Adjusted F Tests for Repeated Measures with the MIXED Procedure
S. Paul Wright, University of Tennessee, Knoxville, Tennessee

Abstract. The traditional univariate mixed-model approach to repeated measures analysis produces valid results only when covariances among the repeated measurements have the "type-H" structure of Huynh & Feldt (1970) which includes the more familiar compound symmetry structure. Alternatives to assuming such a structure include: (1) Greenhouse-Geisser (GG) and Huynh-Feldt (HF) adjustments to the degrees of freedom of the univariate mixed-model F tests, and (2) multivariate test statistics. These methods are available in the GLM procedure using the REPEATED statement, but GLM uses only complete observations. A major attraction of the MIXED procedure is its ability to analyze repeated measures data having missing values. In this paper, I show how the GG, HF, and multivariate approaches can be applied to results from the MIXED procedure. Kleinbaum (1973) showed that the F test produced by MIXED in the case of complete data is, in effect, an approximation to the distribution of the multivariate Lawley-Hotelling trace statistic. It is a very poor approximation in small samples, rejecting the null hypothesis much too frequently. Other F approximations are more accurate. The accuracy of the various F tests is assessed via simulation.

1. Introduction

The traditional univariate mixed-model approach to repeated measures analysis produces valid results only when covariances among the repeated measurements are assumed to have a certain specific pattern, the "type-H" structure of Huynh & Feldt (1970) which includes the more familiar compound symmetry structure. One alternative to relying on this assumption is to use the Greenhouse-Geisser (GG) and Huynh-Feldt (HF) estimates of Box's (1954) adjustment to the degrees of freedom of the univariate mixed-model F tests as given in Greenhouse & Geisser (1959) and Huynh & Feldt (1976). Lecoutre (1991) gives a correction to the HF adjustment. Another approach is to use multivariate test statistics as given, for example, by Morrison (1990, section 5.6). Both of these approaches are implemented in SAS® in the GLM procedure using the REPEATED statement, but GLM uses only complete observations. A major attraction of the MIXED procedure is its ability to analyze repeated measures data having missing values. This paper shows how the GG, HF, and multivariate approaches can be applied to results from the MIXED procedure. From Kleinbaum (1973), it can be seen that the F test formula used by MIXED is, in the case of complete data, a scalar multiple of the multivariate Lawley-Hotelling trace statistic. In effect, the MIXED F is an approximation to the distribution of the Lawley-Hotelling trace — and a very poor one in small samples, tending to reject the null hypothesis much too frequently. Other F approximations are more accurate. The accuracy of the various F tests is assessed via simulation.

2. An Example: SAS/STAT GLM Procedure Example 7

In this example from Cole and Grizzle (1966), also used in the SAS/STAT User's Guide (SAS Institute Inc., 1989, p982-986) as an example of a repeated measures analysis with the GLM procedure, blood histamine levels were measured on sixteen dogs, assigned to four treatment groups, at 0, 1, 3, and 5 minutes after administration of a drug. The four groups formed a 2-by-2 between-subjects design, but for simplicity, it treat it as a one-way 4-group design. There was a missing value at one time point for one of the dogs, so only 15 observations (60 measurements) are available for analysis by GLM. The data are available from the SAS® Sample Library. The GLM analysis is shown below. The within-subjects measurements, Y1–Y4, are the natural logarithms of the histamine levels. For simplicity, the measurements are treated as if they were equally spaced in time.

```
proc glm data=dogs;
  class group;
  model y1 y2 y3 y4-group /nouni;
  repeated time 4 /short;
```

Results are shown below. "G'T" is the Group-by-Time interaction. "Univ" and "Mutt" refer to univariate and multivariate approaches, respectively. The adjustment factors for the univariate approach are GG=.5694, HF=.8475, and Lec=.6636. "Lec" is the corrected version of the H-F adjustment given by Lecoutre (1991). This adjustment is not given by GLM; it was calculated using MIXED and SAS/IML® software (see below). The simulations reported in section 3 generally indicate that GG tends to underestimate the true adjustment factor, HF tends to overestimate it (even when truncated to a maximum of 1), and Lec tends to overestimate slightly but by a much smaller amount.

```
Source   Method  F   DF  Prob
Group
  Univ     9.01  3, 1  11  .03589
  Univ GG  53.44  3, 3  33  9.2E-13
  Univ HF  25.88  3, 9  23  1.8E-12
  Univ Lec 24.03  3, 9      .00012
  Mutt     25.88 9, 33  81  1.8E-12
  Univ     25.88 9, 33  81  1.8E-12
  Univ GG  7.1E-9
  Univ HF  7.1E-9
  Univ Lec 7.5E-11
  Mutt     11.81  9, 23  10  1.0E-6
```
The same results can be obtained using the MIXED and IML procedures as shown below. First the data set must be restructured for use by MIXED. The MEANS procedure is used to obtain and store the total number of measurements. The "where id=6;" statement in the MEANS procedure and in the MIXED procedures below causes the dog which had a missing data value to be omitted.

```
data dogs2; set dogs; id+1;
time=1; y=yl; output;
time=2; y=y2; output;
time=3; y=y3; output;
time=4; y=y4; output;
keep id group time y;
proc means noprint; var y;
where id=6;
output out=totalN(drop=_type__freq__fr)
    n=ntotal, run;
```

For the univariate approach to repeated measures, two MIXED runs are needed, one with a compound symmetry covariance matrix (type-CS) to obtain the univariate results and a second run with an unstructured covariance matrix (type-UN) from which the univariate adjustments are calculated. Multivariate results are also obtained from the second run. The MAKE statements store needed matrices and results for use by IML. The dfbw option on the PROC statement is needed to get correct degrees of freedom for between subjects factors (this will be the default in a future release of MIXED).

```
proc mixed data=dogs2 dfbw;
    where id=6;
    class id group time;
    model y=group time group*time;
    repeated time / subject=id type=CS;
    make 'Tests' out=cs_tests;
    proc mixed data=dogs2 dfbw;
    where id=6;
    class id group time;
    model y=group time group*time;
    repeated time / subject=id type=UN;
    make 'Tests' out=un_tests;
    run;
```

Results from the MIXED runs are then fed into IML for final computations.

```
proc iml;
*** Nr groups, times, data values ***;
    ng=4; nt=4;
    use totalN; read all into ntotal;
    close totalN;
*** Between, within, error df ***;
    nobsn=ntotal/nt; dfb=ng-1;
    dfw=nt-1; dfw=nobs-ng;
    print ntotal nobsn ng dfb dfw dfe;
*** Univariate Adjustments ***;
    use Rmatrix; read all into R;
    close Rmatrix;
    nt=((nt+1)/2); F=sqrsym(R[1:nt,1]);
    rsum=sum(R);
    GG=(trace(R)-rsum/nt)##2;
    GG=GG/(nt-1)##(sqrg(R)-2*sqrg(R[.,1]/nt+(rsum/nt)##2));
    HF=(nobsn+dfw)##G
    /(dfw*(nobsn-ng-dfw)##G));
    Lec=(nobsn+dfw)##G
    /(dfw*(nobsn-ng-dfw)##G));
    print GG HF Lec;
    HF=1<HF; Lec=1<Lec;
    GG=1/((GG/HF-1)/HF/HF);
    Lec=1/Lec/Lec;
    use cs_tests; read all into univ;
    close cs_tests;
    un="(Group" "Time" "Group*Time")
    cn=="OF1" "DF2" "F" "Prob>F" "GG_Prob"
    "HF_Prob" "Lec_Prob";
    F=univ[,3];
    df1=univ[,1]; df2=univ[,2];
    U=1-fProb(f,GG,df1,GG,df2);
    univ=univ/(1-fProb(f,GG,df1,GG,df2));
    univ=univ/(1-fProb(f,HH,df1,HH,df2));
    univ=univ/(1-fProb(f,Lec,df1,Lec,df2));
    reset none;
    print "Univariate Tests",
        univ[rowname=nn colname=cn];
*** Multivariate Tests ***;
    use un_tests; read all into univ;
    close un_tests;
    F=multiv[,3];
    df1=multiv[,1]; df2=multiv[,2];
    U=1-fProb(f,multiv[,3]);
    *** U-Lawley-Hotelling trace ***;
    multiv=U[multiv];
    cn=="Trace" "DF1" "DF2" "F" "Prob>F";
    print "MIXED F Tests",
        multiv[rowname=nn colname=cn];
        dfe=univ[,1]; df1=univ[,2];
        univ=univ/(1-fProb(f,multiv[,3]));
        multiv[,4]=U[multiv[,3]]
        /(df1*df2);
        multiv[,5]=1-probf(
            multiv[,4],multiv[,2],multiv[,3]);
        print "Multivariate Tests"
            "(Pillai-Samson F)"
            multiv[rowname=nn colname=cn];
        B=((dfb+K)##(dfbw+K-1)/((K+1)##(K-2));
        multiv[,3]=4+(2*multiv[,2])/(B-1);
        s=multiv[,3]-2/K;
        multiv[,4]=U[multiv[,3]]
        /(s*multiv[,2]);
        multiv[,5]=1-probf(
            multiv[,4],multiv[,2],multiv[,3]);
        print "Multivariate Tests"
            "(McKeon F)"
            multiv[rowname=nn colname=cn];
        quit;
```

The results are exactly as obtained by GLM.

1155
Additional results for the within-subjects effects are shown below. Method "Mixed" refers to results reported by the MIXED procedure. Two additional F approximations to the Lawley-Hotelling trace statistic are calculated. "Mult-PS" is the traditional F approximation of Pillai and Samson (1959) used by GLM. It is the same as the "Mult" results given above. "Mult-M" uses the more accurate F approximation given by McKeon (1974). For the "Time" effect, in this example, these two approximations are the same.

<table>
<thead>
<tr>
<th>Source</th>
<th>Method</th>
<th>F</th>
<th>DF</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mixed</td>
<td>29.37</td>
<td>3,33</td>
<td>1.9E-9</td>
</tr>
<tr>
<td></td>
<td>Mult</td>
<td>24.03</td>
<td>3,9</td>
<td>.00012</td>
</tr>
<tr>
<td>G*T</td>
<td>Mixed</td>
<td>16.94</td>
<td>9,33</td>
<td>5.4E-10</td>
</tr>
<tr>
<td></td>
<td>Mult-PS</td>
<td>11.81</td>
<td>9,23</td>
<td>1.0E-6</td>
</tr>
<tr>
<td></td>
<td>Mult-M</td>
<td>13.09</td>
<td>9,11.33</td>
<td>9.2E-5</td>
</tr>
</tbody>
</table>

The same analysis was redone without the "where id^=6;" statements so that all available data would be used, that being one of the particularly attractive features of the MIXED procedure. Note that fractional degrees of freedom occur as the result of my defining nobs (the number of multivariate observations) as ntotal/nt (total number of measurements divided by number of times at which measurements were made). For this example, nobs = 63/4 = 11.75.

<table>
<thead>
<tr>
<th>Source</th>
<th>Method</th>
<th>F</th>
<th>DF</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univ</td>
<td>4.31</td>
<td>3,12</td>
<td>.02788</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>4.37</td>
<td>3,12</td>
<td>.02680</td>
</tr>
<tr>
<td></td>
<td>Mult</td>
<td>4.37</td>
<td>3,11.75</td>
<td>0.2742</td>
</tr>
<tr>
<td>Time</td>
<td>Univ</td>
<td>59.32</td>
<td>3,35</td>
<td>8.4E-14</td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lec</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>31.62</td>
<td>3,35</td>
<td>4.5E-10</td>
</tr>
<tr>
<td></td>
<td>Mult</td>
<td>26.24</td>
<td>3,9.75</td>
<td>5.5E-5</td>
</tr>
<tr>
<td>G*T</td>
<td>Univ</td>
<td>28.70</td>
<td>9,35</td>
<td>1.4E-13</td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lec</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>17.82</td>
<td>9,35</td>
<td>1.3E-10</td>
</tr>
<tr>
<td></td>
<td>Mult-PS</td>
<td>12.76</td>
<td>9,25.25</td>
<td>2.3E-7</td>
</tr>
<tr>
<td></td>
<td>Mult-M</td>
<td>13.99</td>
<td>9,12.48</td>
<td>3.4E-5</td>
</tr>
</tbody>
</table>

The univariate adjustment factors are GG = 0.5806, HF = 8.470 and Lec = 6.730.

Notice the three different results for the between-subjects factor "Group." When there is no missing data, the MIXED F test for "Group" will be the same for both type = CS and type = QN covariance structures. When there is missing data, this is no longer the case. Also, when there is no missing data, the MIVQUE0 and REML methods give the same results. When there is missing data, this is no longer the case. The above results were obtained with the default REML method. Some of these differences were investigated by simulation.

3. A Simulation Study

To determine the accuracy of the various tests in a small sample, a simulation study was undertaken. The design consisted of three (3) groups (the between-subjects factor), five (5) observations in each group for a total of n = 15 observations, four (4) occasions or "times" (the within-subjects factor) for a total of N = 60 measurements. The MIXED and IML analysis was conducted on 1000 datasets, and rejections of the null hypothesis were tabulated for significance levels .005, .01, .05, .10, .25, and .50. Four different covariance structures were used to generate the data: compound symmetry (CS), first-order autocorrelation (AR1), Toeplitz (Toep), and unstructured (Unst). When the true covariance structure was AR1 or Toeplitz, multivariate test results were obtained once by having MIXED fit an unstructured covariance matrix and a second time by fitting the correct form of covariance matrix (i.e., AR1 or Toeplitz). The analysis was done for the case of no missing data and again with three (3) missing measurements. Because of the computational demands of the REML method, both REML and MIVQUE0 were used to see whether adequate results would be obtained by the much less demanding MIVQUE0 method. Complete results are available from the author. Only some of the highlights are discussed here.

Conclusions from the Simulation Study: Figures 1 through 10 show some of the simulation results for the G*T interaction. Results for the Time main effect were similar. On the vertical axis is the observed rate of rejection of the null hypothesis for significance levels 0.005, 0.01, 0.05, and 0.10 (plotted on the horizontal axis). A solid black Reference Line marks where the observed error rate equals the nominal rate. In order to use the same vertical scale in all figures while maintaining reasonable resolution, rejection rates larger than 0.20 were omitted from the figures.

The most important conclusion from the study is that the unadjusted F tests reported by the MIXED procedure can have horribly inflated type I error rates in small samples. This is evident in nearly all the accompanying graphs. Since the MIXED F test essentially pretends that the covariance matrix is known, its behavior is more egregious when more covariance terms are estimated, for example, when an unstructured (10 covariance parameters) or Toeplitz (4 parameters) matrix is estimated. The one case (other than CS) where MIXED performs well (at least with REML estimation) is when the true matrix is AR1 (only two parameters) and the estimated matrix is also AR1 (Figure 4).

The good news is that the use of appropriate adjustment procedures largely solves the inflated error rate problem. For the circumstances covered in this study, the adjusted univariate tests controlled the error rate quite well. GG in particular did well, but Lec (not shown in the graphs) did almost as well. The multivariate approach works well when the estimated covariance matrix is unstructured, but this approach falters, generally becoming too conservative, when a structured covariance matrix is fitted (e.g., Figures 4 & 5). This makes sense since the multivariate approach assumes that the estimated
covariance matrix (times a constant) has a Wishart distribution. This will not be true when the estimated covariance matrix is constrained to fit a certain structure.

The bad news is that, for both univariate and multivariate approaches, control of the error rate occurred only with REML estimation, not with MIVQUE0. See Figures 6-10. Unfortunately, in addition to requiring much more computer time, REML rather frequently failed to produce a solution when there was missing data. The number of failures varied from as few as 3 (out of 1000 simulated datasets) when the true covariance was CS and the fitted matrix was unstructured, to as many as 127 when the true matrix was Toeplitz and the fitted matrix was unstructured.

Unanswered questions: Given the frequent failure of REML to obtain estimates, a fruitful line of research might be to investigate alternate methods of estimation, preferably non-iterative ones since, even when REML works, it can take quite a long time.

The MIXED procedure, of course, applies to models much more general than those of repeated measures analysis. In particular, the approaches given here do not address the problem of random effects in the model. A “final solution” to the problem of inference in the mixed model (including repeated measures and multivariate applications of the mixed model) may very well require a Bayesian approach. See Searle, Casella and McCulloch (1992, chapter 9) and Wolfinger, Rosner and Kass (1994).

REFERENCES


ACKNOWLEDGEMENTS

The analyses and simulations were done with the MIXED procedure in SAS/STAT version 6.09 running on an IBM® RS/6000 Model 590 workstation at the Value-Added Research and Assessment Center at the University of Tennessee, Knoxville. Figures were prepared using JMP® version 3.0.2.

JMP, SAS, SAS/IML, and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and other countries. IBM is a registered trademark of International Business Machines Corporation. ® indicates USA registration.

Author Contact: S. Paul Wright, 328 SMC – Statistics Dept., University of Tennessee, Knoxville, TN 37996-0532. Phone: (615) 974-2556. E-mail: wrightsp@utkvx.utk.edu.

FIGURES

In Figures 1-5 all results are for REML estimation. MIVQUE0 gives identical results whenever there is complete data and the estimated covariance matrix is unstructured. In Figures 6-10, (R) indicates REML and (Q) indicates MIVQUE0.

Legend for Figures 1-5:

+ — Mixed
♀ — Univ
♂ — G-G
♀ — McKeon
♂ — Reference Line
Figure 6. G'T: True=CS, Fitted=Unst, 3 Missing

Figure 7. G'T: True=AR1, Fitted=AR1, Complete Data

Figure 8. G'T: True=AR1, Fitted=AR1, 3 Missing

Figure 9. G'T: True=Toep, Fitted=Toep, Complete Data

Figure 10. G'T: True=AR1, Fitted=AR1, 3 Missing

Legend for Figure 10:

- Univ (R)
- Univ (Q)
- G-G (R)
- G-G (Q)
- Reference Line