1

Robust Hierarchical Regression for Calculating Unique Explained Variance with Single and Multiple Indicator Variables

This document describes methods for executing hierarchical regression in Mplus to calculate incremental explained variance. The approach allows you to use an Mplus robust estimator and to adjust for measurement error. I also describe how to use robust MM regression in limited information structural equation modeling (LISEM) to obtain outlier-resistant estimates of incremental explained variance. I assume you are familiar with Mplus programming.

HIERACHICAL REGRESSION IN SEM

To perform hierarchical regression in Mplus, one must use what are known as phantom variables. **Phantom variables** or **phantom factors** are latent variables that are included in a model for their mathematical convenience rather than for their substantive meaning. They may have indicators or causes, but the indicators and causes are matters of convenience. They usually do not affect model fit, and they are applied purely for mathematical reasons. In older iterations of SEM, phantom variables often were needed to incorporate non-linear constraints within models, but this has become less necessary given programing advances.

A common use of hierarchical regression is to provide perspectives on unique explained variance though the analysis of incremental explained variance when a predictor is added to an equation that already includes a set of predictors. There are different ways one can implement hierarchical regression and semi-part correlation estimation in SEM. I describe a method by de Jong (1999), which uses phantom factors. For descriptions of the logic of using phantom factors for hierarchical analysis, see de Jong (1999). de Jong's strategy is tied to Cholesky decompositions of the covariance matrix of variables, the underlying mathematics of which are complex. My focus here is on articulating basic programming strategy rather than explicating the statistical logic of it. I assume you are familiar with the basics of traditional hierarchical regression and semi-partial correlations.

The advantage of using SEM/Mplus to perform hierarchical regression or to estimate semi-partial correlations relative to more traditional regression is that one can use modern methods for taking into account missing data, such as full information maximum likelihood (FIML). One also can use robust estimation to protect against forms of non-normality and

variance heterogeneity. SEM also can be used to provide perspectives on the biasing effects of measurement error in ways not possible with traditional methods.

In hierarchical regression, one enters variables into the prediction equation in steps. To estimate the unique explained variance of a variable relative to all other predictors in the equation, one enters that variable last in the sequence and documents the increase in the squared multiple correlation that occurs by adding it. To implement hierarchical regression in SEM, I need to define *k* phantom variables for the *k* predictors. Ultimately, I will need to conduct *k* computer runs, one for each predictor whose semi-part correlation I want to isolate. Each phantom factor, which I adopt the general practice of labeling with an F followed by the step number it yields information about, is defined in a way that maps onto a sequential step in the hierarchical analysis.

I will use an example where I focus on three types of social support as predictors of parenting satisfaction. One type of support is emotional support, a second is emotional support, and a third is tangible support. I refer to the measures of these constructs as `emot`, `info`, and `tang` and for the outcome, `satis`. There are three phantom factors in the current example because I have three predictors, emotional support, informational support, and tangible support. To illustrate the logic, I decide to enter `emot` at the first step, `info` at the second step, and `tang` at the final step. The key is not so much the order of entry of the first two variables but rather the entry of `tang` lastly. The last entered variable is the one I target for purposes of calculating its unique explained variance over and above the other predictors. The phantom factors are F1, F2 and F3, representing steps 1, 2 and 3 in the hierarchical analysis.

I need to create a "measurement model" for each phantom factor. I will define the observed "indicators" for the phantom factors as follows:

The latent factor F1 influences the indicators `emot`, `info`, `tang`
The latent factor F2 influences the indicators `info`, `tang`
The latent factor F3 influences the indicator `tang`

F1 ultimately will provide information about step 1 in the hierarchical sequence and has as indicators all predictor variables in the equation, ordered by the step in which they will be entered into the equation, in this case `emot`, `info`, and `tang`. F2 provides information about step 2 in the hierarchical sequence and is assumed to influence the last k-1 predictors/indicators listed in F1, in this case `info` and `tang`; F3 provides information about step 3 and is assumed to influence the last k-2 predictors/indicators listed for F1, in this case `tang`. Thus, all predictors are listed for F1 and then you drop a predictor at each subsequent step until you are left with the variable that you want to calculate the squared semi-part

correlation (unique explained variance) for. If there was a fourth predictor, F1 would influence all 4 predictors/indicators, F2 would influence all predictors/indicators except the first one listed in F1, F3 would influence all the predictors/indicators except the first two listed in F1, and F4 would influence all the predictors/indicators except the first 3.

Figure 1 presents the model I will be working with. The correlations between the factors are fixed to 0 and the variance of each factor is fixed to be 1.0. The error variances of the observed "indicators" (i.e., emot, info, tang) are fixed to zero, but not for satis, because it is the outcome variable. The factor means are fixed at 0 and the measurement intercepts are estimated but are not of substantive interest. All of these mathematical manipulations make the phantom variables behave in the desired way (see deJong, 1999, for elaboration).

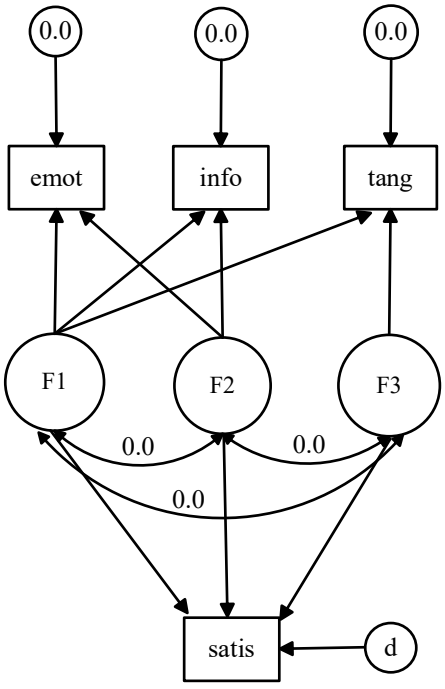

FIGURE 1. Logic model for income intervention

The parameters of interest are the standardized coefficients from the regression of `satis` on F1, F2 and F3. Here are the relevant standardized coefficients from Mplus output:

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 SATIS1   ON
    F1               0.371      0.029     12.645      0.000
    F2               0.184      0.032      5.706      0.000
    F3               0.274      0.030      9.144      0.000
```

Recall the a priori order of entry I specified was `emot` at step 1, then `info` at step2, then `tang` at step 3. The standardized coefficient for F1 (under the column called `Estimate`) is the zero-order correlation for `emot` and `satis`, which is 0.371 because `emot` is entered at the first step. The column labeled `S.E.` is the estimated standard error for the coefficient and the column labeled `Est./S.E.` is analogous to a z test for statistical significance, i.e., it is the critical ratio. The column labeled `Two-Tailed P-Value` is the p value associated with the critical ratio. In this case, the coefficient is statistically significant.

The coefficient for F2 is the semi-part correlation for the variable added at step 2 (`info`) partialling out all the predictors at the prior steps (`emot`). It tests if adding `info` to the equation at step 2 (where `emot` was entered at the prior step) results in statistically significant explained incremental variance. It does so in this case because the critical ratio for the test of the semi-partial correlation is 5.706, $p < 0.05$. The coefficient for F3 is the semi-part correlation for the variable added at step 3 (`tang`) partialling out all the predictors at the prior steps (`emot`, `info`). The test for this incremental explained variance also is statistically significant (critical ratio = 9.14, $p < 0.05$). If I square each of the semi-part correlations in the `Estimate` column, they reflect the proportions of incremental explained variance at each step. The proportion of unique explained variance for `tang` is the semi-part correlation squared, which is (0.27)(0.27) = 0.07. I can repeat the analyses but moving `emot` or `info` to the last position to isolate the semi-part correlation for them. The R square when all predictors are included in the equation is 0.246.

It is straightforward to extend the approach to cases where the predictors represent latent variables with multiple indicators; one simply constructs the phantom variables as underlying the latent predictors rather than the observed predictors. One also can adjust for measurement error for single indicator models using the strategy discussed in the document on the resources tab of my webpage for Chapter 3 on addressing measurement error for single indicator SEM models. The robust estimation strategies in Mplus (e.g., MLR) are not outlier-resistant. To deal with this, one can either use MM regression in a LISEM context as discussed below or one can identify outliers using the *robust outlier analysis* program on

my website and then test for the robustness of results of the above analysis with and without those outliers in the analysis.

OUTLIER-RESISTANT MM REGRESSION AND UNIQUE EXPLAINED VARIANCE

Another option to deal with non-normality, variance heterogeneity, outliers and leverages is to use the MM regression program on my website. The program yields an analog robust R squared. You conduct the regression analysis first using the prediction equation without the target predictor and then again adding the target predictor to the equation. The difference between the two squared Rs will index the incremental explained variance of the target predictor. You will not obtain a significance test for the additional explained variance but you can set an a priori value as a cutoff for meaningfulness, such as an incremental squared R that corresponds to Cohen's medium effect size, which would be a squared R increment of about 0.05 or 0.06 or whatever seems substantively appropriate.

REFERENCES

de Jong, P. F. (1999). Hierarchical regression analysis in structural equation modeling. *Structural Equation Modeling*, 6, 198-211.