1. Two-level formulation with one upper (z) and one lower (x) IV:

\[ \beta_{0j} = \gamma_{00} + \gamma_{01}z_{ij} + U_{0j} \quad \beta_{1j} = \gamma_{10} + \gamma_{11}z_{ij} + U_{1j} \]

\[ Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + R_{ij} \]

2. General formulation for \( Y_{ij} \):

\[ Y_{ij} = \gamma_{00} + \sum_{h=1}^{p} \gamma_{h0}x_{hij} + \sum_{k=1}^{q} \gamma_{0k}z_{kj} + \sum_{k=1}^{q} \sum_{h=1}^{p} \gamma_{hk}z_{kj}x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj}x_{hij} + R_{ij} \]

[Int] [Indiv Fix Eff] [Grp Fix Eff] [Ind/Grp Fixed I/A] [RI] [Rand Slopes] [Err]

Var(\( U_{ij} \))=\( \tau_{h}^{2} \) \quad Cov(\( U_{ij}, U_{ij} \))=\( \tau_{h}^{2} \) \quad Cov(\( U_{ij}, U_{ij}^{\prime} \))=0 \quad Var(\( R_{ij} \))=\( \sigma^{2} \) \quad Cov(\( U_{ij}, R_{ij} \))=0

3. Model Testing

a. t-tests for fixed effects parameters: For a single parameter, say \( \beta_{k} \), test whether it is significantly different from 0 using \( t = \frac{\hat{\beta}_{k}}{SE(\hat{\beta}_{k})} \). This t-statistic will approximately follow a t-distribution with \( M-r-1 \) df for a level 1 variable, where \( M \) is the total number of measurements and \( r \) is the total number of explanatory variables in the model. For a level 2 variable we have \( N-q-1 \) df, where \( N \) is the number of groups and \( q \) is the number of level 2 variables. Don’t use this test for including/excluding categorical explanatory variables with more than \( c=2 \) levels; instead use the “Wald” \( \chi^2 \) test (see below). Although these tests are “exact” for a single level model, they are only approximate for a multilevel model.

There are several available methods to “correct” the denominator df, to make the approximate test closer to correct. Other than the default these are specified as an option after the “/” in the MODEL statement in the form “DDFM=someMethod”. The “default” is chosen automatically based on the kind of model, but may not be best. For a model with a RANDOM statement but no REPEATED statement, the “containment” method is used: df are \( N-q-1 \) for explanatory variables (other than the intercept) if they are mentioned in the RANDOM statement.

The between/within method (DDFM=BW) looks to see which explanatory variables are unchanging within a group, and assuming these are level 2 variables, it uses \( N-q-1 \) df for these. Interactions between level 1 and level 2 variables use \( M-r-1 \) df.
The Satterthwaite method (DDFM=SAT), also called Fai-Cornelius, uses spectral decomposition of the variance/covariance matrix to approximate the denominator df. The Kenwood/Roger method (DDFM=kR) uses a complex multi-step process and also adjusts the standard errors.

With balanced designs (same number of level 1 observations per level 2 group) and with no REPEATED statement, containment or BW are often good. In general, the KR method appears best, based on simulation studies.

See DF.txt for a comparison on the HW2 data. Pay more attention to df than p-values, because we already dropped variables that did not look “statistically significant”.

b. t-tests for covariance parameters: These show up in the “Covariance Parameter Estimates” as Standard Error, Z and Pr(Z), but only if the global option COVTEST is used. These are only reliable when the number of groups is large (but it’s hard to say how large is large enough) and when the estimate is not too close to zero.

c. Wald $\chi^2$ tests: This is a single test that is the square of the t-test for testing a single parameter (usually $=0$ vs. not). In PROC MIXED it is provided for each fixed effects parameter in the “Type-III Tests of Fixed Effects” table. It’s main use is that it gives a single p-value for the any c-level categorical variable, so is good for choosing whether to include vs. exclude any categorical variable.

d. Likelihood ratio (deviance) tests: These can be used to compare any two nested models. A model is nested in another (larger) model if it can be expressed as the same as the larger model but with one or more parameters forced to be zero (or rarely another fixed value). Therefore LR tests work for testing multiple parameters at once, including the set of c-1 indicator variables defining a c-level categorical variable. LR tests have few assumptions, but are only correct asymptotically (for large enough samples). The test is to compare $2*(\text{difference in natural log of the likelihoods})$, which equals the difference in “deviance”, to a chi-square distribution with $m_1-m_0$ df, where $m_1$ is the number of parameters in the larger model, and $m_0$ for the smaller. Be sure to use the same method for both models which should be REML to test covariance parameters or ML for fixed effects.

e. AIC and BIC: These are penalized likelihoods that only are used to compare two models (i.e., they have no meaning for a single model). The models need not be nested. The likelihoods both need to come from ML or both from REML, and you should use REML for random effects and ML for fixed effects. AIC penalizes the likelihood only by the number of parameters, while BIC also incorporates sample size. There are some unresolved issues based on what sample sizes to use for
level 1 vs. level 2 variables. BIC tends to chose more parsimonious models than AIC, especially with larger sample sizes. Like the LR test, AIC and BIC will be unreliable for small sample sizes. As usual, either AIC or BIC suffer from type 1 and type 2 error when choosing the covariance structure, and any errors here adversely affect the fixed effects estimates. AIC may have some advantage for this task. For fixed effects, AIC tends to choose a similar or “larger” model than BIC. If your goal is to make predictions the tradeoff of bias and variance tends to favor AIC. If your goal is to find which IVs truly affect the DV, BIC tends to be better.

4. **Model specification principles and concerns**
   a. Subject matter concerns are primary. Keep theoretically important variables in the model.
   b. If you have lots of subjects/groups and not too many explanatory variables, consider testing the effects of the main treatment variable with all other important variables in the model.
   c. Keep fixed effects, even if not statistically significant, when including the corresponding random effects. Ditto for main effects when interactions are included.
   d. Be careful not to compare models with different numbers of measurements, e.g., when adding a variable with missing data which can cause observations to be dropped.
   e. Consider making upper level mean variables from lower level variables if it make scientific sense.
   f. When the random intercept or the random slope is “large” look for level 2 variables that will explain away some of this random variation. But don’t let a non-significant $\tau^2$ value dissuade you from looking for upper level explanatory variables. When $\sigma^2$ is large, look for level 1 variables that will explain away some of this random variation.
   g. Generally include the covariances of random effects (TYPE=UN).
   h. Avoid over-fitting / lack of parsimony, but worry about type-2 error here.
   i. Be concerned that complex model fitting is analogous to multiple comparison testing. (Be less concerned when the goal is to test a treatment effect and the model fitting is directed elsewhere.)
j. Be aware of the effects of correlated IVs. These cause increase standard errors and each may mask the significance of the others. Also the usual meaning of each $\beta$, which is the effect of a one-unit change in an IV on the average DV holding all other IVs constant, is lost when the correlation means that you can’t change one IV without changing some others.

5. (Modified) Snijders and Bosker approach to model selection
   a. First let random effects account for the level 2 variables (RANDOM statement), and choose good level 1 (x) variables, using forward or backward selection. I recommend BIC with ML, being careful of changing numbers of observations due to different missing data patterns across variables.
   b. Then switch to REML and check if some random effects can be dropped (using AIC rather than Covariance Wald (t) tests.)
   c. Then add in level 2 (z) variables as needed (using BIC and ML again), probably using forward or forward/backward selection.
   d. Then use AIC and REML to drop unneeded random effects.
   e. Then check residual plots.

6. Explained variance
   a. $R^2$ in single level models reflects the fractional drop in variance of prediction when using the explanatory variables vs. when ignoring them.
   b. Level 1 $R^2$: S&B define $R_1^2$ for random intercept models as the fractional drop in variance of prediction of individual values from a model with various explanatory variables and a random intercept to a model with just a random intercept. The prediction variance for both models is the REML estimate of $\sigma^2 + \tau_0^2$, and is (at least in the population) larger for the smaller model. Define $R_1^2$ as 1 minus the ratio (large model over small model) of the estimate prediction variances.
      This represents how much better we can predict the true value of a randomly chosen level 1 measurement when we know the values of the explanatory variables vs. when don’t know them.
   c. Level 2 $R^2$: S&B define $R_2^2$ for random intercept models as the fractional drop in variance of prediction of the group mean from a model with various explanatory variables and a random intercept to a model with just a random intercept. Here the prediction variances are $\sigma^2/n + \tau_0^2$ where n is a group size, ideally a typical one.
   d. S&B recommend approximate $R^2$ values for random slope model be calculated as the $R^2$ values for the corresponding random intercept only models, unless special software is available.