Significance Testing in Multilevel Regression

As with ordinary least squares regression or logistic regression, we can consider significance tests for individual estimates, such as intercepts, slopes, as well as whether the full model accounts for a significant amount of variance in the dependent variable. In between, there is also the possibility of determining whether of subset of predictors contribute significantly. Aside from these fixed effects, we also can test the variance components or random effects (variance of intercepts, variance of slopes, or covariances among them) for significance. Unfortunately, there are several considerations for testing either fixed or random effects that make this an all too complicated topic.

Significance Testing for Fixed Effects

The fixed effects in multilevel regression are typically tested in a familiar way, by creating a ratio of the intercept or slope estimate to the estimate of the standard error. The usual null hypothesis test is whether the coefficient, either intercept or slope, is significantly different from zero (i.e., is the population value zero or not). This kind of ratio, usually assumed to be distributed as a z or t, is used in many statistical tests (referred to as a "Wald" ratio).

Raudenbush and Bryk (2002), and therefore the HLM software, use a *t*-distribution to evaluate this ratio (Fotiu, 1989).

$$t = \frac{\hat{\gamma}_h}{S.E.\left(\hat{\gamma}_h\right)}$$

where γ_h is either the intercept or slope coefficient and $S.E.(\gamma_h)$ is the standard error estimate.[1] The fixed effects hypothesis tests (whether for level-1 or level-2 predictors) used by the HLM software use a degrees of freedom based on the number of level-2 units (i.e., number of groups).

$$df = N - q - 1,$$

in which N refers to the number of groups and q is the number of predictors in the model.[2] In practice, this formula does not seem to be used precisely, and with different runs you will notice somewhat different degrees of freedom listed in the output under "approximate df." In addition, the test of the effects of cross-level interactions often use degrees of freedom based on the number of level-1 units (i.e., total number of individuals in the sample).

In SPSS and R (lme4 by default), the t-test for fixed effects uses an "approximate" or Satterthwaite degrees of freedom (Satterthwaite, 1946; Welch, 1947), appearing in the output with decimal values rather than whole numbers, and are based on the number of individuals in the data set rather than the number of groups. Satterthwaite degrees of freedom are a way of proportionally adjusting the df to provide a more accurate p-value estimate from the family of distributions. The t-test used by SPSS or R and the t-test used by HLM are essentially asymptotically equivalent—with a large number of groups, results will be highly similar. With a small number of groups (e.g., perhaps less than 100; Maas & Hox, 2005), the t-test in HLM will be a more conservative test than the t-test in other software, which may be preferable when there is a smaller number of groups (Fotui, 1989). Manor and Zucker (2004) provide a review and comparison of several of the fixed effects tests and suggest that the Satterthwaite approach works well in most instances. For very small number of groups, (e.g., < 40) use of a Bartlett-correction to the likelihood ratio test will improve Type I error rates. The fixed effects tests are also potentially sensitive to distributional assumptions about the random effects, and robust estimates are sometimes recommended to adjust the standard errors (Raudenbush & Bryk, 2002). There will be more on this topic later.

[1] I have not provided the formula for the standard error, but it is printed in Raudenbush and Bryk (2002) on pages 48 (empty model) and 56 (general formula).

[2] Notation used by Raudenbush and Bryk for the degrees of freedom is $df = J - p - 1$, in which *J* refers to the number of groups and *p* is the number of predictors in the model

Significance Testing for Random Effects

Overview. Individual random effects tests examine hypotheses about whether the variance for each random intercept or slope (and their covariances) are significantly different from zero. Software packages print these estimates under the "random effects" or "covariance tests" portion of the output. Random effects tests are often of theoretical importance to researchers, and, thus, are typically given as much importance as the fixed effects tests. The tests in most software programs (SPSS, SAS, MLWin) use a similar Wald z-test, whereas chi-square test based on a different approach is used in the HLM program. These Wald tests are not always optimal, so other methods are preferred, particularly for small sample sizes. The R packages do not provide significance tests of random effects (probably for this reason), but confidence intervals can be obtained. Likelihood ratio tests are also possible but are difficult or impossible to implement for random slopes without also testing the covariances simultaneously.

Wald test. The Wald random effects tests used by most programs are simply a ratio of the variance estimate divided by its standard error estimate. With large sample sizes, these tests are unlikely to lead to different conclusions that other methods, but with small samples they can be problematic (Snijders & Bosker, 2012). In SPSS, one important precaution is that the significance tests for the intercept or slope variances (but not the covariances) should be interpreted after dividing the p-value from the output in half (i.e., as a one-tailed test; LaHuis & Ferguson, 2009; Snijders & Bosker, 2012, p. 98), following the rationale used for other variance tests (Miller, 1977; Self & Liang, 1987).[3] The Wald test can be improved (Manor & Zucker, 2004) for small number of groups (N or J, depending on the notation) by using Satterthwaite (Satterthwaite, 1946) degrees of freedom (aka Fai-Cornelius degrees of freedom) or the Kenward-Roger (Kenward & Roger, 1997) adjustment (Bell, 2013a; 2013b; McNeish & Stapleton, 2016; see Hox, Moerbeek, & Van de Shoot, 2018 for a discussion).[4] The likelihood ratio test described below or the chi-square approach in HLM are generally preferable approaches to tests of random effects, however, particularly when there are fewer groups.

The chi-square test. The chi-square test used in the HLM package is based on the deviation of group means from the grand mean, given in Raudenbush and Bryk (2002, p.64) as:

$$\chi^2 = \frac{\sum_j \left(\hat{\beta}_{qj} - \hat{\gamma}_{q0} - \sum_{s=1}^{s_q} \hat{\gamma}_{qs} W_{sj} \right)^2}{\hat{V}_{qqj}}.$$

In the above formula, β is the group estimate (intercept or slope), γ is the average estimate (grand mean or average slope), and W is a predictor. The numerator in the equation represents the sum of squared deviations from the average value adjusting for the predictors in the model. The denominator, V_{qqj}, is a variance error estimate (i.e., square of the standard error). Degrees of freedom for this test are $J - S_q - 1$, where J is the number of groups and S_q is the number of predictors in the model (in Snijders & Bosker, 2012, this is $N - q - 1$). Small groups are omitted from the computations (the number omitted is noted in the HLM output).

Recommendations. The Wald variance tests from SPSS (provided the p-values are halved when the Wald variance test is used; Berkhoff & Snijders, 2001), SAS, and the HLM approaches generally give very similar results with sufficient number of groups (perhaps > 100; Hox, 2012) using the default REML estimates. They will also converge with the likelihood ratio test with a large sample size (number of groups), but they may be generally lacking in power when there are fewer groups (Harwell, 1997; Sánchez-Meca & Marín-Martínez, 1997). I have not found a Wald test in any multilevel-related R

[3] With the Mixed procedure in SPSS, at least through version 25, p-values for variance tests in the output should be halved to determine significance (covariance tests do not require halving). SAS 8 and higher uses one-tailed p-values for the variance but not the covariance, so no action is required by the user. HLM uses a different test and no alteration of p-values is needed.

[4] This approach is available in SAS using the DDFM option on the model line, DDFM=SATTERTHWAITE. The option DDFM=KENWARDROGER is a relatively new approach and may provide another promising option for small N circumstances (Luke, 2017).

package (see the likelihood ratio test section below, however), but confidence intervals in the nlme package using a 90% confidence interval can be estimated separately for the random effects (i.e., intercept and slope variance; but use the standard 95% CI for the intercept-slope correlation). The confidence limits for nlme and SPSS are the same in all of the examples I have seen. The lme4 package also provide confidence intervals using the profile likelihood method are available in the lme4 package using the confint() function. The profile likelihood (Bates & DebRoy, 2004) is an iterative method that does not assume a symmetric distribution for the random effect that should perform better than standard Wald tests of random effects. This method as well as several other more recent methods (including MCMC and bootstrapping) are promising and likely preferable to the uncorrected standard Wald test when there are few number groups, but at this date these alternatives do not seem to have received much evaluation in comparative simulations at this point (but see Lahuis & Ferguson, 2009 for a comparison of a few approaches). Finally, there is an alternative approach to the above mentioned traditional hypothesis testing approaches. The Bayes factors approach (Kass & Raftery, 1995) attempts to compare the relative likelihood of two hypotheses rather than compare an obtained sample coefficient to a null hypothesis value. Mplus and MLWin provide a Bayes factors testing approach to variance estimates. The approach can potentially do better than significance testing, but the Bayesian approach requires a number of decisions and assumptions by the user that also may lead to incorrect conclusions or capitalization on chance (Konijn, van de Schoot, Winter and Ferguson, 2015)

Likelihood Ratio Tests for Single or Multiple Parameters, Fixed, Random, or Both

Another approach to significance tests involves a comparison of two “nested models” in the likelihood ratio or "deviance" test. Nested model tests involve comparison of one model to another model that specifies only a subset of the parameters included in the first model (provided the same set of cases are used in both models). Fixed, random, or a combination of fixed and random effects can be tested with this approach. The likelihood ratio test compares the deviance (-2 log likelihood) of two models (see the estimation handout for more information on deviance) by subtracting the smaller deviance (model with more parameters) from the larger deviance (model with larger deviance).[5] A basic comparison might be between the empty model (Snijders & Bosker, 2012, denote the first model with the larger deviance as D_0) and a model with a predictor added (denoted as D_1) with a fixed effect but not a random slope, in which case the model with the smaller deviance (better fit) is subtracted from the larger, $D_0 - D_1$. The difference is a chi-square test with the number of degrees of freedom equal to the number of different parameters in the two models (i.e., $df = 1$ because only one parameter differed in the two models). This would be a test of the fixed effect for the slope and would be testing the same hypothesis as the Wald test described above in which the coefficient is divided by its standard error. The test of a single variance parameter using the likelihood ratio test is asymptotically equivalent to the Wald variance test (p-values should be halved in either case). “Asymptotically” means that with a large number of groups these tests will yield very similar results.

Alternatively, one could compare two models that differ only in the random effects. For example, if one model constrains an intercept variance to be non-varying across groups, it can be compared to a model in which the intercept is allowed to vary. The difference in likelihoods or deviances is again a chi-square, in this instance with $df = 1$ because only one parameter changed. For variance tests, significance should be determined as a one-tailed test as the variance cannot be negative. The one-tailed test seems to produce a good balance in Type I and Type II errors in this case (Snijders & Bosker, 2012; Lahuis & Ferguson, 2009). If only a covariance is tested, a two-tailed test should be used, because the covariance can be negative or positive (i.e., do not adjust the p-value for the intercept-slope covariance test).

Consider another example in which a model with a random effect for the slope (i.e, is the slope is allowed to vary) is compared to a model without the random effect for the slope (i.e., the variance of the slope is constrained). This example would appear to be testing a single parameter, but, in fact, the two models differ by two parameters. The first model will include an estimate of the slope variance, τ^2_1, but also an

[5] I will be covering maximum likelihood estimation and the negative log likelihood values in a subsequent lecture.

estimate of the covariance between the slope and the intercept, τ_{10}, by default. The covariance cannot be estimated when the slope is constrained to be non-varying, however. One would ordinarily expect that the difference between the two models would be compared to the chi-square distribution with $df = 2$, because two parameters differed between the models being compared. But because variance tests should use a one-tailed test and covariance tests are two-tailed tests, a more complicated significance criterion is needed. Snijders and Bosker (2012, p. 99) recommend using a "mixture distribution" (or "chi-bar distribution") by comparing the chi-square difference obtained from subtracting $D_0 - D_1$ to a combination of two critical values. For $\alpha = .05$, the critical values are: one slope $\chi^2_{mix} = 5.14$, two slopes $\chi^2_{mix} = 7.05$, and three slopes $\chi^2_{mix} = 8.76$. The HLM package provides a test of these "multivariate" or "multiparameter" tests preprogrammed, and, though not documented, based on results from a few models, the multiparameter tests in HLM do not seem require any adjustment to the p-value. Any number of parameters can be compared in the two models, of course. In the case where the empty model is compared to a full model, the likelihood ratio test provides information about whether the predictors in the model and the added random effects together account for a significant amount of variance in the dependent variable.

The standard likelihood ratio test in R can be obtained with the `anova(model1,mode2)` function with no mixture adjustment, but the `rand()` function from the `lmerTest` package will provide the appropriate mixture chi-square test (West, Welch, & Galecki, 2014).

An important precaution for likelihood ratio tests in multilevel regression is that whenever the two models compared involve any difference in fixed effects (e.g., a model with a predictor with random slope compared with an empty model without the predictor), the models need to be tested with a full maximum likelihood estimator (FIML) rather than the default restricted maximum likelihood estimator (REML; the estimation handout provides more detail on the distinction). If the difference in the two models involves only a difference in random effects, deviances can be used from the REML estimator.

References

Bell, B., Ene, M., Smiley, W., & Schoeneberger, J. (2013). A multilevel primer using SAS Proc Mixed, SAS Global Forum.

Bell, Schoeneberger, Smiley, Ene, and Leighton (2013). Doubly diminishing returns: an empirical investigation on the impact of sample size and predictor prevalence on point and interval estimates in two-level linear models. Paper presented at the Modern Modeling Methods Conference (M3). Storrs, CT.

Berkhof, J., & Snijders, T. A. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics, 26*, 133-152.

Harwell, M. (1997). An empirical study of Hedges' homogeneity test. *Psychological Methods, 2*, 219–231.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). Multilevel analysis: Techniques and applications (3rd ed.). New York, NY: Routledge.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773–795.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983-997.

Konijn, E., van de Schoot, R., Winter, S., & Ferguson, C.J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. Communication Methods and Measures, 9, 280–302.

Hox, Joop J.. Multilevel Analysis (Quantitative Methodology Series) (p. 332). Taylor and Francis. Kindle Edition.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*(3), 418-435.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. Behavior research methods, 49(4), 1494-1502.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295-314.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics bulletin, 2(6), 110-114.

Snijders, T.A.B., & Bosker, R.J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd Edition). London: Sage

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta analysis: A Monte Carlo comparison of statistical power and type I error. *Quality and Quantity, 31*, 385–399.

Hox, Joop J.. Multilevel Analysis (Quantitative Methodology Series) (p. 338). Taylor and Francis. Kindle Edition.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.