SEM Course practicals
Prac 1 – Introduction and Simple Regression Models

1 Introduction

We will be fitting a range of models using a subsample of the ALSPAC cohort. The data has been restricted to a sample of 1,500 young people (750 boys, 750 girls) who have complete data on all the measures we will use.

It is advisable to use an alternative stats package to derive any data you will require for your analysis as the Mplus approach is a little clunky and long-winded. Both Stata and Mplus have functions that enable Mplus datasets to be created quite easily. Having said that, we will show you how to recode or create new measures in Mplus (using the *define* option) on occasion because it is sometimes quicker to do this than go back to Stata and make a brand new file.

Below is a slightly truncated summary of the variables in the file.

varname	Description and codes
id	A randomly generated ID serving no purpose in the data file
Sex	Male = 1, Female = 2
The file then contains 3 sets of EAS items which are intended to tap into 4 constructs – activity, emotionality, shyness and sociability. Time points are denoted t1/t2/t3 and these refer to the ages of 3yr 2mn, 4yr 9mn & 5yr 9mn. Response options were Not at all like him/her", "Not much like him/her", "Somewhat like him/her", "Quite like him/her" and "Exactly like him/her" for t1; and "Never", "Rarely", "Sometimes", "Often", "Always" for t2 and t3.	
act_t1_1	Always on the go (+ve)
act_t1_2	Moves about slowly (-ve)
act_t1_3	Active on waking (+ve)
act_t1_4	Very energetic (+ve)
act_t1_5	Prefers quiet games (-ve)
emo_t1_1	Cries easily (-ve)
emo_t1_2	Emotional (-ve)
emo_t1_3	Often fusses and cries (-ve)
emo_t1_4	Gets upset easily (-ve)
emo_t1_5	Reacts intensely when upset (-ve)
shy_t1_1	Shy (-ve)
shy_t1_2	Makes friends (+ve)
shy_t1_3	Sociable (+ve)
shy_t1_4	Takes time warming to strangers (-ve)
shy_t1_5	Friendly with strangers (+ve)
soc_t1_1	Likes being with people (+ve)
soc_t1_2	Prefers playing with others (+ve)
soc_t1_3	Finds people stimulating (+ve)
soc_t1_4	Something of a loner (-ve)
soc_t1_5	Isolated when alone (+ve)

	Followed by the 20 items of EAS at t2 and t3
mumage	Maternal age
tenure	Housing tenure (0 = mortgaged, 1 = private rented, 2 = subsidized rented)
crowding	Home overcrowding (> 1 person per room; 0=no, 1=yes)
parity	Parity (0=1st born, 1=2nd born, 2 = 3rd born+)
mumed	maternal educational attainment (0 = A-level+, 1 = O-level, 2 = <O-level)
income	Household income (0 = bottom 20%, 1 = middle 60%, 2 = top 20%)
social	Social class (0 = I/II, 1 = III non-manual or lower)
mumalc	Regular maternal alcohol use in the early postnatal period (0=no, 1=yes)
mumsmk	Maternal cigarette use in the early postnatal period (0=none, 1=low, 2=high)
mdep_pn	Mother exceeding threshold for EPDS in early postnatal period (0=no, 1=yes)
mfq10_*	13 short MFQ depressive symptoms at age 10
mfq18_*	13 short MFQ depressive symptoms at age 18
emotott1	Sum-score for EAS emotionality at time 1
emotott2	Sum-score for EAS emotionality at time 2
emotott3	Sum-score for EAS emotionality at time 3
	etc.

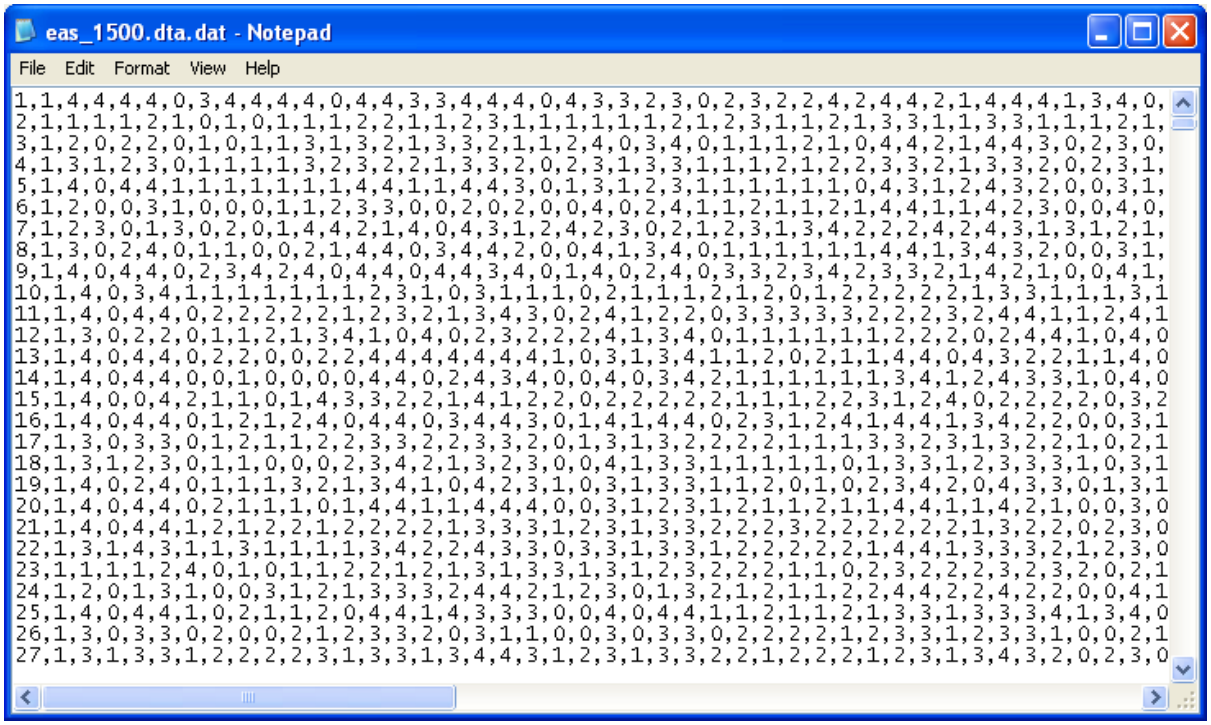
1.1 The input file

Open up the input file called 'prac 1.1.inp'. This should look like this:-

```
Data:
  File is H:\Courses\SEM_2012\data\eas_1500.dta.dat;

Variable:
  Names are id
     sex
     act_t1_1 act_t1_2 act_t1_3 act_t1_4 act_t1_5
     emo_t1_1 emo_t1_2 emo_t1_3 emo_t1_4 emo_t1_5
     shy_t1_1 shy_t1_2 shy_t1_3 shy_t1_4 shy_t1_5
     soc_t1_1 soc_t1_2 soc_t1_3 soc_t1_4 soc_t1_5
     act_t2_1 act_t2_2 act_t2_3 act_t2_4 act_t2_5
     emo_t2_1 emo_t2_2 emo_t2_3 emo_t2_4 emo_t2_5
     shy_t2_1 shy_t2_2 shy_t2_3 shy_t2_4 shy_t2_5
     soc_t2_1 soc_t2_2 soc_t2_3 soc_t2_4 soc_t2_5
     act_t3_1 act_t3_2 act_t3_3 act_t3_4 act_t3_5
     emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5
     shy_t3_1 shy_t3_2 shy_t3_3 shy_t3_4 shy_t3_5
     soc_t3_1 soc_t3_2 soc_t3_3 soc_t3_4 soc_t3_5
     mumage tenure crowding parity mumed income social
     mumalc mumsmk mdep_pn
     mfq10_01 mfq10_02 mfq10_03 mfq10_04 mfq10_05 mfq10_06
     mfq10_07 mfq10_08 mfq10_09 mfq10_10 mfq10_11 mfq10_12 mfq10_13
     mfq18_01 mfq18_02 mfq18_03 mfq18_04 mfq18_05 mfq18_06
     mfq18_07 mfq18_08 mfq18_09 mfq10_10 mfq18_11 mfq18_12 mfq18_13
     emotott1 emotott2 emotott3 acttott1 acttott2 acttott3
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  Missing are all (9999);

Analysis:
  Type = basic;
```

The data section points to the text data file. The variable section lists the names of the variables and the analysis section is currently set up to carry out a basic analysis which will generate sample stats for all the variables.

If you open up the datafile you'll see that its comma delimited text with no variable names. Other delimit options are also accepted. This underlines the utility of something like Stata2mplus to create your dataset and input file. If you had to type all of your variable names in by hand you might get them out of order leading to all sorts of problems.



1.2 A "basic" analysis

Running the input file as it is will swamp you with output including useful stats such as the covariance between emotionality and your ID. We use the "usevariables" option within the variable section to focus on subsets of the data.

Select the EAS sumscores for a basic analysis by adding these lines to the variable section:-

```
usevariables =
   emotott1 emotott2 emotott3 acttott1 acttott2 acttott3
   shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;
```

Note that there is ONE semi-colon at the end of the command. Also note that Mplus has an 80 character limit for lines, hence this is split into three. If you run into problems then open up and use the input file 'prac 1.2.inp' instead.

Click the blue RUN button and have a well earned 1-second rest while the program runs.

Firstly you will see that Mplus outputfiles (.out) contain the input syntax – useful if inp and out are separated.

Secondly we have a summary of the analysis

```
SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        1500

Number of dependent variables                                   12
Number of independent variables                                  0
Number of continuous latent variables                            0

Observed dependent variables

  Continuous
   EMOTOTT1    EMOTOTT2    EMOTOTT3    ACTTOTT1    ACTTOTT2    ACTTOTT3
   SHYTOTT1    SHYTOTT2    SHYTOTT3    SOCTOTT1    SOCTOTT2    SOCTOTT3
```

Indicating that there are 12 continuous variables and 1500 cases.

Next we have a section on missing data issues but this is a complete-case dataset so there is very little to see here. There is a single missing data pattern – denoted by a column of X's, and the covariance coverage for each pair of variables is 1. This output is sometimes useful for flagging if one variable in particular is the cause of a large amount of missing data.

Finally we have the summary statistics. I've doctored these slightly to fit them better on the page:-

```
ESTIMATED SAMPLE STATISTICS

Means

 EMOTOTT1    EMOTOTT2    EMOTOTT3    ACTTOTT1    ACTTOTT2    ACTTOTT3
 ________    ________    ________    ________    ________    ________
    7.327       7.771       7.693       3.630       4.726       4.899


 SHYTOTT1    SHYTOTT2    SHYTOTT3    SOCTOTT1    SOCTOTT2    SOCTOTT3
 ________    ________    ________    ________    ________    ________
    7.279       7.517       7.323       6.835       6.893       6.833
```

No clear pattern for changing means through time for any of the measures. Sum-scales were coded so that a high score indicates being more shy, more emotional, less active or less sociable.

Covariances

	EMOTOTT1	EMOTOTT2	EMOTOTT3	ACTTOTT1	ACTTOTT2	ACTTOTT3
EMOTOTT1	**16.559**					
EMOTOTT2	8.296	**11.408**				
EMOTOTT3	7.664	7.890	**11.518**			
ACTTOTT1	1.844	1.303	1.231	**10.498**		
ACTTOTT2	1.470	1.646	1.511	6.386	**8.884**	
ACTTOTT3	1.481	1.653	1.927	5.659	6.431	**8.799**
SHYTOTT1	1.306	0.839	0.890	1.699	1.060	1.180
SHYTOTT2	0.937	0.963	0.880	0.714	1.059	0.875
SHYTOTT3	1.021	1.005	1.276	0.702	0.885	1.237
SOCTOTT1	-0.192	0.017	0.032	3.776	2.439	2.510
SOCTOTT2	-0.321	-0.280	-0.149	2.025	3.040	2.531
SOCTOTT3	-0.167	0.070	0.070	1.812	2.553	3.128

	SHYTOTT1	SHYTOTT2	SHYTOTT3	SOCTOTT1	SOCTOTT2	SOCTOTT3
SHYTOTT1	**3.657**					
SHYTOTT2	1.088	**2.523**				
SHYTOTT3	1.221	1.310	**2.887**			
SOCTOTT1	1.730	0.636	0.710	**9.336**		
SOCTOTT2	0.737	0.816	0.674	4.121	**6.730**	
SOCTOTT3	0.719	0.736	0.937	3.668	4.176	**6.863**

Correlations

	EMOTOTT1	EMOTOTT2	EMOTOTT3	ACTTOTT1	ACTTOTT2	ACTTOTT3
EMOTOTT1	1.000					
EMOTOTT2	0.604	1.000				
EMOTOTT3	0.555	0.688	1.000			
ACTTOTT1	0.140	0.119	0.112	1.000		
ACTTOTT2	0.121	0.163	0.149	0.661	1.000	
ACTTOTT3	0.123	0.165	0.191	0.589	0.727	1.000
SHYTOTT1	0.168	0.130	0.137	0.274	0.186	0.208
SHYTOTT2	0.145	0.179	0.163	0.139	0.224	0.186
SHYTOTT3	0.148	0.175	0.221	0.128	0.175	0.245
SOCTOTT1	-0.015	0.002	0.003	0.381	0.268	0.277
SOCTOTT2	-0.030	-0.032	-0.017	0.241	0.393	0.329
SOCTOTT3	-0.016	0.008	0.008	0.213	0.327	0.403

	SHYTOTT1	SHYTOTT2	SHYTOTT3	SOCTOTT1	SOCTOTT2	SOCTOTT3
SHYTOTT1	1.000					
SHYTOTT2	0.358	1.000				
SHYTOTT3	0.376	0.485	1.000			
SOCTOTT1	0.296	0.131	0.137	1.000		
SOCTOTT2	0.149	0.198	0.153	0.520	1.000	
SOCTOTT3	0.144	0.177	0.210	0.458	0.614	1.000

There are strong correlations between measures of the same construct as one would expect when repeatedly measuring a scale at yearly intervals, however the correlations between differing scales are generally quite weak, even for scales measured at the same time. Also, it is quite noticeable that the variances (in bold) for the scales measured at time-1 are much higher.

1.3 Simple univariate linear regression

We want to regress the EAS emotionality sum-score from t1 (emotott1) on gender. Emotionality is continuous (ignore skewness for now) and gender is categorical.

In Mplus you much declare any dependent variables that are not continuous so the correct model – logit/probit/poisson – can be fitted. For independent variables such as gender in this example, they must be treated as continuous. This means that for any independent variable with more than two categories they must be converted into dummy indicators, otherwise a linear relationship will be assumed. As gender is a binary variable there is no impact on this model.

Steps

[1] Remove the "type = basic;" command as this will override any additional model commands you make. You can either delete this row or prefix it with an exclamation mark "!". This denotes that row as a comment which is to be ignored. The text should go green to indicate this. For those of you who are colour-blind the line will still be green as it is likely that Mplus is unaware of your condition.

[2] Introduce an additional "model" section with the command to regress emotott1 ON sex. Don't forget the semi-colon at the end of the regression command and a colon after "model". The latter should go blue.

[3] Update the usevariables command so it only contains these two variables. Mplus will quite happily include many more variables that you intend if you don't keep updating the usevariable comand.

Your syntax should look something like this:-

```
  Data:
    File is H:\Courses\SEM_2012\data\eas_1500.dta.dat ;

  Variable:
    Names are id sex
<snip>
       shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

       usevariables = emotott1 sex;


    Missing are all (9999);

  Analysis:
    !Type = basic ;

  Model:
    emotott1 on sex;
```

This syntax file is called ‘prac 1.3.inp’. We’ll use the notation <snip> here in these practicals (but not in the actual syntax) to save having to list all the variables on the file.

In the reams of output that Mplus produces you’ll find the model results:-

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 EMOTOTT1 ON
    SEX               0.790      0.209      3.777      0.000

 Intercepts
    EMOTOTT1          6.143      0.331     18.576      0.000

 Residual Variances
    EMOTOTT1         16.402      0.599     27.386      0.000
```

In other words, girls score on average 0.79 points higher on the emotionality sumscore.

1.4 Still-simple multivariate linear regression

Too many things are misclassified as multivariate nowadays. For a model to be properly multivariate (rather than just multivariable) it must have more than one dependent variable.

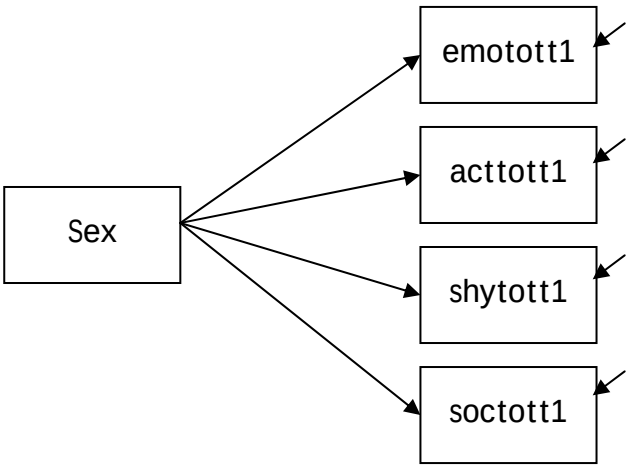
We can easily extend the above gender model to a multivariate one by adding more outcome variables.

Update the usevariable and model statements as follows:-

```
Data:
  File is H:\Courses\SEM_2012\data\eas_1500.dta.dat ;

Variable:
  Names are id sex
<snip>
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  usevariables = emotott1 acttott1 shytott1 soctott1 sex;

  Missing are all (9999);

Analysis:
  !Type = basic;

Model:
  emotott1 acttott1 shytott1 soctott1 on sex;
```

Or use ‘prac 1.4.inp’.

We are now assessing the effect of gender on four outcomes simultaneously. We could draw this model as follows:-



We should anticipate 4 estimated effects for sex, 4 residual variances and a set of additional parameters describing the covariance structure of the residuals. We will also obtain 4 intercepts (the alpha for each of the four regression equations y(i) = alpha(i) + beta(i)*x). The intercepts will have little meaning here as they correspond to the y when sex=0 and here sex is coded as male=1/female=2.

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
EMOTOTT1 ON SEX	0.793	0.209	3.793	0.000
ACTTOTT1 ON SEX	0.345	0.167	2.067	0.039
SHYTOTT1 ON SEX	-0.321	0.098	-3.265	0.001
SOCTOTT1 ON SEX	-0.337	0.158	-2.141	0.032
ACTTOTT1 WITH				
EMOTOTT1	1.776	0.341	5.202	0.000
SHYTOTT1 WITH				
EMOTOTT1	1.370	0.202	6.768	0.000
ACTTOTT1	1.727	0.165	10.447	0.000
SOCTOTT1 WITH				
EMOTOTT1	-0.125	0.319	-0.392	0.695
ACTTOTT1	3.805	0.273	13.930	0.000
SHYTOTT1	1.703	0.156	10.887	0.000
Intercepts				
EMOTOTT1	6.137	0.331	18.560	0.000
ACTTOTT1	3.112	0.264	11.780	0.000
SHYTOTT1	7.761	0.156	49.882	0.000
SOCTOTT1	7.341	0.249	29.471	0.000
Residual Variances				
EMOTOTT1	16.402	0.599	27.386	0.000
ACTTOTT1	10.469	0.382	27.386	0.000
SHYTOTT1	3.631	0.133	27.386	0.000
SOCTOTT1	9.308	0.340	27.386	0.000

We now see that whilst girls score higher than boys on emotionality, boys have higher scores on both shyness and sociability.

1.5 [Advanced] Fitting model 1.3 using the grouping approach

If you'd rather, skip on to exercise 1.7 where we delve into logistic regression models.

The model in 1.3 was simply a t-test. Three parameters were estimated – a difference in means, a residual variance and an intercept. We can estimate the same model by splitting the data into two using a "grouping" and derive our effect-estimate as a difference between the male and female mean scores. Syntax can be found in 'prac 1.5.inp'.

```
Data:
  File is H:\Courses\SEM_2012\data\eas_1500.dta.dat ;

Variable:
  Names are id sex
<snip>
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  usevariables = emotott1 sex;
  grouping = sex (1=male, 2=female);

  Missing are all (9999);

Model:

  model male:
    emotott1    (samevar);
    [emotott1]  (boymean);

  model female:
    emotott1    (samevar);
    [emotott1]  (girlmean);

  model constraint:
    new(diff);
    diff = girlmean - boymean;
```

We have used a grouping command in the variable section to define two groups corresponding to sex=1 (male) and sex=2 (female). Models will now be fit in both groups.

The model section now contains three sections. Note that the regression command (ON) has disappeared as we are now just estimating means and variances. The mean and variance for *emotott1* is estimated for boys and girls. The variances have been constrained to be equal by having the same phrase in brackets at the end of each line ("samevar"). Equal variance is a standard assumption for t-tests.

For the means we refer to those two parameters as ‘boymean’ and ‘girlmean’ using additional bracketing and then in the model constraint section we define a new parameter called “diff” as the difference between these two parameters. This new parameter is not itself part of the model it is estimated afterwards. We will obtain an SE for this parameter (derived using the delta method).

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

Group MALE

 Means
    EMOTOTT1         6.931      0.148     46.867      0.000

 Variances
    EMOTOTT1        16.401      0.599     27.386      0.000

Group FEMALE

 Means
    EMOTOTT1         7.724      0.148     52.231      0.000

 Variances
    EMOTOTT1        16.401      0.599     27.386      0.000

 New/Additional Parameters
    DIFF             0.793      0.209      3.793      0.000
```

You can see from the output that we have two estimated means, a single variance and an estimated differences again indicating the typically higher scores for emotionality for girls.

1.6 [Advanced] Fitting model 1.4 using the grouping approach

This grouping approach extends readily to the multivariate case. See ‘prac1.6.inp’. Notice it is not necessary to specify the residual covariance structure here, only the means and variances.

1.7 Simple logistic model

Here we will dichotomise the emotionality measure using Mplus' define section and fit a logistic regression model with sex as a predictor. Syntax is in 'prac1.7.inp'.

```
Data:
  File is H:\Courses\SEM_2012\data\eas_1500.dta.dat ;

Define:
  emo_bin = emotott1;
  cut emo_bin (10);

Variable:
  Names are id sex
     act_t1_1 act_t1_2 act_t1_3 act_t1_4 act_t1_5
<SNIP>
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  usevariables = sex emo_bin;
  categorical = emo_bin;

  Missing are all (9999);

Analysis:
  link = logit;
  estimator = ML;

Model:
  emo_bin on sex;

Output:
  cint;
```

Steps

[1] Using the define command, create a new variable called "emo_bin" and then dichotomise it – here a "case" is someone with a score of 11 or more.

[2] Add sex and emo_bin to the usevariables section. Variables defined in the define section must come AFTER variables on the datafile.

[3] Tell Mplus that emo_bin is categorical.

[4] In analysis section, request maximum likelihood (ML) estimation and a logit link. If you leave this section blank the results will be a probit model derived using least squares (WLSMV) estimation.

[5] Fit the regression model emo_bin ON sex;

[6] Request confidence intervals with the "cint" command within the output section.

Key output is shown overleaf.

```
UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES

    EMO_BIN
      Category 1    0.795     1192.000
      Category 2    0.205      308.000
```

Our outcome has a prevalence of 20.5%

```
TESTS OF MODEL FIT

Loglikelihood
          H0 Value                         -758.606

Information Criteria

          Number of Free Parameters               2
          Akaike (AIC)                     1521.211
          Bayesian (BIC)                   1531.838
          Sample-Size Adjusted BIC         1525.484
            (n* = (n + 2) / 24)
```

Model fit statistics – useful for model comparison.

```
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 EMO_BIN ON SEX        0.312      0.129      2.424      0.015

 Thresholds
    EMO_BIN$1          1.828      0.209      8.750      0.000

LOGISTIC REGRESSION ODDS RATIO RESULTS

EMO_BIN ON SEX        1.366
```

Log-odds and odds ratios for gender on high emotionality. Odds of high emotionality 36.6% greater for girls compared with boys.

```
CONFIDENCE INTERVALS OF MODEL RESULTS

              Lower .5%  Lower 2.5%  Lower 5%  Estimate  Upper 5% Upper 2.5%  Upper .5%

 EMO_BIN  ON
    SEX          -0.020       0.060     0.100      0.312      0.523      0.564     0.643

 Thresholds
    EMO_BIN$1     1.290       1.419     1.484      1.828      2.172      2.238     2.366

CONFIDENCE INTERVALS FOR THE LOGISTIC REGRESSION ODDS RATIO RESULTS

 EMO_BIN  ON
    SEX           0.981       1.061     1.105      1.366      1.688      1.758     1.902
```

95% confidence interval for odds ratio = 1.37 [1.06, 1.76].

SEM Course practicals
Prac 2 – Confirmatory & Exploratory factor analyses

2 Introduction

In this practical we continue working with the ALSPAC data from the 20-item questionnaire with 4 subscales from Time 3. The questions have 5 response options, so the dependent variables in this practical are categorical.

2.1 CFA with continuous variables

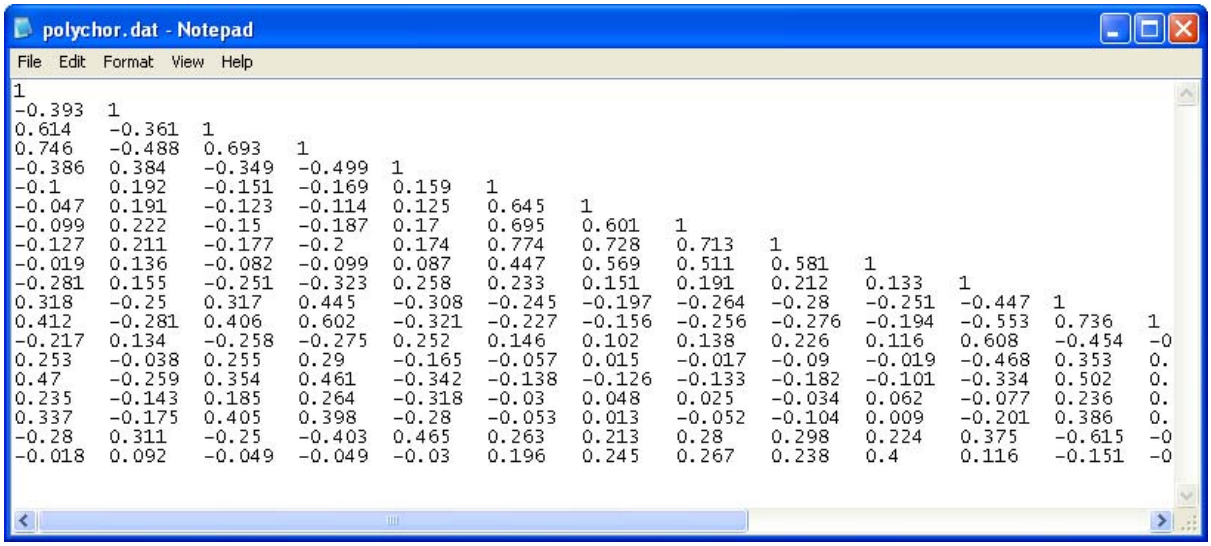
Here we will fit a 4-factor model to the 20 item responses, according to the a priori assignment of items to subscales. The four subscales are allowed to correlate freely. For the purpose of this exercise, we will work with a summary file of polychoric correlations between the 20 item responses. This way we can pretend that the correlations come from continuous variables, and practice working with summary data files, as opposed to full data files.

First put your outcomes on the USEVARIABLE list in the VARIABLE command and add the same set of variables to a CATEGORICAL command. Ask for TYPE = BASIC; in the ANALYSIS command without the MODEL command.

```
Data:
  File is eas_1500.dta.dat ;

Variable:
  Names are id
     sex
<snip>
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  Usevariables = act_t3_1 act_t3_2 act_t3_3 act_t3_4 act_t3_5
     emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5
     shy_t3_1 shy_t3_2 shy_t3_3 shy_t3_4 shy_t3_5
     soc_t3_1 soc_t3_2 soc_t3_3 soc_t3_4 soc_t3_5;

  Categorical = ALL;

Analysis:
  Type = basic;
```

This short program (prac 2.1.inp) will give polychoric correlations as part of the summary statistics. To save time, we have saved these correlations in a separate file ‘polychor.dat’.

We will now use the 'polychor.dat' file as your summary data.



Remember that we are going to be treating our variables as continuous for now!

We will need a new syntax file to read this new datafile. Because this file contains summary data we will need to tell Mplus what sort of data it is (correlations) and also the size of the sample that was used to create these estimates (n = 1500).

The bare bones of this syntax file can be found in (prac 2.1b.inp)

```
 Data:
  File is polychor.dat;
  type is correlation;
  nobservations is 1500;

Variable:
  Names are
     act_t3_1 act_t3_2 act_t3_3 act_t3_4 act_t3_5
     emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5
     shy_t3_1 shy_t3_2 shy_t3_3 shy_t3_4 shy_t3_5
     soc_t3_1 soc_t3_2 soc_t3_3 soc_t3_4 soc_t3_5;
```

Add a model section to this file to estimate 4 freely correlated factors. Use the rules we learnt to set the scales of latent variables. Try to use Mplus defaults, i.e. setting the first loading for each factor to 1; and then override these defaults and set the scale by setting the factor variances to 1. The completed syntax can be found as (prac 2.1c.inp).

Examine: model fit, model parameters and SEs, residuals, and finally modification indices.

Oh dear! It is likely that your program will lead to the following error:-

```
NO CONVERGENCE.  NUMBER OF ITERATIONS EXCEEDED.
```

You can ask for more iterations by adding an extra command to the analysis section (e.g. "iterations 10000;") but this will not help. If you scroll down through your output you'll find a section titled

```
MODEL COMMAND WITH FINAL ESTIMATES USED AS STARTING VALUES
```

along with the various parameters to be estimated by this model. We could copy this whole section as new model syntax and re-run our model. You'll see two kinds of additional symbol here – the at symbol "@" showing that some parameters are fixed to a specific value prior to estimation and an asterisk "*" indicating parameters that are freely estimated but are given specific starting values. We can tweak these starting values to see if the model estimation fairs any better – perhaps the original estimation got stuck somewhere and was unable to converge to a solution.

If you study the various starting values shown, an anomaly should become apparent – there are many unusual values for parameters involving "shy".

```
f_act BY act_t3_1@1;
f_act BY act_t3_2*-0.658;
f_act BY act_t3_3*0.932;
f_act BY act_t3_4*1.208;
f_act BY act_t3_5*-0.675;
f_emo BY emo_t3_1@1;
f_emo BY emo_t3_2*0.943;
f_emo BY emo_t3_3*0.945;
f_emo BY emo_t3_4*1.104;
f_emo BY emo_t3_5*0.752;
f_shy BY shy_t3_1@1;
f_shy BY shy_t3_2*7319510.500;
f_shy BY shy_t3_3*8897723;
f_shy BY shy_t3_4*-5586941;
f_shy BY shy_t3_5*4880746.500;
f_soc BY soc_t3_1@1;
f_soc BY soc_t3_2*0.667;
f_soc BY soc_t3_3*0.804;
f_soc BY soc_t3_4*-0.881;
f_soc BY soc_t3_5*0.136;

f_emo WITH f_act*-0.144;
f_shy WITH f_act*0;
f_shy WITH f_emo*0;
f_soc WITH f_act*0.390;
f_soc WITH f_emo*-0.155;
f_soc WITH f_shy*0;

act_t3_1*0.383;
act_t3_2*0.733;
act_t3_3*0.464;
act_t3_4*0.099;
act_t3_5*0.719;
emo_t3_1*0.303;
emo_t3_2*0.381;
emo_t3_3*0.377;
emo_t3_4*0.151;
emo_t3_5*0.606;
shy_t3_1*0.999;
shy_t3_2*0.395;
shy_t3_3*0.106;
shy_t3_4*0.647;
shy_t3_5*0.731;
soc_t3_1*0.355;
soc_t3_2*0.713;
soc_t3_3*0.582;
soc_t3_4*0.499;
soc_t3_5*0.987;
f_act*0.616;
f_emo*0.696;
f_shy*0;
f_soc*0.645;
```

The model seems to have gotten stuck at a place where the loadings for the shyness factor are all extremely large, the variances of f_shy is zero and the covariances between f_shy and the other factors are also zero.

In this instance it turns out that all we need to do is add some starting values for the estimation of f_shy. We do this as follows:-

```
f_shy by shy_t3_1 shy_t3_2*-1 shy_t3_3*-1 shy_t3_4*1 shy_t3_5*-1;
```

The model should now run properly to convergence (prac 2.1d.inp).

(i) Model fit

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                             2861.696
          Degrees of Freedom                     164
          P-Value                             0.0000

CFI/TLI

          CFI                                 0.835
          TLI                                 0.809

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                            0.105
          90 Percent C.I.                     0.101  0.108
          Probability RMSEA <= .05            0.000

SRMR (Standardized Root Mean Square Residual)

          Value                               0.088
```

The chi-square test and the other traditional fit-statistics suggest this model does not fit that well. Perhaps a structure where each item is only allowed to load on one factor is a little restrictive for these data.

(ii) Parameters

The magnitude of the loadings ranges considerably. Note the very low value for the 5th item on the sociability factor: 0.134 (SE=0.035).

(iii) Residuals

These are obtained with the command "residual" within the output section. These indicate any differences between the observed data (the polychoric correlation matrix) and that implied by the model. In this instance there are a lot of residuals to study! Notice there are a number of extremely large standardized (z-score) residuals.

(iv) Modindices

These are obtained with the command "modindices(3.84)" within the output section. These values indicate additional paths which would improve the chi-square stat by approx 3.84 (the threshold for chi-square with 1 d.f.). Again there are a number of large values here – e.g. allowing the 3rd sociability item to load on the emotionality factor would have a dramatic improvement in model fit.

2.2 CFA with categorical variables

Here we will fit the same 4-factor model to the 20 item responses, but now working with full categorical data rather than the file of summary stats.

Declare the 20 item responses as categorical. In the ANALYSIS section, use ESTIMATOR=WLSMV and PARAMETERIZATION=THETA.

Program a model with 4 freely correlated factors. Use standardized factors: override the Mplus defaults and set the factors' scale by setting their variances to 1. Syntax can be found in (prac 2.2.inp).

Examine: model fit, model parameters and SEs, residuals, and finally modification indices.

```
Data:
  File is eas_1500.dta.dat ;

Variable:
  Names are id
     sex
<snip>
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  Usevariables = act_t3_1 act_t3_2 act_t3_3 act_t3_4 act_t3_5
     emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5
     shy_t3_1 shy_t3_2 shy_t3_3 shy_t3_4 shy_t3_5
     soc_t3_1 soc_t3_2 soc_t3_3 soc_t3_4 soc_t3_5;

  Categorical = ALL;

Analysis:
    estimator = WLSMV;
    parameterization = theta;

Model:
     f_act by act_t3_1* act_t3_2 act_t3_3 act_t3_4 act_t3_5;
     f_emo by emo_t3_1* emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5;
     f_shy by shy_t3_1* shy_t3_2 shy_t3_3 shy_t3_4 shy_t3_5;
     f_soc by soc_t3_1* soc_t3_2 soc_t3_3 soc_t3_4 soc_t3_5;

     f_act@1 f_emo@1 f_shy@1 f_soc@1;

Output:
     residual modindices(3.8);
```

You should notice that the results are similar to the model using the polychoric correlation matrix, but they are not identical. Why might this be?

Examine modification indices. Can you see some troubling problems with this questionnaire? What modifications would you consider based on the largest modification indices?

2.3 EFA with continuous variables

Here we will explore factor solutions for the 20 item responses.

Again, we will work with a summary file of polychoric correlations between the 20 item responses, using the file 'polychor.dat' as summary data. Remember that we are treating our variables as continuous!

In the ANALYSIS section, ask for EFA with solutions ranging from 1 to 4 factors. Explore orthogonal and oblique rotations for your solutions. First, set

```
ROTATION= GEOMIN (OR);
```

- this is only one of several options for orthogonal rotation, you can also use VARIMAX or whatever as an alternative orthogonal rotation. Ask for TYPE=PLOT3; in the PLOT section (this will print a scree plot).

Examine: eigenvalues, model fit for each factor solution, rotated factor loadings, residuals, and the scree plot. Which is your preferred solution?

Now, change to oblique rotation

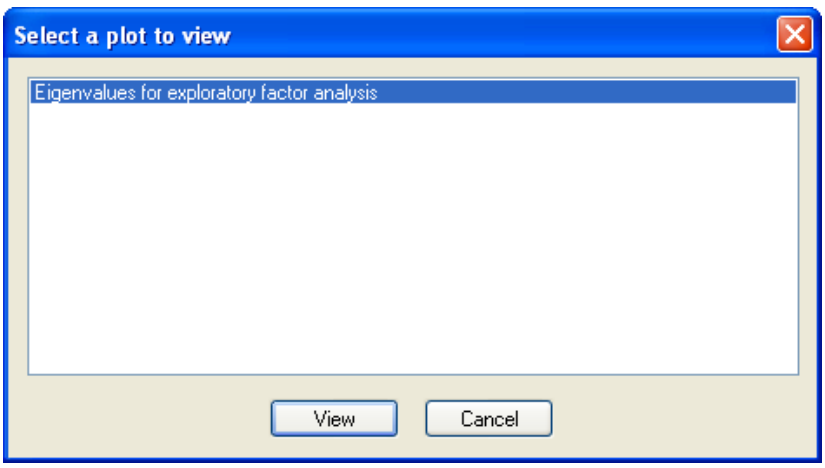
```
ROTATION= GEOMIN (OB);
```

- this is only one of many options for oblique rotation, you can also use PROMAX or whatever. Ask for TYPE=PLOT3; in the PLOT section (this will print a scree plot)

Examine: eigenvalues, model fit for each factor solution, rotated factor loadings, and residuals. Compare factor loadings and residuals of your preferred solution to the ones from the oblique rotation. What are the differences? Interpret the results.

How to view graphs:-

Have the output (.out) window uppermost within Mplus and select graph>view graphs. You may need to locate the graph file or it may already be connected. If the former, the graph file should have the same name as the inp/out files. You often get many graphs to choose from but here the choice is rather limited:-



2.4 (Optional) EFA with categorical variables

You can repeat the above exercise, using the raw rather than the summary data, declaring your variables as categorical and using Mplus default estimator for the EFA analysis.

2.5 Multi-group CFA with categorical variables

Here we will fit the a-priori 4-factor model to the 20 item responses, separately for boys and girls.

In this exercise, we will have to recode some item responses before performing the analysis. This is because some response categories were used so infrequently that they appear in one gender group only, causing Mplus to generate error messages about category coding. To avoid that, rarely endorsed categories should be collapsed prior to the analysis. To do that, use the DEFINE command:

```
Define:
IF (act_t3_1 EQ 0) THEN act_t3_1=1;
IF (act_t3_2 EQ 4) THEN act_t3_2=3;
IF (act_t3_3 EQ 0) THEN act_t3_3=1;
IF (act_t3_4 EQ 0) THEN act_t3_4=1;
IF (act_t3_5 EQ 4) THEN act_t3_5=3;

IF (emo_t3_3 EQ 4) THEN emo_t3_3=3;

IF (shy_t3_1 EQ 4) THEN shy_t3_1=3;
IF (shy_t3_2 EQ 0) THEN shy_t3_2=1;
IF (shy_t3_3 EQ 0) THEN shy_t3_3=1;

IF (soc_t3_1 EQ 0 OR soc_t3_1 EQ 1) THEN soc_t3_1=2;
IF (soc_t3_2 EQ 0) THEN soc_t3_2=1;
IF (soc_t3_3 EQ 0) THEN soc_t3_3=1;
IF (soc_t3_4 EQ 4) THEN soc_t3_4=3;
```

This has been already done in the ‘prac 2.5.inp’ file.

Next, reuse the variables descriptions from previous exercises and declare 20 item responses as categorical.

The only new statement appearing in the VARIABLE section is

```
Variable:
Grouping = sex (1=boys, 2=girls);
```

This defines the two groups.

In the ANALYSIS section, use ESTIMATOR=WLSMV and PARAMETERIZATION=THETA.

Program a model with 4 freely correlated factors. Use the Mplus defaults for setting the factor scales. Request standardised output. Syntax can be found in (prac 2.5.inp).

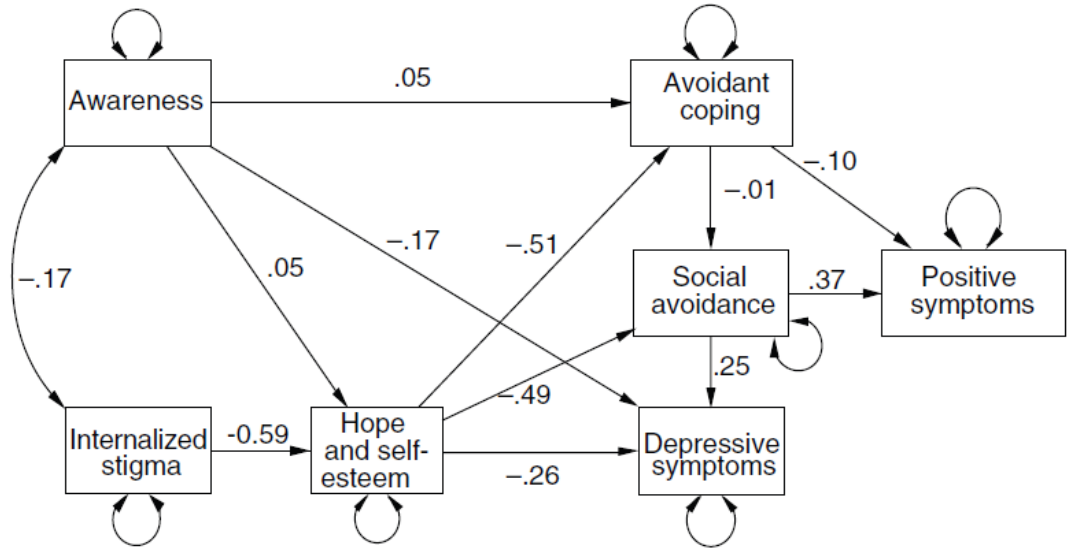
```
Model:
     f_act by act_t3_1 act_t3_2 act_t3_3 act_t3_4 act_t3_5;
     f_emo by emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5;
     f_shy by shy_t3_1 shy_t3_2 shy_t3_3 shy_t3_4 shy_t3_5;
     f_soc by soc_t3_1 soc_t3_2 soc_t3_3 soc_t3_4 soc_t3_5;

Output: stand modindices;
```

Examine: model fit, and model parameters and SEs for each group. Examine carefully the parameters to see which parameters Mplus constrains equal across groups, and which it allows to vary. Examine means and variances for the 4 subscales for boys and girls. What can be said about the two groups? Any significant differences in means or variances?

Practical 3 – Model from Schizophrenia paper

Figure 2

Path model 1, where positive symptoms of schizophrenia are treated as an outcome[a]



[a] N=102. Standardized coefficients are presented.

Key for variables:

`Aware` = awareness
`Stigma` = internalized stigma
`Hope` = hope and self esteem
`Avoidcop` = avoidant coping
`Positive` = positive symptoms
`Socavoid` = social avoidance
`Depress` = depressive symptoms

Write out syntax for:-

(i) associations
(ii) variances of residuals
(iii) variances and covariances of exogenous variables

How many estimates of each type are you expecting?

Be aware that residuals are not correctly indicated in published path diagram

Model:

```
  DATA:
    FILE = "sz input matrix2.txt";
    TYPE = STD CORRELATION;
    NOBSERVATIONS = 102;

  VARIABLE:
    NAMES = aware stigma hope avoidcop socavoid depress positive;
    USEVARIABLES = aware stigma hope avoidcop socavoid depress
positive;

  MODEL:
      positive on avoidcop socavoid;
      avoidcop on aware hope;
      hope on aware stigma;
      depress on hope aware socavoid;
      socavoid on avoidcop hope;

      ! residual variances for endogenous variables
      avoidcop positive socavoid depress hope;

      ! exogenous covariance matrix - unnecessary
      aware stigma;
      aware with stigma;
      positive with depress@0;

  OUTPUT:
      stdyx residual modindices(1.0) sampstat;
```

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit
          Value                             14.508
          Degrees of Freedom                    9
          P-Value                           0.1054

CFI/TLI
          CFI                                0.961
          TLI                                0.914

Loglikelihood
          H0 Value                       -1256.993
          H1 Value                       -1249.739

Information Criteria
          Number of Free Parameters             19
          Akaike (AIC)                    2551.987
          Bayesian (BIC)                  2601.861
          Sample-Size Adjusted BIC        2541.847

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.077
          90 Percent C.I.                    0.000  0.148
          Probability RMSEA <= .05           0.238

SRMR (Standardized Root Mean Square Residual)
          Value                              0.061
```

Note that the results are not perfectly replicated as these are based on the imprecise estimated of the sample stats displayed in the paper. Attempt to match these up with those shown in the figure.

Also note that Mplus may bung in additional parameters that you perhaps weren't expecting. A residual covariance was included in the model between DEPRESS and POSITIVE, it was necessary to constrain this to zero in order to replicate the model shown in the paper. Hence it's a good idea to be on the ball when it comes to each and every parameter you are expecting.

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
POSITIVE ON				
AVOIDCOP	-0.180	0.092	-1.948	0.051
SOCAVOID	0.391	0.087	4.508	0.000
AVOIDCOP ON				
AWARE	0.051	0.087	0.592	0.554
HOPE	-0.508	0.075	-6.771	0.000
HOPE ON				
AWARE	0.056	0.081	0.686	0.493
STIGMA	-0.580	0.067	-8.700	0.000
DEPRESS ON				
HOPE	-0.264	0.097	-2.713	0.007
AWARE	-0.169	0.086	-1.961	0.050
SOCAVOID	0.245	0.097	2.527	0.012
SOCAVOID ON				
AVOIDCOP	-0.020	0.100	-0.201	0.841
HOPE	-0.500	0.090	-5.566	0.000
AWARE WITH				
STIGMA	-0.180	0.096	-1.879	0.060
POSITIVE WITH				
DEPRESS	0.000	0.000	999.000	999.000
Variances				
AWARE	1.000	0.000	999.000	999.000
STIGMA	1.000	0.000	999.000	999.000
Residual Variances				
HOPE	0.649	0.076	8.522	0.000
AVOIDCOP	0.747	0.074	10.048	0.000
SOCAVOID	0.760	0.074	10.299	0.000
DEPRESS	0.757	0.074	10.276	0.000
POSITIVE	0.847	0.066	12.902	0.000

Prac 4 – Fitting a Path Analytical Model

4 Introduction

As you hopefully are aware by now, the difference between Path Analysis and SEM is the presence of latent variables. An SEM model combines the estimation of one or more latent variables (measurement models) with a structural model which describes how these latent variables are hypothesized to be related both to each other and to other non-latent (manifest) variables.

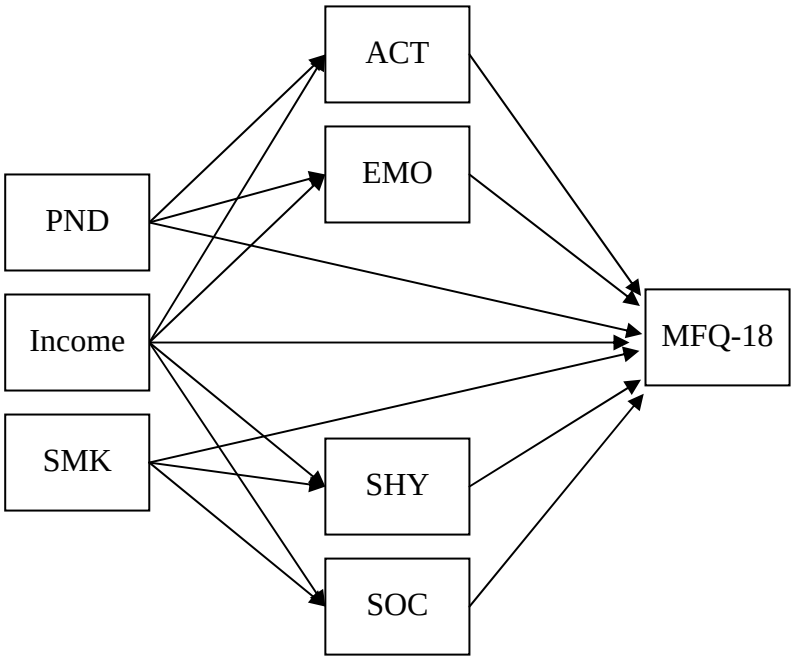
A Path Analysis model on the other hand contains just a structural model – we are describing the relationship between a number of manifest variables.

4.1 The hypothesized model

A vaguely reasonable model is shown below. Clearly there are things missing – there are no residuals shown – but this gives an idea of how we think these measures may be related to each other.

We have three measures from early in the child's life - postnatal depression (a binary measure – yes/no) maternal smoking (a binary measure indicating mums who smoked 20+/day) and an indicator of low family income (binary – bottom quintile versus the rest).

These baseline measures are expected to impact on adolescent depression at age 18 (MFQ). Part of the effect of postnatal depression on MFQ is expected to be mediated through activity and emotionality aspects of temperament, whilst maternal smoking is mediated through shyness and sociability. Income is mediated through all EAS subscales. OK, this doesn't have a solid ground in theory!



4.2 Examine the covariance matrix

Let's not jump in and fit a path model. We know that these models are fit to covariance matrices so we should first examine this information.

If all covariances are negligible then it's clearly not worth carrying on. In addition, this will remind you that it's the variances and covariances that make up the "data" for these models and that adding more variables rather than more cases is the way to provide more degrees of freedom for more complex models.

As you know, we can obtain sample statistics using "type = basic" so this is what we'll do here. Note you can also obtain this information at the same time as fitting an actual model by requesting "sampstat" within the output section.

Here's the syntax you'll need.

```
Data:
  File is "C:\Work\SEM Course\eas_1500.dta.dat" ;

Define:
  smk_hi = (mumsmk EQ 2);   ! mother smokes 20+ per day
  low_inc = (income EQ 0);   ! bottom quintile of income

  mfqsum18 = mfq18_01 + mfq18_02 + mfq18_03 + mfq18_04 + mfq18_05
           + mfq18_06 + mfq18_07 + mfq18_08 + mfq18_09 + mfq18_10
           + mfq18_11 + mfq18_12 + mfq18_13;
Variable:
  Names are id
     sex
<snip>
     emotott1 emotott2 emotott3 acttott1 acttott2 acttott3
     shytott1 shytott2 shytott3 soctott1 soctott2 soctott3;

  Missing are all (9999);

  Usevariables = mdep_pn emotott3 acttott3 shytott3 soctott3
     mfqsum18 smk_hi low_inc;

Analysis:
  Type = basic;
```

Alternatively run "prac 4.2.inp".

Note that I've used the "define" section to create an mfq sumscore from the 13 items. I've also created a measure called "smk_hi" because originally the smoking measure was a 3-level ordinal, and a measure "low_inc" for the same reason. Tare should be taken using DEFINE if your data has missing cases, here we are OK.

The variables we are interested in are added to the usevariable list. I'm using the EAS measures from time point 3. Don't forget that defined variables must be declared at the end of this list, after those that appear on the file.

The output shows means, covariances and correlations. Our model wont be using the means but they might be brought into play were we to plan to fit the same model in parallel for boys and girls.

The variables have a wide range of variances and this is not always a good thing as it can lead to estimation problems. It's often a good idea to rescale measures if possible – e.g. by using cm instead of mm for a head-circumference measure if the variances is much higher than the other variables.

RESULTS FOR BASIC ANALYSIS

Means

MDEP_PN	EMOTOTT3	ACTTOTT3	SHYTOTT3	SOCTOTT3	MFQSUM18	SMK_HI	LOW_INC
0.189	7.693	4.899	7.323	6.833	6.303	0.036	0.101

Covariances

	MDEP_PN	EMOTOTT3	ACTTOTT3	SHYTOTT3	SOCTOTT3	MFQSUM18	SMK_HI	LOW_INC
MDEP_PN	0.153							
EMOTOTT3	0.268	11.518						
ACTTOTT3	0.051	1.927	8.799					
SHYTOTT3	0.054	1.276	1.237	2.887				
SOCTOTT3	-0.009	0.070	3.128	0.937	6.863			
MFQSUM18	0.207	2.781	0.821	0.638	0.147	**26.059**		
SMK_HI	0.005	0.022	-0.020	0.000	0.014	0.066	0.035	
LOW_INC	0.008	0.014	-0.022	-0.007	0.006	0.132	0.006	0.091

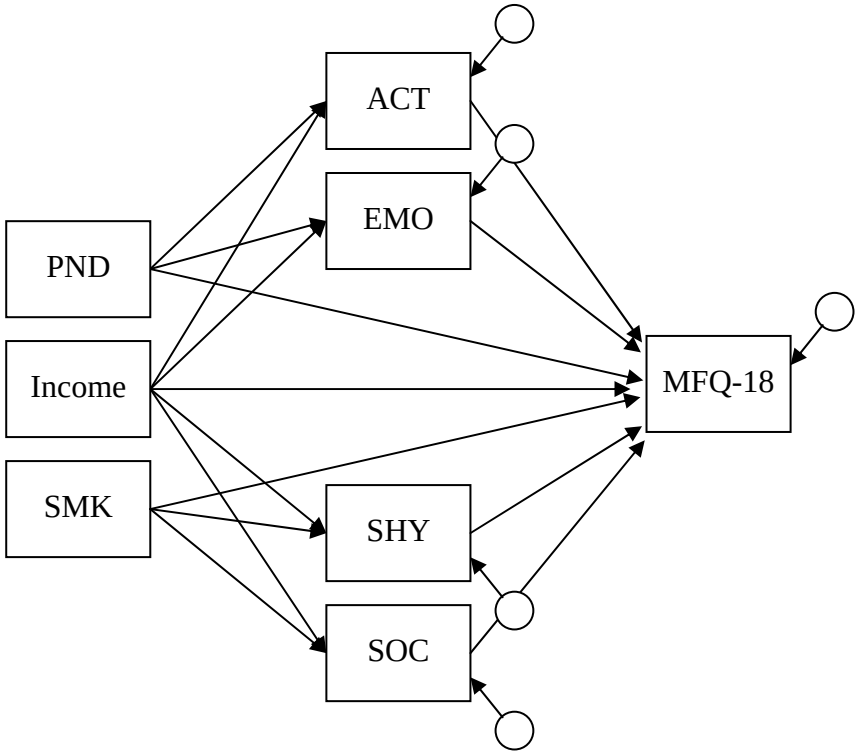
Correlations

	MDEP_PN	EMOTOTT3	ACTTOTT3	SHYTOTT3	SOCTOTT3	MFQSUM18	SMK_HI	LOW_INC
MDEP_PN	1.000							
EMOTOTT3	0.202	1.000						
ACTTOTT3	0.044	0.191	1.000					
SHYTOTT3	0.081	0.221	0.245	1.000				
SOCTOTT3	-0.009	0.008	0.403	0.210	1.000			
MFQSUM18	0.104	0.161	0.054	0.074	0.011	1.000		
SMK_HI	0.062	0.035	-0.036	0.001	0.029	0.070	1.000	
LOW_INC	0.064	0.013	-0.025	-0.014	0.008	0.086	0.101	1.000

The magnitude of covariances/correlations is not always that high. This can also lead to a problem when it comes to estimation. If some covariances are effectively zero then we have less information than we thought. Whilst a model on paper may appear to be identified, it can be turn out to be *empirically unidentified*. This is something you can't assess until you get to look at the data.

4.3 Identifying the model

The covariance matrix contains (8*9)/2 = 36 distinct values, however it might be wise to restrict the number of parameters in our model for reasons discussed previously. The path diagram from earlier has been updated below to reflect actual parameters we intend to estimate. Here we are concentrating on the covariance structure. The output will contain parameters from the mean structure (e.g. the intercept for each dependent variable) but these will not affect the fit of the covariance structure model nor will they steal degrees of freedom from that model. As indicated earlier, we could happily fit this same model to centred data.



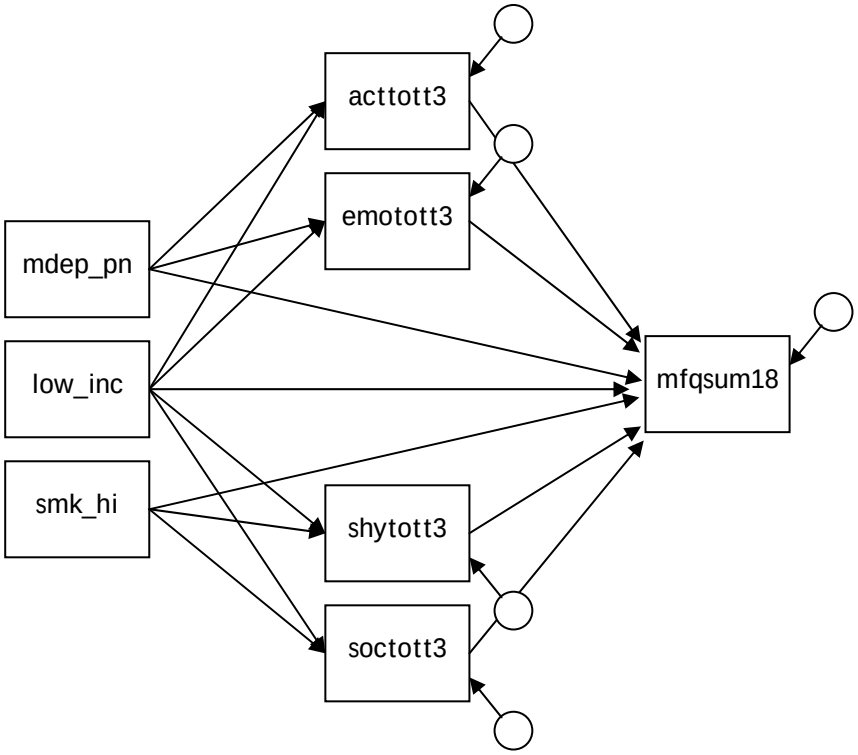
It is (hopefully) clear from the above that we will estimate FIVE residual variances for the five dependent variables and FIFTEEN measures of association connecting the different variables. Note there is an assumed relationship between the baseline (exogenous) measures in the same way that there is with a standard regression analysis. We could derive vars/covars for these three measures but this would just give their sample values and the model itself would not be affected.

This gives 20 parameters which is comfortably lower than the 36 pieces of information in the covariance matrix. There is likely to be scope for adding additional paths through model refinement (perhaps the path from PND to SHY is non-zero), alternatively we may wish to correlate the four EAS residuals to allow for the likely event that we are not fully explaining their relationship with each other using the three baseline measures we currently have in play. (Some of the strongest relationships in our covariance matrix were cross-sectionally within the batch of EAS measures and we are currently claiming that these are all totally explained by three meager binary variables – PND, income and SMK).

4.4 Turning model into syntax

Firstly what kind of variables are we dealing with? All dependent variables are continuous (although skewed) and all independent variables are binary. Recall however that Mplus treats binary independent variables as continuous so as far as Mplus is concerned, all of these measures are continuous. This fact should lead you to expect Maximum Likelihood estimation.

Some of these 20 parameters will be estimated without being specified but it's a good idea to work out how many parameters you expect to highlight if you've specified your model incorrectly. We will, however, need to include a command for each of the fifteen associations in our model. Spell these out as fifteen separate commands and then reduce to a shorter, neater set of commands using shorthand. The model is repeated below using the proper names to help you.



Long-hand commands	Short-hand commands

Answer (there are many options for the short-hand model):-

Long-hand commands	Short-hand commands
mfqsum18 on acttott3; mfqsum18 on emotott3; mfqsum18 on shytott3; mfqsum18 on soctott3; mfqsum18 on mdep_pn; mfqsum18 on low_inc; mfqsum18 on smk_hi; acttott3 on mdep_pn; acttott3 on low_inc; emotott3 on mdep_pn; emotott3 on low_inc; shytott3 on low_inc; shytott3 on smk_hi; soctott3 on low_inc; soctott3 on smk_hi;	mfqsum18 on acttott3 emotott3 shytott3 soctott3; mfqsum18 on mdep_pn low_inc smk_hi; acttott3 emotott3 on mdep_pn low_inc; shytott3 soctott3 on low_inc smk_hi;

4.5 Fitting the model

Amend your earlier syntax file from section 4.2 by removing the "type = basic;" command and adding your model commands. Alternatively, open up "prac 4.5.inp". This is out model statement, complete with some helpful comments:-

```
Model:
   ! effect of EAS temperament on depressive symptoms
   mfqsum18 on acttott3 emotott3 shytott3 soctott3;

   ! effect of baseline factors on depressive symptoms
   mfqsum18 on mdep_pn low_inc smk_hi;

   ! effect of baseline factors on activity and emotionality
   acttott3 emotott3 on mdep_pn low_inc;

   ! effect of baseline factors on shyness and sociability
   shytott3 soctott3 on low_inc smk_hi;
```

Now run this model and check you have estimated 20 parameters for the structural model.

Before we come on to thinking about model fit, let's look at the model parameters (I've removed the intercepts and residual variances from the output):-

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
MFQSUM18 ON				
ACTTOTT3	0.043	0.049	0.868	0.385
EMOTOTT3	0.198	0.040	4.915	0.000
SHYTOTT3	0.108	0.081	1.346	0.178
SOCTOTT3	-0.018	0.055	-0.328	0.743
MDEP_PN	0.847	0.338	2.506	0.012
LOW_INC	1.276	0.431	2.962	0.003
SMK_HI	1.496	0.699	2.140	0.032
ACTTOTT3 ON				
MDEP_PN	0.347	0.196	1.770	0.077
LOW_INC	-0.275	0.254	-1.084	0.278
EMOTOTT3 ON				
MDEP_PN	1.750	0.220	7.960	0.000
LOW_INC	0.006	0.285	0.021	0.984
SHYTOTT3 ON				
LOW_INC	-0.082	0.146	-0.564	0.573
SMK_HI	0.025	0.237	0.104	0.917
SOCTOTT3 ON				
LOW_INC	0.044	0.225	0.197	0.844
SMK_HI	0.397	0.365	1.088	0.277

Things aren't looking great for this model! Emotional temperament is related to maternal postnatal depression and also adolescent depressive symptoms. Maternal postnatal depression is mildly related to adolescent symptoms. Income and maternal smoking also have a strong impact on adolescent depressive symptoms but other pathways are weak. Let's keep going all the same.

4.6 The fit of the model

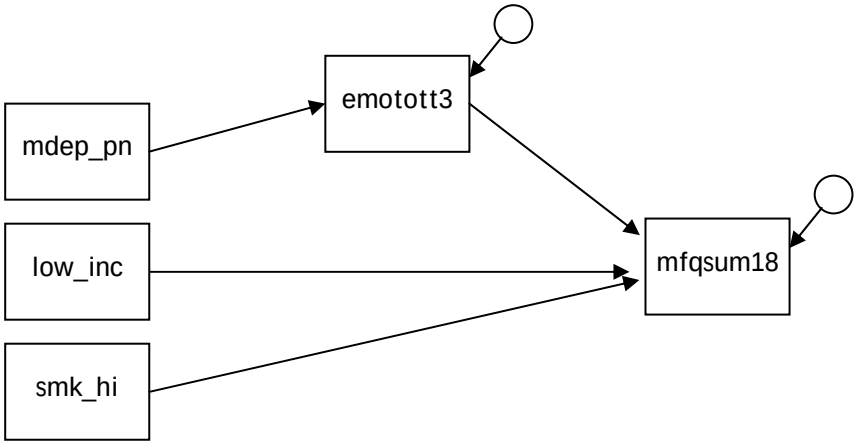
There is fierce debate in the SEM world about the importance of model fit. Some would say that model fit is essential whilst others that model fit statistics are merely alternative estimates of your sample size.

Either way, some consideration of model fit is going to be necessary, as publishing results without it will be an uphill struggle. The stats for this ill-fated model are clearly worse than poor. Whilst the chi-square test usually indicates poor fit when sample sizes are larger, the other measures – CFI/TLI/RMSEA are usually much more amenable. This model in its current guise is clearly beyond salvation.

4.7 A drastic remodelling to illustrate some key points

Setting up a model without a strong theoretical backing is not a good idea. A number of the measures in this model are contributing very little. Shyness, sociability and activity were associated with each other but nothing else and this had dire consequences for the fit. Let's remove them from the model altogether and pretend that was what we wanted to do all along.

Here we have a simpler model suggesting that baseline measures and emotionality are related to adolescent depressive symptoms but that the effect of postnatal depression on adolescent depressive symptoms is wholly through emotionality and there is no direct effect.



Revise your usevariables list and model statements accordingly, or use "prac 4.7.inp".

```
  Usevariables = mdep_pn emotott3 mfqsum18 smk_hi low_inc;

Model:
   ! effect of EAS temperament on depressive symptoms
   mfqsum18 on emotott3;

   ! effect of baseline factors on depressive symptoms
   mfqsum18 on low_inc smk_hi;

   ! effect of postnatal depression on emotionality
   emotott3 on mdep_pn;
```

Our fit measures are much improved – that's a relief!!!

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit
          Value                             7.440
          Degrees of Freedom                    3
          P-Value                          0.0591

CFI/TLI
          CFI                               0.962
          TLI                               0.912

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                          0.031
          90 Percent C.I.                   0.000  0.061
          Probability RMSEA <= .05          0.832

SRMR (Standardized Root Mean Square Residual)

          Value                             0.015
```

Notice however that some of the parameter estimates have barely changed:-

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 MFQSUM18 ON
    EMOTOTT3         0.237      0.038      6.210      0.000
    LOW_INC          1.317      0.431      3.055      0.002
    SMK_HI           1.545      0.699      2.211      0.027

 EMOTOTT3 ON
    MDEP_PN          1.750      0.219      7.979      0.000
```

Where to go from here?

[1] Whilst this model does appear to fit by simply eyeballing the fit stats we could examine the modification indices to see if further improvement could be made.

[2] It would be good if we could formally test whether some of out paths are actually zero – in particular the direct path from postnatal depression to adolescent symptoms.

4.8 Modification Indices (MI)

Modification indices indicate the approximate improvement in fit were an additional path to be included which is currently constrained to be zero.

Add the following command to the previous model

```
Output:
    modindices(3.84);
```

or use "prac 4.8.inp"

This will list any new pathways that would decrease the chi-square model fit statistic by 3.84 or more, i.e. a change which would be deemed significant at the 5% level.

The MI results will appear at the bottom of the output file. The actual model will be unchanged.

```
MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index     3.840

                          M.I.     E.P.C.  Std E.P.C.  StdYX E.P.C.

ON Statements

EMOTOTT3 ON MFQSUM18      4.262    -0.157     -0.157       -0.236
MFQSUM18 ON MDEP_PN       6.605     0.871      0.871        0.067

WITH Statements

MFQSUM18 WITH EMOTOTT3    6.674    -5.527     -5.527       -0.332
MDEP_PN  WITH MFQSUM18    6.604     0.132      0.132        0.067
SMK_HI   WITH MFQSUM18    6.558    -1.113     -1.113       -1.192
LOW_INC  WITH MFQSUM18    6.686    -1.764     -1.764       -1.166
```

We see there are a number of pathways which would result in a similar improvement in model fit. Let's add the additional path we already considered – from postnatal depression to adolescent symptoms.

Note that improving model fit on the basis of modification indices should only be done with strong theoretical justification. Simulation studies have shown that an stepwise approach to model revision purely based on statistics is unlikely to lead you to the correct model.

4.9 A revised model

Tweak your model to include a direct effect from postnatal depression to adolescent symptoms.

```
Model:
   ! effect of EAS temperament on depressive symptoms
   mfqsum18 on emotott3;

   ! effect of baseline factors on depressive symptoms
   mfqsum18 on low_inc smk_hi mdep_pn;

   ! effect of postnatal depression on emotionality
   emotott3 on mdep_pn;
```

Keep your modification indices command in the model for reasons that will soon become clear.

Things to notice in your new output

[1] The chi-square model fit has improved – from 7.440 to 0.822. This change of 6.618 is only approximately the same as the expected change of 6.605 reported by the modification indices output from the previous model.

[2] We can use this change of 6.618 to formally test the null hypothesis that this direct pathway is zero. P = 0.010 so there is moderate evidence for the inclusion of this path.

[3] In the output for this revised model, all the other modification indices have gone away. This is because there are often a number of different model revisions that can be made which would remove the same bit of model misfit. Many of the pathways suggested in the previous output would have yielded the same result – allowing postnatal depression to affect adolescent symptoms by another route not involving emotionality. If one is planning to use MI to make more than one model revision then new MI values should be generated at each step.

4.10 Quantifying direct and indirect effects

With our new model we have two pathways from postnatal depression to adolescent symptoms – one direct and one indirect. It would be useful if we could comment on which pathway is more dominant. To do this, add a further model command:-

```
Model indirect:
   mfqsum18 IND mdep_pn;
```

This will give some additional information, but will not effect the estimated model.

New output relating to direct and indirect effects:-

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Effects from MDEP_PN to MFQSUM18

  Total                1.250      0.335      3.736      0.000
  Total indirect       0.380      0.083      4.573      0.000

  Specific indirect

    MFQSUM18
    EMOTOTT3
    MDEP_PN            0.380      0.083      4.573      0.000

  Direct
    MFQSUM18
    MDEP_PN            0.871      0.338      2.576      0.010
```

One can see from this new output that there is a substantial, non-zero pathway from postnatal depression to adolescent symptoms through emotionality. Were we to have fitted a more complex model with more mediators (but a sensible model, not the one from earlier) then we could use this output to study the different indirect pathways.

Notice that the indirect effect here is the product of two terms from the standard output:

```
MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 MFQSUM18 ON
    EMOTOTT3           0.217      0.039      5.581      0.000
    LOW_INC            1.253      0.431      2.908      0.004
    SMK_HI             1.455      0.698      2.084      0.037
    MDEP_PN            0.871      0.338      2.576      0.010

 EMOTOTT3 ON
    MDEP_PN            1.750      0.219      7.977      0.000

 Intercepts
    EMOTOTT3           7.363      0.095     77.276      0.000
    MFQSUM18           4.291      0.323     13.304      0.000

 Residual Variances
    EMOTOTT3          11.049      0.403     27.386      0.000
    MFQSUM18          25.014      0.913     27.386      0.000
```

i.e. you multiply the coefficients for the various paths along the route from exposure to outcome. Of course this only works if the measures along the path are continuous (either measured or latent).

4.11 What we have learned

[1] There is a simple transition from a properly drawn path diagram to the Mplus syntax that would be needed to model it. Note that these are not proper DAG's but are still a useful way to picture the relationship between your variables.

[2] There are a number of model fit statistics we can use to get a quick idea regarding the adequacy of out models. We would encourage you to follow this up with a more thorough examination of key areas of misfit by studying the estimated covariance matrix and the resulting residuals. This can convey much more information than a single model fit statistic. We will be covering these issues in the lectures.

[3] We can use modification indices to make small changes (improvements?) to our models, but sometimes it is necessary to return to the drawing board.

[4] It's the old adage of garbage-in, garbage-out. If there doesn't appear to be a great deal of information in your sample covariance matrix, don't be surprised if your path model is less than fruitful.

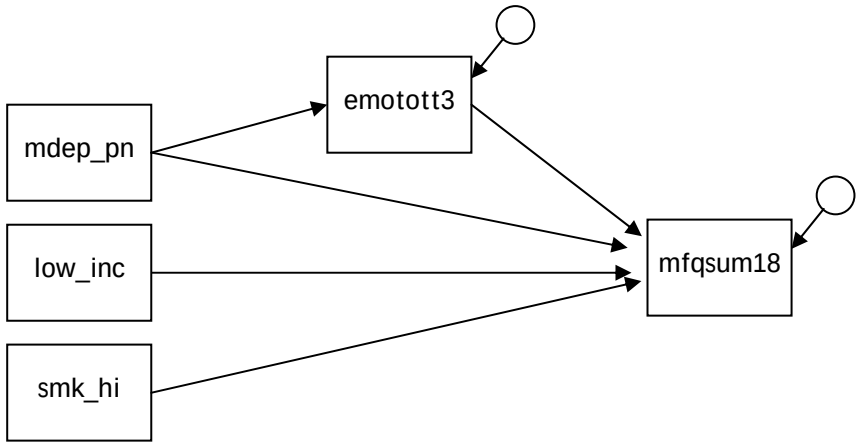
Prac 5 – Fitting a proper Structural Equation Model

5 Introduction

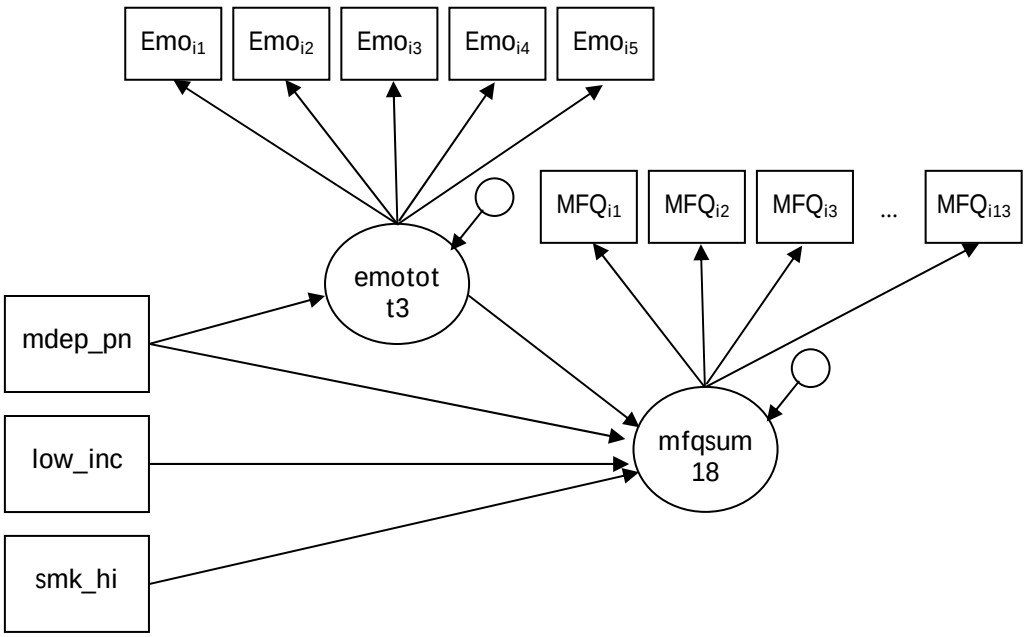
The aim of this session is for you to fit a model which combines a structural component (mainly using ON statements) with two measurement components (using BY statement). This will be an amended version of the smaller/more-successful EAS model from yesterday.

5.1 The hypothesized model

The path model from yesterday (complete with the direct path we tested):-



An SEM model along the same lines:-

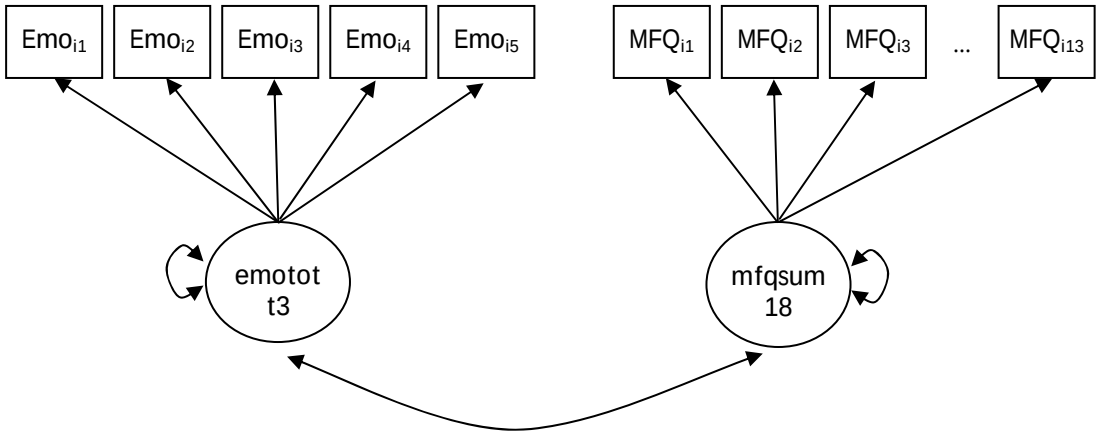


These two models are structurally the same however the second model contains two measurement models which are used to derive latent variables for emotionality and adolescent depressive symptoms.

5.2 A simpler CFA model

You should never jump straight into to a complicated model. It's much better to build up the model gradually and check that each component is working as you intended. For instance, we could fit the model shown on the page overleaf but if the model fit stats suggest it is inadequate we would have an awful job tracking down the source of the problem.

To save a little time, let's join the action halfway through the model building process and fit a model without any structural component – both latent variables along with a covariance between them.



Note that these latent variables are no longer *dependent* variables hence they now have estimated variances rather than residual variances. The important commands are shown below. Alternatively, open prac 5.2.

```
  Usevariables = emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5
    mfq18_01 mfq18_02 mfq18_03 mfq18_04 mfq18_05 mfq18_06
    mfq18_07 mfq18_08 mfq18_09 mfq18_10 mfq18_11 mfq18_12 mfq18_13;

  Categorical = emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5
    mfq18_01 mfq18_02 mfq18_03 mfq18_04 mfq18_05 mfq18_06
    mfq18_07 mfq18_08 mfq18_09 mfq18_10 mfq18_11 mfq18_12 mfq18_13;

Model:
   emotion by emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5;

   mfq_18 by mfq18_01 mfq18_02 mfq18_03 mfq18_04 mfq18_05 mfq18_06
    mfq18_07 mfq18_08 mfq18_09 mfq18_10 mfq18_11 mfq18_12 mfq18_13;

   emotion with mfq_18;

Output:
   stdyx;
```

Here we have defined two latent variables – emotion, which is *measured by* the five items of the emotionality subscale from time point 3, and mfq_18 which is *measured by* the 13 MFQ items from age 18.

Notice that all the *manifest* items in these models have been declared as categorical variables. Therefore, the default approach in Mplus will be to estimate using least squares (WLSMV). Here all categorical items will be assumed to be imperfect measures of underlying continuous and normally distributed variables. The correlations between these underlying continuous measures will be estimated (polychorics) and the measurement models estimated using this information.

Things to check/observe

[1] You are putting in and getting out what you expect:-

```
Binary and ordered categorical (ordinal)
  EMO_T3_1    EMO_T3_2    EMO_T3_3    EMO_T3_4    EMO_T3_5    MFQ18_01
  MFQ18_02    MFQ18_03    MFQ18_04    MFQ18_05    MFQ18_06    MFQ18_07
  MFQ18_08    MFQ18_09    MFQ18_10    MFQ18_11    MFQ18_12    MFQ18_13

Continuous latent variables
   EMOTION      MFQ_18
```

[2] You are using the estimator you intended to use

```
Estimator                                                WLSMV
Maximum number of iterations                              1000
Convergence criterion                                0.500D-04
Maximum number of steepest descent iterations               20
Maximum number of iterations for H1                       2000
Convergence criterion for H1                         0.100D-03
Parameterization                                          DELTA
```

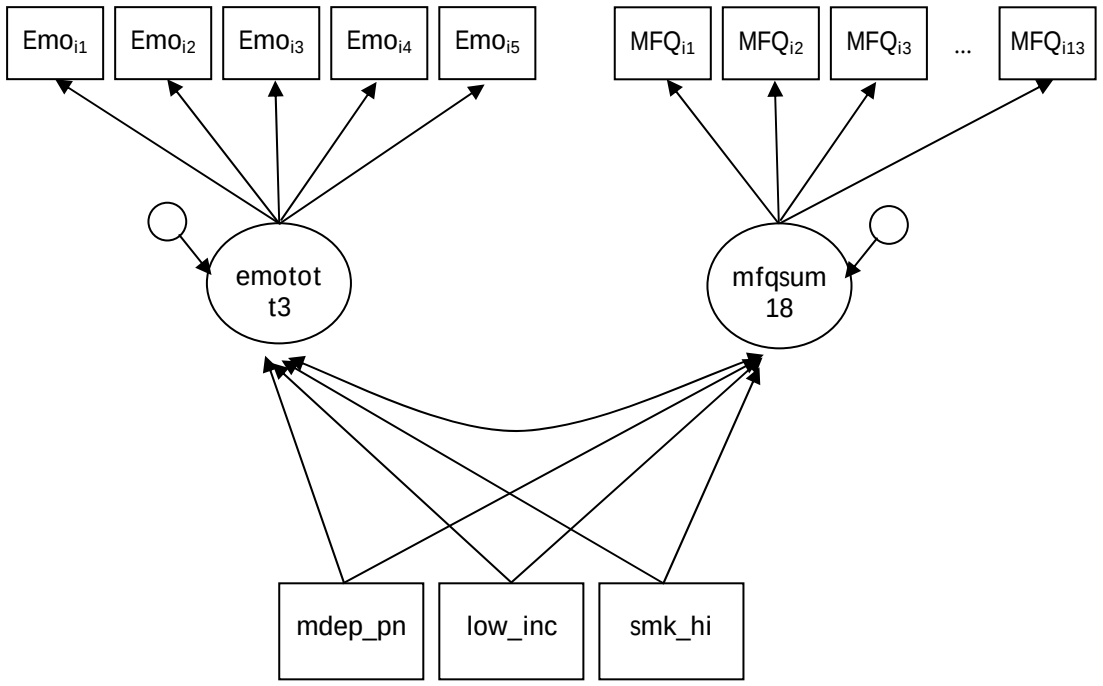
[3] Model fit - in the model fit section you will observe that the chi-square fit statistic is high (again!) however the other measures are as we would hope: CFI = 0.981, TLI = 0.978. RMSEA 0.051 (a little high but not excessive).

[4] The model results indicate a covariance of 0.118 (SE=0.019) between the two factors. If you scroll down further to the standardized output you a moderate correlation of 0.190 (SE=0.029)>

[5] The items do not all load on the two factors to the same extent. There is a relatively weak loading for the 5th item of the EAS emotionality trait. A number of the MFQ items also load quite weakly.

5.3 MIMIC Models

MIMIC (Multiple Indicator Multiple Cause) models are measurement models fitting along with covariates. We can briefly look at one of these models before fitting the final SEM model.



Add two model lines to the previous model statement

```
   emotion on smk_hi low_inc mdep_pn;
   mfq_18 on smk_hi low_inc mdep_pn;
```

and don't forget to also add these new variables to the usevariable list (defined ones last).

```
EMOTION  ON
    SMK_HI             0.108      0.121      0.891      0.373
    LOW_INC           -0.014      0.075     -0.182      0.855
    MDEP_PN            0.443      0.058      7.622      0.000

 MFQ_18   ON
    SMK_HI             0.205      0.113      1.816      0.069
    LOW_INC            0.201      0.072      2.792      0.005
    MDEP_PN            0.195      0.055      3.537      0.000
```

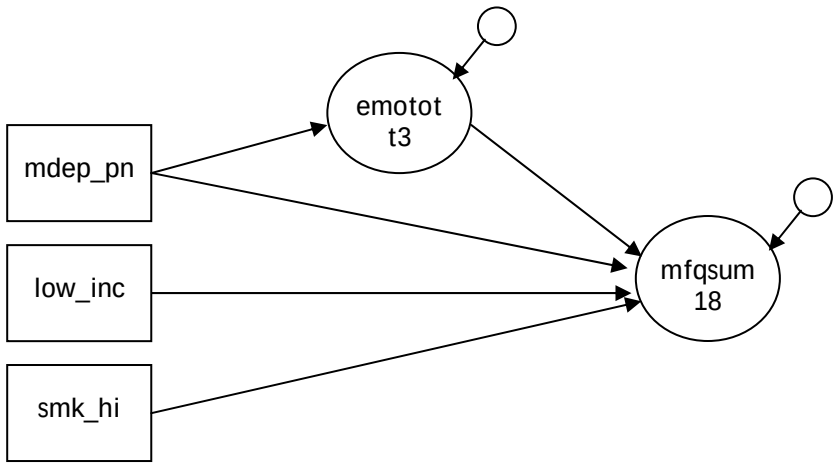
From the model results there is a strong relationship between postnatal depression and both latent variables and also a strong effect of income on adolescent symptom but no apparent effect on emotionality.

5.4 The SEM model

If you've been paying attention you shouldn't be at all surprised about the form of the syntax needed for the SEM model.

```
Model:
   emotion by emo_t3_1 emo_t3_2 emo_t3_3 emo_t3_4 emo_t3_5;

   mfq_18 by mfq18_01 mfq18_02 mfq18_03 mfq18_04 mfq18_05 mfq18_06
    mfq18_07 mfq18_08 mfq18_09 mfq18_10 mfq18_11 mfq18_12 mfq18_13;

   ! effect of EAS temperament on depressive symptoms
   mfq_18 on emotion;

   ! effect of baseline factors on depressive symptoms
   mfq_18 on low_inc smk_hi mdep_pn;

   ! effect of postnatal depression on emotionality
   emotion on mdep_pn;
```

This looks like a combination of the CFA model from earlier with the structural model from yesterday. Note that we've removed the factor covariance (emotion with mfq_18). Had we left it in this would have specified that the residuals for emotott3 and mfqsum18 were correlated setting up a pathway from the outcome back to the mediator (a non-recursive model).




```
MFQ_18   ON
    EMOTION            0.153      0.027      5.666      0.000

 MFQ_18   ON
    LOW_INC            0.201      0.072      2.792      0.005
    SMK_HI             0.205      0.113      1.817      0.069
    MDEP_PN            0.127      0.056      2.277      0.023

 EMOTION  ON
    MDEP_PN            0.443      0.058      7.622      0.000
```

As before we can add an addition command to allow us to partition the total effect of postnatal depression on adolescent symptoms

```
Model indirect:

   mfq_18 IND mdep_pn;
```

This gives the additional output:-

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Effects from MDEP_PN to MFQ_18

  Total                0.195      0.055      3.537      0.000
  Total indirect       0.068      0.015      4.577      0.000

  Specific indirect
    MFQ_18
    EMOTION
    MDEP_PN            0.068      0.015      4.577      0.000

  Direct
    MFQ_18
    MDEP_PN            0.127      0.056      2.277      0.023
```

So the total effect of postnatal depression on adolescent symptoms is 0.195 (SE = 0.055). In the fitted model this total effect is partitioned into an indirect effect of 0.068 and a direct effect of 0.127, i.e. approximately 35% of the effect of postnatal depression is mediated through childhood emotionality.

Interpreting the magnitude of the total effect is not easy as we do not know the variance of MFQ_18 in this model. If you change the measurement models to use the alternative formulation of freeing the loading and fixing the variances at one then things become a little clearer:-

```
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Effects from MDEP_PN to MFQ_18

  Total                0.265      0.075      3.539      0.000
  Total indirect       0.092      0.020      4.527      0.000
```

We can now see that the total effect of postnatal depression on adolescent symptoms is of moderate size – those with and without postnatal depression have adolescents who differ on average by 0.27 SD's.