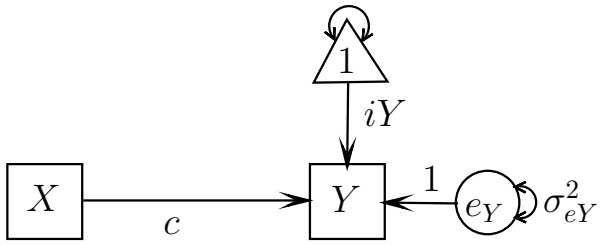
BMEM Manual

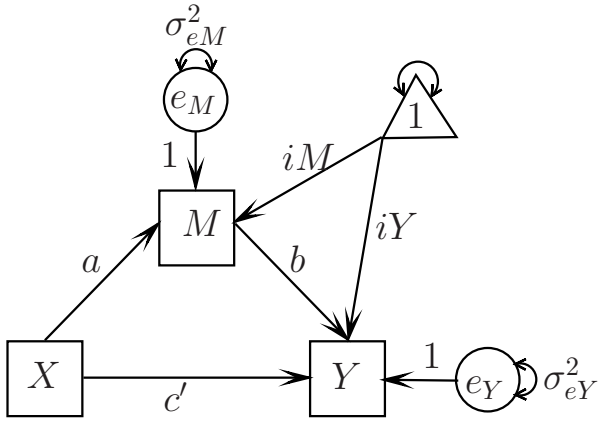
April 6, 2009

1 Background

The software BMEM estimates a mediation model specified in Figure 1.



(a) Without mediator



(b) With mediator

Figure 1: Mediation model

In the figure, X , M , and Y are input, mediator, and output variables. The unique feature of BMEM is to analyze the mediation effect with missing data. Data can be missed or unobserved on any of the three variables. However, for obvious reasons, there are two requirements for the data:

1. for each case, at least one variable is observed,
2. and at least 10 or more cases are complete for all three variables.

BMEM implements two types of missing data handling techniques - the pairwise deletion and the Expectation-Maximization (EM) algorithm together with the MLE estimation method.

To assess the mediation effect, either the stratified bootstrap method or the direct bootstrap method can be used to calculate three types of confidence intervals - the percentile interval, the bias-corrected (BC) interval, and the bias corrected and accelerated interval (BCa).

2 Files

The following files are included in the provided package. (If the package you recieved has the files BMEM.ex_ and run.ba_, please change the file names to BMEM.exe and run.bat to properly run the program.)

1. BMEM.exe: the executable program
2. Manual.pdf: the current file
3. active.txt: a subset of data from the ACTIVE study
4. comp100.txt and comp1000.txt: simulated complete data with sample sizes 100 and 1000 ($a = b = .39$ and $c = 0$)
5. mar100.txt and mar1000.txt: simulated MAR data with sample sizes 100 and 1000 ($a = b = .39$ and $c = 0$)
6. mcar100.txt and mcar1000.txt: simulated MCAR data with sample sizes 100 and 1000 ($a = b = .39$ and $c = 0$)
7. batch.txt: an example batch file to run BMEM
8. run.bat: a command file to run BMEM in batch mode.

3 How to use

There are two ways to use BMEM – the step by step method and the batch method. For both methods, we suggest putting the data file in the folder where the program file is saved.

3.1 Step by step method

After double clicking on the executable file BMEM.exe, a DOS window will pop out and require the following 7 parameters to run the program.

1. The output file name: the name of the file in which one wants to save the results of the analysis.
2. The data file name: the name of the data file. The data file should be a text file with the following sequence of variables, X , M , and Y. Variable names are not allowed in the data file. Missing data should be denoted by 99999, five 9s.
3. The α level for confidence intervals. It should be a number between 0 and 1. There is no difference to input .05 or .95. BMEM automatically recognizes the confidence level.
4. The random number seed. It should also be between 0 and 1. The random number seeds determine the bootstrap samples. The same analysis can be replicated using the identical random number seed.
5. The bootstrap sample size. We suggest a bootstrap sample size of at least 1000.
6. The missing data handling method. 1 for the EM algorithm and 2 for the pairwise deletion method.
7. The bootstrap method. 1 for the stratified bootstrap and 2 for the direct bootstrap.

3.2 Batch method

To use the batch method, one can put the seven parameters in the step by step section into a file with each parameter on one line. An example can be to create a file called `batch.txt` with the following contents

```
output.txt
input.txt
.95
.5
1000
1
1
```

There are two ways to use the batch method. The first way is to use the DOS command. Open the DOS windows through `start --> Run... --> cmd`. In the DOS window, change the directory to where BMEM is located. Then use the command `BMEM.exe < batch.txt` to run the analysis. The batch file here is saved in the same directory as BMEM. The above batch file conducts the analysis using the EM algorithm and the stratified bootstrap and constructs the 95% CIs. All the output is saved in the file `output.txt`.

The second way is more convenient and is recommended. To help the use of the batch method, a command file `run.bat` is included. To use it, one first change the parameters in the `batch.txt` file and then double click the `run.bat` file. The output is saved in the file specified in the batch file. To use the method, the data file, `BMEM.exe`, `batch.txt`, and `run.bat` should be in the same folder on your computer.

4 Example output

The following output is from the analysis of the attached ACTIVE data (active1.txt). There are several important parts of the output.

The first part includes the missing data patterns and the sample size of each pattern.

The second part includes the results from a logistic regression on the test of missing mechanism. For any pair of variables AB, it tests whether A can predict the missingness of B. If an absolute number obtained is larger than 2, one may say the missing data are not MCAR. However, one may not be able to conclude MAR, either. In addition, even if all numbers are less than 2, one still cannot conclude MCAR.

The third part includes the estimated parameters when the mediator is not considered.

The fourth part includes the estimated parameters when the mediator is included in the model.

```
-----------------------------------------------------------
| Program name: BMEM.exe (V3.0)                           |
| See manual.pdf for more information                     |
-----------------------------------------------------------


The BMEM program is run on Sun Apr 05 12:20:41 2009

The output file is: output.txt
The data file is: active.txt

--------------------------------
|     Missing Data Pattern     |
--------------------------------

Pattern  X  M  Y      Size
1        o  o  o      1932
2        o  o  x         3
3        o  x  o       195
4        o  x  x       672
The overall missing percentage is 18.34%.
NOTE: o: observed; x: missing
```

```
The confidence level is: 0.95
The random number seed is: .5


-----------------------------------------------------------
|           Testing Missing Mechanism                     |
-----------------------------------------------------------

       X         M         Y
X               -6.629    -4.685
M      N/A             -0.06249
Y      N/A       2.728
NOTE: Row variables predict missingness in column variables.
      N/A means no missing data for B in the pair AB.
      Magnitudes larger than 2 can be considered as not MCAR


The bootstrap sample size is: 1000
The missing data are handled by the EM algorithm.
The boostrap method is the stratified bootstrap.


-----------------------------------------------------------
|     Estimated parameters and confidence intervals       |
|               EM & STRATIFIED BOOTSTRAP                 |
-----------------------------------------------------------


-----------------------------------------------------------
|               Model without the mediator                |
-----------------------------------------------------------

                                  Percentile                BC                BCa
 Parameter  Estimate      S.E.       L         U         L         U         L         U
        iY   36.4006    1.9542   33.1868   39.5712   33.3245   39.6159   33.3245   39.6159
         c   -0.2337    0.0227   -0.2775   -0.1905   -0.2781   -0.1925   -0.2781   -0.1925
       eY2   29.6108    1.3187   27.7499   31.3741   27.7721   31.4097   27.8184   31.4841



-----------------------------------------------------------
|                  Model with the mediator                |
-----------------------------------------------------------

                                  Percentile                BC                BCa
 Parameter  Estimate      S.E.       L         U         L         U         L         U
        iM   87.9124    4.4688   81.3919   94.7234   81.6041   95.0530   81.6041   95.0530
        iY   10.1932    1.5230    7.3708   13.1822    7.4677   13.2625    7.4834   13.3274
         a   -0.7812    0.0532   -0.8730   -0.6949   -0.8768   -0.6974   -0.8768   -0.6974
         b    0.2981    0.0120    0.2839    0.3120    0.2839    0.3120    0.2839    0.3120
        c'   -0.0008    0.0185   -0.0387    0.0335   -0.0401    0.0325   -0.0401    0.0324
       eM2  148.7585    6.3842  140.0712  156.7297  140.3451  157.3551  140.3984  157.4002
       eY2   16.3909    0.7433   15.3558   17.3710   15.3942   17.4421   15.4336   17.4469
       a*b   -0.2329    0.0168   -0.2630   -0.2048   -0.2637   -0.2055   -0.2633   -0.2052
NOTE. See the manual for the meaning of each parameter.

The total running time is 111.9210 seconds.
```

5 Testing

Simulations have been conducted to test BMEM. Most of the functions provided in BMEM are also cross-validated. Here, we present several examples for some naive testing of BMEM. In all examples, the EM algorithm with stratified bootstrap method is used.

5.1 Complete data analysis

BMEM certainly can analyze the complete data sets. Simulated data with $a = b = .39; c = 0$ are estimated and the results are given in Table 1. Note that the results are based on one single replication of data simulation. Thus, the parameter estimates may be different from the true values.

Table 1: Results from complete data analysis

			N=100				N=1000			
		TRUE	Est.	S.E.	BCa L	BCa U	Est.	S.E.	BCa L	BCa U
No Mediator	iY		0.045	0.102	-0.168	0.242	0.010	0.034	-0.056	0.076
	c		0.344	0.112	0.137	0.575	0.151	0.034	0.087	0.218
	eY2		1.059	0.140	0.835	1.385	1.150	0.063	1.063	1.268
With Mediator	iM	0	0.051	0.111	-0.175	0.257	-0.021	0.031	-0.083	0.036
	iY	0	0.025	0.092	-0.157	0.195	0.018	0.032	-0.045	0.080
	a	0.39	0.430	0.121	0.197	0.676	0.370	0.031	0.316	0.428
	b	0.39	0.398	0.087	0.213	0.557	0.389	0.034	0.330	0.454
	c'	0	0.173	0.108	-0.037	0.383	0.007	0.034	-0.060	0.070
	eM2	1	1.238	0.160	0.995	1.602	0.966	0.054	0.889	1.064
	eY2	1	0.864	0.111	0.697	1.141	1.004	0.056	0.923	1.110
	a*b	0.152	0.171	0.064	0.064	0.313	0.144	0.017	0.116	0.180

Note. Only BCa intervals are provided here.

5.2 Missing data patterns

BMEM can generate missing data patterns. Table 2 summarizes the missing data pattern for the included example data.

Figure 2: Missing data patterns

				MCAR		MAR	
Pattern	X	M	Y	N=100	N=1000	N=100	N=1000
1	o	o	o	57	507	43	438
2	o	o	x	7	145	17	113
3	o	x	o	14	121	11	128
4	o	x	x	5	39	6	87
5	x	o	o	13	138	18	166
6	x	o	x	3	23	2	34
7	x	x	o	1	27	3	34

5.3 Logistic regression test for missing data mechanisms

Data ($N = 100$ and $N = 1000$) are generated under both MCAR and MAR. Table 2 lists the test results for the logistic regression coefficients. In general, if there is an element that is larger than 2, the missing dta are not MCAR. In simulating the MAR data, the missingness of M and Y depends on X and the missingness of X depends on both M and Y. Note that when the data are MAR, BMEM correctly identifies the missingness is not MCAR. When the data are MCAR, none of the regression coefficients is significant.

Table 2: Examples for logistic regression test

		N=100			N=1000		
		X	M	Y	X	M	Y
MCAR	X		0.128809	-0.122878		1.82052	0.0788706
	M	-1.27454		-1.41961	-0.027626		-0.534997
	Y	-1.21032	0.496444		0.194686	1.05025	
MAR	X		-2.72085	-2.64622		-10.2803	-9.22239
	M	-3.12824		-0.0996647	-11.0627		-1.54608
	Y	-4.05013	-1.29419		-10.6598	-1.1646	

5.4 Analysis with missing data

The simulated MCAR and MAR are then analyzed in BMEM. The results for $N = 1000$ are given in Table 3. Note that the results are based on only one simulation.

Table 3: Results from missing data analysis (MAR)

			MCAR				MAR			
		TRUE	Est.	S.E.	BCa L	BCa U	Est.	S.E.	BCa L	BCa U
No Mediator	iY		-0.023	0.037	-0.097	0.050	0.033	0.036	-0.039	0.101
	c		0.208	0.043	0.122	0.291	0.201	0.042	0.116	0.281
	eY2		1.094	0.069	0.986	1.216	1.105	0.066	1.007	1.225
With Mediator	iM	0	0.061	0.035	-0.006	0.129	-0.038	0.033	-0.099	0.031
	iY	0	-0.043	0.037	-0.121	0.023	0.045	0.036	-0.022	0.115
	a	0.39	0.432	0.039	0.361	0.506	0.489	0.042	0.418	0.571
	b	0.39	0.315	0.039	0.243	0.388	0.322	0.043	0.240	0.405
	c'	0	0.072	0.044	-0.014	0.160	0.043	0.048	-0.054	0.134
	eM2	1	0.973	0.058	0.884	1.075	0.943	0.062	0.852	1.062
	eY2	1	0.997	0.065	0.895	1.123	1.008	0.060	0.919	1.115
	a*b	0.152	0.136	0.020	0.101	0.175	0.158	0.024	0.113	0.205

Note. Only BCa intervals are provided here.

5.5 Computing time of BEME

The EM algorithm and bootstrap method can be very time-consuming. Table 4 gives the time used in running the examples included in the BMEM package. Note that the time consumed is related to the sample size and the percentage of missing data.

Table 4: Time consumed in seconds

	Time			Missing percentage	
	Complete	MCAR	MAR	MCAR	MAR
N=100	0.94	3.07	6.07	17.3%	22.7%
N=1000	10.44	33.75	70.58	19.4%	23.9%

6 Licence

BMEM is free software; you can redistribute it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

7 Advanced users

Researchers who are interested in obtaining the source codes for BMEM may contact xxxx. The sources codes may not be copied or distributed without permission.

8 To-do list

The future development of BMEM will focus on

- Graphical interface for BMEM,
- General mediation model using path analysis,
- Mediator and moderator model,
- and Estimation methods for MNAR data.

If you are familiar with C++ and are willing to contribute to the development of BMEM, please contact us at xxxx.

9 Citation

It is appreciated for citing the software in the following way if you are willing to use it.

```
XXXX (xxxx). Mediation analysis with missing data using EM
    algorithm and bootstrap. xxxx
```

10 Acknowledgment

The program is written in C++ to analyze the mediation effects with missing data. Two libraries are used: Newmat and newran.

1. Davies, R.B. (1994) Writing a matrix package in C++. In OON-SKI'94: The second annual object-oriented numerics conference, pp 207-213. Rogue Wave Software, Corvallis.
2. Eddelbuttel, Dirk (1996) Object-oriented econometrics: matrix programming in C++ using GCC and Newmat. Journal of Applied Econometrics, Vol 11, No 2, pp 199-209.

11 Questions or comments?

Please direct questions or comments to `xxxx`.