UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

Introduction to Mplus: Latent variables, traits and classes

Peterhouse College, Cambridge

24th -25th January 2011

This course is prepared by

Anna Brown, PhD ab936@medschl.cam.ac.uk
Research Associate

Tim Croudace, PhD tjc39@cam.ac.uk
Senior Lecturer in Psychometric Epidemiology

UNIVERSITY OF CAMBRIDGE
School of Clinical Medicine > Department of Psychiatry

This course is funded by the ESRC RDI and hosted by The Psychometrics Centre

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

RDI Researcher Development Initiative

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Programme

- Monday 1 pm – 6 pm
  - 1 pm Lunch
  - 1:30 pm Introduction to M*plus* (EFA, regression, CFA)
  - 6 pm Finish
- Tuesday 9 am – 2 pm
  - 9 am Breakfast
  - 9:30 am Latent traits (IRT models) and Latent classes (LCA)
  - 1 pm Lunch

Day 1

- We introduce Mplus modelling environment and show how to describe your data and variables.
- We then move on to modelling, introducing Mplus capabilities, commands and outputs gradually.
- We cover Exploratory Factor Analysis (EFA) with different rotations, Confirmatory Factor Analysis (CFA), regression and path analysis.
- We will work with different types of observed variables – continuous and categorical.

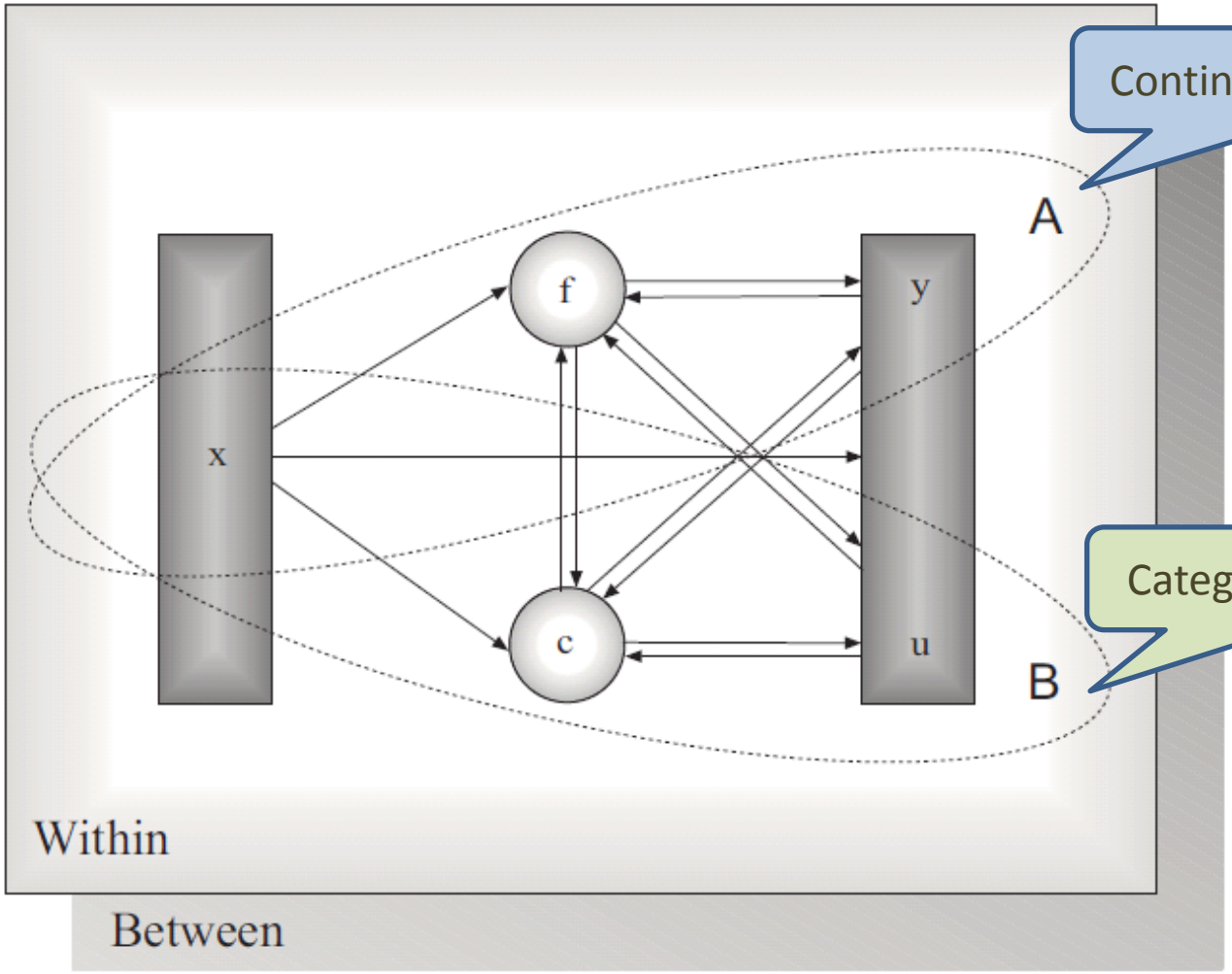
Introduction

WHAT IS M*PLUS*

Mplus strengths

- Comprehensive modelling capabilities
  - Regression and path analysis
  - Exploratory factor analysis
  - Confirmatory factor analysis and SEM
  - Growth modelling
  - Mixture modelling
  - Multilevel modelling
  - Missing data modelling
  - Monte Carlo simulation studies
- Comprehensive set of estimators

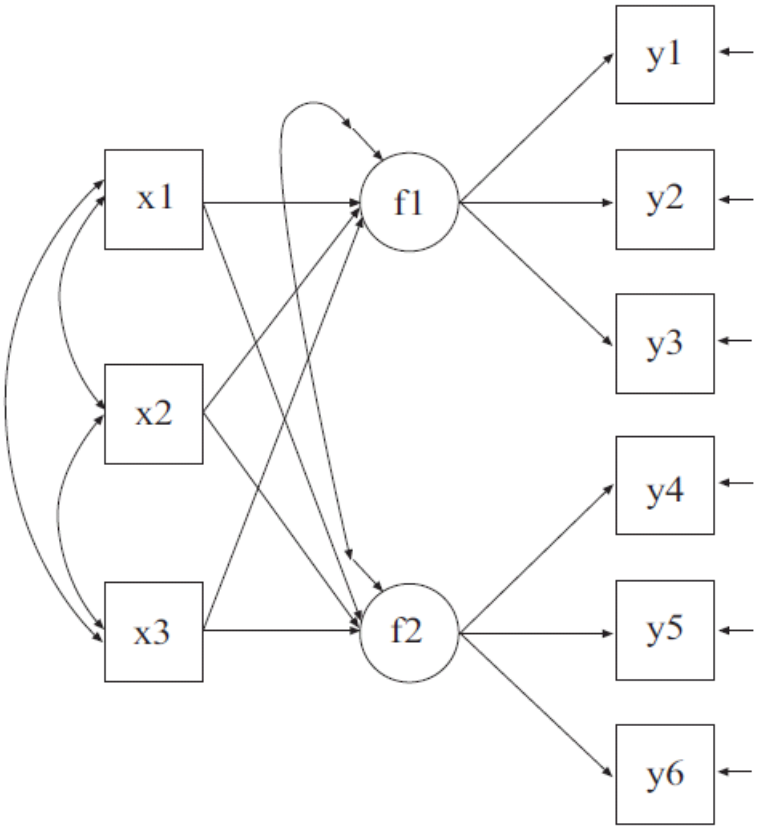
The M*plus* modelling framework



Mplus features

- Strong statistics
- Command language
  - Only very simple wizard to help with writing basic syntax
- No graphical input or output for model spec
- No data import facilities from other statistical packages

No graphical input or output for model spec…



MODEL:

f1 BY y1-y3;

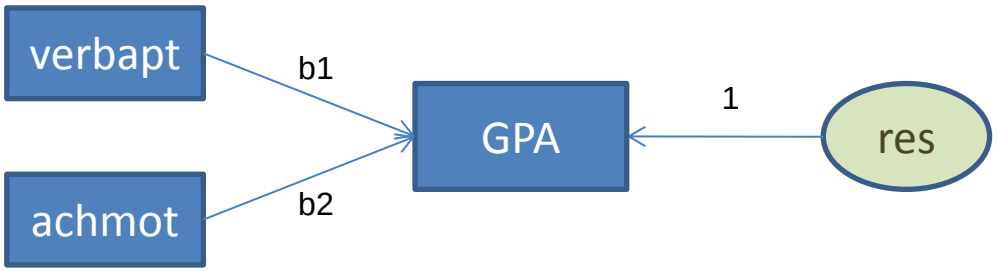
f2 BY y4-y6;

f1 f2 ON x1-x3;

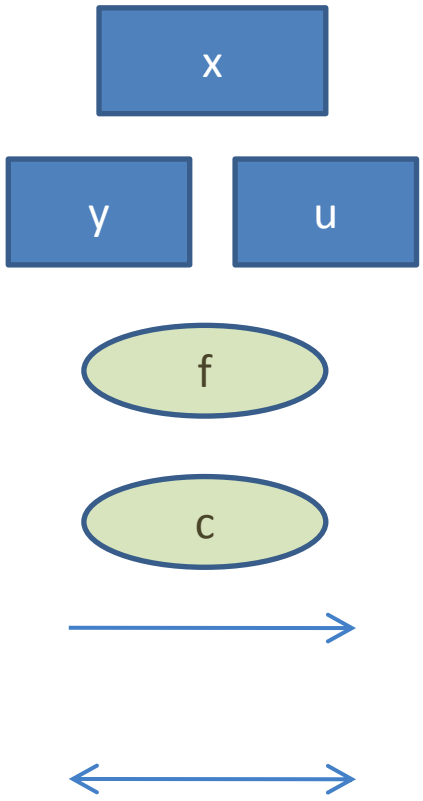
Structural equation modelling

- Models describing relations (covariation) among observed variables using (in simple cases linear) equations
- May include latent (unobserved) variables (factors)
- Confirmatory method: model is tested for fit
- For example, multiple regression is an SEM model

  GPA = b0 + b1*verbapt + b2*achmoti + residual



SEM conventions



- Observed variables
  - Outcome (y or u) / Background (x)
  - Continuous (y) / Categorical (u)
- Latent (unobserved) variables
  - Continuous (f; factors)
  - Categorical (c; classes)
- Causal effects
- Covariance (no causal effect)

Mplus basic commands

GETTING STARTED

Syntax structure

- TITLE: a title for the analysis (not part of the syntax)
- DATA: (*required*) information about the data set
- VARIABLE: (*required*) information about the variables in the data set
- DEFINE: transform existing variables and create new variables
- ANALYSIS: technical details of the analysis
- MODEL: describe the model to be estimated
- OUTPUT: request additional output
- SAVEDATA: save the analysis data, auxiliary data, and results
- PLOT: request graphical displays of observed data and results
- MONTECARLO: details of a simulation study

Some conventions

- Order of syntax sections can be any
- The records in the input setup must be no longer than 90 columns
- Each command finishes with "**;**"
- Not case sensitive (but capital letters are often used for readability)
- A comment is anything followed by an exclamation mark, like this **!** This is a comment
- Clever with expanding names:

  item1-item100 is understood to be item1 item2 … item100

Learning plan

1. How to describe data and variables (required for all types of analysis)
2. We will start with simple analyses, introducing syntax gradually
    1. EFA (ANALYSIS command)
    2. CFA and Path Analyses (MODEL command)

Data files

- Individual data (*default*)
  - Data must be in external ASCII file
  - No more than 500 variables
  - The maximum record length is 5000
  - Each case starts on new line
  - Free format (*default*)
    - Data values separated by <tab> <space> or comma
    - Note: do not use blanks to indicate missing values, or commas to indicate decimal points!
  - Fixed format (FORTRAN-like, e.g. 5F4.0, 10x, 6F1.0;)
- Matrix input
  - Sequence is means, standard deviations, correlations
  - Default is lower triangle only for correlations

DATA command (basic)

DATA:

FILE IS *filename*; full path or just name if in the same folder, in ‘ ’ if contains spaces

FORMAT IS 5F4.0, 10x, 6F1.0; not needed if *free*

TYPE IS covariance; Or *corr*, or *means* etc.
not needed if *individual*

NOBSERVATIONS ARE 581; only needed for summary data

- With summary data
  - means come first, then SDs, and then entries of the lower triangular correlation matrix
- Note that **IS**, **ARE** and **“=“** can be used interchangeably (apart from using “=“ in arithmetic operations)

VARIABLE command

VARIABLE:

NAMES ARE names of variables in the data set

USEVARIABLES ARE names of analysis variables; default is all variables in NAMES

USEOBSERVATIONS ARE conditional statement to select observations, default is all

MISSING ARE variable (#); or . ; * ; **BLANK;**

- And many more commands declaring type of variables, e.g.

CATEGORICAL ARE binary and ordinal dependent variables;

First analyses

- Convert your data into one of the above formats
- Describe your data file (DATA command)
- Describe your variables (VARIABLE command)
- And we can already do some simple analyses

ANALYSIS command

ANALYSIS:

TYPE = GENERAL; (*default*, classical SEM)

BASIC; (compute basic statistics)

MEANSTRUCTURE; (*default*, models with means)

MISSING, H1; (*default*, MAR analysis incomplete data)

COMPLEX; (complex data)

EFA; (exploratory factor analysis)

Combinations apply, e.g. TYPE = COMPLEX MISSING;

ESTIMATOR =

- Choice of estimator depends on type of data and model
- Usually Maximum Likelihood (ML) or robust ML (MLR/MLM/MLMV)
- Also limited information ULS or DWLS (in Mplus ULSMV, WLS, WLSM, WLSMV)
- Bayes

OUTPUT command

OUTPUT:

SAMPSTAT; (sample statistics)

STANDARDIZED; (standardized solution)

RESIDUAL; (residuals)

MODINDICES; (modification indices > 10)

MODINDICES (#); (modification indices > #)

CINTERVAL; (confidence interval)

CINTERVAL (BOOTSTRAP / BCBOOTSTRAP);

FSCOEFFICIENT; (factor score coefficients)

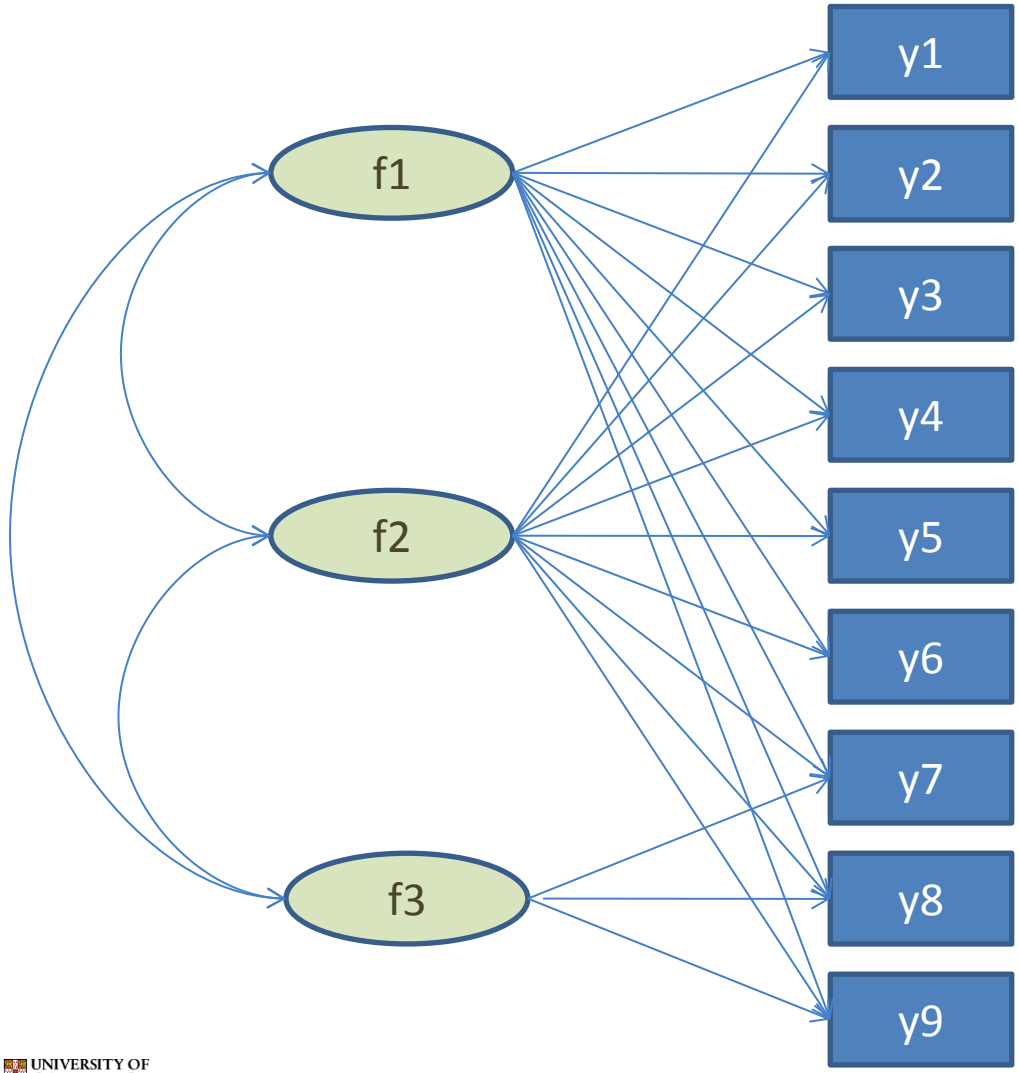
TECH#; (various technical outputs, often used for finding problems)

Learning to perform and interpret EFA

EXPLORATORY FACTOR ANALYSIS

EFA model

- It is useful to think of a model underlying EFA
- Factors are indicated by all observed variables
- Factors can be correlated or not



EFA command

ANALYSIS:

TYPE = EFA # #;

ROTATION = **GEOMIN**; ! (OBLIQUE) - default or (ORTHOGONAL)

QUARTIMIN !oblique only

CF-VARIMAX

CF-QUARTIMAX

CF-EQUAMAX

CF-PARSIMAX

CF-FACPARSIM

CRAWFER

OBLIMIN

PROMAX !oblique only

VARIMAX !orthogonal only

TARGET

Estimation methods for EFA

- For continuous variables default is ML
- Also available ESTIMATOR=MLM or MLMV for robust estimation (non-normal continuous data)
  - MLM = ML – Mean corrected (Satorra/Bentler)
  - MLMV = ML – Mean & Variance corrected (Muthen)
  - Both also produce (corrected) χ2-test and RMSEA
- ULS can also be used

Thurstone's data

- We will use this simple example to practice EFA (and later CFA) with continuous variables
- Classic study of "primary mental abilities" by Thurstone
- We have 9 subtests (continuous variables) measuring 3 mental abilities
  - Subtest1-subtest3 measure Verbal Ability
  - Subtest4-subtest6 measure Word Fluency
  - Subtest7-subtest9 measure Reasoning Ability

Thurstone's data – cont.

- We will analyse a correlation matrix (THUR.dat), *n*=215

```
 1
.828  1
.776 .779  1
.439 .493 .460  1
.432 .464 .425 .674  1
.447 .489 .443 .590 .541  1
.447 .432 .401 .381 .402 .288  1
.541 .537 .534 .350 .367 .320 .555  1
.380 .358 .359 .424 .446 .325 .598 .452  1
```

1= sentences
2= vocabulary
3= sentence completion
4= first letters
5= four-letter words
6= suffixes
7= letter series
8= pedigrees
9= letter grouping

Thurstone data – syntax for EFA

TITLE: EFA of Thurstone correlation matrix of Primary mental abilities – subtests

DATA: **FILE IS** THUR.dat**;**

TYPE IS CORRELATION;

NOBSERVATIONS = 215**;**

VARIABLE: **NAMES ARE** subtest1-subtest9**;**

ANALYSIS:

TYPE IS EFA 1 3**;** !we will fit 1, 2 and 3 factor models

ROTATION=CF-VARIMAX (ORTHOGONAL); !we will try different rotations

!ROTATION=CF-VARIMAX (OBLIQUE);

OUTPUT: **RESIDUALS;** !optional, but very useful in model assessment

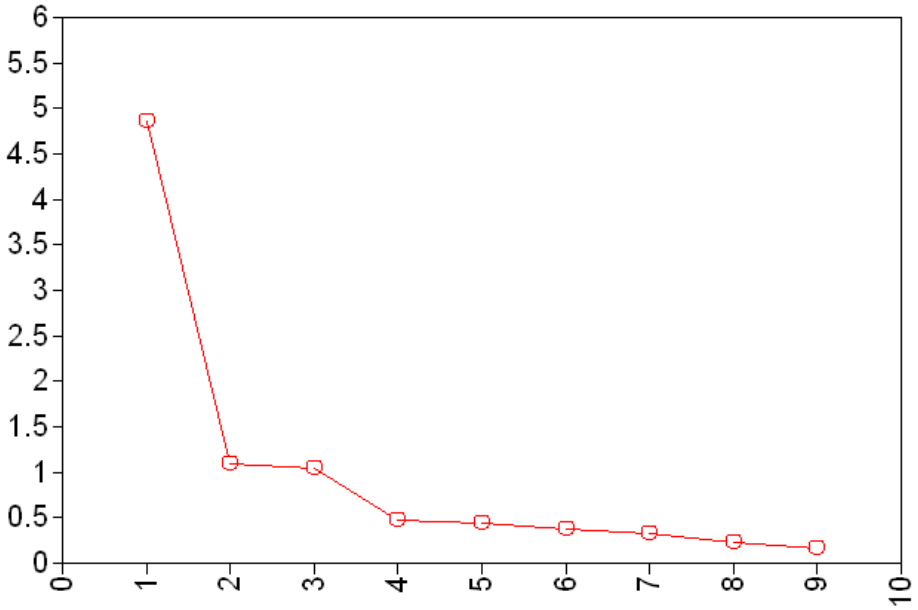
PLOT: **TYPE = PLOT2;** !optional, will produce a scree plot

Eigenvalues

EIGENVALUES FOR SAMPLE CORRELATION MATRIX

1	2	3	4	5	6	7	8	9
4.851	1.090	1.038	0.475	0.448	0.375	0.321	0.234	0.168

- Scree plot (**PLOT** command)



Importance of checking residuals

- Residuals are not printed by default - ask for them

  OUTPUT: RESIDUALS**;**

- Looking at the 1-factor model and 2-factor model residuals it is easy to see where the areas of misfit are
- For instance, in the 2-factor model correlations between the last 3 subtests are not explained well

	SUBTEST6	SUBTEST7	SUBTEST8	SUBTEST9
SUBTEST6	0.000			
SUBTEST7	-0.086	0.000		
SUBTEST8	-0.048	0.217	0.000	
SUBTEST9	-0.062	0.284	0.143	0.000

- 3-factor model has near-0 residuals
- We will proceed with 3 factors for this data

Goodness of fit

- Residuals are a good way of assessing fit
- Global fit indices exist summarising the overall goodness of fit
- Exact fit – chi-square statistics
  - For continuous data it is based on discrepancies between the observed and the model-based covariance matrices
  - Can reject good models if the sample is very large, and accept bad models if the sample is small
- Global fit indices not sensitive* to sample size
  - Comparative indices, relative fit compared to the baseline model (CFI, TLI > 0.95 for good fit)
  - Absolute indices, index of discrepancy between model and population (RMSEA < 0.05 for good fit)
    - ❖ Many indices have been found to be dependent on sample size after all

Fit for different models

	1 factor	2 factors	3 factors
Chi square	236.848	86.112	2.944
df	27	19	12
CFI	.806	.938	1
RMSEA	.190	.128	0

- Extraction method – Maximum Likelihood
- 3 factor model is over fitting but 2 factor model is clearly not acceptable
- Check standard errors – are they of magnitude $1/\sqrt{n}$ (is the model identified?)
  - Sample size is *n*=215, so SE should be of order 0.07

Examining orthogonal rotated loadings

	1	2	3
SUBTEST1	0.858	0.196	0.223
SUBTEST2	0.854	0.270	0.180
SUBTEST3	0.800	0.240	0.187
SUBTEST4	0.287	0.782	0.197
SUBTEST5	0.269	0.698	0.261
SUBTEST6	0.358	0.598	0.103
SUBTEST7	0.277	0.185	0.779
SUBTEST8	0.478	0.151	0.503
SUBTEST9	0.200	0.317	0.622

- Factor loadings are largely in line with expectations, however, there are many non-zero loadings

Examining oblique rotated loadings

	1	2	3
SUBTEST1	0.824	0.044	0.121
SUBTEST2	0.811	0.139	0.058
SUBTEST3	0.758	0.111	0.078
SUBTEST4	0.025	0.817	0.053
SUBTEST5	0.011	0.709	0.145
SUBTEST6	0.187	0.614	-0.031
SUBTEST7	0.016	-0.003	0.842
SUBTEST8	0.332	-0.012	0.501
SUBTEST9	-0.061	0.198	0.643

- Factor loadings are much closer to an independent clusters solution

Factor correlations

- In the oblique solution, factors are correlated

	1	2	3
1	1.000		
2	0.463	1.000	
3	0.455	0.464	1.000

  - We would expect mental abilities to be correlated
- We are happy with the solution with 3 correlated factors

Learning to describe and analyse simple CFA models

CONFIRMATORY FACTOR ANALYSIS

Common factor models

- Latent (unobserved) factors are indicated BY observed variables
- Latent factors do not have a scale, it needs to be set either by
  1. Setting one factor loading – this is default:

     F1 BY y1 y2 y3; !means F1 BY y1@1 y2 y3;
  2. Setting factor variance (to 1) and freeing the first factor loading:

     F1 BY y1* y2 y3; F1@1;
- Continuous observed variables have their own scale which they can pass to the latent factors; categorical do not

MODEL command

MODEL: <specification>

- This is where the SEM model is specified
- Important keywords are BY, ON, WITH

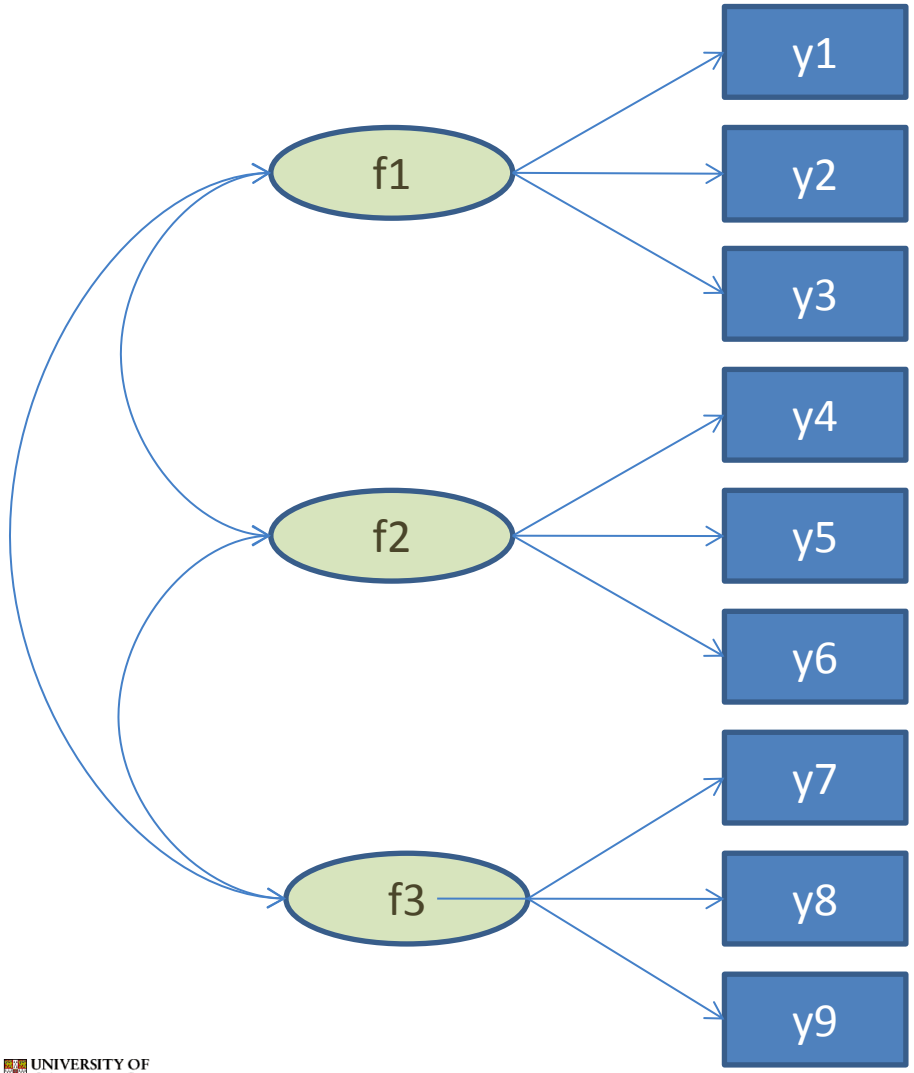
  <factor> *Measured* BY <indicator>

  <outcome> *Regressed* ON <predictor>

  <(latent) variable> *Correlated* WITH <(latent) variable>
- @ fix parameter (specify a constraint)
- * free up parameter (if previously constrained)

Thurstone's data

- We have 9 subtests (continuous variables) measuring 3 mental abilities
  - Subtest1-subtest3 measure Verbal Ability
  - Subtest4-subtest6 measure Word Fluency
  - Subtest7-subtest9 measure Reasoning Ability



CFA syntax

```
TITLE: CFA of Thurstone correlation matrix
DATA: FILE IS THUR.dat;
      TYPE IS CORRELATION;
      NOBSERVATIONS = 215;
VARIABLE:  NAMES ARE subtest1-subtest9;
ANALYSIS: !defaults are ok
MODEL:
test1 BY subtest1-subtest3*;
test2 BY subtest4-subtest6*;
test3 BY subtest7-subtest9*;
test1-test3@1;
test1 WITH test2@0 test3@0;          !we will try orthogonal solution first
test2 WITH test3@0;                  ! but then will relax these constraints
OUTPUT: RES;
PLOT: TYPE=PLOT2;
```

Uncorrelated factors - model fit

Chi-Square Test of Model Fit

Value	219.484
Degrees of Freedom	27
P-Value	0.0000
CFI	0.822

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.182
90 Percent C.I.	0.160 0.205

SRMR (Standardized Root Mean Square Residual)

Value	0.330

- Model fits very poorly
- Standard errors of estimates are of order 0.07 or below (model is identified)

Uncorrelated factors – model results

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
TEST1 BY				
SUBTEST1	0.906	0.054	16.802	0.000
SUBTEST2	0.910	0.054	16.906	0.000
SUBTEST3	0.852	0.056	15.296	0.000
TEST2 BY				
SUBTEST4	0.855	0.064	13.452	0.000
SUBTEST5	0.784	0.064	12.195	0.000
SUBTEST6	0.687	0.065	10.529	0.000
TEST3 BY				
SUBTEST7	0.855	0.070	12.190	0.000
SUBTEST8	0.646	0.069	9.332	0.000
SUBTEST9	0.696	0.069	10.028	0.000
TEST1 WITH				
TEST2	0.000	0.000	999.000	999.000
TEST3	0.000	0.000	999.000	999.000
TEST2 WITH				
TEST3	0.000	0.000	999.000	999.000

Uncorrelated factors – residuals

- Model fails to explain correlations *between* clusters

	1	2	3	4	5	6	7	8	9
SUBTEST1	0.000								
SUBTEST2	0.000	0.000							
SUBTEST3	0.000	0.000	0.000						
SUBTEST4	0.437	0.491	0.458	0.000					
SUBTEST5	0.430	0.462	0.423	0.000	0.000				
SUBTEST6	0.445	0.487	0.441	0.000	0.000	0.000			
SUBTEST7	0.445	0.430	0.399	0.379	0.400	0.287	0.000		
SUBTEST8	0.538	0.535	0.532	0.348	0.365	0.319	0.000	0.000	
SUBTEST9	0.378	0.356	0.357	0.422	0.444	0.323	0.000	0.000	0.000

Correlated factors - model fit

Chi-Square Test of Model Fit

Value	38.737
Degrees of Freedom	24
P-Value	0.0291
CFI	0.986
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.053
90 Percent C.I.	0.017 0.083
SRMR (Standardized Root Mean Square Residual)	
Value	0.044

- Model fits well
- Standard errors of estimates are of order 0.07 or below

Correlated factors – model results

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
TEST1 BY				
SUBTEST1	0.903	0.054	16.805	0.000
SUBTEST2	0.912	0.053	17.084	0.000
SUBTEST3	0.854	0.056	15.388	0.000
TEST2 BY				
SUBTEST4	0.834	0.060	13.847	0.000
SUBTEST5	0.795	0.061	12.998	0.000
SUBTEST6	0.701	0.064	11.012	0.000
TEST3 BY				
SUBTEST7	0.779	0.064	12.231	0.000
SUBTEST8	0.718	0.065	11.050	0.000
SUBTEST9	0.702	0.065	10.729	0.000
TEST2 WITH				
TEST1	0.643	0.050	12.815	0.000
TEST3 WITH				
TEST1	0.670	0.051	13.215	0.000
TEST2	0.637	0.058	10.951	0.000

Correlated factors – residuals

- Model explains all correlations quite well

	1	2	3	4	5	6	7	8	9
SUBTEST1	0.000								
SUBTEST2	0.001	0.000							
SUBTEST3	0.001	-0.003	0.000						
SUBTEST4	-0.047	0.002	0.000	0.000					
SUBTEST5	-0.031	-0.004	-0.014	0.008	0.000				
SUBTEST6	0.038	0.076	0.056	0.003	-0.019	0.000			
SUBTEST7	-0.026	-0.046	-0.047	-0.035	0.005	-0.061	0.000		
SUBTEST8	0.104	0.096	0.120	-0.033	0.001	-0.002	-0.007	0.000	
SUBTEST9	-0.046	-0.072	-0.044	0.049	0.088	0.010	0.048	-0.054	0.000

Modification Indices

- Useful to guide modification of the model
- Modification index (M.I.) is the value by which chi-square will drop if the parameter currently fixed or constrained was freely estimated
- To request modification indices

  OUTPUT: MOD (<*min.value*>);

- E.P.C. is expected parameter change index
  - Expected value of the parameter if it was freely estimated

Extension to categorical variables

- Often the observed variables are binary or ordinal (for example test items)
- Mplus provides straightforward extension of EFA techniques to this type of data
  - Either using full information estimators (ML) – *full information factor analysis*, suitable for models with few factors (not more than 4)
  - Or limited information estimators (ULS, WLS) – based on polychoric correlations and can handle large models with many factors and variables
- The only modification to syntax needed is to declare variables as categorical

```
CATEGORICAL ARE i1-i10;
```

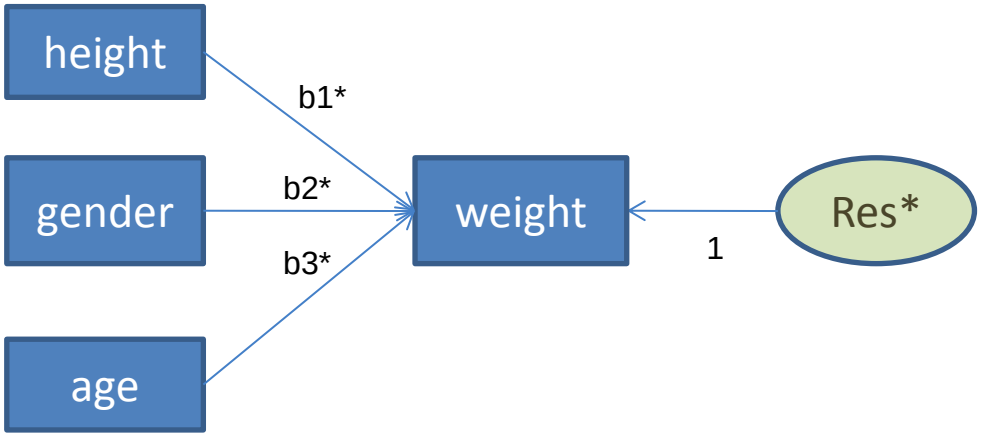
Continuous, categorical, count variables

MODELLING REGRESSION

Multiple regression model

- Multiple regression is an SEM model
- Dependent variable is a (linear) combination of independent variables

weight = b0 + b1*height + b2*gender + b3*age + residual



Different types of dependent variables

- Continuous variables
  - Linear regression
    - *Predicting human body weight from height, gender and age*
- Binary and ordinal variables
  - Logistic regression
    - *Predicting probability of answering an ability item correct from the overall ability score*
- Count variables
  - Poisson regression
    - *Predicting the number of migrations from education and the number of years since leaving school*

Regression syntax

```
TITLE: Multiple regression with body measurements
DATA: FILE IS body.dat;
VARIABLE:
    age             !Age (in years)
    weight          !Weight (in kg)
    height          !Height (in cm)
    gender;         !Gender; 1 for males and 0 for females
USEVARIABLES ARE age weight gender height;

MODEL:
weight ON age gender height;

OUTPUT: SAMPSTAT; STD;
```

It is not necessary to refer to the means, variances, and covariances among the **x** variables because they are not part of the model estimation, and no restriction on them is imposed.

Continuous variables: regressing weight on height, age and gender

- Data from N=507 individuals on various body measurements**

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
WEIGHT ON				
AGE	0.180	0.040	4.458	0.000
GENDER	7.687	1.060	7.249	0.000
HEIGHT	0.725	0.056	12.977	0.000
Intercepts				
WEIGHT	-64.115	9.356	-6.853	0.000
Residual Variances				
WEIGHT	74.098	4.654	15.923	0.000

Older people, men and tall people are heavier, controlling for all other variables

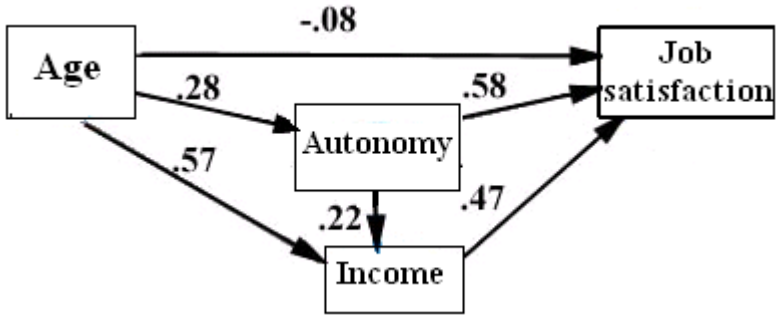
- Regression models are saturated, nothing to test
- R-square = 0.583 (0.028)

**Source
http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html

PATH ANALYSIS

Path models

- Path analysis is an extension of the (multiple) regression model – same equations apply
- Independent (*exogenous*), intermediary and dependent (*endogenous*) variables
- In path analysis, a variable can be a dependent in one relationship and an independent in another. These variables are referred to as *mediating* variables.
- There could be undirected (covariance) and directed paths (causal relationships)



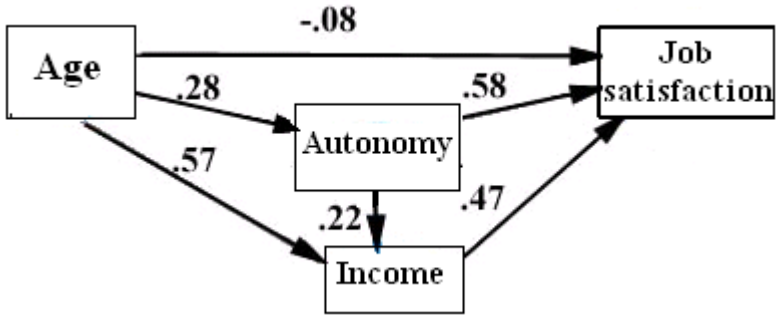
Types of relationships between two variables

Direct effect: X causes Y or Y causes X, or both.

Indirect effect: The relationship between X and Y is *indirect* if X causes Z which in turn causes Y.

Spuriousness: The relationship between X and Y is *spurious* if Z causes both X and Y.

Unexplained covariation: Both X and Y are exogenous and so variation between them is not explained by the model.



Decomposition of correlation

- Purpose of Path analysis is to decompose observed correlations between variables
- Correlation between two endogenous variables:

  Correlation = Direct Effect + Indirect Effects + Spuriousness
- Correlation between an endogenous variable and an exogenous variable:

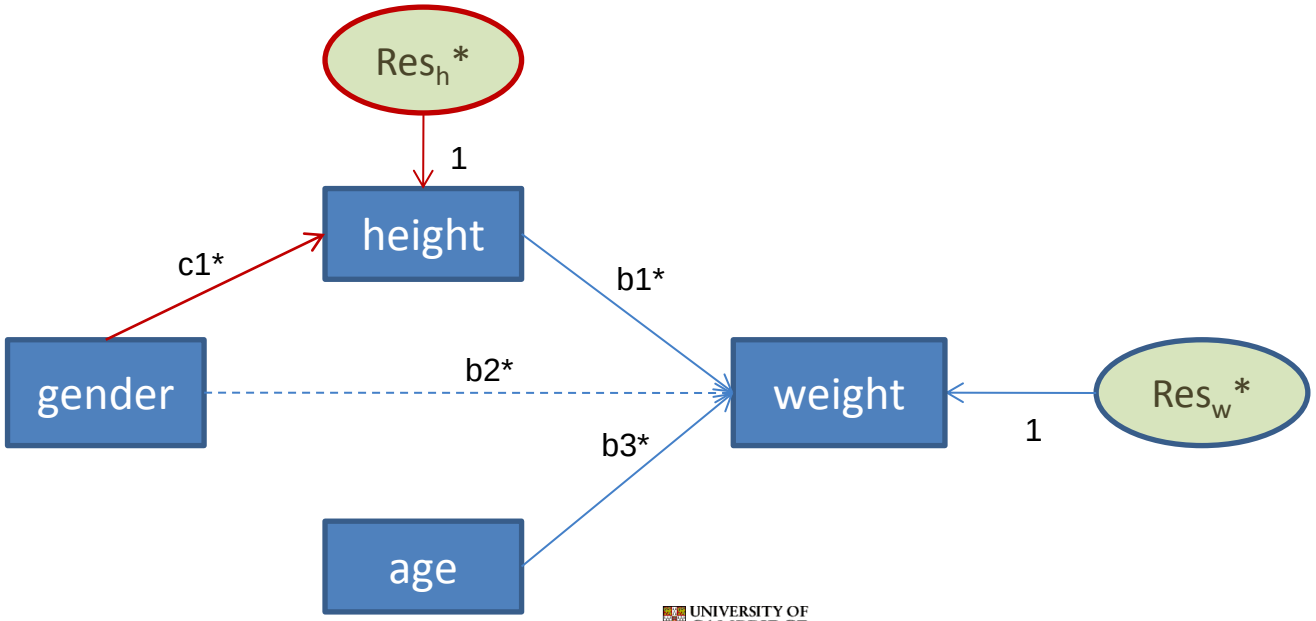
  Correlation = Direct Effect + Indirect Effects + Unspecified Covariance

Testing path models

- The purpose is to estimate model parameters using observed data
- Collect scores on all variables in a (large) sample
- Observed data is covariances between all variables
  - If include all covariances in the model, it is *saturated* – there is nothing to test
  - The aim is to test hypotheses with more restrictions on relationships between variables
- Estimate path coefficients (regression coefficients) and (co)variances

Example: weight, height, gender and age

- This is the same data we used for multiple regression
- This time we test a hypothesis that weight in only influenced by gender indirectly, through gender differences in height
  - We know from the regression results that this is incorrect



Testing alternative path models

MODEL: ! No direct path from gender to weight

weight ON age height;

height ON gender;

- This model does not fit
  - Chi-square = 51.277 (df=2)

MODEL: ! There is direct path from gender to weight

weight ON age height;

height ON gender;

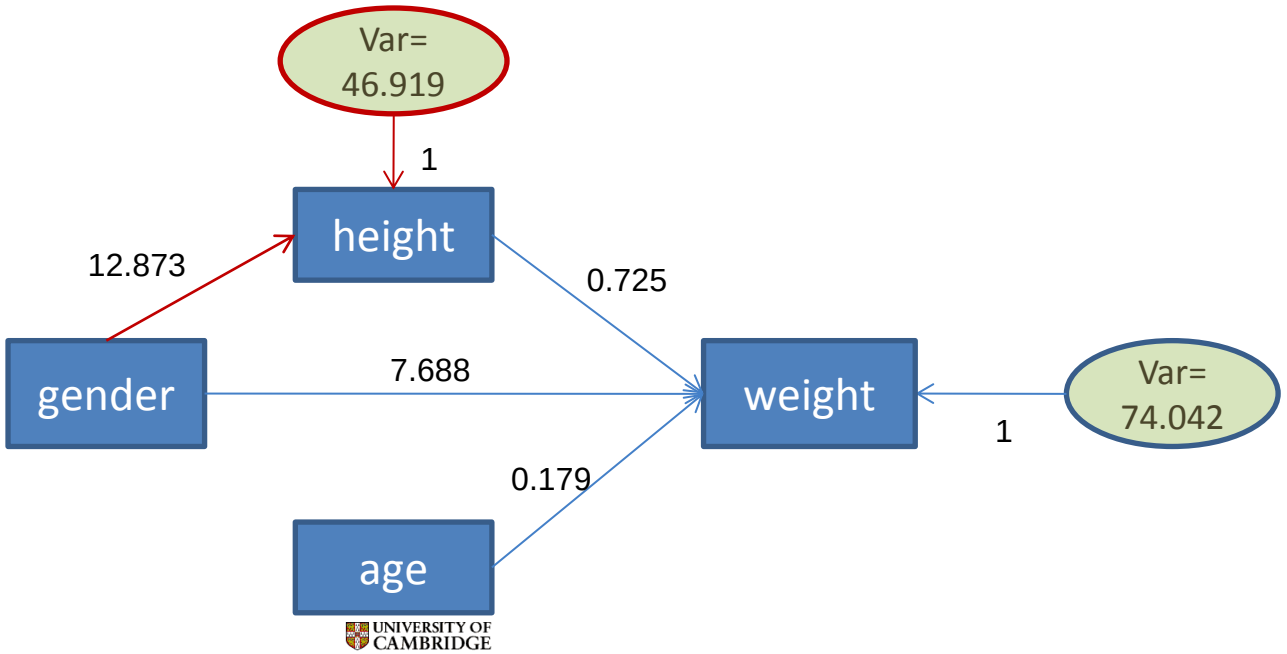
- This model fits very well
  - Chi-square = 1.214 (df=1)

Path analysis estimates

- R-square

WEIGHT	0.585	(0.028)
HEIGHT	0.469	(0.032)

- Unstandardised estimates



Summary and plan for Day 2

- Covered today
  - Introduced Mplus modelling framework and syntax
  - Practiced EFA and CFA with continuous variables
  - Discussed the extension to categorical variables
  - Practised multiple regression and path analysis with continuous variables
- Tomorrow
  - Will start with logistic regression
  - Analyse data with multiple groups (using MIMIC models and multi-group setup in Mplus)
  - Analyse data with unobserved grouping (LCA)