COMPARISONS BETWEEN THE LIKELIHOOD RATIO TEST (LRT) AND THE F-TEST IN REPEATED MEASURES DATA

Jeremy Meyer
Brigham Young University
Department of Statistics

ABSTRACT

Repeated measures experiments are typically used to asses the effects of treatments over time. The effects of the treatments over time are fixed effects that are frequently assesed for significance. However, there are two different ways to accomplish this. Either an F test or a Likelihood Ratio Test (LRT) could be performed. Both of these will yield slightly different results in a mixed model. We construct a simulation study that compares statistical power of the F-test and LRT across CS, AR1, and RC covariance structure data. We consider different degrees of freedom for the F-test and compare two different hypothesis tests. Both hypotheses are tested under the null for the correct type I error. We then compare the results of our simulations and make conclusions about the power of the LRT vs the F-test.

1 Introduction

In many experiments, different treatments are often compared over time for various subjects. These are called repeated measurement experiments. These types of experiments can be analyzed using mixed model methods or with non-diagonal covariance structures. Mixed models make use of both fixed effects, or well-defined categories that are desired to make inference on, and random effects. Random effects are variables that are only sampled from the population. Including random effects in the model will allow us to make inferences on levels outside the experiment. Repeated measures experiments typically have treatments as fixed effects and subjects as random effects. Additionally, each subject has the same correlation structure over all time measurements. Traditional ordinary least squares (OLS) methods fall short in that they don't capture the correlation of subjects over time well. For this reason, we will explore different mixed model methods to analyze the repeated measures data.

It is often helpful to examine the fixed effects for significance. The fixed effects are usually covariates that correspond to questions of interest. For example, in medical experiments we may be interested to see if different treatments affect the health of different patients. The fixed effects can be evaluated for significance using both an F-test and a likelihood ratio test (LRT). The F-test tests linear combinations of effects (β) against some null hypothesis. The linear combinations are stored in a *rxp* matrix $\mathbf{C}$ and $\hat{\boldsymbol{\beta}}$ is a *px1* vector of the estimated effects. The F-test statistic is calculated as shown in equation 1. Assuming the residuals of the linear model are normal and independently identically distributed (iid), this follows an $F_{r,n-p}$ distribution, with n being the total number of observations.

$$F = \frac{1}{r}(\boldsymbol{C}\hat{\boldsymbol{\beta}})'[\mathbf{C}\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}) \tag{1}$$

The likelihood ratio value, $\lambda(\boldsymbol{x})$, compares a full and reduced model for significance. One requirement for this test is that one model is contained in another. It compares likelihoods between the 2 models and if it is significant, the full model is needed. This has an exact distribution under the same normal iid conditions that were mentioned earlier for the F-test.

Due to the correlated nature of the repeated measures data, the assumption of iid error terms does not hold. As a result, both the F and LR test will only be approximate. In an OLS setting, these both yield the exact same p-value. Since the data has some dependence structure, the LRT and F-test do not yield the same value, although they are usually close. Additionally, finding the denominator degrees of freedom for the F-test is not as straightforward as it is in the OLS case and must be estimated.

Determining which of the 2 tests is optimal to use is not often clear and obvious. In practice, the test that ends up being used usually depends on which one is easier to access. The goal of this study is to compare the performance of both the LRT and F-test on various repeated measures datasets. We will simulate repeated measures data from various variance structures, fit appropriate models abd perform perform both tests. Since the results of the F and LR test could depend on the hypothesis chosen, we will compute statistical power for 2 different hypotheses:

1. Do the first two treatments have an effect over time?
2. Do all the treatments have different effects over time?

2 Data

To accomplish the goal as stated previously, we will consider the Compound Symmetric, Auto-Regressive(1), and Random Coefficients covariance structures. All 3 of these structures allow for some type of dependence within each subject. Data was simulated using 4 treatments, with 6 separate subjects per treatment. Each subject had 4 equally spaced time measurements, so there were a total of 96 measurements. To address the first hypothesis, each treatment was given a starting response value of 0, and slopes 1,1,0, and 0 for the respective treatments. While there are several ways to test power for the second hypothesis, each treatment was given an intercept of 0 and slopes 1,1,0, and -1 respectively.

Each of the 24 subjects were simulated as a multivariate normal draw using each treatment slope and intercept as the mean and the the matrices in Figure 1 for the covariance. The Compound symmetric (CS) data was given a subject variance of 8 and error variance of 10. This type of structure has a constant correlation over all time intervals. The AR(1) data was given an overall variance of 20 and a correlation of 0.6 for each time interval difference. Thus, the correlation between measurements 1 and 3 should be $.6^2 = 0.36$. This particular structure has a decaying correlation at larger time intervals. The final structure used to generate the data was the random coefficients (RC) covariance structure. This structure fits a random intercept and slope for each subject and can model greater complexity. We simulated the RC data with a slope variance of 10, slope/intercept covariance of -1, intercept variance of 2, and independent random error of $\sigma^2 = 7$. The resulting RC covariance matrix is shown in Figure 1.

The determinants of the three covariance structures are 42000 for CS, 41943.04 for AR1, and 40229 for RC, so they all account for roughly the same amount of variability. Data was generated under both hypotheses for all three structures, making for a subtotal of 6 data sets. Separate datasets were also generated under null conditions, where the slopes were all 0 for both hypothesis cases.

$$\begin{pmatrix} 18 & 8 & 8 & 8 \\ 8 & 18 & 8 & 8 \\ 8 & 8 & 18 & 8 \\ 8 & 8 & 8 & 18 \end{pmatrix} \quad \begin{pmatrix} 20 & 12 & 7.2 & 4.32 \\ 12 & 20 & 12 & 7.2 \\ 7.2 & 12 & 20 & 12 \\ 4.32 & 7.2 & 12 & 20 \end{pmatrix} \quad \begin{pmatrix} 27 & 9 & 8 & 7 \\ 9 & 17 & 11 & 12 \\ 8 & 11 & 21 & 17 \\ 7 & 12 & 17 & 29 \end{pmatrix}$$

CS **AR(1)** **RC**

Figure 1: Covariance matrices used for each subject under the three different models. Each row/column represents the different time values. Shown are the Compound symmetric (left), Autoregressive-1 (middle) and Random Coefficients (right).

3 Methods

After the data was generated, we fit a general least squares model by including treatment and treatment:time interaction as fixed effects. We did not include an overall intercept and fit the data to the general covariance structure it was generated from. In total, there were 4 intercepts or treatment effects, and 4 treatment slopes. Since many experiments are interested in questions regarding how treatments affect subjects over time, we decided to examine how the LRT and F-test perform with the slope coefficients. For future reference, the slope coefficients for each treatment will be referred to as $\beta_{s1}, \beta_{s2}, \beta_{s3}$, and β_{s4}.

We are interested to see if the F-test and LRT perform better under the different hypotheses listed below. These were informally mentioned in the introduction.

1. $H_0 : \beta_{s1} = \beta_{s2} = 0$
   $H_A : \beta_{s1} \neq 0$ or $\beta_{s2} \neq 0$

2. $H_0 : \beta_{s1} = \beta_{s2} = \beta_{s3} = \beta_{s4}$
   H_A : Not all β_s are equal

We tested the hypothesis that matched the hypothesis the data were generated under. For example, if the underlying treatment slope values were 1,1,0, and -1, then we tested hypothesis 2. If the underlying slopes were 1,1,0,0, we tested hypothesis 1. Since the value for the denominator degrees of freedom is not obvious in the F-test, we determined significance using three different degrees of freedom. The degrees of freedom are a measure of the effective sample size and so a conservative estimate (24) would be the total number of subjects. We also used a relaxed estimate (88), which was the number of observations minus the number of fixed effects, and the average of the two estimates (56).

In summary, separate datasets were simulated for each hypothesis and covariance structure combination. To check if the test was at the appropriate type I error (α) level, data was simulated with all zero treatment slopes for each variance structure. All of these datasets were generated 2000 times and fitted to their appropriate models. The percentage of times the null hypothesis was rejected is reported in Table Groups 1 and 2.

4 Results

Table Group 1 shows the results from the null tests. These are expected to be around 5% with responses between [4.04%, 5.96%] still being reasonable. In general, the F and LR tests did a reasonable job at maintaining the appropriate type I error rate for both the CS and AR(1) data. The LRT tended to be a little higher than 5% and the F-tests tended to be slightly lower than 5%. The conservative degrees of freedom resulted in low type I error rates for the CS and AR(1) data. Both tests struggled maintaining the appropriate type I error rate with the RC model. In fact, the type I error rate more than doubled in some cases for the RC data.

Table Group 1: Null Cases

$\beta_{s1} = \beta_{s2} = 0$

Test	CS	AR1	RC
F-24	3.40	3.45	8.5
F-56	4.20	4.45	9.6
F-88	4.35	4.65	10.0
LRT	5.30	5.85	8.6

$\beta_{s1} = \beta_{s2} = \beta_{s3} = \beta_{s4}$

Test	CS	AR1	RC
F-24	3.00	3.00	9.4
F-56	3.75	4.05	11.5
F-88	4.25	4.30	12.4
LRT	5.60	6.05	9.1

These tables show the results of the simulation. The rows indicate the specific test, where the F-24 indicates an F-test using 24 as the denominator degrees of freedom. The columns represent the underlying data structure and model that was used fit to the data. The values on the top of each table represent the true values used for the slope coefficients.

Table Group 2: Power Tests

$\beta_{s1} = \beta_{s2} = 1$

Test	CS	AR1	RC
F-24	51.10	31.85	41.70
F-56	54.95	35.40	44.45
F-88	56.00	36.40	45.20
LRT	59.50	40.55	41.85

$\beta_{s1} = 1, \beta_{s2} = 1, \beta_{s3} = 0, \beta_{s4} = -1$

Test	CS	AR1	RC
F-24	56.70	36.30	44.8
F-56	61.90	41.40	49.3
F-88	63.05	42.95	50.6
LRT	66.40	47.60	42.9

Table Group 2 shows statistical power with the specific parameter values used on the top of each table. These were tested against the null hypotheses in the respective tables in Table Group 1. The F test using the conservative degrees of freedom was reluctant to reject, and so had the lowest power. The data generated under the Compound Symmetric structure had the best power, with the AR1 or RC having the least depending on the hypothesis. The LRT outperforms the F-test on CS and AR1 data, but the F-test wins out on the RC data. Although, the power results from the RC data may be questionable since its type 1 error rate was inflated. It is worth noting that as the parameters get further away from their null values and as the size of the dataset increases, the statistical power for all cases will increase.

5 Discussion

Under the null hypothesis for the CS and AR1 data, the LRT tended to have a type I error rate slightly above 0.05, and the F-test tended to have a type I error slightly below 0.05. However, in the RC data, both tests were inflated and some

F-test type I error values were about twice the original rate. This suggests that some data structures like RC do not work as well with these tests. For more complicated hypothesis tests, the LRT type I error rates were the same relative to each other, but the differences were exaggerated. The fact that the inflated RC type I error rates did not improve between hypothesis tests suggests that the F and LR tests may not work well with the RC data.

The frequency at which the LRT and F-test reject the null hypothesis is similar, but not the same across all types of data. Overall, the LRT did a better job at detecting true differences the null hypothesis. For both the CS and AR1 data, the LRT had higher power. Although the F-test did have more power than the LRT for the RC data, the type I error rate was inflated to about twice the appropriate value for both tests in the RC data, so the the RC power measurements are questionable. A more complicated hypothesis test had the same results for the F and LRT tests, but the power for the AR1 dataset was larger than the RC data. This suggests that the power of the hypothesis tests do not always behave consistently for each data structure. However, both tests were able to pick up differences the best in the CS dataset.

To address the degrees of freedom for the F test, the conservative estimate was as expected, more reluctant to reject. It resulted in significantly lower power for all three data structures. The average of the conservative and relaxed degrees of freedom (df) yielded the best results and performed more similar to the relaxed df. However, for the RC data, the conservative estimate for the df had the least inflated type I error of all three df estimates.

In practice, the choice of test used usually depends on which one is more accessible. For example, sometimes it is easier to think about the hypothesis test in terms of a full and reduced model, rather than trying to construct an appropriate C matrix. Conversely, sometimes it is easier to construct a C matrix that tests a few coefficients rather than to try and restructure data so a reduced model can be built. One big advantage of the LRT is that it doesn't depend on choosing the denominator degrees of freedom or constructing a complicated C matrix. However, the F-test may be more straightforward when only a few specific levels of a factor need to be tested. The results of our simulation show that although there may be special cases where one test outperforms another, they tend to perform similar enough so that the choice does not matter too much.

One limitation to this study is that we only take into account situations where the correct model was fit to the data. It may be interesting to consider the F-test and LRT power when the incorrect model is fit. We have also only considered three different covariance structures, so it's difficult to generalize to all types of data. It may be worth looking into how the LRT and F-test perform across a wider variety of mixed models. Since the choice of alternative hypotheses were arbitrary, With more computing resources, it may be interesting to plot power curves for both the LRT and F-test and see the effect of power as the alternatives deviate more from the null hypotheses.