[Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences]_Cohen_2003

Chapter 14 (pp. 563)

14.11.1 Fixed Effects

Tests of the fixed effects are made against the standard error of the fixed effect, resulting in a z test. Alternatively, a t test is computed, as is given in both SAS PROC MIXED and the specialized multilevel software package HLM (Raudenbush, Bryk, Cheong, & Congdon, 2001). Degrees of freedom for the test depend on whether the predictor is a level 2 predictor or a level 1predictor. **For level 1 predictors, the df depend on the numbers of individual cases, groups, and level 1 predictors. For level 2 predictors, the df depend on number of groups and number of level 2 predictors, and are specifically (g - S_q - 1) df, where g is the number of contexts (groups) and S_q is the number of level 2 predictors.**

14.11.2 Variance Components

Each variance component may be tested for significance of difference from zero in one of several ways. First is a chi square test, based on OLS estimates of within group coefficients, which contrasts within group estimates with the fixed population estimate. Use of this test requires that most or all contexts be of sufficient size to yield OLS estimates. The result is distributed approximately as **χ^2 with (g - S_q - 1) df**, where **g is the number of contexts (groups) from which OLS estimates can be obtained and S_q is the number of level 2 predictors (this test is reported in HLM output)**. Second is a z test based on large sample theory, reported in SAS PROC MIXED. Both Raudenbush and Bryk (2002) and Singer (1998) express caution concerning this latter test because of the skew of the sampling distribution of the variance components and because of the dependence on large sample theory (asymptotic normality is assumed but not achieved). There is a third approach to examining variance components, a model comparison approach, based on likelihood ratio tests of nested models. This is the same form of test as in the testing of nested models in logistic regression explained in Section 13.2.14. In the RC context we specify a model that allows a particular variance component to be nonzero, for example, the variance of the slopes. We then specify a second, more restrictive, model that forces this variance component to zero. A likelihood ratio χ^2 test is used to test whether the model fit is significantly worse when the variance component is forced to zero. If so, we conclude that the variance component is nonzero.

Bolker_2009_Generalized linear mixed models_ A practical guide for ecology and evolution

Box 3. Inference details

Drawing inferences (e.g. testing hypotheses) from the results of GLMM analyses can be challenging, and in some cases statisticians still disagree on appropriate procedures. Here we highlight two particular challenges, boundary effects and calculating degrees of freedom.

Calculating degrees of freedom

The degrees of freedom (df) for random effects, needed for Wald t or F tests or AIC, **must be between 1 and N – 1 (where N is the number of random-effect levels)**. **Software packages vary enormously in their approach to computing df** [61]. The simplest approach (the default in SAS) uses the minimum number of df contributed by random effects that affect the term being tested [29]. The Satterthwaite and Kenward-Roger (KR) approximations [29,62] use more complicated rules to approximate the degrees of freedom and adjust the standard errors. KR, only available in SAS, generally performs best (at least for LMMs [63]). In our literature review, most SAS analyses (63%, n = 102) used the default method (which is 'at best approximate, and can be unpredictable' [64]). An alternative approach uses the hat matrix, which can be derived from GLMM estimates. **The sample size n minus the trace t (i.e. the sum of the diagonal elements) of the hat matrix provides an estimate of the residual degrees of freedom** [43,51]. **If the adjusted residual df are >25, these details are less important.**

Accounting for boundary effects and computing appropriate degrees of freedom is still difficult. **Researchers should use appropriate corrections when they are available, and understand the biases that occur in cases where such corrections are not feasible** (e.g. ignoring boundary effects makes tests of random effects conservative).

[Multilevel Analysis (2nd ed)]_Snijders_2012

Chapter 6 (pp. 94-95)

Under the null hypothesis, T (γ_h) has approximately a t distribution, but the number of degrees of freedom (df) is somewhat more complicated than in multiple linear regression, because of the presence of the two levels. The approximation by the t distribution is not exact even if the normality assumption for the random coefficients holds. Suppose first that we are testing the coefficient of a level-one variable. If the total number of level-one units is M and the total number of explanatory variables is r, then we can take **df = M - r - 1**. To test the coefficient of a level-two variable when there are N level-two units and q explanatory variables at level two, we take **df = N - q - 1**. To test the coefficient of the cross-level interaction between level-one variable X and level-two variable Z, when the model contains a total of q other level-two variables also interacting with this variable X, we also use **df = N - q - 1**.

If the number of degrees of freedom is large enough (say, larger than 40), the t distribution can be replaced by a standard normal distribution.

This rule for the degrees of freedom in the t-test is simple and has good properties. The literature contains proposals for various other rules. Manor and Zucker (2004) give a review and a simulation study. Their conclusion is, first, that **these tests should use the standard errors obtained from REML estimation**, not from ML estimation. Furthermore, they found that the so-called **Satterthwaite approximation** as well as the **simple approximation**

mentioned here give a good control of the type I error rate, and most other methods give inflated type I error rates for fewer than 30 groups. Maas and Hox (2004) confirmed that for 30 or more groups, the Wald tests of fixed effects using REML standard errors have reliable type I error rates.

[Multilevel Analysis_Techniques and Applications]_Hox_2010

Chapter 3 (pp. 45-47)

3.2.1 Testing regression coefficients and variance components

Maximum likelihood estimation produces parameter estimates and corresponding standard errors. These can be used to carry out a significance test of the form **Z = (estimate)/(standard error of estimate)**, where Z is referred to the standard normal distribution. This test is known as the Wald test (Wald, 1943). The standard errors are asymptotic, which means that they are **valid for large samples**. As usual, it is not precisely known when a sample is large enough to give confidence about the precision of the estimates. Simulation research suggests that **for accurate standard errors for level 2 variances, a relatively large level 2 sample size is needed**. For instance, simulations by van der Leeden, Busing, and Meijer (1997) suggest that with fewer than 100 groups, ML estimates of variances and their standard errors are not very accurate. In ordinary regression analysis, a rule of thumb is to require 104 + p observations if the interest is in estimating and interpreting regression coefficients, where p is the number of explanatory variables (Green, 1991). If the interest is in interpreting (explained) variance, the rule of thumb is to require 50 + 8p observations. In multilevel regression, the **relevant sample size** for higher-level coefficients and variance components is the **number of groups**, which is often not very large. Green's rule of thumb and van der Leeden et al.'s simulation results agree on a preferred group-level sample size of at least 100. Additional simulation research (Maas & Hox, 2005) suggests that if the interest lies mostly in the fixed part of the model, far fewer groups are sufficient, especially for the lowest-level regression coefficients. The issue of the sample sizes needed to produce accurate estimates and standard errors is taken up in more detail in Chapter 12.

It should be noted that **the p-values and confidence intervals produced by the program HLM (Raudenbush, Yang, & Yosef, 2000) differ from those obtained from most other programs**. As part of their output, most multilevel analysis programs produce parameter estimates and asymptotic standard errors for these estimates, all obtained from the maximum likelihood estimation procedure. The usual significance test is the Wald test, with Z evaluated against the standard normal distribution. Bryk and Raudenbush (1992, p. 50), referring to a simulation study by Fotiu (1989), argue that **for the fixed effects it is better to refer this ratio to a t-distribution on J − p − 1 degrees of freedom, where J is the number of second level units, and p is the total number of explanatory variables in the model**. The p-values produced by the program HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004) are based on these tests rather than on the more common Wald tests. When the number of groups J is large, the difference between the asymptotic Wald test and the alternative Student t-test is very small.

However, **when the number of groups is small, the differences may become important**. Since referring the result of the Z-test on the regression coefficients to a Student t-distribution is conservative, this procedure should provide a better protection against type I errors. **A better choice for the degrees of freedom in multilevel models is provided by the Satterthwaite approximation (Sattertwaithe, 1946). This estimates the number of degrees of freedom using the values of the residual variances.** Simulation research (Manor & Zucker, 2004) shows that the **Satterthwaite approximation works better than the Wald test when the sample size is small (e.g., smaller than 30).**

Several authors (e.g., Berkhof & Snijders, 2001; Raudenbush & Bryk, 2002) argue that the Z-test is not appropriate for the variances, because it assumes a normal distribution, whereas the sampling distribution of variances is skewed, especially when the variance is small. Especially if we have both a small sample of groups and a variance component close to zero, the distribution of the Wald statistic is clearly non-normal. Raudenbush and Bryk propose to test variance components using a chi-square test on the residuals. This chi-square is computed by: $\chi^2 = \sum(\hat{\beta}_j - \beta)^2 / \hat{V}_j$,

where βˆj is the OLS estimate of a regression coefficient computed separately in group j, β its overall estimate, and Vˆj its estimated sampling variance in group j. The number of degrees of freedom is given by **df = J − p − 1**, where **J is the number of second-level units, and p is the total number of explanatory variables in the model**. Groups that have a small number of cases are passed over in this test, because their OLS estimates are not sufficiently accurate.

Simulation studies on the Wald test for a variance component (van der Leeden et al., 1997) and the alternative chi-square test (Harwell, 1997; Sánchez-Meca & MarínMartínez, 1997) suggest that with small numbers of groups, both tests suffer from a very low power. The test that compares a model with and without the parameters under consideration, using the chi-square model test described in the next section, is generally better (Berkhof & Snijders, 2001; Goldstein, 2003). Only if the likelihood is determined with a low precision, which is the case for some approaches to modeling non-normal data, is the Wald test preferred. Note that if the Wald test is used to test a variance component, a one-sided test is the appropriate one.

[Multilevel and Longitudinal Modeling with IBM SPSS]_Heck_2013

Chapter 3 (pp. 91)

The default covariance structure is Variance Components (VC). VC is the default covariance structure for random effects. This specifies a diagonal covariance matrix for the random effects; that is, it provides a separate variance estimate for each random effect, but not covariances between random effects. In this case, there is only one **random effect (the intercept)**, so we can use the default **VC**. For models with **random intercepts and slopes**, a common choice is an **“unstructured” (UN)**, or a completely general, covariance matrix, which fits all variances and covariances between random effects.

Chapter 3 (pp. 98-99)

Once again, the **t test of the significance of the intercept is not really interesting**, since it is a test of whether the intercept is equal to 0. The degrees of freedom reported for each fixed effect, which reflect the **Satterthwaite (1946) correction for approximating the denominator degrees of freedom** for significance tests of fixed effects in models where there are **unequal variances and group sizes**, are useful in determining at what level each variable is measured in the model. For example, we know there are 419 schools in the sample. This is consistent with the 375.7 degrees of freedom reported in Table 3.10. In contrast, we know that SES is an individual-level variable. There are 6,871 individuals in the sample, so the degrees of freedom of 3,914.6 are consistent with a variable measured at Level 1. We note in passing that **if we wish to estimate a model without using the Satterthwaite correction for degrees of freedom, we can use GENLINMIXED (instead of MIXED) and specify a continuous dependent variable with normal distribution and identity link function**. In the Build Options dialog box, for estimating degrees of freedom it is possible to select **Residual method** instead of **Satterthwaite approximation**. It is also possible to select robust (rather than model-based) standard errors. In this case, for example, with **robust standard errors**, the standard error of the SES parameter would be **more conservatively estimated** at 0.145 (not tabled) as opposed to 0.137, as in Table 3.10. Analysts should keep in mind, however, that presently only REML estimation is available with that routine.

Conclusion: How to compute *df* in multilevel models

T-test should use the standard errors obtained from REML estimation, not from ML.

t = (estimate)/(standard error of estimate) = b/se

For large sample size (i.e., number of groups; >= 30 groups):

Level-1 predictor (fixed slopes; without cross-level interaction):

$$df = N_{\text{sample size}} - p_{\text{level-1 predictors}} - q_{\text{level-2 predictors}} - 1$$

$$(= N_{\text{level-1-sample-size}} - p_{\text{all-parameters}})$$

Level-2 predictor, Level-1 predictor (random slopes; with cross-level interaction), Intercept, Cross-level interaction term:

$$df = K_{\text{groups}} - q_{\text{level-2 predictors}} - 1$$

$$(= K_{\text{level-2-sample-size}} - q_{\text{all-level-2-parameters}})$$

For small sample size (i.e., number of groups; < 30 groups):

Use the Satterthwaite approximation (Sattertwaithe, 1946; available in the R package *lmerTest*, or in SPSS MIXED):

***df* is estimated by the values of the residual variances**
(correction for unequal variances and group sizes)

Table 1. Computation of Degree of Freedom (*df*) in Multilevel Modeling

Predictor	Effect	*df*	Software
1. Large group sample size (*K* >= 30 groups)			
Level-1 predictor (γ_{10})	Fixed slope	$N - p - q - 1$	HLM
	Random slope	$K - q - 1$	
Level-2 predictor (γ_{01})			
Cross-level interaction (γ_{11})			
Intercept (γ_{00})			
2. Small group sample size (*K* < 30 groups)			
All predictors		Estimated by Satterthwaite's approximation	R (lmerTest), SPSS (MIXED), jamovi (GAMLj)

Note.

N = number of observations (individual-level sample size)

K = number of groups (group-level sample size)

p = number of level-1 predictors in level-1 sub-model

q = number of level-2 predictors in level-2 sub-model (*specific* for different equations)

Table 2. Common Examples

Level-1 Model	Level-2 Model	Type	Slope (X)	$df_{intercept}$	$df_{level\text{-}1\ (X)}$	$df_{level\text{-}2\ (W)}$
$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$	$\beta_{0j} = \gamma_{00} + u_{0j}$	One-way ANOVA with random effects[1]	—	$K - 1$	—	—
	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j}$	Intercepts-as-outcomes model	—	$K - 2$	—	$K - 2$
$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij}$	$\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10}$	One-way ANCOVA with random effects	Fixed	$K - 1$	$N - 2$	—
	$\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$	Random-coefficients regression model	Random	$K - 1$	$K - 1$	—
	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j}$ $\beta_{1j} = \gamma_{10}$	Contextual model (fixed effects)	Fixed	$K - 2$	$N - 3$	$K - 2$
	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$	Contextual model (random effects)	Random	$K - 2$	$K - 1$	$K - 2$
	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}W_{1j} + u_{1j}$	Full model[2]	Random	$K - 2$	$K - 2$	$K - 2$

Note.

1. Also called: "null model", "intercept model", "variance component model".
2. Also called: "intercepts-and-slopes-as-outcomes model".