Structural equation models with a binary outcome using STATA and Mplus

Richard Woodman

Centre for Epidemiology and Biostatistics

Flinders University

Flinders University
Centre for Epidemiology
and Biostatistics

Motivation

- Structural equation modelling (SEM) provides a framework for assessing likely causal pathways
- Specific research question: Is Homocysteine (HCY) an independent risk factor for **CAD** or is it merely a marker of increased risk?
- Which software offers most flexibility for SEM analysis with **binary outcomes**?

Flinders University
Centre for Epidemiology and Biostatistics

Study dataset

- Elderly Chinese population (76±7 years age)
- Case-control data: 460 individuals with (50%) and without (50%) hypertension
- Cross-sectional data: Individuals with (53%) and without (47%) CAD
- 1 binary variable
  - Coronary artery disease (CAD) status
- 9 continuous variables
  - Lipids (LDL, HDL-cholesterol, Triglycerides (TG))
  - Body mass index (BMI)
  - Systolic Blood pressure (SBP)
  - Homocysteine (HCY)
  - Kidney function (Blood urea nitrogen: BUN)
  - Inflammation (C-reactive protein (CRP))
  - Oxidative stress (Uric acid (UA))

Flinders University
Centre for Epidemiology and Biostatistics

Structural Equation Modelling (SEM)

- Allows estimation of
  - Underlying “latent” factors
  - Multiple regression models
  - Direction of causal pathways
  - Strength of causal pathways
  - Direct and indirect effects
  - Tests of Mediation
- Traditionally used by the Social Sciences
- Gaining acceptance within the Health Sciences

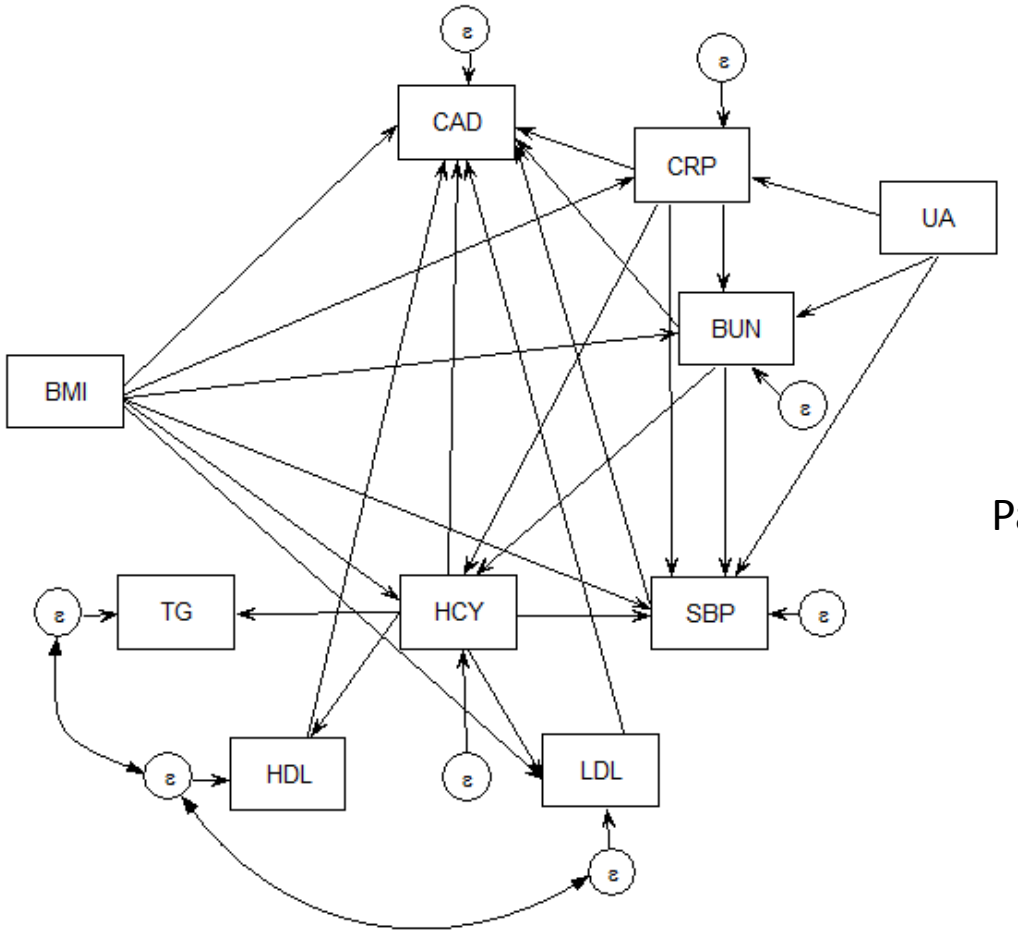
Flinders University
Centre for Epidemiology
and Biostatistics

Research objectives

- **Obtain parameter estimates**
  - Determine the **direct effect** of HCY on CAD
  - Determine **explained variance** (R^2) of each variable
  - Determine the **indirect effects** of HCY on CAD
- **Mediation**
  - Through which variables are the indirect effects mediated?
    - Blood pressure
  - Are there indirect effects of other factors via HCY?
    - Insulin sensitivity
    - Inflammation
    - Oxidative stress
- **Model fit**
  - Does the proposed causal pathway model fit?
  - Is the model the same across genders?

Flinders University
Centre for Epidemiology and Biostatistics

Hypothesised causal pathway for CAD and risk factors



Path diagram for analysis

Flinders University
Centre for Epidemiology and Biostatistics

Software for SEM

- Software packages
  - STATA
  - Mplus
  - LISREL (Joreskog, 1986)
  - EQS (Bender, 1985)
  - AMOS (SPSS add-on)
  - R (libraries: sem and semPlot)
  - SmartPLS
- Analysis of binary outcomes available in
  - STATA (since version 13; 2013)
  - Mplus (since version 2; 2001)

Flinders University
Centre for Epidemiology
and Biostatistics

SEM estimation with categorical outcomes

- ML estimation requires numerical integration for combination of
  - Categorical outcomes and
    - Continuous latent variables
    - Missing data
- Numerical integration available in
  - STATA
  - Mplus
- Mplus has 2 additional estimation options
  - Weighted least squares (WLS)
  - Bayesian

Flinders University
Centre for Epidemiology
and Biostatistics

Mplus estimation methods with categorical outcomes

- Default method for categorical outcomes is means and variance adjusted weighted least squares
  - (Estimator=WLSMV)
  - Uses probit regression (CDF for CAD treated as a latent variable)
  - Computationally demanding
- ML estimation
  - (Estimator=ML)
  - Rectangular, Gauss-Hermite or Monte Carlo integration
  - With or without adaptive quadrature
- Bayes estimation

Flinders University
Centre for Epidemiology and Biostatistics

STATA estimation methods with categorical outcomes

- GSEM
  - ML with numerical integration is default for GSEM
  - The **only** estimator option for categorical outcomes
- Integration methods
  - Mean-variance adaptive gauss hermite (mvaghermite) (the default)
  - Mcaghermite (computationally intensive but better convergence)
  - Ghermite
  - Laplace (less accurate but less computationally intensive)
- Technique (for VCE)
  - Observed information matrix (OIM)

Flinders University
Centre for Epidemiology and Biostatistics

STATA code

```
gsem (CAD <- HCY CRP SBP LDL HDL BUN BMI, family(binomial) link(logit)) ///
     (BUN <- BMI CRP UA) ///
     (CRP <- BMI UA) ///
     (SBP <- BUN HCY BMI UA CRP) ///
     (HCY <- BMI BUN CRP) ///
     (LDL <- BMI HCY) ///
     (TG <- HCY) ///
     (HDL <- HCY) ///
     if sex==0, cov(e.TG*e.HDL e.HDL*e.LDL) nocapslatent ///

     method(ml) ///
     vce(oim) ///
     intmethod(mvaghermite) ///
     iterate(1001)
```

Flinders University
Centre for Epidemiology
and Biostatistics

Mplus code (for ML)

```
VARIABLE:
  Names are
   sex age HCY TG HDL LDL BUN CR UA CRP BS SBP DBP CAD BMI group;
   Missing are all (-9999);
   Usevariables are HCY TG HDL LDL BUN SBP  CAD BMI CRP UA;
   Categorical is CAD;
   Useobservations are sex==0;

ANALYSIS:
estimator=ml;
iter=200000;
Algorithm=int;
integration=GAUSSHERMITE;
Adaptive=on;

MODEL:
  CAD on BUN SBP HCY HDL LDL CRP BMI;
  BUN on BMI CRP UA;
  CRP on BMI UA;
  SBP on BUN HCY BMI UA CRP;
  HCY on BMI BUN CRP;
  LDL on BMI HCY;
  TG on HCY;
  HDL on HCY;
  TG with HDL; LDL with HDL;

OUTPUT:stdyx;tech1 tech2;modindices(3)

Model indirect:
          CAD ind HCY;
          CAD ind BUN;
          CAD ind BMI;
          CAD ind SBP;
          CAD ind LDL;
          CAD ind HDL;
          CAD ind CRP;
          CAD ind UA;
```

To obtain indirect effects on CAD with 95% CI's

Flinders University
Centre for Epidemiology and Biostatistics

Typical SEM research questions

- Parameter estimates
  - Non-standardised
  - Standardised
- Model fit
  - Absolute fit (χ^2 for proposed model versus saturated model)
  - Relative fit (AIC/BIC)
- Test for group invariance of parameter estimates
  - i.e. can the same parameter estimates be used for different groups?
  - E.g. Males versus females, race
  - Typically uses
    - χ^2 difference testing of constrained and unconstrained models
    - Difference in -2 LL
- Estimate indirect effects

Flinders University
Centre for Epidemiology and Biostatistics

Non-standardised β's

	β STATA GSEM (Logit coefficient)	β Mplus ML (Logit coefficient)	β Mplus WLSMV (Probit coefficient)	β Mplus Bayes (Probit coefficient)
Males				
CAD				
HCY	**0.311±0.046**	**0.311±0.046**	**0.108±0.019**	**0.187±0.024**
CRP	0.119±0.131	0.119±0.132	0.047±0.050	0.059±0.069
SBP	**0.034±0.014**	**0.034±0.014**	**0.013±0.004**	**0.014±0.007**
LDL	-0.28±0.299	-0.28±0.299	-0.048±0.094	-0.202±0.158
HDL	0.527±00.681	0.527±0.681	0.158±0.226	0.119±0.353
BUN	0.114±0.122	0.114±0.122	**0.089±0.045**	0.049±0.064
BMI	0.037±0.057	0.037±0.057	0.020±0.019	-0.001±0.029

Flinders University
Centre for Epidemiology and Biostatistics

Standardised β's

	β STATA GSEM	β Mplus ML	β Mplus WLSMV	β Mplus Bayes
Males				
CAD				
HCY	N/A	**0.62±0.08**	**0.58±0.08**	**0.68±0.07**
CRP	N/A	0.07±0.07	0.07±0.08	0.06±0.07
SBP	N/A	**0.20±0.07**	**0.21±0.07**	**0.15±0.07**
LDL	N/A	-0.06±0.07	0.030±0.058	-0.08±0.06
HDL	N/A	0.05±0.06	0.04±0.06	0.02±0.06
BUN	N/A	0.07±0.07	**0.15±0.08**	0.05±0.07
BMI	N/A	0.04±0.07	0.07±0.06	-0.001±0.06

Flinders University
Centre for Epidemiology and Biostatistics

CAD as continuous - standardised β's

Males	β STATA SEM	β Mplus ML	β Mplus Bayes
CAD			
HCY	**0.65±0.06**	**0.64±0.05**	**0.63±0.05**
CRP	0.07±0.05	0.07±0.05	0.07±0.05
SBP	**0.11±0.04**	**0.11±0.05**	**0.11±0.05**
LDL	-0.037±0.039	-0.032±0.04	-0.032±0.04
HDL	0.038±0.041	0.038±0.04	0.39±0.04
BUN	0.026±0.04	0.024±0.04	0.02±0.04
BMI	0.02±0.04	0.02±0.04	0.022±0.04
χ^2	49.2 (38df); p=0.11	48.8 (37df); p=0.09	
Satorra-Bentler χ^2	46.3 (38df); p=0.17	47.9 (37df); p=0.11	

Flinders University
Centre for Epidemiology and Biostatistics

Model fit - Mplus

Absolute fit (χ^2 test of model fit) with WLSMV

```
Value                              32.717*
Degrees of Freedom                 36
P-Value                            0.6255
χ² Contribution From Each Group
MALES                              12.877
FEMALES                            19.839
```

Relative Fit (AIC/BIC) with ML (single groups only)

```
Loglikelihood    H0 Value    -2567.236
Akaike (AIC)                  5216.472
Bayesian (BIC)                5348.727
Sample-Size Adjusted BIC      5218.866
```

Nested model comparisons

WLSMV: Use **difftest** option

```
SAVEDATA:
difftest is mydiff.dat;
ANALYSIS:
difftest is mydiff.dat;

Chi-Square Test for Difference Testing
          Value                28.409
          Degrees of Freedom   22
          P-Value              0.1625
```

ML: Apply with and without **model constraint** option and compare -2LL e.g:

```
MODEL CONSTRAINT:
0 = b1;

Loglikelihood  H0 Value  -2567.854
Loglikelihood  H0 Value  -2567.236
```

Flinders University
Centre for Epidemiology and Biostatistics

Testing group invariance - Mplus

WLSMVχ^2 test of model fit

Unconstrained model

```
VARIABLE:
Grouping is sex (0=males, 1=females)
SAVEDATA:
difftest is mydiff.dat;

Value                    32.717*
Degrees of Freedom       36
P-Value                  0.6255
χ² Contribution From Each Group
MALES                    12.877
FEMALES                  19.839
```

Constrained model

```
ANALYSIS:
estimator=wlsmv;
iter=20000;
difftest is mydiff.dat;
MODEL:
BUN on BMI(b1); etc.

Chi-Square Test for Difference Testing
Value                    28.409
Degrees of Freedom       22
P-Value                  0.1625
```

ML: Mixture models

```
VARIABLE:
  Categorical is CAD;
  classes=sex(2);
  knownclass= sex (sex=0, sex=1);
ANALYSIS:
 type=mixture;
 estimator=ml;
 iter=20000;
 algorithm=integration;
```

Unconstrained model

```
MODEL:
%overall%
Model code
%sex#1%
Model code
%sex#2%
Model code
```

Constrained Model

```
MODEL:
%OVERALL%
Model code

Number of Free Parameters   76
Loglikelihood H0 Value      -6589.617
Number of Free Parameters   50
Loglikelihood H0 Value      -6572.265

di chi2(34.7, 26)
.14339388
```

Flinders University
Centre for Epidemiology and Biostatistics

Mplus versus STATA for categorical outcomes

	Mplus (WLSMV)	Mplus (ML)	STATA (GSEM)
Estimates			
Non-standardised	✓		✓
Standardised	✓		×
Model fit			
Absolute fit (χ^2 test of model fit)	✓		×
Relative fit (AIC/BIC)	✓		✓
Nested models (χ^2 diff testing with LL)		✓	✓
Test for group invariance			
with χ^2 difference testing	✓		×
with -2 x Log Likelihood difference testing		✓ (ML Mixture model)	×
Test of indirect effects	✓		×
R^2 for CAD	✓	✓	×

Flinders University
Centre for Epidemiology and Biostatistics

Summary of results

- Treating binary variables as continuous can produce quite biased results although substantive conclusions remain
- Mplus allows 3 estimation options versus 1 for STATA
  - WLSMV more accurate? (Psychological Methods, 17(3): 354-373)
- Mplus provides
  - tests of absolute fit
  - tests of indirect effects for ML
  - testing for group invariance using WLSMV (`difftest`)
  - Testing for group invariance using ML (mixture model)
  - standardised estimates for ML
  - R^2 estimates

Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From syntax to diagram

Step 1: Run from syntax file

Mplus - [hcy_wlsmv_M&Funconstrained.inp]

File Edit View Mplus Plot Diagram Window Help

```
TITLE:
   Hcy_ml_Males
DATA:
  File is hcy.dat;
  LISTWISE=ON;
VARIABLE:
  Names are
      sex age HCY TG HDL LDL BUN CR UA CRP BS SBP DBP CAD BMI group;
  Missing are all (-9999);
  Usevariables are HCY TG HDL LDL BUN SBP  CAD BMI CRP UA;
  Categorical is CAD;
   grouping is sex (0=males, 1=females);
 !   USEOBSERVATIONS ARE SEX==0;

ANALYSIS:
! estimator=wls;
 iter=30000;
MODEL:
  BUN on BMI CRP UA;
  CRP on BMI UA;
  SBP on BUN HCY BMI UA CRP;

  CAD on BUN SBP HCY HDL LDL CRP BMI;

  HCY on BMI BUN CRP;
  LDL on BMI HCY;
  TG on HCY;
  HDL on HCY;

  TG with HDL; LDL with HDL;
  CAD with TG@0;
savedata:
 ! difftest is mydiff2.dat;
OUTPUT:stdyx;tech1 tech2;modindices(3)

Model indirect:
          CAD ind HCY;
          CAD ind BUN;
          CAD ind BMI;
          CAD ind SBP;
          CAD ind LDL;
          CAD ind HDL;
          CAD ind CRP;
          CAD ind UA;
```

Ready Ln 32, Col 1 NUM

Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From syntax to diagram

Step 2: In the output file, click: Diagram - View diagram

Mplus - [hcy_wlsmv_m&funconstrained.out]

File Edit View Mplus Plot Diagram Window Help

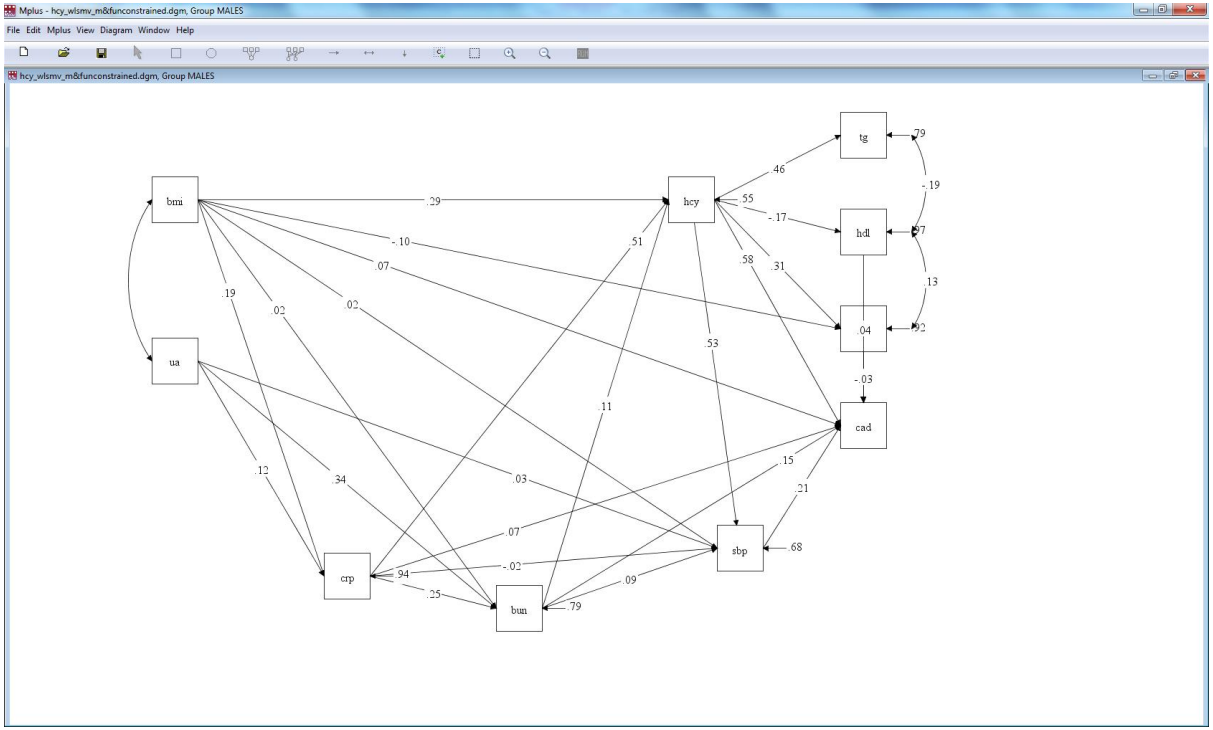
```
Mplus VERSION 7.31
MUTHEN & MUTHEN
08/12/2015  12:48 PM

INPUT INSTRUCTIONS

  TITLE:
     Hcy_ml_Males
  DATA:
    File is hcy.dat;
    LISTWISE=ON;
  VARIABLE:
    Names are
       sex age HCY TG HDL LDL BUN CR UA CRP BS SBP DBP CAD BMI group;
    Missing are all (-9999);
    Usevariables are HCY TG HDL LDL BUN SBP  CAD BMI CRP UA;
    Categorical is CAD;
     grouping is sex (0=males, 1=females);
   !   USEOBSERVATIONS ARE SEX==0;

  ANALYSIS:
  ! estimator=wls;
   iter=30000;
  MODEL:
    BUN on BMI CRP UA;
    CRP on BMI UA;
    SBP on BUN HCY BMI UA CRP;

    CAD on BUN SBP HCY HDL LDL CRP BMI;

    HCY on BMI BUN CRP;
    LDL on BMI HCY;
    TG on HCY;
    HDL on HCY;

    TG with HDL; LDL with HDL;
    CAD with TG@0;
  savedata:
   ! difftest is mydiff2.dat;
  OUTPUT:stdyx;tech1 tech2;modindices(3)

  Model indirect:
            CAD ind HCY;
            CAD ind BUN;
            CAD ind BMI;
            CAD ind SBP;
            CAD ind LDL;
            CAD ind HDL;
            CAD ind CRP;
            CAD ind UA;


INPUT READING TERMINATED NORMALLY
```

Ready Ln 1, Col 1 NUM

Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From syntax to diagram

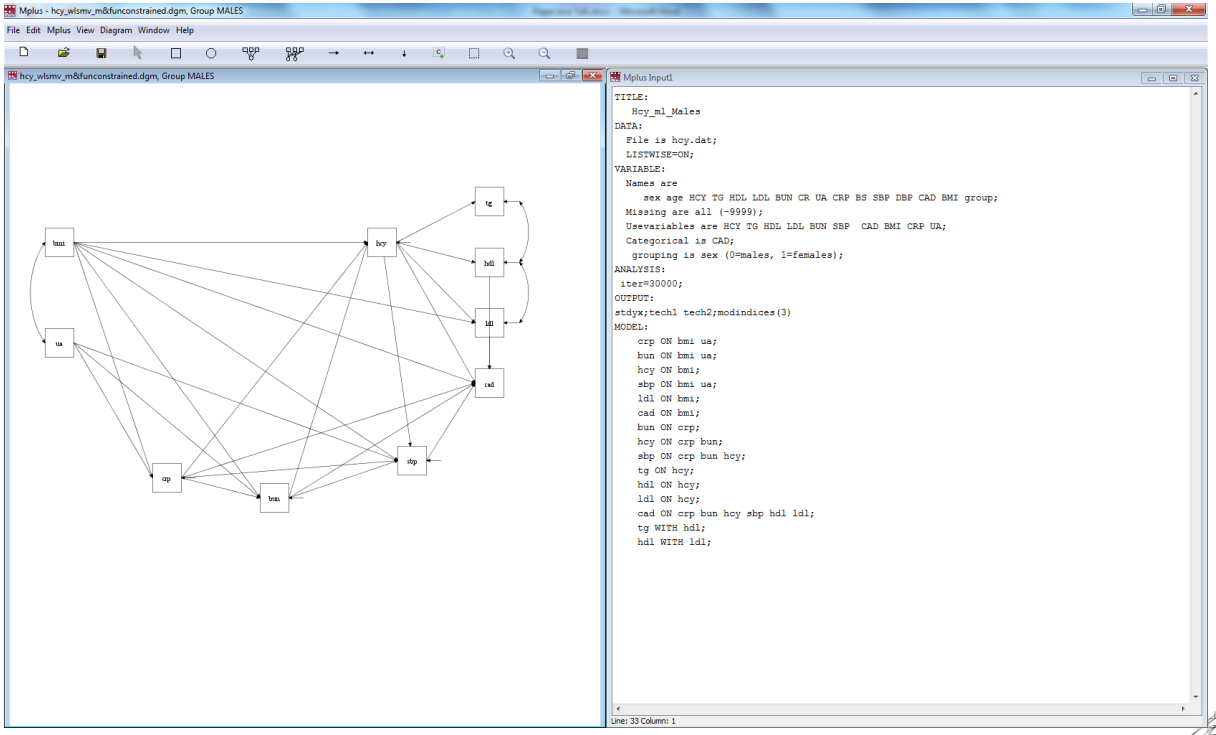
Step 3: This brings up the model with the estimates (.dgm file)



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From syntax to diagram

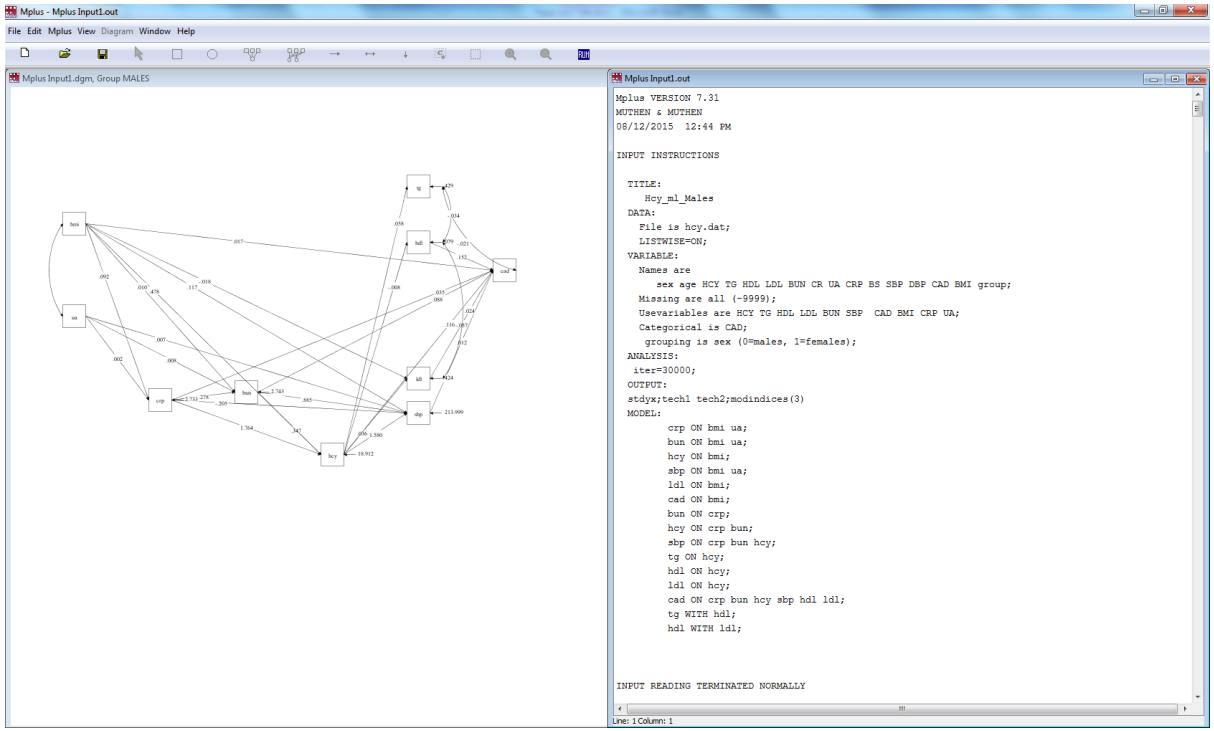
Step 4: Go to Input mode (click on Diagram-Input), and either alter the syntax in the newly written Input file, or alter the path diagram (.mdg file) (this will automatically alter the syntax). Save input file and click "Run"



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From syntax to diagram

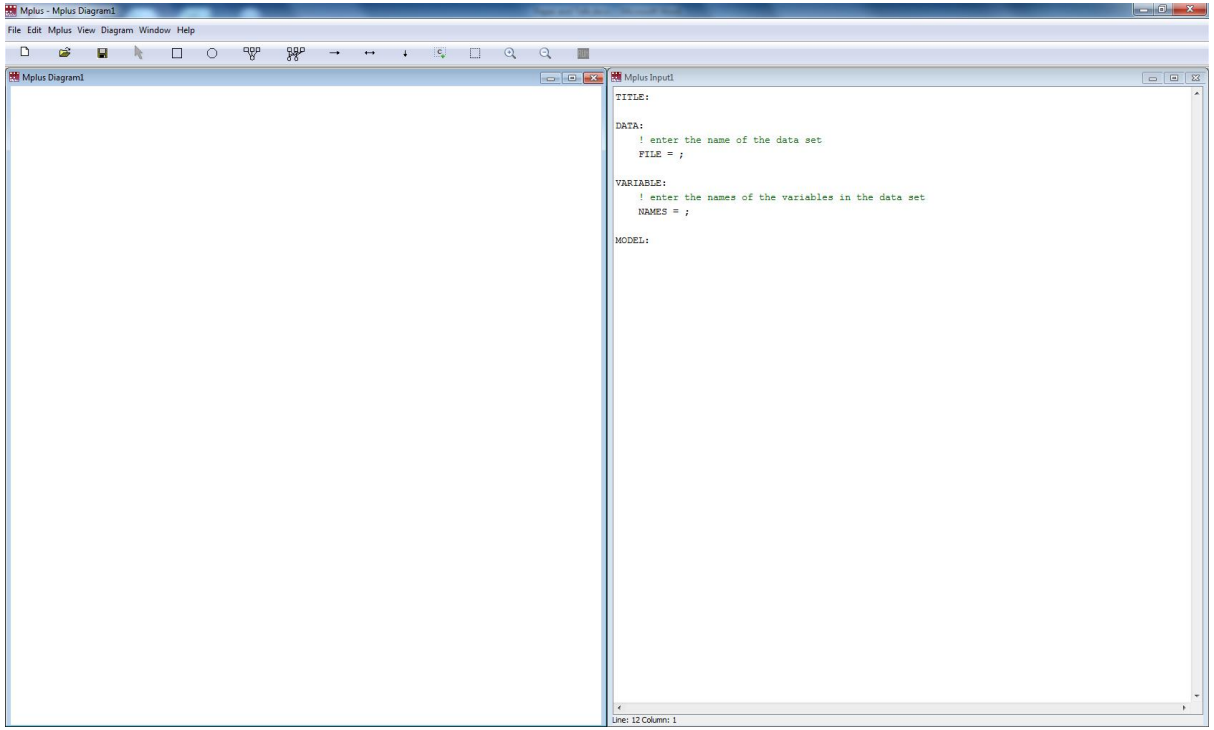
Step 5: View output and new path diagram



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From diagram to syntax

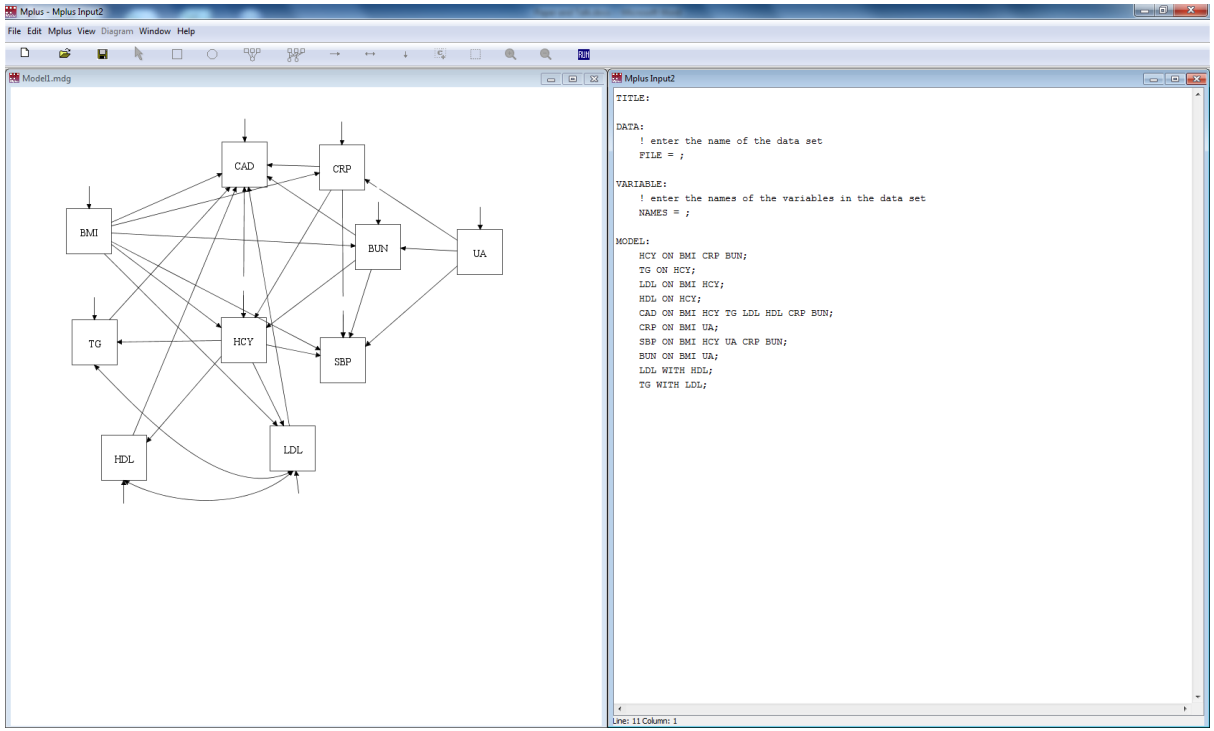
Step 1: Open up Diagrammer from within Mplus Editor (Diagram – Open Diagrammer)



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From diagram to syntax

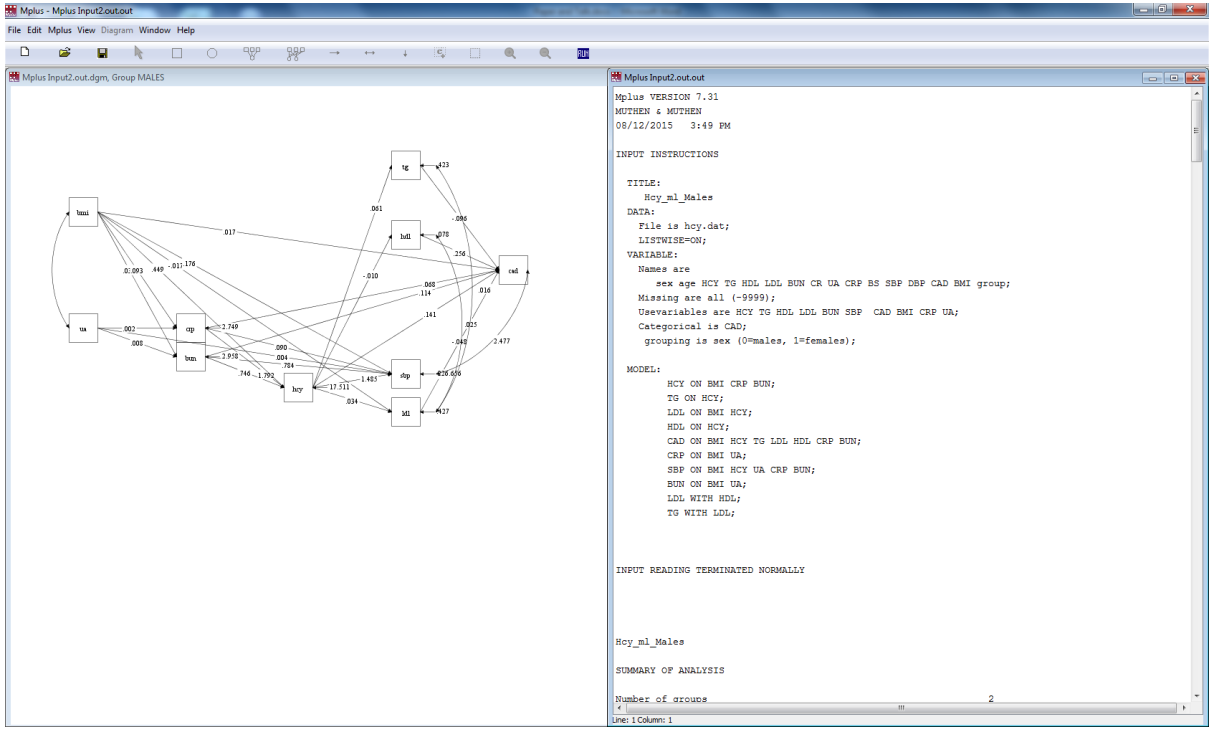
Step 2: Create path diagram. The model part of the syntax will appear on the RH side but not other aspects of the syntax. The path diagram is a .mdg file.
The syntax file is a .inp file.



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – Mplus: From diagram to syntax

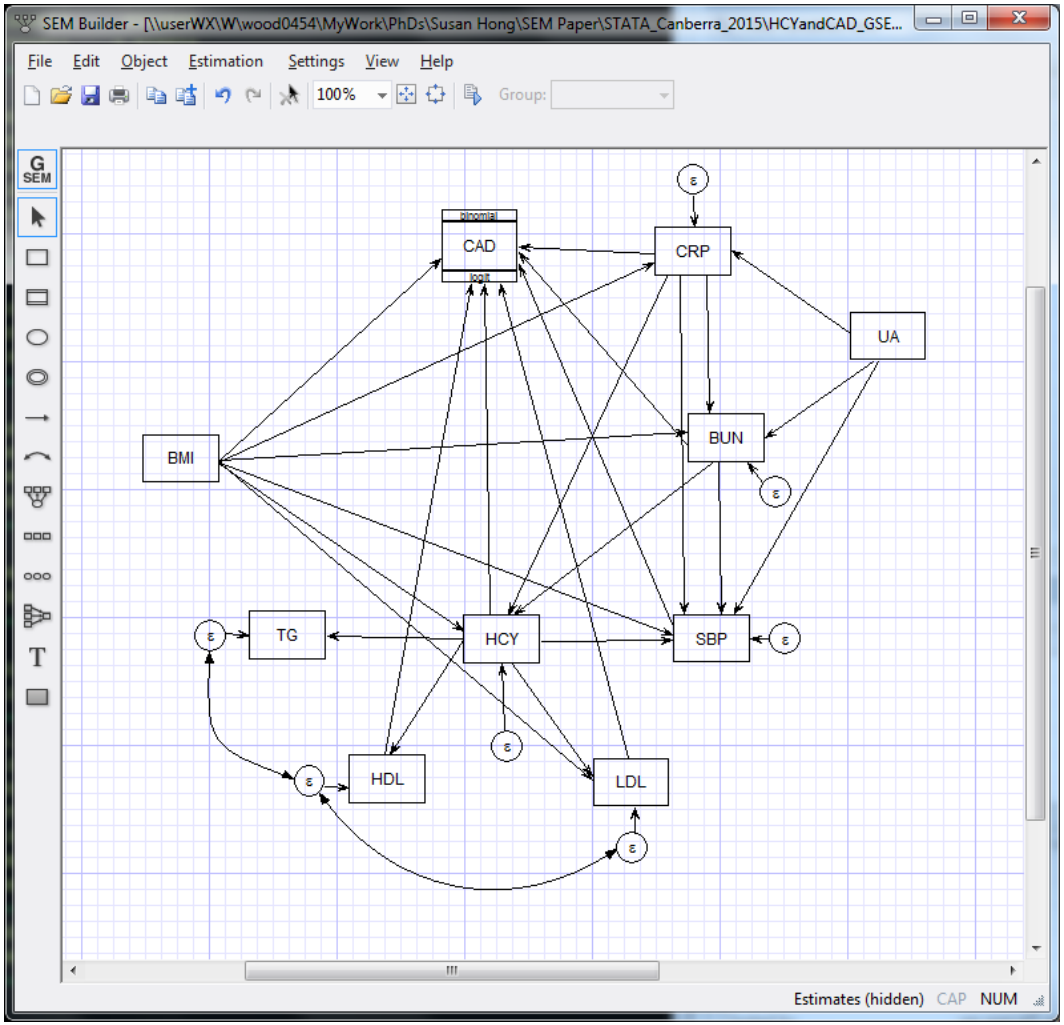
Step 3: Save the Input file and click Run. This will produce a path diagram (.dgm file) with estimates and some output. This is the equivalent of step 5 for option 1



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – STATA

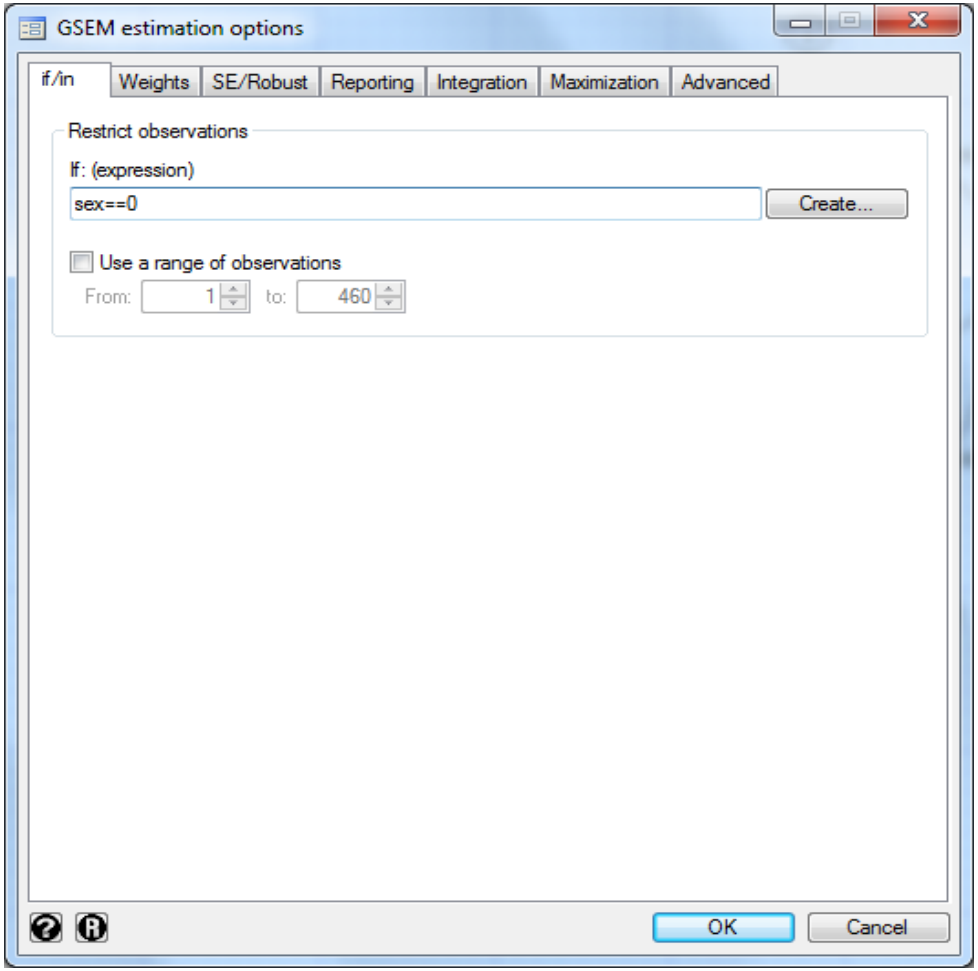
Step 1: Draw diagram



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – STATA

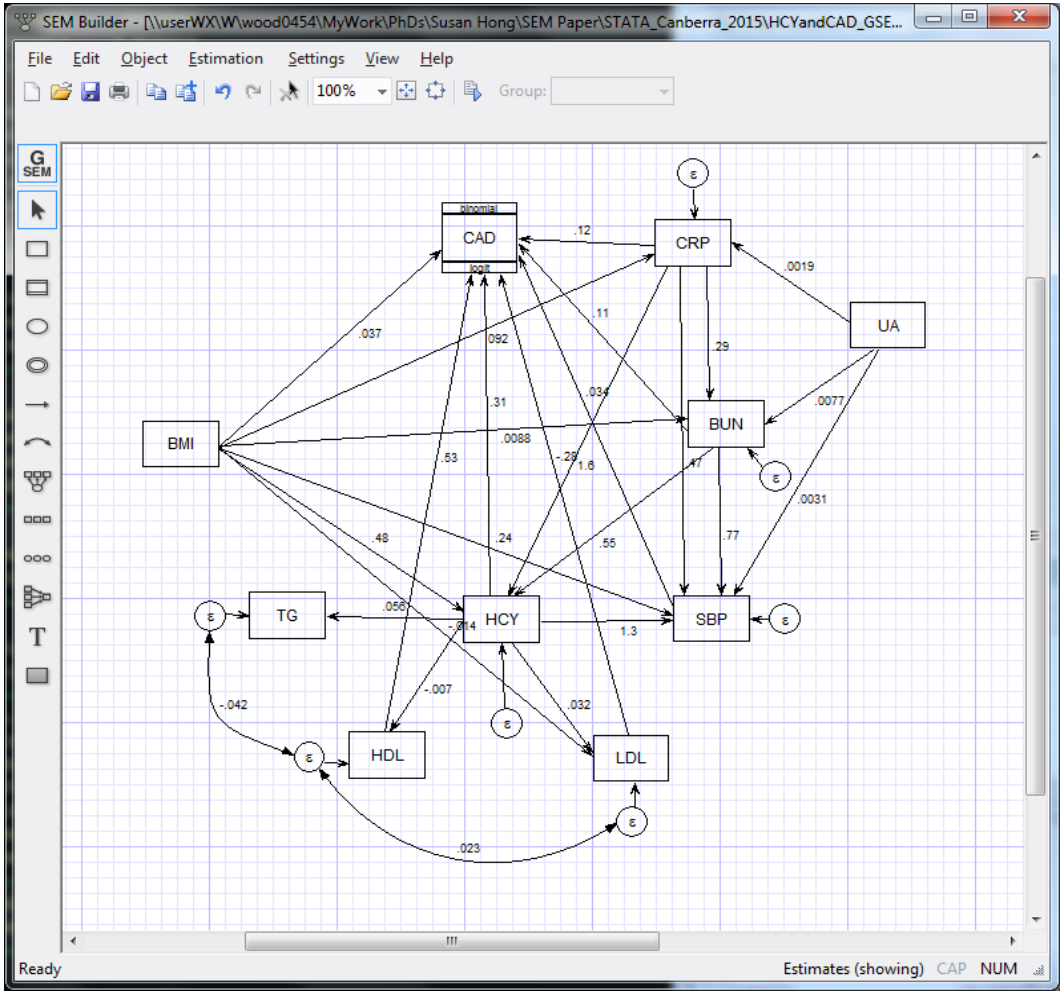
Step 2: Select options and click OK



Flinders University
Centre for Epidemiology
and Biostatistics

Diagrammer – STATA

Step 3: View results and Output



Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer – STATA

Step 4: Copy syntax from Output window

```
. gsem (BMI -> HCY, ) (BMI -> CAD, family(binomial) link(logit))
(BMI -> CRP, ) (BMI -> SBP, ) (BMI -> LDL, ) (BMI -> BUN, )
> (HCY -> TG, ) (HCY -> CAD, family(binomial) link(logit)) (HCY ->
SBP, ) (HCY -> LDL, ) (HCY -> HDL, ) (CRP -> HCY, ) (CRP -
> > CAD, family(binomial) link(logit)) (CRP -> SBP, ) (CRP -> BUN,
) (SBP -> CAD, family(binomial) link(logit)) (LDL -> CAD,
> family(binomial) link(logit)) (HDL -> CAD, family(binomial)
link(logit)) (UA -> CRP, ) (UA -> SBP, ) (UA -> BUN, ) (BUN ->
> HCY, ) (BUN -> CAD, family(binomial) link(logit)) (BUN -> SBP, )
if sex==0, cov( e.TG*e.HDL e.HDL*e.LDL) nocapslatent
```

Flinders University
Centre for Epidemiology
and Biostatistics

Diagrammer - STATA

- PROS
  - Simple to create
    - observed variables, factors, paths, variable names
  - Path diagram (.stem) files can be
    - saved and modified
    - converted to other file forms (.pdf, .tiff etc.)
  - Additional estimation options easy to apply via a GUI
  - Writes out the corresponding syntax when run
- CONS
  - Some aspects of drawing are a bit tricky
    - Resizing
    - Variances and co-variance arrows are hard work to get just right
  - Cannot produce a diagram from syntax

Flinders University
Centre for Epidemiology and Biostatistics

Diagrammer - Mplus

- PROS
  - Writes syntax as a diagram is drawn
  - Provides a diagram from syntax
- CONS
  - Automatic xxx.dmg output files often ugly
  - Dealing with 2 rather than 1 file type
    - .mdg (the hand drawn diagram file from scratch)
    - .dmg (the automatically produced diagram from syntax estimation)

Flinders University
Centre for Epidemiology
and Biostatistics

Diagrammer comparison

	Mplus	STATA
Run a diagram to produce syntax	✓	✓✓
Run syntax to produce a diagram	✓	×
Run syntax to produce a **nice** diagram	×	×
Diagrams simple to create	✓	✓✓
Diagrams convert to .pdf, .tiff	✓	✓
Wizard option to improve appearance (available in some packages e.g. AMOS)	×	×

Flinders University
Centre for Epidemiology and Biostatistics

Overall Summary of results

- PRO's for Mplus
  - 3 estimation options (ML, WLS, Bayes)
  - Provides
    - Tests of model fit (WLS estimator)
    - Indirect effects (ML and WLS)
    - Standardised estimates (ML and WLS)
    - Testing for group invariance (ML and WLS)
    - R^2 estimate

- PRO's for STATA
  - Only one estimation option to choose from!
  - Better path diagrammer
    - Diagrams easier to draw
    - For saving diagrams - pdf's **and** tiff's
    - For obtaining the syntax from the diagram
    - HELP menu

Flinders University
Centre for Epidemiology and Biostatistics

Acknowledgements

Dr Susan Hong (PhD student), School of Public Health, Central South University, China

Prof Shuiyuan Xiao (PhD supervisor), Institute of Gerontology, Hunan Geriatric Hospital, China

Prof Arduino A Mangoni: Clinical Pharmacology, Flinders University