Using R to Find a Minimal Model for a Factorial ANOVA

What I will demonstrate in this section is how to perform an ANOVA where we will not try to fit the (one) model to the data, but instead fit the data to the model by searching for the best fit of the data. This procedure is identical to what was seen in Section 7.4.5 in the book on regression, since ANOVA in R is modeled as a regression equation, and then commands like Anova() reformat the results into an ANOVA table.

As with regression, it makes more sense to find the minimally adequate model for the data and only later examine plots that evaluate the suitability of assumptions such as homogeneity of variances, normality of distribution, etc. I will assume that my reader has already looked at the online documents for Chapter 7 on finding the minimally adequate model in regression but, as a review, the steps I will follow to find the minimally adequate model are (following Crawley, 2007):

1. Create a full factorial model.
2. Examine the output for statistical terms.
3. Create a new model that deletes non-statistical entries, beginning with largest terms first and working backwards to simpler terms.
4. Compare the two models, and retain the new, simpler model if it does not cause a statistical increase in deviance.

Using the Obarow (2004) data contained in Obarow.Story1 (imported as obarow into R for ease of writing here), I will first create the full factorial model. Remember that, instead of writing out the entire seven-parameter model (with 1 three-way interaction, 3 two-way interactions, and 3 main effects), using the asterisk (*) notation creates the full model.

```
attach(obarow)
model=aov(gnsc1.1~gender*musict1*picturest1)
summary(model)
```

The summary() command produces an ANOVA table. This summary shows that there is only one statistical entry, the main effect of gender. The choice of aov() instead of lm() for the regression model means that the summary() command will return an ANOVA table instead of regression output.

Although only one variable is statistical, just as we saw with regression modeling, the best way to go about our ANOVA analysis is to perform stepwise deletion, working backwards by removing the largest interactions first, and testing each model along the way. We could simply stop after this first pass and proclaim that gender is the only variable that is statistical, but that would not be finding the minimally adequate model. I suggested in the documents for Chapter 7 on regression and I will also suggest here that finding the minimally adequate model is a much more informative analysis than a simple full factorial model with reporting on which main effects and interactions are statistical. What we will find is that, by performing the minimally adequate model, the stepwise removal of variables may change the analysis along the way, since

order matters in a dataset where the number of persons in groups is not strictly equal (as is true in this case).

Now we will simplify the model by removing the least significant terms, starting with the highest-order terms first (in our case, we'll remove the third-order interaction first).

```
model2=update(model,~.-gender:musict1:picturest1)
```

Remember, the syntax here must be exactly right! The syntax in the parentheses that comes after the name of the model to update (here, called just model) is "comma tilde period minus." This syntax means that model2 will take the first argument (model) as it is and then subtract the three-way interaction. Now we compare the updated model to the original model using anova().

```
anova(model,model2)
```

The fact that the ANOVA is not statistical tells us that our newer model2 does not have statistically higher deviance than the original model, and thus it is not worse at explaining what is going on than the original model. We prefer model2 because it is simpler. We can now look at a summary of model2 and decide on the next least statistical argument to take out from among the two-way interactions:

```
summary(model2)
```

The two-way interaction with the highest *p*-value is gender:picturest1, so we will remove that in the next model.

```
model3=update(model2,~.- gender:picturest1)
anova(model2,model3)
```

Again, there is no difference in deviance between the models, so we will leave this interaction term out. Examining the summary of model3:

```
summary(model3)
```

we see that the next interaction term with the highest *p*-value is musict1:picturest1. I will leave it to the reader to verify that, by continuing this procedure, you will reach model7, which contains only the variable of gender. At this point, we should compare a model with only gender to the null model, which contains only the overall average score and total deviance. The way to represent the null model is simply to put the number 1 as the explanatory variable after the tilde. If there is no difference between these two models, then what we have found is that none of our variables do a good job of explaining what is going on!

```
model8=aov(gnsc1.1~1)
anova(model7,model8)
```

Fortunately, here we find a statistical difference in the amount of deviance each model explains, so we can stop and claim that the best model for the data is one with only one explanatory variable, that of gender.

```
> summary(model7)
            Df Sum Sq Mean Sq F value Pr(>F)
gender       1  11.86  11.863   5.286 0.0249 *
Residuals   62 139.14   2.244
```

The regression summary will give more information:

```
> summary.lm(model7)

Call:
aov(formula = gnsc1.1 ~ gender)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5294 -0.8824  0.3333  1.3333  4.4706

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.5294     0.2569   5.953 1.34e-07 ***
gender[T.female] -0.8627     0.3752  -2.299   0.0249 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.498 on 62 degrees of freedom
Multiple R-squared:  0.07856,   Adjusted R-squared:  0.0637
F-statistic: 5.286 on 1 and 62 DF,  p-value: 0.02488
```

The multiple R^2 value shows that the variable of gender accounted for about 8% of the variance in scores, a fairly small amount. Notice that the F-statistic and *p*-value given in the regression output are exactly the same as in the ANOVA table, but this is only because there is only one term in the equation.

A shortcut to doing the stepwise deletion by hand is the command boot.stepAIC( ) from the bootStepAIC package with the original full factorial model (as mentioned in the Chapter 7 online materials, this function is much more accurate and parsimonious than the step command found in R's base package):

```
library(bootStepAIC)
boot.stepAIC(model,data=obarow) #need to specify dataset even though attached!
```

The full model has an **AIC** (Akaike information criterion) score of 60.66. As long as this measure of fit goes *down*, the step procedure will continue to accept the more simplified model, as we did by hand. In other words, the lower the AIC, the better (the AIC is a measure of the tradeoff between degrees of freedom and the fit of the model). The step procedure's last step is to stop when there are still three terms in the equation. You can always check out the AIC value of any model by using the command AIC(model). Doing this I find that the AIC for model 7 is 237.3, while for model 9 it is 239.0.

Let's check out the difference between this model that boot.stepAIC kept (I'll create a model9 for this) and the model we retained by hand (model7).

```
model9=aov(gnsc1.1~gender+musict1+gender:musict1)
summary.lm(model9)
```

data deleted . . .

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.2500     0.3343   3.739 0.000414 ***
gender[T.female]                  -0.3333     0.5459  -0.611 0.543751
musict1[T.music]                   0.6786     0.5209   1.303 0.197696
gender[T.female]:musict1[T.music] -1.0952     0.7628  -1.436 0.156223
```

data deleted . . .

```
Residual standard error: 1.495 on 60 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.06755
F-statistic: 2.521 on 3 and 60 DF,  p-value: 0.06631
```

We find that model9 has a slightly smaller residual standard error on fewer degrees of freedom than model7 (1.495 for model9, 1.498 for model7), and a higher R2 value (11.2% for model9, 7.9% for model8), but a model with more parameters will always have a higher R2 (there are 64 participants in the Obarow study, and if we had 64 parameters the R2 fit would be 100%), so this is not surprising.

Just to make sure what we should do, let's compare model7 to model9 using anova().

```
anova(model7,model9)
```

An ANOVA finds no statistical difference between these models. All other things being equal, we prefer the model that is simpler, so in the end I will retain the model with only the main effect of gender.

The message here is that, although boot.stepAIC can be a useful tool, you will also want to understand and know how to hand-calculate a stepwise deletion procedure. Crawley (2007) notes that step procedures can be generous and leave in terms that are not statistical, and you will want to be able to check by hand and take out any terms which are not statistical. Another situation where a stepping command is not useful is when you have too many parameters relative to your sample size, called overparameterization (see the section "Further steps in finding the best fit: Overparameterization and polynomial regression" in the Chapter 7 online document "Finding the Best Fit in Multiple Regression" for more information on this). In this case, you should work by hand and test only as many parameters as your data size will allow at a time.

Now that the best fit of the data (so far!) has been determined, it is time to examine ANOVA (regression) assumptions by calling for diagnostic plots:

```
plot(model7)
detach(obarow)
```

Plotting the model will return four different diagnostic plots, but we've been over these in Section 10.5.2 of the book, so I won't explain them here.

Using R to find a Minimal Model for a Factorial ANOVA

There is no simple formula for finding the minimal model, but there are steps to take:

1 Create a full factorial model of your data using either aov() or lm()
(N.B. items in red should be replaced with your own data name)
model1=aov(gnsc1.1~gender*musict1*picturest1, data=obarow)

2 Begin to look for the best model that fits the data. You might start with the step function from the bootStepAIC library:
boot.stepAIC (model1)

3 Confirm boot.stepAIC's choices with your own stepwise deletion procedure:

a. Look at ANOVA table of model (use anova()) and decide which one term to delete in the next model.

b. Choose the highest-order term to delete first. If there is more than one, choose the one with the highest *p*-value. If an interaction term is statistical, always retain its component main effects (for example, if gender*musict1 is statistical, you must keep both gender and musict1 main effects).

c. Create a new model by using update:
model2=update(model1,~.-gender:musict1:picturest1, data=obarow)

d. Compare the new and old model using anova:
anova(model1, model2)

e. If there is no statistical difference between models, retain the newer model and continue deleting more terms. If you find a statistical difference, keep the older model and you are finished.

f. Finally, you may want to compare your model with the null model, just to make sure your model has more explanative value than just the overall mean score:
model.null=lm(gnsc1.1~1,data=obarow)

4 4. When you have found the best fit, check the model's assumptions of normality and constant variances:
plot(maximalModel)

Bibliography

Crawley, M. J. (2007). *The R book*. New York: Wiley.

Obarow, S. (2004). *The impact of music on the vocabulary acquisition of kindergarten and first grade students*. Unpublished Ph.D., Widener University, Chester, PA.