**Lab Day x: Model selection**

Today’s lab meeting consists of three exercises, each illustrating the evaluation of theory-based hypotheses. The first exercise does hypotheses in a two-factor confirmatory factor analysis; the second in a latent regression; and the third in a multigroup laten regression.

Practical information:

* All the files for these exercises can be found at the LLL platform.

Make sure to unzip the files. The folder containing these files will be your working directory.

* Solutions to the exercises can be found in the Solutions folder.

We provided R scripts for doing each of the exercises with the R packages lavaan and restrictor**.**

Data:

The examples below will all be applied to the same data set: a simulated data set based

on the Sesame Street data (Stevens, 1996), which is included as the dataset ‘**sesamsim**’ in the

gorica package. Thus, when you install and load the gorica package, the object sesamsim exists, which contains all the data.

Since the data sets contains simulated data, there are no missing values. All the variables are numeric, which does not affect the results in these exercises since there are no grouping variables with more than two level. We need only one grouping variable, namely sex, which consists of two levels. It can be helpful though to make it a factor such that labels can be assigned to the two levels:

sesamesim$sex <- factor(sesamesim$sex, labels = c("boy", "girl"))

The exercises concern the effect of watching one year of the tv-series “Sesame Street” on the knowledge of numbers of N = 240 children aged between 34 to 69 months. Several variables have been measured before and after watching Sesame Street for one year: Knowledge of numbers before (Bn) and after (An) watching, and analogously, knowledge of body parts (Bb and Ab), letters (Bl and Al), forms (Bf and Af), relationships (Br and Ar), and classifications (Bc and Ac). The score ranges on these variables ranges from ‘1 to 20’ to ’1 to 70‘. In the exercises, we will use these variables as well as the following ones: biological age in months (age; score range: 34 to 69), the Peabody test measuring the mental age of children (peabody; score range: 15 to 89), and gender (sex; 1 = boy, 2 = girl).

***Exercise 1: Confirmatory Factor Analysis***

In this exercise, we will inspect the two-factor confirmatory factor analysis (CFA), in which the A(fter) measurements (i.e., Ab, Al, Af, An, Ar, and Ac) load on the factor A, and the B(efore) measurements load on the factor B.

***1a)***

It is a good exercise to draw the model you plan to estimate for yourself.

**1b)**

Specify the two-factor CFA model (such that it can be used in the lavaan function later on).

For ease of specifying the hypotheses (done next), label the parameters; e.g. use for the factor Ab: A1\*Ab

**1c)**

Come up with one or more hypotheses and a failsafe hypothesis.

* On paper, hypotheses are specified in terms of population parameters (often Greek letters). In R, hypotheses are specified in terms of the model parameters used in 1b).
* Make sure that the comparison of factor loadings or comparison of factor loadings to a pre-specified value is fair. Hint: You probably have hypotheses w.r.t. standardized model parameters.
* In the solution file, we evaluate one theory-based hypothesis against its complement. The hypothesis states that each standardized factor loading is larger than .6, meaning that the indicators are strongly related to the factors to which they are assigned.

**1d)**

Run the two-factor CFA specified in 1b).

**1e)**

Use the goric() function in the restriktor package to evaluate the hypotheses from 1c) with the GORICA (type = "gorica").

* In case there is one hypothesis of interest, then your failsafe or competing hypothesis is its complement. Then, use: comparison = "complement"
* In case the hypotheses address standardized model parameters, also include the argument: standardized = TRUE

*Interpret the output. What are your conclusions?*

***Exercise 2: Latent regression***

In this exercise, we will inspect a latent regression model. The factors B and A have the same indicators as in Exercise 1. The difference is the addition of a latent regression in which A is regressed on B, age, and peabody, to investigate whether children’s knowledge after watching Sesame Street for a year is predicted by their knowledge one year before, as well as by their biological and mental age (which have a correlation of .24, so there is no multicollinearity).

Now, repeat the steps in Exercise 1 but now for this model, summarized by:

***2a)***

Draw the model.

**2b)**

Specify the latent regression model, using labels.

**2c)**

Come up with one or more hypotheses and a failsafe hypothesis.

* In the solution file, we evaluate three theory-based hypotheses (H1-H3) on standardized regression coefficients. Since these three do not overlap all the possible theories, we include the unconstrained as failsafe.

H1: specifies that a larger score on B corresponds to a larger score on A (i.e., a positive relation between B and A) and that age and peabody do not predict A.

H2: specifies that the positive relation between B and A is stronger than the positive relation between peabody and A and that age cannot be used to predict A.

H3: specifies that the predictive power of B is larger than that of peabody, which, in turn, is larger than that of age which in turn is positive.

Notably, only in case all these hypotheses are of interest, these should all be included in the set; especially if there is overlap like here.

**2d)**

Run the latent regression model.

**2e)**

Evaluate the hypotheses with the GORICA.

*Interpret the output. What are your conclusions?*

***Exercise 3: Multigroup latent regression***

In this exercise, we will inspect a multi-group regression model, by including the grouping variable gender (sex) in a regression model. This means that there is one regression model for girls and one for boys, where the standardized model parameter estimates may differ between girls and boys.

In the regression, postnumb is regressed on prenumb, to investigate whether children’s knowledge of numbers after watching Sesame Street for a year is predicted by their knowledge of numbers one year before.

Now, repeat the steps in Exercise 1 but now for this model, summarized by:

***3a)***

Draw the model.

**3b)**

Specify the multigroup regression model, using labels.

**3c)**

Come up with one or more hypotheses and a failsafe hypothesis.

* In the solution file, we evaluate one theory-based hypothesis against its complement. The hypothesis states that the standardized relationship between postnumb and prenumb is higher for girls than for boys. In this case, the complement states that the standardized relationship between postnumb and prenumb is lower for girls than for boys.

**3d)**

Run the multigroup regression model.

**3e)**

Evaluate the hypotheses with the GORICA.

Note

*Interpret the output. What are your conclusions?*

**Extra)**

Do the same evaluation but now use the standardized estimates and their covariance matrix as input to the goric function (instead of the lavaan object). E.g.,

goric(fit, H1, comparison = "complement", type = "gorica")

instead of

goric(est, VCOV = vcov, H1, comparison = "complement", type = "gorica ")