UCT MA Psychology
Multivariate Statistics
Test 2, November 2018
==========

Instructions

A. Submit answers as html or word doc or pdf, as compiled from R markdown or R script. Submit zip of entire project folder, in the usual way. Upload all to the VULA assignment page. Your report should attempt to attain full reproducibility. Please note that there is an allocation of marks for well formatted reports.

B. The deadline – strict - for submission is 12 November 2018 at 12h55.

C. Work entirely on your own. Do not consult other class members or indeed any other person. In all other respects this is an open book test.

D. Page length – ***strictly*** a maximum of 15 pages, 1.5 line spacing, font size 12.

Questions

The data sets for this test are located on VULA, under the Test 2 Assignment tab.

Question 1 – Structural Equation Modeling

The data set peru2.dat contains variables from earlier and later rounds of measurement in the Peru component of the Young Lives Study. We want to build a Structural Equation Model that predicts anti-social, or risk-taking behaviour at Round 5. There are several possible predictors in the peru2.dat dataset, measured at earlier and later rounds, and sometimes throughout. Construct a model that seems plausible to you, and test it. You are not expected to construct a Latent Growth Model, and that might be somewhat difficult to do. In order to arrive at a plausible model, you might want to consult one or more of the references on Vula.

This question is deliberately open-ended. Here are some things you might want to take into account when building and testing your model. Measures of fit? Significance and strength of individual paths? Unstandardized or standardized coefficients? Residuals? Modification indices? Stability of model (is there some way of cross-validating it)? Alternate models? Diagram of proposed model? Diagram of model after fitting, path pruning, etc? Measurement model adequacy? Full structural model adequacy?

List of manifest variables in the data set, as well as potential latent variables.

Variable	***type***	***Definition - notes***	
childid	<chr>	Child identification number – unique	
hq_r12	<dbl>	Housing quality	**These variables are indicators of a latent variable, 'wealth_index'. They are averaged over rounds 1 and 2**
sv_r12	<dbl>	Access to services	
cd_r12	<dbl>	Consumer durables	
round	<dbl>	Round of study	2002, 2005, 2009, 2012, 2015

agemon	<dbl>	Age in months	
sex	<dbl+lbl>	Sex of child	
schtype	<dbl>	Type of school	Government or Private
typesite	<dbl+lbl>	Rural or Urban status of research site	Urban or Rural
bmi	<dbl>	Body Mass Index	Height x weight
bwght	<dbl+lbl>	Weight of child at birth	
zhfa	<dbl>	Standardized height for Age	Standardized against WHO norms
maths_perco	<dbl>	Percentage score in maths test	
ppvt	<dbl>	Receptive vobabulary score	Peabody Picture Vocab Test
self_eff1	<dbl>		
self_eff2	<dbl>		
self_eff3	<dbl>		
self_eff4	<dbl>		**These 8 items are indicators for a latent variable measuring self-efficacy**
self_eff5	<dbl>		
self_eff6	<dbl>		
self_eff7	<dbl>		
self_eff8	<dbl>		
agency1	<dbl+lbl>		
agency2	<dbl+lbl>		**These 5 items are indicators for a latent variable measuring agency**
agency3	<dbl+lbl>		
agency4	<dbl+lbl>		
agency5	<dbl+lbl>		
sdq1	<dbl>		
sdq2	<dbl>		**These 5 items are indicators for a latent variable measuring mental wellness**
sdq3	<dbl>		
sdq4	<dbl>		
sdq5	<dbl>		
FRNSMKR5	<dbl>	Have friends who smoke, R5 (reversed)	
FRNALCR5	<dbl>	Have friends who use alcohol, R5 (reversed)	**These are candidate indicators for a latent variable measuring anti-social or risk taking behaviour. It is not known if they form a coherent latent variable. Many items have been reversed, or had unusual categories removed. You can accept them as is without needing to modify them before analysis.**
YOUALCR5	<dbl>	Uses alcohol, R5 (reversed)	
BEATEN	<dbl>	Composite score, beaten up by friends, strangers, teachers, parents (more = more beatings)	
ARRSTDR5	<dbl>	Has been arrested, R5	
FRNGNGR5	<dbl>	Has friends in gangs, R5 (reversed)	
MEMGNGR5	<dbl>	Is a member of a gang, R5	
CRYWPNR5	<dbl>	Has carried a weapon, R5	
NUMPRTR5	<dbl>	Number of sex partners had, R5	

Note: questionnaire items have been included under the assignment tab on VULA. In many cases items have been reversed, some categories removed, so use only as a rough guide. There are also two articles that you might wish to consult for ideas on what to put in your model.

Question 2: Mixed-Effects Models (100 marks)

The Young Lives is a research organisation dedicated to investigating various aspects of childhood and youth. One focus of the YL team has been to investigate a range of potential early childhood predictors of cognitive development, including poverty & inequality, nutrition, education and various other sociodemographic factors. Using the data and information provided below, construct a mixed-effects model that predicts receptive vocabulary scores on the Peabody Picture Vocabulary Test at round 5 for children in Ethiopia, India, Peru and Vietnam. When building the model, you may choose to use information from all 5 rounds, or you may decide to only include information from particular rounds. Whichever you choose, be sure to justify your decision.

The data sets for question 2 are labelled "YL_June2017_Ethiopiadata_2017-09-08", "YL_June2017_Indiadata_2017-11-22", "YL_June2017_perudata_2017-11-12", and "YL_June2017_Vietnamdata_2017-11-01".

1. Import the data files into R, creating a single dataset. **(5 marks)**
2. Using the information from the articles below (and others, if you wish) as well as exploratory data analysis, create a separate dataset containing all variables you think may be theoretically important for predicting receptive vocabulary. **(30 marks)**
    a. Use summary statistics, tables and graphs to examine the data. Which variables seem most important?
    b. Using this information, does there seem to be a need to include random effects in your model?
    c. If so, estimate the intraclass correlation coefficient for your intended model.
3. Using the top-down approach, construct a mixed-effects model from your chosen variables. **(40 marks)**
    a. Build and compare several different mixed-effects models.
    b. For bonus marks, construct a linear model for comparison.
    c. Consider using information criteria, bootstrapping and Kenward Roger simulation to decide which random and fixed effects to retain in the final model.
4. Do there appear to be any problems with your model? If so, how could these be addressed? **(10 marks)**
5. Provide a thorough interpretation of the fixed and random effects in your final model. **(10 marks)**
6. Be sure to annotate your code where appropriate and ensure the output is formatted neatly. **(5 marks)**

Young Lives Articles

https://www.ejhd.org/index.php/ejhd/article/view/1234

https://www.younglives.org.uk/node/7121

https://www.younglives.org.uk/node/8319

(see following page)

Some variables that may be of interest:

- wi – wealth index
- round – round identifier
- agemon – age at interview date
- sex – child's sex (1 = male; 2 = female)
- typesite – urban/rural residence (1 = urban; 2 = rural)
- bmi - calculated BMI=weight/squared(height)
- bwght – child's birth weight
- zhfa - Height-for-age z-score
- ppvtraw - Peabody Picture Vocabulary Test

Young Lives Data Codebook

https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8357#!/documentation

You may want to include several variables in your model not suggested above. If so, you can find information for those variables in the round 5 codebooks for each country. Note that in some cases the variable names in the datasets provided do not match those in the codebooks. The variables in the datasets do have fairly self-explanatory labels, but if you feel you need more information it may be worth searching through the data dictionaries. Also note that several variables included in the data dictionaries are not in the datasets provided. You do not need to worry about these.