  1:  D & I  |  C A B F G M
```

Figure 12. **REDUCE DI**. The conditioning set is defined by decomposition of regression graph, Figure 9a

Having derived the hypotheses using one of the three commands described above, the next thing to do is to create the table. One way to do that is to use that TABULATE command, but this means that you have to redefine the hypotheses for the table, Instead you may use the

CHOOSE <hypothesis no.>

which creates both the table and the hypothesis. This is illustrated in Figure 13 for the second hypothesis generated by separation in the regression graph.

Obviously, you do not have to include the hypothesis number together with the CHOOSE command if the list of hypotheses only contains one hypothesis.

Having defined the hypothesis, the next you need to do is to invoke the TEST command. You probably do not want to see the complete 7-dimensional table, but remember that only global test results will be reported unless you invoke the LOCAL command before the TEST command. You should also insist on Monte Carlo estimates of exact p-values using either the EXACT, SEQUENTIAL or REPEAT command.

```
Separation hypotheses:

2 Hypotheses:

HYPOTHESIS  1:  D & I  |  C B F G M
HYPOTHESIS  2:  D & I  |  A B F G M
The marginal DABFGIM model:

Variables   DABFGIM
        : D **+*+ +
        : A **+++
        : B ++*+  +
        : F *++****
        : G ++ ****
        : I    ****
        : M + +****

  +-----------------------------+
  |                             |
  | Report on missing responses |
  |                             |
  +-----------------------------+

D   Income  Observed =   2342  Missing =    809
A ChronDis  Observed =   2668  Missing =    483
B   Unempl  Observed =   2667  Missing =    484
F  VocEduc  Observed =   2669  Missing =    482
G   School  Observed =   2857  Missing =    294
I Intellig  Observed =   3151  Missing =      0
M      Sex  Observed =   3151  Missing =      0

Number of cases with complete responses = 2219 out of 3151

Marginal loglinear model DBFGM,ABFGIM,DAF   Fixed: DAF,ABFGIM

1 Hypothesis:

HYPOTHESIS  1:  D & I  |  A B F G M
```

Figure 13. **SEPARATE DI** and **CHOOSE 2**

The test result is shown in Figure 14. There are several things you should notice here:

1) The difference between the asymptotic and the exact p- values for the χ^2 statistic. This is a very sparse table and differences like this turn up all the time. The asymptotic p-value of the γ coefficient looks better. This is often the case, but we nevertheless suggest that you always use Monte Carlo tests rather than asymptotic tests for analysis of multidimensional contingency tables.

2) The conditional association between intelligence and Income is very weak (γ = 0.07), but nevertheless significant. We have used 1-sided p-values here, since we expect a positive correlation between the variables, but the 2-sided p-value would also have been significant.

3) The local test results do not disclose significant evidence of an effect of intelligence on income among women and among persons who has been unemployed for some time after having finished their education. You have to be very careful forming final conclusions on these test statistics. Methods that may help you there will be described later in these notes.

```
-------------------------------------------------------------------------------------
                         p-values                  p-values (1-sided)
Hypothesis       X²  df asymp exact 99% conf.int. Gamma asymp exact 99% conf.int. nsim
-------------------------------------------------------------------------------------
 1:D&I|ABFGM 1086.81018 0.066 0.639 0.599 - 0.677  0.07 0.014 0.012 0.006 - 0.025 1000    +
-------------------------------------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by ChronDis (A) **
                          p-values      p-values (1-sided)
 A: ChronDis   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    None 708.19  656 0.0774 0.3440   0.07 0.0192 0.0200
 2:      1+ 378.59  362 0.2637 0.9580   0.07 0.1163 0.1460
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by   Unempl (B) **
                          p-values      p-values (1-sided)
 B:   Unempl   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:< 1 year 659.26  624 0.1590 0.5980   0.08 0.0175 0.0130
 2:1+ years 427.51  394 0.1180 0.5940   0.04 0.2416 0.2420
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by  VocEduc (F) **
                          p-values      p-values (1-sided)
 F:  VocEduc   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    LANG  40.25   54 0.9178 0.9080   0.18 0.1539 0.1890
 2:MELLEMLA 127.03  121 0.3357 0.5630   0.04 0.3387 0.3380
 3:    KORT 162.24  137 0.0695 0.3120   0.09 0.1393 0.1420
 4:LæRLINGE 381.65  370 0.3270 0.6330   0.08 0.0357 0.0400
 5:   INGEN 375.61  336 0.0671 0.3970   0.04 0.2850 0.2670
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by   School (G) **
                          p-values      p-values (1-sided)
 G:   School   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:   0 - 2 235.72  191 0.0152 0.0470   0.07 0.2188 0.2170
 2:   3 - 4 267.92  230 0.0436 0.1410   0.04 0.2784 0.2680
 3:   5 - 8 354.35  380 0.8233 0.9680   0.08 0.0343 0.0350
 4:  9 - 12 228.78  217 0.2785 0.6770   0.09 0.1384 0.1390
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by      Sex (M) **
                          p-values      p-values (1-sided)
 M:      Sex   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    Male 610.72  564 0.0847 0.5640   0.11 0.0091 0.0080
 2:  Female 476.05  454 0.2290 0.6370   0.05 0.1525 0.1720
-----------------------------------------------------------

+------------------------------------------------------+
|                                                      |
|   Summary of gamma coefficients in separate strata   |
|                                                      |
| Significance evaluated by asymptotic 2-sided p-values |
|                                                      |
+------------------------------------------------------+

  gamma    p >0.05      0.01<p<=0.05      p<0.01
--------------------------------------------------------------
Negative      43              3             2
Positive      41              7             2
```

Figure 14. Test of the hypothesis defined in Figure 13.

Testing model based hypotheses without tabulating

If you only need global test statistics for the model based hypotheses there an easier way to do it. A simple

TEST <variable pairs>

or

GTEST <variable pairs>

will do. Following these commands, DIGRAM identifies the hypotheses derived by separation and decomposition in the regression graph and tests the hypotheses. The tables are not saved, however, for which reason the local test results and the other special features described later in these notes will not be available.

If a table already exists, DIGRAM may have problems deciding whether the parameters of the TEST command are parameters requesting special types of analyses or whether they refer to variables. To avoid misunderstandings we have added the GTEST command (for tests defined by the Graphical model) where the parameters always refer to variables.

Figure 15 shows the result of a TEST DI command, assuming that no table exists.

```
Test of separation and core hypotheses

-------------------------------------------------------------------------------------
                         p-values                 p-values (1-sided)
Hypothesis       X²  df asymp exact               Gamma asymp exact               nsim
-------------------------------------------------------------------------------------
 1:D&I|CBFGM 1174.21040 0.002 0.304 (0.268-0.343)  0.07 0.030 0.042 (0.028-0.062) 1000
 2:D&I|ABFGM 1086.81018 0.066 0.656 (0.616-0.694)  0.07 0.014 0.011 (0.005-0.023) 1000
 3:D&I|CABFGM1220.01082 0.002 0.676 (0.637-0.713)  0.07 0.043 0.047 (0.033-0.067) 1000
-------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.008 for FDR = 0.05
                          and p <  0.002 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------------
```

Figure 15: **TEST DI** or **GTEST DI:** The hypothesis defined by decomposition is referred to as a core hypothesis since the marginal ABCDFGIM model contains the core of the DI problem.

Model building

There are two ways to build models in DIGRAM. You can either build the model yourself based on subject matter knowledge or you can ask DIGRAM to do it for you using procedures for semi-automatic model search. The first approach leads to a confirmatory analysis and the second to an exploratory analysis. In practice you will probably use both approaches during the analysis.

The easiest way to build the model based on subject matter knowledge is to define the Markov graph of the model in DIGRAM's Graph module[2], but you can also do it in the DIGRAM module where you analyze data using one of the commands below.

ADD var1 var2	Adds an edge/arrow between var1 and var2 to the graph being shown in the graph window.
DELETE var1 var2	Delete the edge/arrow between var1 and var2 in the graph being shown in the graph window.
FIX var1 var2	Fixes the edge/arrow between var1 and var2 in the graph being shown in the graph window. Fixed edges between variables will are shown with thick lines in the graph and will not be removed by any of DIGRAM's model selection procedures.
PREVENT var1 var2	Prevents inclusion of an edge/arrow between var1 and var2 in the graph by DIGRAM's model selection procedures.
NEW <status>	Initiates the graph with connections between variables defined by the status in the following way: Status 1 : all variables are unconnected. Status 2: all variables are connected by unfixed edges.

[2] The guided tour through the Graph module tells you how to do this by adding or deleting edges of the model.

XPLANATORY Fixes edges between all variables in the final

Two different commands are available for exploratory model search. The first, **SCREEN**, creates an initial graphical model that is meant to serve as a useful starting point for a more careful stepwise exploratory model search. To invoke this analysis you must use the **MODELSEARCH** command. During this analysis we distinguish between backwards model search where edges are deleted from the model and forwards model search where edges are added to the model.

Of the two procedures, only screening is fully automatic, requiring no help from the user after the command has been used. The stepwise model search is semi-automatic. During each step, DIGRAM suggests ways to improve the model, but the decision on what to do is always up to the user.

Screening for an initial model

Screening is described by Kreiner (1986).

SCREEN **<parameters>** Creates an initial model based on analysis of 2- and 3-way tables.

Invoke a "**SCREEN ?"** if you want to see the complete list of parameters that can be included among the arguments to this command. Here we only show the default screening without parameters and the extended screening following a "SCREEN X".

The default screening consists of three steps:

1) Tests of marginal independence for all pairs of variables. Edges between marginally independent variables are removed from the model.

2) Tests of conditional independence of marginally independent variables in 3-way tables given variables that are marginally associated with both variables. An edge will be included in the model if conditional independence is rejected. Associations disclosed during this step are called hidden associations.

3) Tests of conditional independence of variables that are connected in the graph after step 2. The test is performed in 3-way tables that include variables that are connected with both variables after step 2. The edge is excluded if conditional independence is accepted.

```
--------------------------------
* Analysis of
* Twoway tables

  DCABFGIJKLM
 D*++++++++++
 C+*+++++ ++
 A++*++++++ +
 B+++*+++++++
 F++++*++++++
 G+++++*+++++
 I++++++*++++
 J+ +++++*++
 K++++++++*++
 L++ ++++++*
 M+ +++++ + *

 0.0500 level of significance
     *: fixed
     +: undecided
     -: conditional independence
     h: hidden interaction
     o: cond. ind. was not used
```

```
------------------------------
* Analysis of
* hidden association

  DCABFGIJKLM
 D*++++++++++
 C+*+++++h++h
 A++*++++++h+
 B+++*+++++++
 F++++*++++++
 G+++++*+++++
 I++++++*++++
 J+h+++++*++
 K++++++++*++
 L++h++++++*
 M+h+++++ + *

 0.0500 level of significance
     *: fixed
     +: undecided
     -: conditional independence
     h: hidden interaction
     o: cond. ind. was not used
```

(a) (b)

```
The final SCREEN model:
Level of significance:  0.0500

Variables   DCABFGIJKLM
  Income: D *+ ++++++++
     SRH: C +*++++
ChronDis: A  +*+++  +
  Unempl: B +++*+++   +
 VocEduc: F ++++*++ +++
  School: G +++++*+ +++
Intellig: I +  +++* +++
Urbaniza: J +      *++
  FamSES: K + + ++++*++
 FamEduc: L +   +++++*
     Sex: M +  ++++ + *

 h : Hidden interaction
 o : unused conditional independence
```

Figure 16. **SCREEN**. (a) Tests of marginal independence. (b) After test of conditional independence of marginally independent variables. A ‘h’ refer to hidden association that

was disclosed during this step of the screening procedure. (c) After elimination of association between conditionally independent variables (step 3).

SCREEN X extends the default screening with fully automatic backward elimination of edges until all tests for conditional independence in separation hypotheses are significant at a 0.01 level followed by a fully automatic forward inclusion of edges until all tests are insignificant at the same level (p > 0.01). The results are shown in Figure 17.

```
+-----------------------------------------------------+
|                                                     |
| Automatic model search: Backwards from current model |
|                                                     |
+-----------------------------------------------------+

Critical level of significance =  0.010
P-values are twosided
Nonmonotonous relationships will be disregarded
Repeated MC tests if p-asymptotic <  0.100

Deleted: DG  p =  0.020
Deleted: FL  p =  0.333
Deleted: FM  p =  0.015


+---------------------------------------------------+
|                                                   |
| Automatic model search: Forward from current model |
|                                                   |
+---------------------------------------------------+

Critical level of significance =  0.010
P-values are twosided
Nonmonotonous relationships will be disregarded
Repeated MC tests if p-asymptotic <  0.100

Included: BG  p =  0.007
```

Figure 17 SCREEN X. The first part of the initial screening corresponds to the default screening sho0wn in Figure 16. This figure shows results of the automatic model search procedures included in the extended screening. Three edges have been removed.

Notes on technicalities during screening:

First, the default screening uses asymptotic tests only, since all tests are calculated in 2- or 3-way tables, where we do not expect to have any problems with asymptotics. Tests calculated during the automatic backwards and forwards procedures following the default analysis are 2-sided repeated Monte Carlo tests with a maximum of 400 random tables per test.

Second, you must not expect the model defined by screening to be the final model. In particular, the model defined by default screening is expected to contain far more edges than necessary. For this reason you should only regard the model defined by one of the screening procedures as the model where proper model search starts. Screening is useful since the final model – in our experience – is often quite close to the final model, but anything can (and will eventually) happen. So be prepared. Model search may be time consuming.

Stepwise model search

Stepwise model search is invoked by either of the following three commands.

MODELSEARCH <anchor variables>
BACKWARDS <anchor variables>
FORWARDS <anchor variables>

The only difference between the effects of the three commands is that model search starts with forwards edge inclusion in the third command, whereas the first two starts with backwards edge elimination.

The main idea of the stepwise model search procedures is that you, the user, is totally responsible for what happens to the model during the analysis and that DIGRAM never changes the model on its own. DIGRAM analyzed the model during the analysis, derive separation hypotheses to be tested and makes suggestions. But you make the decisions. These decisions are supposed to rely on three factors:

1) The strength of the associations among variables as measured by the partial γ coefficients,
2) The level of significance (the p-values) of the tests of conditional independence.
3) The subject matter knowledge of the user.

Most automatic model and variable selection methods in standard programs only use p-values to make decisions. You should recall, however, that p-values are uniformly distributed for all tests of true null-hypotheses. A p-value of 0.46 is therefore not evidence of a better fit of data to the model under the null-hypotheses than a p-value of 0.14. They are both expressions of adequate fit between data and null-hypothesis model. For this and other reasons we prefer to make decisions based primarily on subject matter knowledge if at all possible – if it is impossible or difficult to see why two variables should be directly associated while subject matter actually suggests that two other variables are connected, then we suggest that you remove the first edge before the other if both tests of conditional independence are insignificant, even though the p-value of the second is larger than the first. If the are no subject matter arguments supporting one edge in favor of the other we suggest that you look at the strength of the associations and remove the edge with the smallest γ coefficient before the edge with the stronger coefficient.

Since you are making the decisions you are supposed to be able to overlook and evaluate all the test results and γ coefficients calculated during each step. This can be very difficult since the number of potential edges is very large in high-dimensional models. For this reason, DIGRAM's model search procedures are restricted to edges connecting to a subset[3] of anchor variables. Model search in this way is itself a meta-stepwise procedure where you start with one anchor variable or a small set of anchor variables and then proceed to other anchor variables until you have been through the complete structure. How to select anchor variables can be discussed, of course. In general we prefer to start

[3] You can of course use all variables as anchor variables if you think that you can keep track of all that goes on during a global model search for the complete structure.

with the least interesting variables and end with the primary variables of interest, but that of course is up to you to decide.

The model generated by extended screening is shown in Figure 18. Note that partial γ coefficients have bee added to each edge. These coefficients are automatically calculated after screening. Recall that they depend on the model structure. They therefore have to be recalculated if and when you change the model. Use the default **GAMMA** command without additional parameters for this purpose.

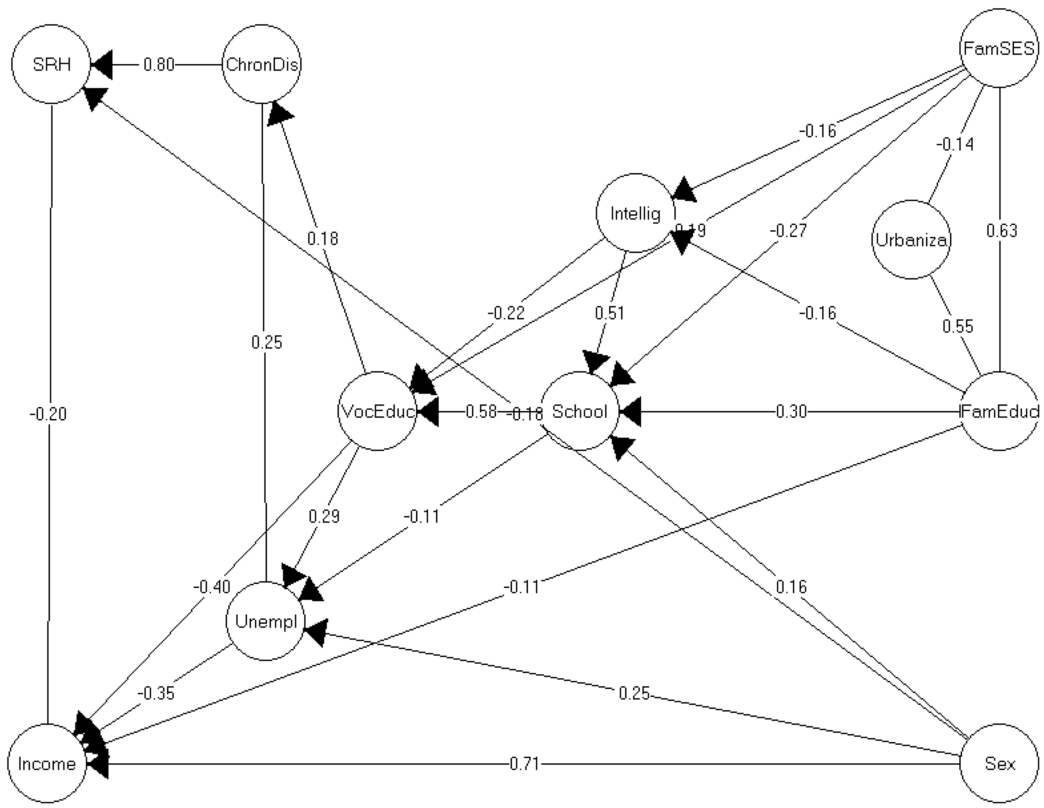


Figure 18. The model derived by extended screening.

After screening we need to see whether some of the edges in Figure 18 can be excluded or whether screening has eliminated too many edges. We therefore invoke the MODELSEARCH command having first decided that we want to use 2-sided repeated Monte Carlo tests during the analysis. Since Intelligence is of particular interest in this example we show the analysis with an exploratory analysis of the model structure attached to I. **MODELSEARCH I** opens the model search dialog shown in Figure 19.

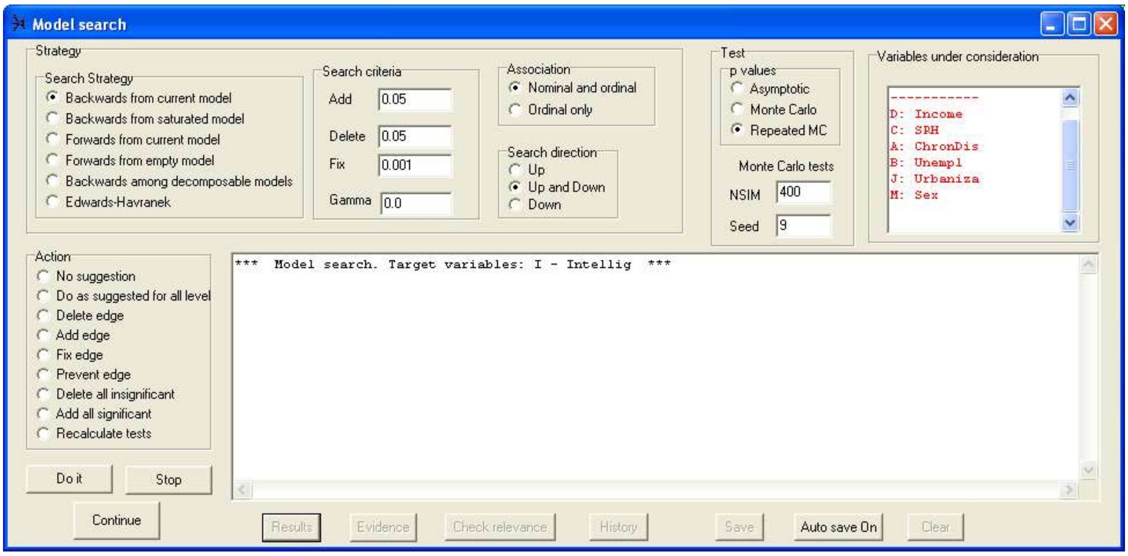


Figure 19. **MODELSEARCH I.** The model search dialog

There are a lot of things you have to decide during model search, but it is easier than it looks at first glance and you will get used to it.

You have to decide upon a model search strategy. In most cases the first and the third option on the list of strategies will suffice.

You have to decide upon search criteria: critical levels for adding, deleting and fixing edges during model search, and a minimal γ value that you will consider of interest to the model. (Recall that DIGRAM only uses these criteria to make suggestions for you).

You have to decide whether you will only consider γ coefficients (ordinal only) or whether DIGRAM should also take notice of significant χ^2 statistics (Nominal and ordinal).

The search directions are upwards, up- and downwards, or downwards. During upwards model search, DIGRAM only considers variables that the anchor variable has an effect upon (the anchor variable is an independent variable). Downwards considers variables that are current with or behind the anchor variable in the recursive model structure. (The anchor is a dependent variable).

Next, you may change the way the statistical tests are performed.

And finally, you have to decide what to do after DIGRAM have calculated test statistics and made a suggestion. We return to this after the first tests.

We select ordinal tests only and click on “Start”. DIGRAM derives separation hypotheses for all edges connected to L (FamEduc) and report the results as shown in Figure 20.

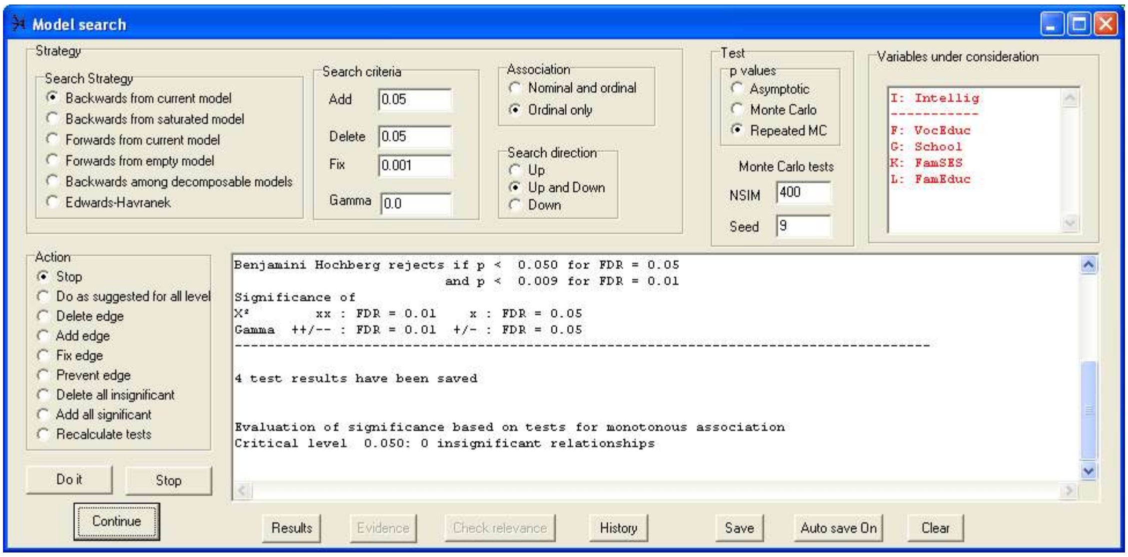


Figure 20. The model search dialog after calculation of test results

Since DIGRAM does not find any insignificant test results it suggest that we stop the search for a new model. There is a risk that the initial screening has overlooked something. For that reason we change the strategy to “Forwards from current model” and click on Continue.

The results are shown in Figure 21. DIGRAM discloses evidence of direct association between Income (D) and Intelligence (I) and between Intelligence (I) and Urbanization (J). In both cases the association is weak. If we had imposed a 0.10 limit on the γ coefficients, DIGRAM would have suggested, that we stop the analysis. Since this was not the case, DIGRAM prefer the DI edge to the DJ. Since the DI edge is of particular importance to the analysis and since we see no obvious reason that intelligence among school children should be *directly* related to urbanization, we follow DIGRAM's advice and click on continue. The result is shown in Figure 22.

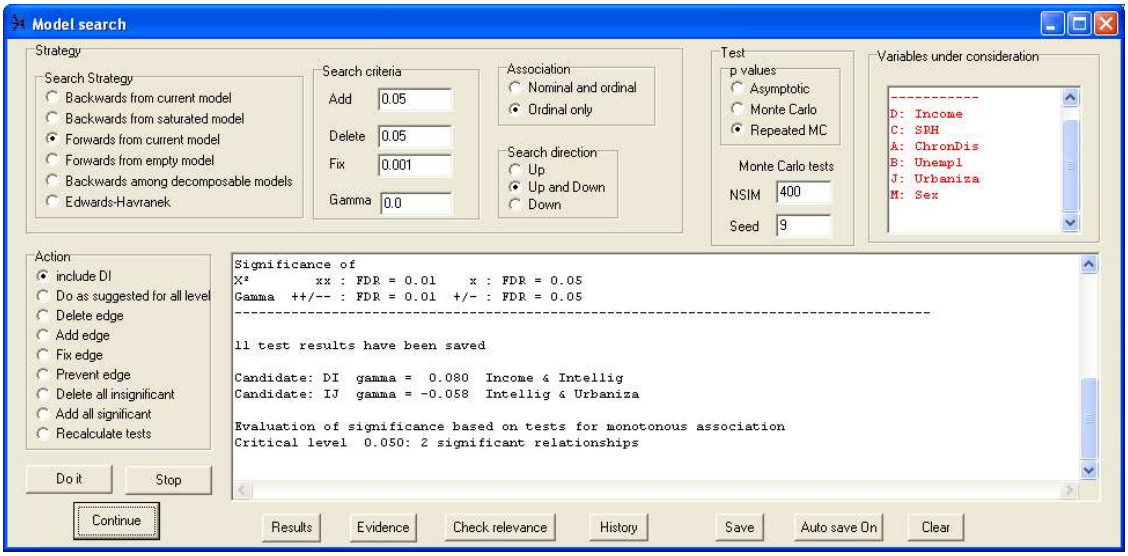


Figure 21. Results after the first step forward.

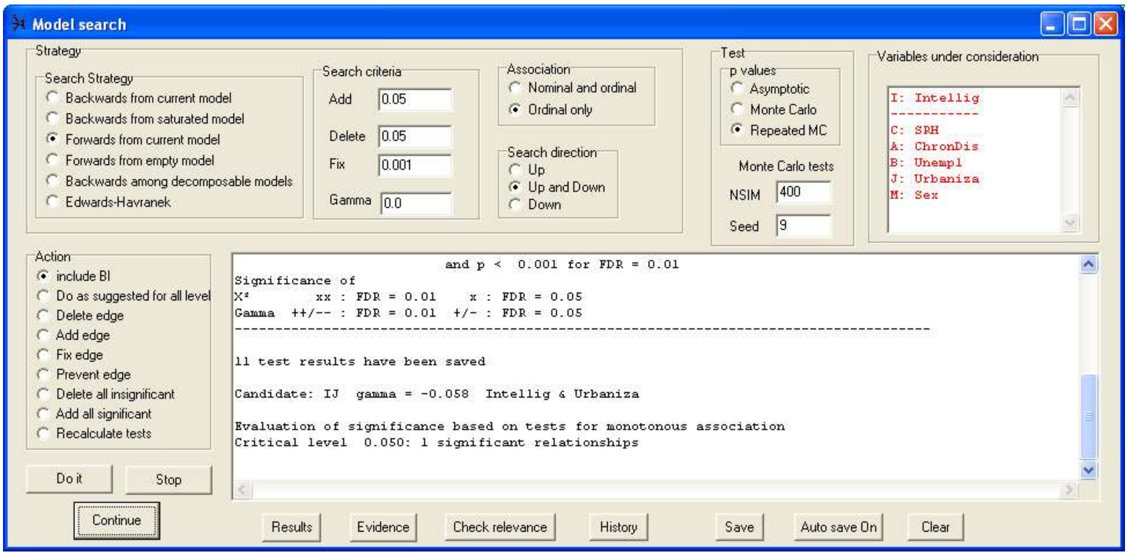


Figure 22. Results of the second step forward

Since the association between urbanization and intelligence is weak and since we do not think such an association makes much sense we stop for now, intending to return to this association when we have had a closer look at the rest of the model.

A click on the “STOP” button leads to another question since DIGRAM has noticed that we are about to change the model and want to make sure that that is what we intende to do:

Since we do intend to include the Income-Intelligence association in the model we click on yes, following which the model search dialog disappears and the following model search report is added to the DIGRAM output window.

```
The model was revised during model search

Search history.

Step number:                                        3
Number of edges that have been changed since start:   1
Number of edges that were changed in the last step:   1
The current model were first encountered in step:     2


        step no
   init  1  2  3

DI:   -  +

  +-----------------------------------------+
  |                                         |
  | ****  Description of I - Intellig **** |
  |                                         |
  +-----------------------------------------+

Intellig  has a direct effect on

   D - Income
   F - VocEduc
   G - School
```

Figure 23. Model search report.

Model search buttons

The model search dialog has a number of buttons that will be useful during the model search. You need some of these buttons to (re)activate the model search procedure and while other buttons will help you handle output and keep track of the model search history.

The buttons are,

Start/Continue	Take the action indicated in the Action list and take the next model search step
Do it	Take the action indicated in the Action list without continuing the model search procedure
Stop	Finish searching for models
Results	Print the test results of the recent model search step
Evidence	Print all significant or all insignificant test results (depending on whether you are going backwards or forwards)
Check relevance	To be used if test result suggest that two or more edges should be either removed or added to the model. For each edge, the test results of the other edges will be recalculated under the assumption that the edge status was changed.
History	Show the search history so far
Save	Save the current output on DIGRAM's output field.
Auto save On	Save all output obtained during model search on DIGRAM's output field
Clear	Erase the current model search output.

Model Checking

A model check in DIGRAM consists of tests of all separation hypotheses of missing edges and all separation hypotheses of existing edges. "**CHECK"** without parameters gives you the first part while "**CHECK +".**

The check of the model shown in Figure 1 is shown below. Note that significance is assessed after adjustment for multiple testing by the Benjamini-Hochberg procedure

Hypothesis	X²	df	p-values asymp	p-values exact		p-values (2-sided) Gamma	asymp	exact		nsim	
1:D&A|CBFGM	328.8	318	0.327	0.762	(0.481-0.917)	0.01	0.937	1.000	(0.760-1.000)	21	
2:D&I|CBFGM	1174.2	1040	0.002	0.315	(0.259-0.377)	0.07	0.060	0.078	(0.050-0.119)	400	
3:D&I|ABFGM	1086.8	1018	0.066	0.660	(0.597-0.718)	0.07	0.027	0.015	(0.005-0.040)	400	
4:D&J|CBFGM	866.6	817	0.112	0.865	(0.815-0.903)	-0.11	0.004	0.015	(0.005-0.040)	400	
5:D&J|ABFGM	828.9	805	0.272	0.950	(0.914-0.971)	-0.13	0.001	0.000	(0.000-0.016)	400	--
6:D&K|CBFGM	1077.5	962	0.005	0.476	(0.237-0.727)	-0.04	0.335	0.429	(0.203-0.689)	21	
7:D&K|ABFGM	983.7	914	0.054	0.560	(0.383-0.723)	-0.04	0.264	0.200	(0.094-0.376)	50	
8:D&L|CBFGM	985.9	896	0.019	0.452	(0.390-0.517)	-0.07	0.087	0.093	(0.062-0.137)	400	
9:D&L|ABFGM	911.0	869	0.157	0.710	(0.649-0.765)	-0.10	0.011	0.013	(0.004-0.037)	400	
10:C&B|DAF	112.6	98	0.149	0.265	(0.212-0.325)	0.17	0.015	0.013	(0.004-0.037)	400	
11:C&G|DAF	244.9	228	0.211	0.242	(0.105-0.466)	-0.06	0.341	0.303	(0.145-0.527)	33	
12:C&I|DAF	362.4	344	0.238	0.571	(0.311-0.797)	0.02	0.714	0.667	(0.392-0.861)	21	
13:C&J|DAF	277.8	277	0.474	0.857	(0.580-0.963)	-0.02	0.682	0.810	(0.529-0.941)	21	
14:C&K|DAF	351.4	342	0.352	0.538	(0.306-0.756)	-0.05	0.351	0.269	(0.111-0.522)	26	
15:C&L|DAF	357.3	364	0.590	0.810	(0.529-0.941)	0.05	0.429	0.476	(0.237-0.727)	21	
16:C&M|DAF	122.2	98	0.049	0.115	(0.080-0.163)	-0.19	0.008	0.007	(0.002-0.029)	400	
17:A&I|BFG	122.6	111	0.212	0.286	(0.110-0.564)	0.00	0.989	1.000	(0.760-1.000)	21	
18:A&I|FGM	120.8	115	0.337	0.242	(0.105-0.466)	-0.01	0.752	0.848	(0.632-0.948)	33	
19:A&J|BFG	119.7	90	0.020	0.043	(0.023-0.077)	-0.06	0.254	0.278	(0.224-0.338)	400	
20:A&J|FGM	90.8	92	0.516	0.587	(0.461-0.702)	-0.07	0.120	0.154	(0.084-0.265)	104	
21:A&K|BFG	142.3	114	0.037	0.060	(0.036-0.098)	0.05	0.391	0.407	(0.346-0.472)	400	
22:A&K|FGM	115.3	114	0.447	0.550	(0.478-0.620)	0.08	0.113	0.110	(0.073-0.163)	318	
23:A&L|BFG	126.5	122	0.372	0.571	(0.311-0.797)	-0.01	0.901	0.905	(0.634-0.981)	21	
24:A&L|FGM	119.7	122	0.543	0.571	(0.311-0.797)	-0.02	0.787	0.619	(0.351-0.830)	21	
25:A&M|BFG	36.2	33	0.320	0.286	(0.110-0.564)	-0.04	0.574	0.333	(0.139-0.608)	21	
26:B&G|AFM	65.3	52	0.103	0.105	(0.071-0.154)	-0.07	0.157	0.154	(0.112-0.208)	370	
27:B&I|AFM	90.2	75	0.112	0.127	(0.091-0.177)	-0.07	0.064	0.072	(0.046-0.113)	400	
28:B&I|FGM	111.8	117	0.619	0.744	(0.656-0.815)	-0.07	0.096	0.128	(0.079-0.202)	195	
29:B&J|AFM	75.9	58	0.057	0.070	(0.044-0.110)	0.01	0.881	0.897	(0.852-0.930)	400	
30:B&J|FGM	93.1	93	0.476	0.619	(0.351-0.830)	0.01	0.743	0.714	(0.436-0.890)	21	
31:B&K|AFM	84.0	79	0.329	0.231	(0.105-0.435)	0.04	0.397	0.410	(0.235-0.611)	39	
32:B&K|FGM	99.2	116	0.867	0.861	(0.657-0.952)	0.05	0.292	0.250	(0.114-0.464)	36	
33:B&L|AFM	92.6	93	0.492	0.429	(0.203-0.689)	0.03	0.464	0.714	(0.436-0.890)	21	
34:B&L|FGM	132.0	124	0.294	0.374	(0.256-0.509)	0.07	0.162	0.165	(0.088-0.287)	91	
35:F&J|GIKM	810.9	697	0.002	0.137	(0.081-0.221)	-0.00	0.974	0.975	(0.920-0.993)	161	
36:F&L|GIKM	646.5	600	0.092	0.524	(0.273-0.763)	0.04	0.312	0.333	(0.139-0.608)	21	
37:G&J|IKLM	655.6	599	0.054	0.567	(0.503-0.630)	0.08	0.043	0.050	(0.029-0.086)	400	
38:I&J|K	78.5	60	0.055	0.058	(0.034-0.095)	-0.09	0.000	0.000	(0.000-0.016)	400	--
39:I&L|K	124.4	88	0.006	0.013	(0.004-0.037)	-0.16	0.000	0.000	(0.000-0.016)	400	--
40:I&M|K	34.2	20	0.025	0.020	(0.008-0.047)	0.04	0.157	0.160	(0.118-0.213)	400	
41:J&M	4.7	3	0.192	0.193	(0.094-0.356)	0.01	0.846	0.807	(0.644-0.906)	57	
42:K&M	9.6	4	0.047	0.030	(0.015-0.061)	0.03	0.264	0.245	(0.194-0.304)	400	
43:L&M	7.2	5	0.207	0.128	(0.079-0.202)	0.04	0.260	0.251	(0.180-0.339)	195	

Benjamini Hochberg rejects if p < 0.002 for FDR = 0.05
and p < 0.000 for FDR = 0.01

Figure 24. **CHECK.** Tests of all separation hypotheses relating to missing edges

30 Separation hypotheses related to existing edges

Hypothesis	X²	df	p-values asymp	p-values exact		Gamma	p-values (2-sided) asymp	p-values (2-sided) exact		nsim		
1:D&C|AF	272.2	156	0.000	0.000	(0.000-0.016)	-0.13	0.002	0.000	(0.000-0.016)	400	xx	--
2:D&B|CFGM	471.1	276	0.000	0.000	(0.000-0.016)	-0.40	0.000	0.000	(0.000-0.016)	400	xx	--
3:D&B|AFGM	425.4	239	0.000	0.000	(0.000-0.016)	-0.40	0.000	0.000	(0.000-0.016)	400	xx	--
4:D&F|CBGM	905.6	544	0.000	0.000	(0.000-0.016)	-0.31	0.000	0.000	(0.000-0.016)	400	xx	--
5:D&G|CBFM	576.5	468	0.000	0.018	(0.007-0.044)	0.16	0.000	0.000	(0.000-0.016)	400	x	++
6:D&G|ABFM	565.0	418	0.000	0.000	(0.000-0.016)	0.17	0.000	0.000	(0.000-0.016)	400	xx	++
7:D&M|CBFG	732.8	278	0.000	0.000	(0.000-0.016)	-0.72	0.000	0.000	(0.000-0.016)	400	xx	--
8:D&M|ABFG	700.2	241	0.000	0.000	(0.000-0.016)	-0.71	0.000	0.000	(0.000-0.016)	400	xx	--
9:C&A|DF	663.1	70	0.000	0.000	(0.000-0.016)	0.82	0.000	0.000	(0.000-0.016)	400	xx	++
10:C&F|DA	194.5	128	0.000	0.002	(0.000-0.021)	0.07	0.110	0.102	(0.070-0.148)	400	xx	
11:A&B|FG	42.0	19	0.002	0.002	(0.000-0.021)	0.23	0.001	0.000	(0.000-0.016)	400	xx	++
12:A&B|FM	28.7	10	0.001	0.002	(0.000-0.021)	0.26	0.000	0.000	(0.000-0.016)	400	xx	++
13:A&F|BG	64.9	31	0.000	0.000	(0.000-0.016)	0.19	0.000	0.000	(0.000-0.016)	400	xx	++
14:A&G|BF	49.7	29	0.010	0.015	(0.005-0.040)	-0.07	0.149	0.170	(0.127-0.224)	400	x	
15:A&G|FM	51.8	29	0.006	0.007	(0.002-0.029)	-0.09	0.057	0.058	(0.034-0.095)	400	xx	
16:B&F|AM	101.1	16	0.000	0.000	(0.000-0.016)	0.27	0.000	0.000	(0.000-0.016)	400	xx	++
17:B&M|AF	28.1	10	0.002	0.000	(0.000-0.016)	0.23	0.000	0.000	(0.000-0.016)	400	xx	++
18:B&M|FG	43.9	18	0.001	0.000	(0.000-0.016)	0.25	0.000	0.000	(0.000-0.016)	400	xx	++
19:F&G|IKM	1388.3	414	0.000	0.000	(0.000-0.016)	-0.60	0.000	0.000	(0.000-0.016)	400	xx	--
20:F&I|GKM	548.3	441	0.000	0.005	(0.001-0.025)	-0.23	0.000	0.000	(0.000-0.016)	400	xx	--
21:F&K|GIM	516.2	402	0.000	0.010	(0.003-0.033)	0.19	0.000	0.000	(0.000-0.016)	400	x	++
22:F&M|GIK	310.0	192	0.000	0.000	(0.000-0.016)	0.11	0.011	0.015	(0.005-0.040)	400	xx	+
23:G&I|KLM	852.6	364	0.000	0.000	(0.000-0.016)	0.51	0.000	0.000	(0.000-0.016)	400	xx	++
24:G&K|ILM	480.5	344	0.000	0.000	(0.000-0.016)	-0.27	0.000	0.000	(0.000-0.016)	400	xx	--
25:G&L|IKM	513.9	391	0.000	0.010	(0.003-0.033)	-0.30	0.000	0.000	(0.000-0.016)	400	x	--
26:G&M|IKL	237.7	167	0.000	0.000	(0.000-0.016)	0.16	0.001	0.005	(0.001-0.025)	400	xx	++
27:I&K	205.0	16	0.000	0.000	(0.000-0.016)	-0.25	0.000	0.000	(0.000-0.016)	400	xx	--
28:J&K|L	157.0	63	0.000	0.000	(0.000-0.016)	-0.14	0.000	0.000	(0.000-0.016)	400	xx	--
29:J&L|K	429.5	66	0.000	0.000	(0.000-0.016)	0.55	0.000	0.000	(0.000-0.016)	400	xx	++
30:K&L|J	1884.8	80	0.000	0.000	(0.000-0.016)	0.63	0.000	0.000	(0.000-0.016)	400	xx	++

Benjamini Hochberg rejects if p < 0.048 for FDR = 0.05
and p < 0.009 for FDR = 0.01

Significance of
X² xx : FDR = 0.01 x : FDR = 0.05
Gamma ++/-- : FDR = 0.01 +/- : FDR = 0.05

Figure 25. **CHECK +.** Tests of all separation hypotheses relating to existing edges

The check of the model in Figure 1 disclosed a few problems. First, Figure 24 suggests that income (D) depends on urbanization during childhood (J) and the Intelligence depends on Urbanization and family education (L). Second, Figure 25 shows that the dependence of Chronically Diseases (A) on School education (G) is not supported by data. The model shown in Figure 2 is therefore not the final model. But it is close.

Description of relationships in graphical models.

At the end of the day, we need at least to describe the important relationships in the model. To do this you must use the following command

DESCRIBE **<variable pairs>**

Let us take a look at the dependence of Income on Intelligence (**DESCRIBE DI**)

Quite a lot of output is generated – probably more than you think you need – but also (we hope) all that you need. We will go through this output, one piece at a time.

```
   +-----------------------------------+
   |                                   |
   |      Relationship between         |
   |                                   |
   |        D: Income                  |
   |        I: Intellig                |
   |                                   |
   | Status = conditional independence |
   |                                   |
   +-----------------------------------+



Table 1. The DI distribution.

     +Intellig
     | | D:--Income                              |
     I | < 100 100-2 150-2 200-2 250-3 300-4  400+ | TOTAL |
-------+-------------------------------------------+-------+
     - |    22    42    82    56    20     9     5 |   236 |
   row%|   9.3  17.8  34.7  23.7   8.5   3.8   2.1 | 100.0 |
 26-30 |    23    55    99    81    26    21     1 |   306 |
   row%|   7.5  18.0  32.4  26.5   8.5   6.9   0.3 | 100.0 |
 31-35 |    37   100   147   113    40    28    14 |   479 |
   row%|   7.7  20.9  30.7  23.6   8.4   5.8   2.9 | 100.0 |
 36-40 |    35   108   148   137    44    52    16 |   540 |
   row%|   6.5  20.0  27.4  25.4   8.1   9.6   3.0 | 100.0 |
   41+ |    43    92   160   213   109    96    68 |   781 |
   row%|   5.5  11.8  20.5  27.3  14.0  12.3   8.7 | 100.0 |  X² = 141.9
-----------------------------------------------------------+  df =  24
 TOTAL |   160   397   636   600   239   206   104 |  2342 |   p = 0.000
   row%|   6.8  17.0  27.2  25.6  10.2   8.8   4.4 | 100.0 | Gam =  0.20
-----------------------------------------------------------+   p = 0.000
```

Figure 26. **DESCRIBE DI**. The marginal distribution

```
Separation hypotheses:

2 Hypotheses:

HYPOTHESIS  1:  D & I  |  C B F G M
HYPOTHESIS  2:  D & I  |  A B F G M
```

Figure 27. **DESCRIBE DI**. The separation hypotheses

The is a moderate, highly significant *marginal* association between Income and Intelligence (Figure 26). To estimate the conditional dependence we need to control for the conditioning sets of variables in separation hypotheses. These are shown in Figure 27. This leads to two separate analyses. One for each of the hypotheses. Here we only show the first one.

Next follows (Figure 28) information on the loglinear structure of the model containing DI and the conditioning (separating) set of variables.

```
  +-------------------------+
  |                         |
  | Marginal model: DC|BFGIM |
  |                         |
  +-------------------------+


The marginal model is not graphical

Cliques of the marginal graph: DCBFGM,CBFGIM
Log linear generators        : DBFGM,DCF,CBFGIM
Fixed interactions           : CBFGIM
Collapsibility:                Parametric.

Estimable parameters         : DB,DBF,DG,DBG,DFG,
                               DBFG,DM,DBM,DFM,
                               DBFM,DGM,DBGM,DFGM,
                               DBFGM,D,DC,DF,DCF
```

Figure 28. **DESCRIBE DI**. The marginal distribution

The collapsed model is loglinear with a simpler structure than a graphical model.

```
------------------------------------------------------------------------------------------
                         p-values                 p-values (2-sided)
Hypothesis       X²  df asymp exact 99% conf.int. Gamma asymp exact 99% conf.int. nsim
------------------------------------------------------------------------------------------
 1:D&I|CBFGM 1174.21040 0.002 0.338 0.280 - 0.401  0.07 0.060 0.052 0.030 - 0.089  400
------------------------------------------------------------------------------------------
** Local testresults for strata defined by      SRH (C) **
                          p-values      p-values (1-sided)
 C:      SRH   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:VeryGood 726.58  648 0.0171 0.1350   0.06 0.0453 0.0350
 2:    Fair 331.63  311 0.2015 0.8625   0.09 0.1074 0.1150
 3:LessFair  65.08   42 0.0127 0.1025   0.15 0.1650 0.2250
 4:     Bad  50.92   39 0.0958 0.8200   0.00 0.5000 1.0000
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by   Unempl (B) **
                          p-values      p-values (1-sided)
 B:   Unempl   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:< 1 year 738.46  658 0.0157 0.3275   0.07 0.0408 0.0275
 2:1+ years 435.74  382 0.0298 0.4400   0.06 0.1768 0.2075
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by  VocEduc (F) **
                          p-values      p-values (1-sided)
 F:  VocEduc   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    LANG  33.18   45 0.9038 0.9350   0.15 0.1900 0.2000
 2:MELLEMLA 157.30  143 0.1954 0.3950  -0.03 0.3698 0.3725
 3:    KORT 170.99  152 0.1389 0.4700   0.07 0.2084 0.2200
 4:LæRLINGE 461.78  415 0.0560 0.3450   0.07 0.0475 0.0375
 5:   INGEN 350.96  285 0.0046 0.0900   0.07 0.1764 0.1850
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by   School (G) **
                          p-values      p-values (1-sided)
 G:   School   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:   0 - 2 212.38  177 0.0357 0.2400   0.14 0.0556 0.0375
 2:   3 - 4 274.80  221 0.0080 0.0575  -0.01 0.4422 0.4575
 3:   5 - 8 433.37  425 0.3790 0.8350   0.08 0.0348 0.0275
 4:  9 - 12 253.65  217 0.0445 0.2975   0.02 0.3931 0.3925
-----------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by      Sex (M) **
                          p-values      p-values (1-sided)
 M:      Sex   X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:    Male 622.06  545 0.0122 0.3150   0.12 0.0082 0.0050
 2:  Female 552.14  495 0.0382 0.4400   0.03 0.2604 0.3275
-----------------------------------------------------------

+-----------------------------------------------------+
|                                                     |
|   Summary of gamma coefficients in separate strata  |
|                                                     |
| Significance evaluated by asymptotic 2-sided p-values |
|                                                     |
+-----------------------------------------------------+

  gamma    p >0.05      0.01<p<=0.05      p<0.01
--------------------------------------------------------------
Negative      46             9             1
Positive      58             5             8
```

Figure 29. **DESCRIBE DI**. Local tests of conditional independence

The global test in Figure 29 accepts that Income is not directly dependent on Intelligence. The local tests suggest, however, that there are strata, where intelligence has an effect. The most striking is the difference between test results for men and women. Among men there is a significant effect ($\gamma = 0.12$, $p = 0.005$), whereas the effect is insignificant for women. There appear, in other words, to be an interaction between the effect of Sex and the effect of Intelligence on income.

We have to be very careful, however. Following the global test results, DIGRAM produces results comparing the γ coefficient in different strata. Figure 30 shows the result of a test of whether the coefficients for men and women are different. No significant difference is found.

```
   +---------------------------------------+
   |                                       |
   | Partial Gamma coefficients in M-strata |
   |                                       |
   +---------------------------------------+

Least square estimate:  Gamma =  0.0749 s.e.  =  0.0345

M:      Sex  Gamma variance     s.e. weight  residual
-----------------------------------------------------
 1:    Male   0.12   0.0023   0.0482  0.510   1.228
 2:  Female   0.03   0.0024   0.0493  0.490  -1.228
-----------------------------------------------------

Test for partial association: X² =    1.5 df = 1 p =  0.220
```

Figure 30. **DESCRIBE DI**. Homogeneity of γ coefficient in strata defined by Sex

Figure 31 shows a similar analysis comparing γ coefficients in strata defined by School education. Again, there is no significant difference between the coefficient from different strata even though the first stratum has a markedly stronger γ coefficient. The analysis here includes at stepwise pairwise comparison procedure where it is once again concluded, that there is no difference between that strata.

```
  +---------------------------------------+
  |                                       |
  | Partial Gamma coefficients in G-strata |
  |                                       |
  +---------------------------------------+

Least square estimate:  Gamma =  0.0614 s.e.  =  0.0321

G:   School  Gamma variance     s.e. weight  residual
-----------------------------------------------------
 1:   0 - 2   0.14   0.0079   0.0891  0.130   0.991
 2:   3 - 4  -0.01   0.0048   0.0696  0.213  -1.160
 3:   5 - 8   0.08   0.0020   0.0451  0.506   0.647
 4:  9 - 12   0.02   0.0068   0.0826  0.151  -0.508
-----------------------------------------------------

Test for partial association: X² =    2.3 df = 3 p =  0.505


Analysis of collapsibility by pairwise comparison of 4 ordinal categories

Significance evaluated at each step controlling the FDR at 0.050

Collapsed:

      3 p-values. Crit. level = 0.01667  Max(p) = 0.52972  Collapsed:3 & 4
      4 p-values. Crit. level = 0.01250  Max(p) = 0.32667  Collapsed:2 & 3+4

2 groups after collaps

Groups       Mean  s.e.

1            0.14 0.089
2+3+4        0.05 0.034

      5 p-values. Crit. level = 0.01000  Max(p) = 0.32163  Collapsed:1 & 2+3+4

All groups have been collapsed
```

Figure 31. **DESCRIBE DI**. Homogeneity of γ coefficient in strata defined by School

Following these analyses, DIGRAM produce results from loglinear model fitting assuming no higher order interaction between intelligence and income and other variables. These results will not be discussed in this version of the guided tour. At the end the analysis is summarized as shown in Figure 32.

```
 Summary of analysis of conditional relationship between
Income and Intellig

 C:     SRH     Potential confounder
 B:  Unempl     Potential confounder
 F: VocEduc     Potential confounder
 G:  School     Potential confounder
 M:     Sex     Potential confounder


Summary statistics

Marginal Gamma (all cases)         =  0.20   n =   2342
Marginal Gamma (missing excluded) =  0.20   n =   2218
Partial Gamma                      =  0.07  df =   1040
```

Figure 32. **DESCRIBE DI**. Summary

“Causal” pathways

According to the model there is no *direct* effect of Intelligence on Income. There are, of course important indirect effects the description of which should be part of the description of the effect of intelligence. To disclose these invoke the

CAUSALPATH <variable pairs>

The result is shown in Figure 33. Intelligence has direct effect both on the length of your school education and on the vocational education after school, and both of these have direct effects on Income.

```
  +------------------------------------+
  |                                    |
  | Overview of paths between D and I |
  |                                    |
  +------------------------------------+

    =========================
    Paths in truncated graphs
    =========================

Indirect (mediated) effects
--------------------------------
Intellig->VocEduc->Income
Intellig->School->Income

Spurious (confounded) effects
----------------------------------
```

Figure 32. **DESCRIBE DI**. Summary