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 or 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 can be assumed to be distributed as a z or t. When the z distribution is used (often referred to as a "Wald" test), the test assumes a large enough sample that the test behaves regularly but with smaller sample sizes the z-test approach will suffer from inflated Type I errors (van der Leeden, Busing & Meijer. 1997). A safer approach with a smaller number of groups is to use the Student's t distribution. With a large sample size (e.g., over 120), it does not matter which test is used, because the use of the Wald z and t-test will give similar p-values (i.e., they are asymptotically equivalent).

Raudenbush and Bryk (2002), and therefore the HLM software, use a *t*-distribution to evaluate the ratio of the regression coefficient to the standard error (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] With different runs you may notice somewhat different degrees of freedom listed in the output under "approximate df." Note that in the HLM package, the test of the effects of cross-level interactions uses degrees of freedom based on the number of level-1 units (i.e., total number of individuals in the sample) rather than the number of level-2 units.

SPSS and R (`nlme` and `lme4`) also use t-tests for fixed effects. SPSS and the `lme4` package in R use Satterthwaite degrees of freedom (Satterthwaite, 1946; Welch, 1947) which appear in the output under "df" or "approximate df" with decimal values rather than whole numbers. The Satterthwaite degrees of freedom adjust the degrees of freedom based on the number of individuals in the data set rather than the number of groups. Satterthwaite degrees of freedom (sometimes "Fai-Cornelius" degrees of freedom) are a way of proportionally adjusting the df to provide a more accurate p-value estimate from the family of t distributions. The Satterthwaite approach is an improvement over traditional z or t-test (Manor & Zucker, 2004) for small number of groups (N or J, depending on the notation) but the Kenward-Roger (Kenward & Roger, 1997) adjustment, which corrects standard errors as well as degrees of freedom,

[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

offers further improvement with small sample sizes (Bell, 2013a; 2013b; Luke, 2017; McNeish & Stapleton, 2016; see McNeish, 2017, and Hox, Moerbeek, & Van de Shoot, 2018, Chapter 3, for good overview discussions).[3] With larger number of groups (e.g., > 50; McNeish, 2017; Snijders & Bosker, 2012), HLM or the Satterthwaite and Kenward-Roger *t*-test approaches will likely all have adequate control of Type I errors. With a small number of groups (e.g., < 50 but particularly 25 or fewer; McNeish, 2017), the *t*-test in HLM (Fotui, 1989) and the Satterthwaite and Kenward-Roger will be a more conservative tests with preferrable Type I error rates than *z*-tests (such as those used by Mplus) or unadjusted *t*-tests (like those used in `nlme` in R by default). Simulation studies suggest that Satterthwaite corrections have better Type I error rates (Manor & Zucker, 2004) with a small number of groups and the Kenward-Roger corrections provide even better control of Type I error at the smallest number of groups (McNeish & Stapleton, 2016). When group sizes vary in the small number of groups case, the Satterthwaite and Kenward-Roger approaches may lack statistical power and an alternative denominator degrees of freedom, suggested by Schluchter and Elashoff (1990) seems to perform better (Li & Redden, 2015). More work is needed to better understand the conditions the performance of these corrections under various conditions such as unequal group sizes, ICC levels, and nonnormality.[4]

A Bayesian estimation via Markov Chain Monte Carlo (MCMC) sampling is becoming more commonly employed and available. It is sometimes difficult to characterize the performance of Bayesian estimation because the process can involve different settings for distributional priors (e.g., uninformative, weakly informative, strongly informative). With strongly informative priors, results can be misleading unless they are correct (which is difficult to know in practice). Although an earlier, widely-cited article by Stegmueller (2013) found that Bayesian estimation had less coefficient bias and superior confidence intervals than full maximum likelihood estimation for a small number of groups (15-20), Elff and colleagues (2012), using a reanalysis, make the case that this result was in error and that with restricted maximum likelihood and degrees of freedom correction approaches described above for fixed effects tests are accurate with traditional (“frequentist”) approaches. McNeish (2017) makes a similar case, arguing that superiority of Bayesian estimation superiority may rely on having good, strongly informative priors.

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 number of groups. Lower power can be expected for either approach when the number of groups is small (e.g., < 50 groups; Harwell, 1997; van der Leeden et al., 1997), although this differs considerably for intercept variance (more powerful) and slope variance (less powerful) due to differences in reliability (more on this topic later). Moreover, downwardly biased standard errors, and thus inflated Type I error rate, have been shown to be occur with small number of cases per group (Maas & Hox, 2005). 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 (see subsequent section below).

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

[3] The Kenward-Roger approach is now available in SPSS using the subcommand /CRITERIA=DFMETHOD(KENWARDROGER), in R using the `lmertest` package with the `lme4` package using the `lmer` function to test the model and then modifying the summary statement, `summary(model, ddf=c("Kenward-Roger"))`, and in SAS using the DDFM option on the model line, DDFM=KENWARDROGER.

[4] An additional issue is that fixed effects tests are also potentially sensitive to distributional assumptions about the errors, and robust estimates are sometimes recommended to adjust the standard errors (Raudenbush & Bryk, 2002). There will be more on this topic later. Another robust standard error approach that has been recently studied (Huang, Wiedermann, & Zhang, 2023), called bias reduced linearization method or CR2 (Bell and McCaffrey, 2002), also holds promise, including for heteroscedastic data.

to different conclusions than other methods, but with small samples (i.e., small number of groups primarily) they can be problematic (Snijders & Bosker, 2012). In SPSS and in R with the `nlme` package, 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) or using a 90% confidence interval, following the rationale used for other variance tests (Miller, 1977; Self & Liang, 1987).[5] 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).

Profile Likelihood Confidence Intervals. The `lme4` package also provides confidence intervals using the profile likelihood (using penalized least squares) method with the `confint()` function. The profile likelihood approach (Bates & DebRoy, 2004; Bates et al., 2015) is an iterative method that does not assume a symmetric sampling distribution for the random effect and should perform better than standard, symmetric Wald tests of random effects. The approach is related to the likelihood ratio testing approach discussed below, and, given this derivation and the asymmetrical intervals, it should theoretically have advantages over Wald or t-test approaches to variance tests. Despite the incorporation into the variance testing in the R package `lme4`, it does not appear to have been extensively compared to other variance testing approaches in simulation studies to date.

Recommendations. The Wald variance tests from SPSS and R (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 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. 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

[5] For the MIXED procedure in SPSS, the 90% CI can also be requested in SPSS, with `/CRITERIA=CIN(90)`, or, in R, adding `intervals(model,.90)`. Be sure to use these intervals only for intercept and slope variance, not for their covariance or for fixed effects. SAS 8 and higher uses one-tailed p-values for the variance but not the covariance, so no action is required by the user. Because HLM uses a different test, no alteration of p-values is needed either.

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)

Tests for 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). This type of test is most commonly used for testing whether or not all of the predictors together account for a significant amount of variance, which is akin to the test of the multiple R-square in OLS regression analysis. In multilevel models, fixed, random, or a combination of fixed and random effects can be tested with the likelihood ratio 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).[6] 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). For a comparison of two models with and without a single predictor, this would be a test of the fixed effect for the slope and would be testing the same hypothesis (but not with the same method or necessarily same result) as the Wald variance test described above in which the estimate is divided by its standard error. The are asymptotically equivalent in 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. 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). For example, if one model constrains a slope variance to be non-varying across groups, it can be compared to a model in which the slope 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 this example further, however. In 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), on the surface it would appear to be testing a single parameter, when, 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 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. 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

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

"multivariate" or "multiparameter" tests preprogrammed. The documentation does not discuss the methodology, but based on results from a few models, the multiparameter tests in HLM do not seem require any adjustment to the p-value. Although it may not be permitted with the software program, a nested test of this sort with only the variance of the slope differing can theoretically be tested if the program allows a multilevel model to be tested in which the covariance between the intercept and slope can be constrained to be zero. Then a model with the slope variance estimated but the covariance constrained to be zero could be compared to a model that constrains both the slope variance and then intercept-covariance parameter to be zero, and, thus, the two models will differ only in the slope variance. The potential problem with this test is that the model constraining the slope-intercept covariance to zero may be incorrect and lead to estimation problems or a model with other estimates affected by an unreasonable constraint.

Any number of parameters can be compared in the two models, so that a test of a full model can be compared to the empty model without any predictors in order to test the significance of a set of the variance accounted for by a set of predictors entered together. Such tests are complicated, however, by the inclusion of random slopes in the full model, because the empty model will differ in the fixed and random effects. An important precaution for these likelihood ratio tests in multilevel regression is that whenever the two models compared involve any difference in fixed effects (whether or not they differ in random effects also), 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.

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). The HLM package has a feature through the menus under “Other Settings” to conduct multivariate tests comparing two models. Likelihood ratio tests with SPSS would need to be conducted by hand.

References

Bates, D.,Maechler, M.M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bates, D. M., & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis, 91*(1), 1-17.

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*(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.

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.

Li, P., & Redden, D. T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology, 15*, 1-12.

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

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Manor, O., & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. Computational statistics & data analysis, 46(4), 801-817.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*(5), 661-670.

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.

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.

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

Schluchter, M. D., & Elashoff, J. T. (1990). Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. Journal of Statistical Computation and Simulation, 37(1-2), 69-87.

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

Van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (April, 1997). Applications of bootstrap methods for two-level models. Unpublished paper, In Multilevel Conference, Amsterdam.

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