Methods for tests of fixed effects in small and nonnormal samples

In chapter 1, we outlined the basics of analyzing longitudinal data and introduced linear mixed models. Next, we will examine inference of linear mixed models, and how methods such as Kenward-Roger (KR) and Satterthwaite can be used in situations where standard procedures for inference may produce questionable results.

Inference

In statistical inference, the goal is to make conclusions about the underlying characteristics of a set of data and establish a relationship between certain variables. Hypothesis testing is one of the primary examples of inference, and is carried out in order to assess the true value of a population parameter. In linear models, the significance of a slope parameter, β_k, is often assessed, where the null hypothesis, $H_0 $ is $\beta_k = 0$, and the alternative hypothesis H_a is $\beta_k \neq 0$. A test of the null hypothesis involves using a Wald statistic in the form

$$Z = \frac{\hat{\beta}_k}{\sqrt{Var(\hat{\beta}_k)}},$$

which is then compared to the normal distribution, and a subsequent p-value is obtained.

Building on foundations of a general linear hypothesis test, given a matrix L of size $q \times p$, where q represents the number of estimable functions of β,

$$(L\hat{\beta} - L\beta)'[L(X'\widehat{Cov}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta} - L\beta)$$

is approximately $\chi^2(q)$ (Rencher and Schaalje, 2008). For a null hypothesis $H_0 : L\beta = 0$, the test statistic G is

$$(L\hat{\beta})'[L(X'\widehat{Cov}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta}).$$

Aside from using the Wald statistic, likelihood ratio tests are another method to make inferences about β, and involves comparing two models: (1) a nested model, which assumes that β_k is 0, and (2) a full model, that allows β_k to vary without constraint. The difference in the maximized log-likelihood of the two models, $\hat{l}_{reduced}$ and $\hat{l}_{full}$ are compared. This difference is represented by the statistic

$$G^2 = 2(\hat{l}_{full} - \hat{l}_{reduced}),$$

which is compared to a chi-square distribution. The larger the difference, the more likely we are to conclude that the nested model is insufficient, and that β is not zero. While there are benefits to using the likelihood ratio test, the rest of this study will focus on method of using the Wald statistic.

Inference in small sample sizes

One crucial assumption when conducting inference using the ML estimate for β is that the sample size is sufficient enough where it does not affect the estimate for Σ_i. However, what happens when the sample size is too small? This causes $\hat{\Sigma}_i$ to underestimate the true variance, which in turn causes $\widehat{\text{Cov}}(\hat{\beta})$ to be too small since it relies on covariance estimator. If $\widehat{\text{Cov}}(\hat{\beta})$ is too small, the denominator of the test statistic is inflated, leading to increased Type I error. One can see that the bias of the covariance estimator weakens the entire foundation of estimation and inference.

In very limited cases, where data are complete, balanced, and produce nonnegative values in REML estimation, is it possible to perform exact small-sample inferences. If $[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^{-}L']^{-1}$ with g degrees of freedom can be rewritten such that

$$\frac{(L\hat{\beta})'Q(L\hat{\beta})}{g}\frac{w}{d} = \frac{(L\hat{\beta})'[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta})}{g},$$

where w is a chi-square random variable with d degrees of freedom. If so, this statistic is F-distributed.

However, in most scenarios, an approximate small-sample method must be used, in which the statistic

$$F = \frac{(L\hat{\beta})'[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta})}{g}$$

follows a distribution with numerator degrees of freedom g, and unknown denominator degrees of freedom (DDF). There are several ways to approximate the DDF.

Both Satterthwaite and KR are proposed methods of reductions to the DDF when conducting tests in order to account for the uncertainty of the covariance estimator. The KR method goes one step forward to also adjust the test statistic itself.

Satterthwaite

Sattherthwaite approximation was developed by Fai & Cornelius (1996), with the F statistic following the form:

$$F = \frac{1}{l}\hat{\beta}'L'(L\widehat{\text{Cov}}(\hat{\beta})L')^{-1}L\hat{\beta}.$$

For the denominator degrees of freedom we perform spectral decomposition on $L'\widehat{\text{Cov}}(\hat{\beta})L = P'DP$, where D is a diagonal matrix of eigenvalues and P is an orthogonal matrix of eigenvectors. When r represents the r^{th} row of $P'L$, we have $v_r = \frac{2(d_r)^2}{g_r'Wg_r}$, where g_r is a gradient vector, d_r is the r^{th} diagonal element of D, and W is the covariance matrix of $\hat{\sigma}^2$. The denominator degrees of freedom is calculated by:

$$\frac{2E}{E-l},$$

where $E = \sum_{r=1}^{l} \frac{v_r}{v_r-2} I(v_r > 2)$ if $E > l$, otherwise $DF = 1$.

When $l = 1$ the KR and Satterthwaite approximation will produce the same denominator degrees of freedom. However, since the statistic used for the two methods are not the same, the results for inference will not be the same. It is important to note that both methods are only valid when using REML.

Kenward-Roger

In Kenward-Roger (1997), a Wald statistic is proposed in the form of:

$$F = 1/l(\hat{\beta}-\beta)^T L(L^T\hat{\Phi}_A L)^{-1}L^T(\hat{\beta}-\beta),$$

where l represents the number of linear combinations of the elements in β, L is a fixed matrix, and $\hat{\Phi}_A$ is the adjusted estimator for the covariance matrix of $\hat{\beta}$. As mentioned previously, $\widehat{\text{Cov}}(\hat{\beta})$ is a biased estimator of $\text{Cov}(\hat{\beta})$ when samples are small, and underestimates. This adjusted estimator is broken down into $\hat{\Phi}_A = \widehat{\text{Cov}}(\hat{\beta}) + 2\hat{\Lambda}$, where $\hat{\Lambda}$ accounts for the amount of variation that was underestimated by the original estimator of covariance of $\hat{\beta}$. The value Λ is approximated using a Taylor series expansion around σ, to be

$$\Lambda\text{Cov}(\hat{\beta})[\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}(Q_{ij} - P_i\text{Cov}(\hat{\beta})P_j)]\text{Cov}(\hat{\beta}),$$

where $P_i = X^T\frac{\partial\Sigma^{-1}}{\partial\sigma_i}X$ $Q_{ij} = X^T\frac{\partial\Sigma^{-1}}{\partial\sigma_i}\Sigma\frac{\partial\Sigma^{-1}}{\partial\sigma_j}X$, and W_{ij} is the (i,j)th element of W $= V[\hat{\sigma}]$.

This Wald statistic that uses the adjusted estimator is scaled in the form:

$$F^* = \frac{m}{m+l-1}\lambda F,$$

where m is the denominator degrees of freedom, and λ is a scale factor. Using the expectation and variance of the Wald statistic, F Both m and λ need to be calculated from the data, such that:

$$m = 4 + \frac{l+2}{l\rho - 1},$$

where $\rho = \frac{V[F]}{2E[F]^2}$ and $\lambda = \frac{m}{E[F](m-2)}$. This statistic will ultimately follow an exact $F_{l,m}$ distribution.

Other methods

Residual DDF: The DDF is calculated as $N - rank[X]$, where N is the total number of individuals in the dataset. This method is only suitable for data that are independent and identically distributed, so it is not typically used in linear mixed models.

Containment Method (cite SAS) In the containment method, random effects that contain the fixed effect of interest are isolated. The smallest rank contribution to the $[XZ]$ matrix among these random effects becomes the DDF. If there are no effects found, then the DDF is equal to the residual DDF.

Between-Within Method Schluchter and Elashoff (1990) propose a DDF method where residual DDF are calculated for both between-subject and within-subject subgroups. If there are changes in the fixed effect within subjects, then the within-subject DDF is used, otherwise the between-subject DDF is used.

Existing literature

Both methods are frequently used and compared, and its performance is highly dependent on the structure of the data. A majority of studies focusing on DF method comparison in mixed models use split-plot design, as small sample sizes are more common in agricultural and biological fields. Schaalje, et al. (2002) found that in comparison to other degrees of freedom-adjusting methods like Satterthaite, KR was the most suitable for small sample data. Using factors such as imbalance, covariance structure, and sample size, they demonstrated that the KR method produced simulated Type I error rates closest to target values. However, their focus was primarily on complexity of covariance structure, and they found that more complicated structures, such as ante-dependence, produced inflated error rates when coupled with small sample size. Arnau (2009) found that KR produces more robust results compared to Satterthwaite and Between-Within approaches, especially in cases where larger sample size was paired with covariance matricies with larger values.

These studies are conducted with data drawn from normal distributions. However, real-world data used in fields such as psychometrics have distributions that are nonnormal. In Arnau et. al's 2012 paper, the authors extend their evaluation of KR for split-plot data that follow a log-normal or exponential distribution, and for when the kurtosis and skewness values are manipulated. They found that, compared to normal distribution, the test is less robust for log-normal distributions, but that there is no signficant difference in performance between exponential and normal distributions. In addition, they suggest that skewness has a bigger effect on robustness of KR compared to kurtosis.

Existing research evaluating the performance of methods that reduce Type I error rate in small samples are thorough, however, the differences in simulation setup and structure of data used make generalizations difficult. Although the KR method has been shown as a viable option for analysis of small samples in many occasions, it should continue to evaluated against other methods. To date, there is no literature on the performance of Satterthwaite for nonnormal longitudinal data design. Given the prevalence of nonnormal and small data samples, it is important to continue exploring methods that ensure robust results.

Goals of this study:

In this study, we aim to expand on previous simulations, evaluating how methods for evaluated fixed effects perform under different nonnormal distributions and sample sizes. The aforementioned studies often use a

split-plot design and impose a covariance structure, but goal of this study will be to compare performance of KR and Satterthwaite methods for repeated measures longitudinal data fitted with a linear mixed effects model, and no imposed covariance structure. Since most mixed models use unstructured covariance structure, it would be beneficial to see how these methods perform without considering covariance structure as a factor.

Simulation Set up:

Generating data: Sample size

In this study, we consider a linear mixed effects model with two discrete covariates: time and treatment. The range of possible values that time takes on depends on how many number of measurements per individual, which can be 4 or 8. The treatment covariate takes on values of 0 or 1, and each assigned to half of the sample. The number of individuals take on possible values of 10, 18, and 26. These were chosen to reflect possible samples that would not hold under the common assumption that the sample size must be at least 30 for it to be considered sufficient enough for the Central Limit Theorem to hold.

Generating data: Fixed Effects

We have three fixed effects: the intercept value and the covariates time and treatment. The intercept, an arbitrary value, is set at 3.1. Time and treatment have a value of 0, and the Type I error rates of treatment will be evaluated.

Generating data: Random effects

In order to generate a continuous response variable that is nonnormal, we generate our random effects values from nonnormal distributions, which are either exponential or lognormal. Previous research shows that many data used in social and health sciences follow nonnormal distributions (Limpert, Stahel, & Abbt, 2001). More specifically many follow lognormal distributions, such as age of onset of Alzheimer's disease (Horner, 1987), or exponential distribution to model survival data. In order to cover a wide range of exponential and lognormal distributions, parameters were chosen to model distinct distributions. For exponential distributions, lambda values of .2, and .9 were used, (DO I NEED TO INSERT GRAPH?). For lognormal distribution, mean and standard deviation parameter combinations were (0,.25), and (1,.5).

Using the `SimMultiCorrData` package, we derive kurtosis and skewness values based on the distributions specified above. The table below shows the range of skewness and kurtosis values for the Lognormal distribution. In the intercept only model, only one non-normal continuous variable is generated for the random effect, so the function `SimMultiCorrData::nonnormvar1` is used. Values are generated through Fleishman's method for simulating nonnormal data by matching moments using mean, variance, skew, and kurtosis and then transforming normally distributed values.

Kurtosis and skew values for the distributions used in this simulation are shown below.

```
##          mean        sd      skew   kurtosis        fifth        sixth
## [1,] 1.031743 0.2620191 0.7782516  1.0959313    2.2963724 6.478484e+00
## [2,] 1.656986 0.1661137 0.3017591  0.1623239    0.1282109 1.346775e-01
## [3,] 3.080217 1.6415718 1.7501897  5.8984457   31.4320590 2.400042e+02
## [4,] 1.499303 1.6748679 4.7453294 57.4107553 1530.1579066 8.692474e+04
```

In the case of the linear model that has both random effects for intercept and slope, we want to generate random effects values that are correlated. Using `SimMultiCorrData::rcorrvar`, we use a similar process for generating one nonnormal continuous variable, but extend it to generating variables from multivariate normal distribution that take in to account a specified correlation matrix, and are then transformed to be nonnormal.

KR_err	0.0468333
S_err	0.0546667
Z_err	0.0826667
t_err	0.0760000

We use a correlation value of -.38 to generate the random effects, which is based off the correlation observed when fitting a linear mixed effects model from the dataset used in the application portion of this study.

Lastly, to account for measurement/sampling error, we assume that the error is random and drawn from a $N\tilde{(}0, .2)$. The standard deviation value was chosen to minimize the variation of the errors in relation to the random effects of the intercept and the covariate.

Linear mixed effects model

In a linear mixed effects model, the amount of random effects that will be modeled depends on the research question at hand. Here, we will examine both a random intercepts-only model, where the intercept of the model is assumed to have a random effects structure, as well as a random intercept and slope model, where in addition to intercept, the covariate time will also have a random effects structure.

We use the `lmerTest` package to fit the linear mixed effects model, and evaluate the significance of the covariate in the model. To evaluate significance, we compare both the KR and Satterthwaite method for adjusting denominator degrees of freedom and its resulting p-value. Because the value of the covariate in our model is fixed at 0 in order to identify Type I error, we expect to see that the p-value for the covariate time to not be significant ($p > .05$) in an ideal scenario.

Evaluating and Results

After performing 1,000 replications of each condition at a significance level of .05, we evaluate robustness using Bradley's criterion,which considers a test to robust if the empirical error rate is between .025 and .075. In the following section, we will compare Type I error rates produced from KR and Satterthwaite methods as well as t-as-z and using the standard DF formula, further stratified by distribution and other manipulated parameters. T-as-z and standard DF formula are not adjustments to account for smaller sample sizes, and are used as comparison to Satterthwaite and KR, since they are expected to be anti-conservative.

Error rate

Average Type I error rates are shown in (TABLE 1?) across all simulations. Despite a wide range of distributions, performance of methods align closely with previous studies examining only normal populations (CITE). T-as-z and standard degrees of freedom methods are much more anti-conservative compared to KR and Satterthwaite at these smaller sample sizes. Type I error rates are closest to .05 when using KR and Satterthwaite. After confirming that the simulation results follow similar patterns to what is expected, we proceed to examining more specific differences in performance of these methods among the nonnormal distributions.

Error rate by distribution / skewness/kurtosis

(TABLE 2?) shows Type I error rates for each DF method by distribution. Across all distributions, t-as-z and standard DF produce the highest Type I error rates compared to KR and Satterthwaite. However, they do perform relatively better and produce less liberal rates in nonnormal distributions compared to the normal distribution. KR and Satterthwaite methods are overall more robust and produce Type I error rates closer to .05; in comparison to Satterthwaite, KR appears to be more conservative, with an error rate of .03

distribution	skew	kurtosis	KR_type1	S_type1	Z_type1	t_type1
Gaussian	0.0000000	0.000000	0.0308333	0.0375000	0.0875000	0.0841667
Lognormal	0.7782516	1.095931	0.0391667	0.0500000	0.0816667	0.0700000
Lognormal	1.7501897	5.898446	0.0508333	0.0650000	0.0733333	0.0716667
Exponential	2.0000000	6.000000	0.0566667	0.0604167	0.0854167	0.0770833

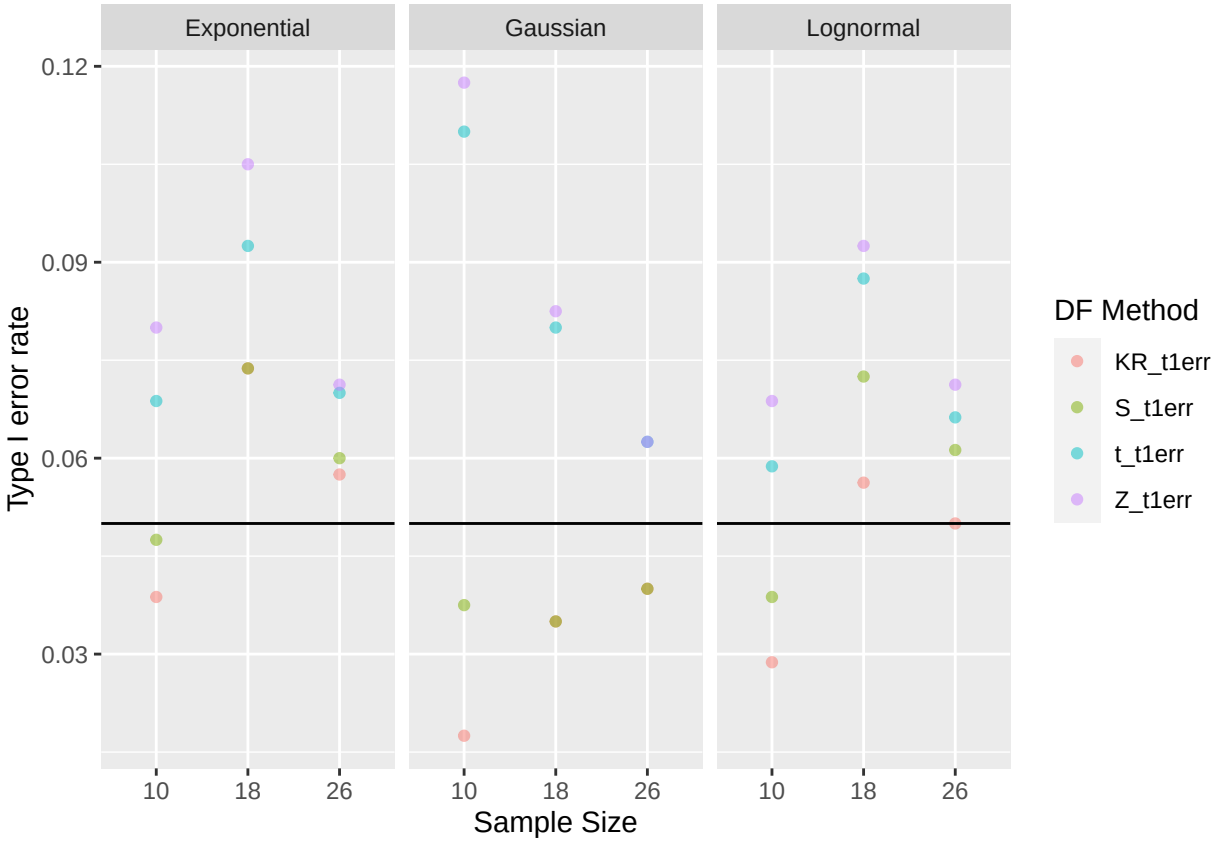
distribution	number_measurements	KR_type1	S_type1	Z_type1	t_type1
Exponential	4	0.0583333	0.0600000	0.0841667	0.0783333
Exponential	8	0.0550000	0.0608333	0.0866667	0.0758333
Gaussian	4	0.0283333	0.0316667	0.0866667	0.0816667
Gaussian	8	0.0333333	0.0433333	0.0883333	0.0866667
Lognormal	4	0.0483333	0.0608333	0.0816667	0.0716667
Lognormal	8	0.0416667	0.0541667	0.0733333	0.0700000

for the normal distribution compared to .04 for Satterthwaite. Even as skewness and kurtosis increases, KR produces error rates that are slightly more conservative: for a lognormal distribution with skewness of 1.75 and kursosis of 5.9, the error rate for KR is .05 and .07 for Satterthwaite.

Error rate by distribution + number of measurements

Increasing the number of measurements per individual is generally thought to increase robustness, but sensitivity of number of measurements across methods may vary. (TABLE 3?) captures the Type I error rates stratified by distribution and number of measurements. It appears that t-as-z and standard DF methods are not sensitive to increases in measurement, as Type I error rates do not significantly improve from 4 measurements to 8. On the other hand, both KR and Satterthwaite see a slight increase in robustness as a result of increasing measurements, with more significant improvements in lognormal and normal distributions.

Type 1 error rate, by number of subjects



Apart from number of measurements, sample size also plays a crucial role in inference. Type I error rates across 3 different sample sizes and distributions are displayed in (FIGURE 1). The relationship between sample size and robustness differs across distributions and methods. Similar to examining other parameters, t-as-z and standard DF have a significantly higher Type I error rate regardless of distribution or sample size when comparing to KR or Satterthwaite methods.

In the normal distribution, we see a linear relationship between sample size and Type error rate across all 4 methods; this relationship follows common knowledge that increasing the sample size will increase robustness. However, in the case of lognormal and exponential distribution, a sample size of 18 tends to yield less conservative error rates compared to a sample of size 10. It may be plausible that at sizes this small, an additional handful of individuals in the sample may not contribute significantly to improving robustness in these nonnormal distributions.

While KR and Satterthwaite adjustments typically both perform better in comparison to the other two methods, there are slight differences between them in terms of sample size. In samples of size 10, KR is significantly more conservative and produces error rates further from .05 compared to Satterthwaite. However, the difference decreases once the sample size is 26, with KR being a slightly more robust method once the sample size has increased.

KR VS Satterthwaite

			Sample Size					
			10		18		26	
			Random Intercept	Random Slope	Random Intercept	Random Slope	Random Intercept	Ra
skew	kurtosis	DF_method	random_intercept_10	random_slope_10	random_intercept_18	random_slope_18	random_intercept_26	rand
Normal								
0.0000000	0.000000	KR_t1err	0.0200	0.0150	0.03	0.0400	0.0400	
0.0000000	0.000000	S_t1err	0.0200	0.0550	0.03	0.0400	0.0400	
Lognormal								
0.7782516	1.095931	KR_t1err	0.0200	0.0300	0.04	0.0500	0.0350	
0.7782516	1.095931	S_t1err	0.0200	0.0500	0.04	0.0800	0.0350	
1.7501897	5.898446	KR_t1err	0.0250	0.0400	0.02	0.1150	0.0500	
1.7501897	5.898446	S_t1err	0.0250	0.0600	0.02	0.1500	0.0500	
Exponential								
2.0000000	6.000000	KR_t1err	0.0275	0.0500	0.04	0.1075	0.0575	
2.0000000	6.000000	S_t1err	0.0275	0.0675	0.04	0.1075	0.0575	

KR Only

				Sample Size					
				10		18		26	
				Random Intercept	Random Slope	Random Intercept	Random Slope	Random Intercept	Rand
params	number_measurements	skew	kurtosis	10_FALSE	10_TRUE	18_FALSE	18_TRUE	26_FALSE	2
Exponential									
0.2	4	2.0000000	6.000000	0.05	0.05	0.01	0.07	0.07	
0.2	8	2.0000000	6.000000	0.04	0.06	0.04	0.12	0.02	
0.9	4	2.0000000	6.000000	0.01	0.03	0.03	0.14	0.12	
0.9	8	2.0000000	6.000000	0.01	0.06	0.08	0.10	0.02	
Normal									
0,0.2	4	0.0000000	0.000000	0.02	0.01	0.02	0.07	0.04	
0,0.2	8	0.0000000	0.000000	0.02	0.02	0.04	0.01	0.04	
Lognormal									
0,0.25	4	0.7782516	1.095931	0.00	0.02	0.04	0.10	0.04	
0,0.25	8	0.7782516	1.095931	0.04	0.04	0.04	0.00	0.03	
1,0.5	4	1.7501897	5.898446	0.03	0.02	0.00	0.09	0.06	
1,0.5	8	1.7501897	5.898446	0.02	0.06	0.04	0.14	0.04	

Discussion