

Portland State EDA Software Package Documentation

Version 0.9

Thaddeus Shannon

Program Operation

The current EDA executable runs as a console application under DOS and Windows based operating systems. When called from a command line, optional arguments specifying control, data, and results filenames may be used.

Syntax:

eda <control file> <data file> <results file>

Default file names are “control.in”, “data.in”, and “results.out”. The user may specify all three file names, specify only *control file*, or omit the filenames all together and use the defaults. In addition to the results file, the program will also produce a file named “data.out”, which contains the data from *data file* in contingency table format.

Program Options

User specified options include the size of the dependency set, the type of statistical test to use for evaluating model significance, and the significance level to use in the test. The dependency set size limits the potential size of the best model found. More importantly, it controls the breadth of the search through the space of models. Even when the best significant model is likely to contain only a few variables, it may be useful to specify a larger dependency set size so as to search through larger sets of interacting variables.

The types of significance tests include an aggregate test, a cumulative test, and an information test. The aggregate test compares the uncertainty reduction of the proposed model to that of the next simplest model. It is the appropriate test to use when searching for an explanatory model. The cumulative test compares the uncertainty reduction of the proposed model to that of the independence model. It is appropriate when searching for a predictive model. The information option is appropriate for variable selection in contexts where regularization or structural risk minimization will be included in later stage of analysis. In the information case, the α value specifies the percentage of information loss acceptable for a reduction in model complexity. The information option is not fully supported in the current EDA release.

Control File Structure

The control file specifies the search options to use and format and handling of the data file. The control file format is:

```
Independent Variables: 26
Dependency Set Size: 9
DA Search: Cum
Dependent Variable: 4
Data File Type: Obs
Use alpha = 0.05
Omit 3 variables: 1 2 3
```

The first line specifies the number of independent variables to include in the analysis. The second line sets the size of the dependency set. The third line specifies the type of significance test to use (currently supported options are **Agg** and **Cum**). The fourth line determines which column of the data file is to be treated as the dependent variable. The fifth line indicates whether the data file is in observation format (**Obs**), or contingency table format (**Tab**). The sixth line sets the α level for the significance test (currently supported values are **0.05** and **0.01**). The final line indicates which columns (if any) of the data file should be ignored in the current analysis.

Data File Format

Two data file formats are supported: observation based or contingency table based. The observation based format treats each line as a single observation. Each character specifies a variable value, numerals, characters and punctuation marks are all valid. For example:

```
001100010100000001013010000011
010100110001010000003010000011
111001010011000101003010000011
111111100101001100014011000022
```

The table based format treats each line as a table entry with the last column (separated by a space) treated as the count for the specified cell. For example:

```
0000000000010000081111101 1
0000000000010000091011100 7
00000000000100000a0001100 4
00000000000100000b0000100 6
00000000000100000b0000101 1
0000000000010000120000000 1
0000000000010000130100000 1
```

In addition to the results file described below, the EDA program will output a contingency table of the data (i.e. the independent variables actually included plus the specified dependent variable) used in the analysis in the file “data.out”.

Results File Format

The results file contains four sections: a listing of the control file options used to generate the results, a complete listing of the most explanatory three-way interaction term for each variable found during the initial heuristic search, the listing of the initial dependency set and its evaluation, and the listing of the final dependency set and its evaluation.

control file options

Independent Variables: 29
Active Independent Variables: 22
Dependency Set Size: 11
DA Search: Cumulative
Dependent Variable: 4
Data File Type: Observation
using alpha= 0.05
Omit 7 Variables: 1 2 3 21 27 28 29

initial search results

{var, best companion, U reduction, significance (?>1), df, % reduction in U}

Uncertainty of dependent variable: 0.982000

6, 8: 0.173737, 88.759057, 3, 17.692
8, 6: 0.173737, 88.759057, 3, 17.692
30, 8: 0.173309, 49.189062, 7, 17.649
7, 8: 0.168033, 85.845084, 3, 17.111
10, 8: 0.166248, 84.933031, 3, 16.930
12, 8: 0.161701, 82.610064, 3, 16.466
16, 8: 0.160936, 82.219483, 3, 16.389
26, 8: 0.158599, 81.025635, 3, 16.151
20, 8: 0.158087, 80.764071, 3, 16.099
11, 8: 0.157428, 80.427321, 3, 16.031
14, 8: 0.156768, 80.090154, 3, 15.964
9, 8: 0.153215, 78.274625, 3, 15.602
18, 8: 0.152985, 78.157112, 3, 15.579
5, 8: 0.152950, 78.139517, 3, 15.575
15, 8: 0.151427, 77.361337, 3, 15.420
23, 8: 0.150301, 76.786300, 3, 15.306
13, 8: 0.150191, 76.730027, 3, 15.294
19, 8: 0.148194, 75.709459, 3, 15.091
17, 8: 0.147753, 75.484266, 3, 15.046
25, 8: 0.146697, 74.944760, 3, 14.939
22, 8: 0.145542, 74.354998, 3, 14.821
24, 8: 0.145358, 74.260987, 3, 14.802

initial dependency set results

Dependency set size: 11, variables: 6 8 30 7 10 12 16 26 20 11 14
degrees of freedom :4095
transmission of the dependency set is :0.570916
uncertainty reduction of: 58.138 percent
significance of the dependency set is :0.542696

final dependency set results

Dependency set size: 10, variables: 6 8 30 7 12 16 20 14 15 18
degrees of freedom :2047

transmission of the dependency set is :0.543892
uncertainty reduction of: 55.386 percent
significance of the dependency set is :1.008994