Notes on Error Structures in Multilevel and Growth Curve Models

Nested Designs

With nested data, the standard assumption of homogeneity of variance (or constant variance) is that the within-group variance estimate, σ^2, is constant across groups or values of a predictor. The violation of this assumption is heterogeneity of variance (or nonconstant variance). A simple case in which this is violated is when a binary predictor representing groups, such as treatment and control or public and private, do not have the same estimates of residual variance. This heterogeneity may be a sign of a model with inadequacies, such as an omitted variable that if included would account for the differences in variance. For some software (e.g., SPSS, R, HLM, SAS), the researcher can relax the homogeneity assumption with sufficient data, but they differ in their level of flexibility. These types of analyses require additional degrees of freedom and in some circumstances (e.g., low reliability, highly variable group sizes, many very small group sizes, other ill-behaved data), estimation difficulties may arise, such as failure of the model to converge.

SPSS. For nested data, heterogeneity of within-group variances across groups can be estimated by using DIAG instead of UN. SPSS, however, does not have any built-in features that would allow different error variances within levels of a grouping variable, such as treatment and control. In the examples below, I obtain separate within-group variance estimates for male and female students in the HSB data with `mathach` as the outcome and `ses` as the predictor. Weaver and Black (2015) present a work around that exploits the REPEATED subcommand usually used for repeated measures. For any binary variable (I will use the variable `female` from the HSB data), a second variable is be computed which reverses the codes. The variable `id` below is an index variable for the student (which I created for this analysis).

```
MIXED mathach BY id WITH ses female male
  /METHOD = REML
  /PRINT = SOLUTION TESTCOV
  /FIXED =  ses | SSTYPE(3)
  /CRITERIA=DFMETHOD(SATTERTHWAITE)
  /RANDOM = INTERCEPT | SUBJECT(schoolid) COVTYPE(ID)
 /RANDOM = INTERCEPT male | SUBJECT(id) COVTYPE(DIAG)
  /REPEATED = male | SUBJECT(id*schoolid) COVTYPE(DIAG).
```

R. The lme4 package does not currently have features that allow heterogeneous variances. So, the easiest way to estimate heterogeneous variances is to include the `VarIdent` function using the `nlme` package. Here, the model estimates different within-group variances for males and females:

```
model <- lme(mathach ~ ses, random = ~ 1|schoolid, data = mydata, weights = varIdent(form =~ 1 | female),
method="REML")
```

The `anova` function could then be used to compare the two nested models. You should expect the fit of the model to be better when heterogeneity is allowed.

HLM. HLM has built in features that allow heterogeneity across a grouping variable. Using MATHACH as the dependent variable and SES (uncentered) as the predictor variable. To estimate this model in HLM, go to **Other Settings** -> **Estimation Settings** and check the **Heterogeneous sigma^2** button. Then, double click on FEMALE in the "Possible Choices" box to move over the variable you wish to stratify variances by. Check "ok", save, and run your model. The output gives both homogeneous and heterogeneous solutions. The latter output provides separate estimates and tests for the within-group variance by male and female. HLM also automatically prints a nested test of the two models. In this case, the results indicated a significant difference in the within-group variance for males in females because the model fit was significantly improved by allowing heterogeneous variances, $\chi^2(1) = 11.48560$, $p < 0.001$.

These models examine within-group variance heterogeneity across two or more values of an independent variable, but more general heterogeneity of variances models have also been discussed, using a two-step approach (Kasim & Raudenbush, 1998), a log-linear approach (Hedeker, Mermelstein, & Demirtas, 2008; Lee & Nelder, 2006), or quasi-likelihood/pseudo-likelihood approach (Lin, Raz, & Harlow, 1997). Leckie and colleagues review these approaches, demonstrate with a simulation that ignoring variability across Level-2 units can lead to biased results, and illustrate several software approaches (Leckie, French, Charlton, & Browne, 2014).

Longitudinal Designs

Although heterogeneity can be conceived of similarly in the longitudinal context, discussion of heterogeneity and covariance structures typically revolve around a somewhat different conceptualization for growth curve models. Most commonly, discussion of heterogeneity for growth curve models concerns different residual variances at each time point, where this conceptualization represents a special case of the within-group residual variances differing across levels of the independent variable, in this case the time variable. For example, residual variance may be larger at later time points than at earlier time points. This type of pattern would imply the model fits better or better predicts y_{ti} at earlier time points than later time points. The cause of such a pattern may very well be the omission of a predictor, such as a time-varying covariate, and its inclusion might resolve the problem. It is also the case that heterogeneous variance of this sort may be difficult to distinguish from a need to model a nonlinear effect (Bauer & Cai, 2009). Imagine an increasing trajectory over time which curves (decelerates) at later time points. Fitting a linear model would result in larger residual variance at the later time points, because the linear slope does not account for the data as well at those time points.

SPSS. Growth curve models can be estimated to allow residual variances to be unequal over time. In R and HLM, the same procedure as described above for nested models, where the time variable is the predictor by which the residual variance is allowed to vary. This pattern of variances can also be described in matrix terms. The diagonal of the variance-covariance matrix contains the variances. So a "diagonal matrix" is one in which the variances are estimated and allowed to differ across time points (and all of the off-diagonal elements, or covariances, are zero). So, in SPSS, to specify variances that are allowed to vary across time points, the diagonal pattern is indicated on the `REPEATED` subcommand: `/REPEATED = time | SUBJECT(id) COVTYPE(DIAG)`. Note that this is not the same as indicating `DIAG` on the `/RANDOM` command line, which serves a different purpose (i.e., variances and no covariances among intercepts and slopes).

A fairly common modification is to allow covariances among residuals over time. Such a variance structure involves some autocorrelation in which an association exists among the dependent variable (residual) values after accounting for the time predictor and other predictors in the model. One possibility is to assume a "lag-1" or "AR1" autocorrelation pattern which implies the highest correlation among the adjacent time points and decreasing correlations among higher lags following a specific mathematical pattern, using `COVTYPE(AR1)` on the `REPEATED` subcommand. The covariance patterns can be of other forms too, such as unstructured, which allows residuals covariances to be freely estimated at any value across any lag, using `COVTYPE(UN)`,[1] or toplitz using `COVTYPE(TP)` which estimates one value for lag-1 autocorrelations, another value for lag-2 autocorrelations, and so on. Autocorrelations among errors may obscure nonlinear effects (Bianconcini, 2012; Voelkle, 2008), so careful analysis of the data is needed before incorporating them.

R.

The `lme4` package does not allow for heterogeneous variances over time, but this can be accomplished with `nlme`.

```
model2 = lme(depress ~ time, random = ~ 1 + time | rid,  data = mydata,
```

[1] This use of UN on the REPEATED subcommand is again a different function than on the RANDOM command.

```
            weights = varIdent(form = ~1|time))
```

HLM. HLM has built in features that allow heterogeneity across a time points. You need to set up the data file using the repeated measures construction (then the output uses π and β for the within and between level equations). Then, to estimate a growth curve model with heterogenous residuals over time, go to **Other Settings** -> **Estimation Settings** and check the **Heterogeneous sigma^2** button. Then, double click on the time variable (e.g., TIME) in the "Possible Choices" box to move over the variable.

References

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*, 97-114.

Bianconcini, S. (2012). Nonlinear and quasi-simplex patterns in latent growth models. *Multivariate Behavioral Research, 47*, 88–114.

Voelkle, M. C. (2008). Reconsidering the use of autoregressive latent trajectory (ALT) models. *Multivariate Behavioral Research, 43*, 564–591.