0. Abstract

In a two factor (treatment and investigator) clinical trial design, "pooling" usually refers to a model reduction from the full cross classification to a "main effects" model. The removal of the treatment by investigator interaction terms is typically decided by the size of the p-value for the test of no interaction. This decision process historically is due to Type I and Type II error considerations. In clinical trials, possible bias incurred by model reduction is of prime concern. In this paper, parameterization corresponding to SAS® Type I, Type II, and Type III sums of squares is illustrated and the decision to pool is examined from a bias perspective. An inspection of mean square errors in the estimation of treatment differences suggests that when the F statistic for the interaction hypothesis test exceeds unity, pooling should not be done.

1. Introduction

Early work on the question of pooling in Analysis of Variance detailed its impact on Type I and Type II errors [3,7]. These papers focused on the equal sample size situation. When sample sizes are equal in the R×G cross classification, all point estimators (SAS® Type I, II, III) are equal and bias in estimation is not a concern. When pooling in this case, one simply trades off the risk of an inflated $\sigma^2$ for additional degrees of freedom. Of course the impact of a two step approach on p-values is also a concern. This early work led to recommendations for pooling when the p-value for that preliminary test was 0.5 [1], 0.25 [8], or even as low as 0.10 [6]. In the clinical trial setting however, the sample sizes among "cells" in the two factor classification (treatment by investigator) are almost always unbalanced. The usual situation is near balance within investigator (due to blocked randomization) and disparity among investigators. In this situation, the question of whether or not to pool becomes more complex. Type I, II, and III estimators of treatment differences are usually different and possible bias becomes a question. Predictive bias in a regression context was examined in the variable selection problem by Allen [2]. Here, we will examine the use of M.S.E. as a criterion for pooling in the clinical trial setting.

A SAS® LSMEAN estimate of treatment difference under the full two factor (treatment and investigator) model gives equal weight to the treatment difference at each center. If there were no interaction term in the model, then the least squares estimator becomes a weighted average of the center differences with the larger centers having correspondingly larger weights. This weighted average may have substantially lower variance than the unweighted average. If, however, the weighted average is used when interaction is present, the resultant estimator is biased. In §2 below, the model parameters corresponding to SAS® Type I, Type II, and Type III sums of squares are defined for the two treatment, two investigator case along with the "treatment difference" of interest. In §3, the treatment estimators, their expectations, and their variances are provided for each parameterization. In §4, the mean squared error of the weighted estimator is contrasted with that of the unweighted estimator when interaction is present. The results of this calculation are discussed in §5.

2. Parameterizations

In this section, parameterizations for the two treatment, two investigator clinical trial are provided which describe the model under assumptions equivalent to those provided under SAS® Type I, Type II, and Type III sums of squares. To be more precise, when we specify the model,

```
MODEL OUTCOME: TRT INV TRT*INV / SS1 SS2 SS3;
```

three tests of no "treatment effect" are provided, one corresponding to each of Type I-III sums of squares. Each test statistic is of the form:

$$F = \frac{\hat{\beta}_{i}^{2}}{\text{var} (\hat{\beta}_{i})},$$

for $i = I, II, or III$.

Where $\hat{\beta}_{i}$ corresponds to the estimated treatment difference under model Type I. A parameterization for each model type and the definitions of $\hat{\beta}_{I}, \hat{\beta}_{II}, and \hat{\beta}_{III}$ are provided below:

A. Type III Model (No assumptions)

<table>
<thead>
<tr>
<th>INV1</th>
<th>INV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRT1</td>
<td>$\mu_{11}$</td>
</tr>
<tr>
<td>TRT2</td>
<td>$\mu_{11} - \beta_{1}$</td>
</tr>
</tbody>
</table>

TRT1-TRT2 $= \hat{\beta}_{1}$

Parameters: $\mu_{11}, \mu_{12}, \beta_{1}, \beta_{2}$

In this model, we define as an overall measure of treatment difference:

$$\Delta_{III} = \frac{1}{2} (\hat{\beta}_{1} + \hat{\beta}_{2})$$

Of course, if interactions are sizable, the treatment differences among clinics should be examined for qualitative versus quantitative
differences [4] and in any case, means for each investigator should be presented [5]. Here, $\Delta_{III}$ provides a single number to represent a "treatment difference". As it is an unweighted average of (population) treatment differences among centers and does not favor any one center more than another, it is a reasonable measure of the "difference" in the response between treatments [6].

B. Type II Model (No TRT*INV interaction)

<table>
<thead>
<tr>
<th>TRT</th>
<th>INV1</th>
<th>INV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRT1</td>
<td>$\mu_{11}$</td>
<td>$\mu_{12}$</td>
</tr>
<tr>
<td>TRT2</td>
<td>$\mu_{12} - \Delta_{II}$</td>
<td>$\mu_{12}$</td>
</tr>
</tbody>
</table>

$\Delta_{II} = \frac{n_{11} \bar{x}_{11} + n_{12} \bar{x}_{12} - n_{21} \bar{x}_{21} - n_{22} \bar{x}_{22}}{n_{11} + n_{12} - n_{21} - n_{22}}$

Parameters: $\mu_{11}, \mu_{12}, \Delta_{II}$

C. Type I Model (No INV effect and no TRT*INV interaction)

<table>
<thead>
<tr>
<th>TRT</th>
<th>INV1</th>
<th>INV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRT1</td>
<td>$\mu_1$</td>
<td>$\mu_1$</td>
</tr>
<tr>
<td>TRT2</td>
<td>$\mu_1 - \Delta_1$</td>
<td>$\mu_1$</td>
</tr>
</tbody>
</table>

$\Delta_1 = \frac{n_{11} \bar{x}_{11} + n_{12} \bar{x}_{12} - n_{21} \bar{x}_{21} - n_{22} \bar{x}_{22}}{n_{11} + n_{12} - n_{21} - n_{22}}$

Parameters: $\mu_1, \Delta_1$

3. Estimators, Expectations, and Variances

The least squares estimates of $\Delta_1, \Delta_{II},$ and $\Delta_{III}$ are provided below under the Type I, Type II, and Type III parameterizations respectively. Thus, these (squared) form the numerator of our test statistic under the corresponding SAS® Type sums of squares. They can be obtained from SAS® LSMEANS when the model statement indicates the appropriately reduced model. Note, however, that the F statistic for treatment under the reduced model will (whenever the sample sizes among the four groups vary) in general differ from the F statistic obtained under the full model with the sums of squares corresponding to the reduced model assumptions; this is because $\Delta_2$ in the denominator of the F statistic is estimated differently.

$\hat{\Delta}_1 = \left( \frac{n_{11} \bar{x}_{11} + n_{12} \bar{x}_{12} - n_{21} \bar{x}_{21} - n_{22} \bar{x}_{22}}{n_{11} + n_{12} - n_{21} - n_{22}} \right)$

$\hat{\Delta}_{II} = \frac{\bar{x}_{11} - \bar{x}_{21}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}}}} + \frac{(1-w) \left( \bar{x}_{12} - \bar{x}_{22} \right)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}}}}$

where $n_{ij}$ is the number of investigator j's patients receiving treatment i.

and $w = \frac{\frac{1}{n_{12}} + \frac{1}{n_{22}}}{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

$\hat{\Delta}_{III} = \left( \frac{\bar{x}_{11} - \bar{x}_{21}}{2} + \frac{\bar{x}_{12} - \bar{x}_{22}}{2} \right)$

Notes: a) When $n_{11} = n_{21} = n_1$ and $n_{12} = n_{22} = n_2$, then $w = \frac{n_1}{n_1+n_2}$ and $\hat{\Delta}_1 = \hat{\Delta}_{II}$. Also, $\hat{\Delta}_1 = \hat{\Delta}_{II} = \hat{\Delta}_{III}$.

The variances of the above estimators (assuming homoscedasticity) are easily calculated as:

$\text{var}(\hat{\Delta}_1) = \frac{\bar{x}_{11} - \bar{x}_{21}}{\sqrt{n_{11} + n_{12} - n_{21} - n_{22}}}$

$\text{var}(\hat{\Delta}_{II}) = \frac{\bar{x}_{11} - \bar{x}_{21}}{\sqrt{n_{11} + n_{12} - n_{21} - n_{22}}}$

$\text{var}(\hat{\Delta}_{III}) = \frac{\bar{x}_{11} - \bar{x}_{21}}{\sqrt{n_{11} + n_{12} - n_{21} - n_{22}}}$

I suspect in most cases

$\text{Var}(\hat{\Delta}_1) \leq \text{Var}(\hat{\Delta}_{II}) \leq \text{Var}(\hat{\Delta}_{III})$

When $n_{11} = n_{21} = n_1$, say, and $n_{12} = n_{22} = n_2$, then

$\text{var}(\hat{\Delta}_1) = \text{var}(\hat{\Delta}_{II}) = \frac{n_2^2}{(n_1 + n_2)^2} \leq \text{var}(\hat{\Delta}_{III}) = \frac{n_1^2 + n_2^2}{(n_1 + n_2)^2}$

Thus, in this case, when $n_1 = 2n_2$, the Type II standard error is only 6% less than the Type III standard error. When the ratio of sample sizes increases to 5:1, the reduction increases to 25%.

Expectations of the $\hat{\Delta}_i$ ($i= I, II, or III$) under the three parameterizations are provided below in Table 1.

4. MSE Calculation

In this section we compare $\hat{\Delta}_{II}$ and $\hat{\Delta}_{III}$ when the interaction is nonzero (i.e., in the Type III situation). Here, we compare mean squared errors since $\hat{\Delta}_{II}$ is a biased estimator of $\hat{\Delta}_{III}$. In this section, while not explicitly indicated, all MSE's and expectations (E's) are conditional on the Type III model. We have:

$\text{MSE}(\hat{\Delta}_1) = \text{MSE}(\hat{\Delta}_{II}) = \text{MSE}(\hat{\Delta}_{III})$

When $n_{11} = n_{21} = n_1$, say, and $n_{12} = n_{22} = n_2$, then

$\text{MSE}(\hat{\Delta}_1) = \text{MSE}(\hat{\Delta}_{II}) = \frac{n_2^2}{(n_1 + n_2)^2} \leq \text{MSE}(\hat{\Delta}_{III}) = \frac{n_1^2 + n_2^2}{(n_1 + n_2)^2}$

Thus, in this case, when $n_1 = 2n_2$, the Type II standard error is only 6% less than the Type III standard error. When the ratio of sample sizes increases to 5:1, the reduction increases to 25%.

Expectations of the $\hat{\Delta}_1$ ($i= I, II, or III$) under the three parameterizations are provided below in Table 1.
Through algebraic manipulation it can be shown that:

\[
\text{MSE} \left( \delta_{111} \right) = \text{var} \left( \delta_{111} \right) + \left( \delta_{111} - E \left( \delta_{111} \right) \right)^2
\]

\[
= \sigma^2 \left[ \frac{1}{n_{11}} + \frac{1}{n_{21}} \right] + (1-\omega)^2 \left( \frac{1}{n_{12}} + \frac{1}{n_{22}} \right) + (\omega - \theta)^2 \left( \delta_{1-2} \right)^2
\]

and

\[
\text{MSE} \left( \delta_{111} \right) - \text{var} \left( \delta_{111} \right) = \sigma^2 \left[ \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right].
\]

Through algebraic manipulation it can be shown that:

\[
\frac{\text{MSE} \left( \delta_{111} \right)}{\text{MSE} \left( \delta_{11} \right)} = \frac{1}{1 + (2\omega-1)^2 \left[ \frac{\left( \delta_{1-2} \right)^2}{\text{var} \left( \delta_{1-2} \right)} - 1 \right]}
\]

where \( \delta_{1-2} \) is the interaction and \( \delta_{1-2} \) is its least squares estimator under the Type III model.

For estimation, we should keep the interaction term when

\[
\frac{\text{MSE} \left( \delta_{111} \right)}{\text{MSE} \left( \delta_{11} \right)} < 1
\]

or, equivalently, when

\[
\frac{\left( \delta_{1-2} \right)^2}{\text{var} \left( \delta_{1-2} \right)} > 1
\]

or approximately, when the F statistic for interaction exceeds unity.

**TABLE 1**

EXPECTATIONS OF THE \( \delta_i \) WITH DIFFERENT UNDERLYING MODELS

<table>
<thead>
<tr>
<th>Underlying Model</th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_1 )</td>
<td>( \Delta_1 )</td>
<td>( \Delta_{11} + (w_1-w_2)(\mu_{11}-\mu_{12}) )</td>
<td>( \Delta_{11} + (w_1-w_2)(\mu_{11}-\mu_{12}) + (w_2-1/2)(\delta_{1-2}) )</td>
</tr>
<tr>
<td>( \delta_{11} )</td>
<td>( \Delta_{11} )</td>
<td>( \Delta_{11} )</td>
<td>( \Delta_{11} + (w-1/2)(\delta_{1-2}) )</td>
</tr>
<tr>
<td>( \delta_{111} )</td>
<td>( \Delta_{11} )</td>
<td>( \Delta_{11} )</td>
<td>( \Delta_{11} )</td>
</tr>
</tbody>
</table>
Historically, the question of whether to drop an interaction term from an ANOVA model centered on power considerations. This early work led to the use of significance levels for the interaction test as a basis for the decision to pool the interaction sums of squares with the error sums of squares. In these studies, bias in the numerator was not a concern since only the balanced case was considered. Today, it is common to pool if the p-value exceeds 0.15 or some other moderate value even in the unbalanced situation. The results from this paper are:

1. For clinical trials which are 'block randomized' within center (and thus have approximately equal sample sizes among treatments within centers) there is no gain in attempting to remove investigator as an effect from a Type II model. Any 'improvement' in the statistical significance level from such a model reduction is likely due to bias.

2. There is little to be gained in removing the interaction term from a Type III model unless the total sample sizes among investigators exceeds a 2:1 ratio.

3. When the sample size ratio between investigators exceeds 2:1 (and thus a potential for non-negligible variance reduction exists), it is recommended that pooling be done when the F statistic for no interaction is less than unity. This is nearly the same as using a p-value of 0.7 for the pool decision in the two treatment, two investigator situation.

The results above (especially number 3) are at odds with common practice today. Work should be done exploring the impact of deviating from these results and in extending these results to the k-investigator case.

References


SAS® is the registered trademark of SAS Institute, Inc., Cary, North Carolina, USA

Author Contact

Gregory T. Schwemer
Boehringer Mannheim Pharmaceuticals Corporation
1301 Piccard Drive
Rockville, Maryland 20850