Two-way ANOVA

Analysis of variance models can be generalized to more than one grouping variable. Say there are two such variables: one representing rows having \( I \) categories, and one representing columns having \( J \) categories. The two-way ANOVA model has the form

\[
y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, n_{ij}.
\]

The \( \alpha_i \) and \( \beta_j \) parameters represent the main effects of rows and columns, respectively, and have the same general interpretation as the effect in a one-way ANOVA does. The \((\alpha\beta)_{ij}\) represents an interaction effect. What does the interaction effect mean? Consider the following hypothetical example. Say it was desired to investigate the relationship between the mileage of new automobiles (in miles per gallon) and the year the auto was built (2008, 2013, or 2018) and the size of the auto (Small, Midsize, or Large). An experiment is run where five autos of each (Year, Size) pair have their mileage tested (that is, five small 2008 autos; five small 2013 autos; five small 2018 autos; five midsize 2008 autos; etc.). The following table gives the average mileage figures for the five cars of each pair:

<table>
<thead>
<tr>
<th>Size</th>
<th>2008</th>
<th>2013</th>
<th>2018</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>26</td>
<td>29</td>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>↓6</td>
<td>↑0</td>
<td>↑1</td>
<td></td>
</tr>
<tr>
<td>Midsize</td>
<td>20</td>
<td>29</td>
<td>31</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>↑0</td>
<td>↑8</td>
<td>↑1</td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>20</td>
<td>21</td>
<td>30</td>
<td>23.7</td>
</tr>
<tr>
<td></td>
<td>↑1</td>
<td>↑9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>22</td>
<td>26.3</td>
<td>31</td>
<td>26.4</td>
</tr>
</tbody>
</table>

Broadly speaking, the main effects of size and year are clear: smaller autos get better mileage, as do newer autos. The interaction effect, however, says that this is not really the case; rather, the year effect is different for different sized autos; equivalently, it says that the size effect is different for different year autos. That is, knowing the size of an auto tells you something about the year effect, or knowing the year of an auto tells you something about the size effect (in this sense, they interact with each other).
very real sense, therefore, the main effects don’t have any real meaning, since each changes
derpending on the level of the other variable. The numbers above the horizontal arrows and
next to the vertical arrows reflect those *conditional* effects. What would the interaction
effect look like for these data? An easy way to see this is by constructing an *interaction
plot*. In this plot the vertical axis represents the group averages of the target variable, with
the categories of one effect along the horizontal axis, and different lines representing the
categories of the other effect. Here is an interaction plot for the table above:

If no interaction effect is present in the data, the lines in the interaction plot should be
roughly parallel, so lack of parallelism in the lines supports the existence of an interaction
effect. The plot makes clear what the interaction effect is saying. Small autos have steadily
increasing mileage from 2008 to 2013 to 2018; midsize autos exhibit a big jump in mileage
from 2008 to 2013, followed by a slight increase to 2018; and large autos exhibit a slight
increase from 2008 to 2013, followed by a big jump from 2013 to 2018.

While examination of main effect and interaction plots (that is, groups means) helps
to understand what the different effects mean, they are not adequate to test whether those
effects could just be due to random chance. For that, we need formal tests of hypotheses.
These tests are just partial *F*-tests, based on the appropriate regression models (models
that are based on effect codings). ANOVA packages or functions will generally do this

© 2019, Jeffrey S. Simonoff
automatically, but it's worth knowing what tests are being constructed in case you are using a package that does not include the general linear model. The three hypotheses of interest here are whether there is a row effect,

\[ H_0 : \alpha_1 = \cdots = \alpha_I = 0, \]

whether there is a column effect,

\[ H_0 : \beta_1 = \cdots = \beta_J = 0, \]

and whether there is an interaction effect,

\[ H_0 : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{IJ} = 0. \]

The appropriate regression model has as predicting variables effect codings as follows:

**Year effect**

\[
\begin{align*}
Y08 &= \begin{cases} 1 & \text{if 2008 auto} \\
0 & \text{if 2013 auto} \\
-1 & \text{if 2018 auto} \end{cases} \\
Y13 &= \begin{cases} 0 & \text{if 2008 auto} \\
1 & \text{if 2013 auto} \\
-1 & \text{if 2018 auto} \end{cases}
\end{align*}
\]

**Size effect**

\[
\begin{align*}
SS &= \begin{cases} 1 & \text{if small auto} \\
0 & \text{if midsize auto} \\
-1 & \text{if large auto} \end{cases} \\
SM &= \begin{cases} 0 & \text{if small auto} \\
1 & \text{if midsize auto} \\
-1 & \text{if large auto} \end{cases}
\end{align*}
\]
Interaction effect

Four variables that correspond to the pairwise products of the effect codings above; that is,

\[ Y_{08SS} = Y_{08} \times SS \]
\[ Y_{08SM} = Y_{08} \times SM \]
\[ Y_{13SS} = Y_{13} \times SS \]
\[ Y_{13SM} = Y_{13} \times SM \]

In general, there are \( I - 1 \) variables to represent the row effect, \( J - 1 \) variables to represent the column effect, and \( (I - 1)(J - 1) \) variables to represent the interaction effect.

Once these variables are created, the partial \( F \)-tests are constructed in the usual way. That is, the full model is a regression on all 8 variables above, while the subset models omit \( Y_{08} \) and \( Y_{13} \) (to test the year effect), \( SS \) and \( SM \) (to test the size effect), and \( Y_{08SS} \), \( Y_{08SM} \), \( Y_{13SS} \), and \( Y_{13SM} \) (to test the interaction effect), respectively. Usually the \( t \)-statistics for the individual effect coding variables are not of interest, since it is the entire row or column effect that is being tested (remember: it takes \( K - 1 \) effect codings to represent one effect that involves \( K \) categories).

The first partial \( F \)-test to examine is that for the interaction effect. Tests for main effects should not be examined in a model that includes an interaction effect (even if the partial \( F \)-test for that interaction is not statistically significant), as it is not meaningful; tests for main effects should only be examined if the interaction effect is not included as part of the fitted model.

What if the interaction effect is not statistically significant? If only one of the main effects is needed, the problem has collapsed down to a one-way ANOVA; otherwise, it is a two-way ANOVA where the main effects are meaningful and are viewed conditionally on the presence of the other. So, in this hypothetical example a model that included both main effects and no interaction effect would be referring to differences between the mileages of 2008, 2013, and 2018 autos given the size of the auto, and differences between the mileages of small, midsize, and large autos given the model year of the auto. You should not fit a model that includes an interaction without the corresponding main effects, since then the fitted values for the observations in the \( (i, j) \)th cell will not be the mean target value for that cell (most packages will not let you fit such a model, since there is no
unique way to define the fitted values in this case).

Are multiple comparisons still an issue? Yes and no. As noted above, if the interaction effect is not in the model, comparisons of the (adjusted) means for each factor are meaningful, as they represent effects given the other factor is held fixed, and should be examined. If the interaction is needed, however, the main effects aren’t meaningful (since an interaction says that a main effect changes depending on the level of the other factor), so neither are multiple comparisons between the adjusted means. It is possible to compare in a pairwise manner all $IJ$ combinations of rows $\times$ columns, but this isn’t in my opinion a very evocative way of describing the interaction (compared to an interaction plot).

Interestingly enough, in this context an $F$-statistic that is unusually small (that is, has a $p$-value very close to 1.00) can be worth investigating. The paper appended to this handout (“Small $F$-ratios: Red flags in the linear model,” by G.E. Meek, C. Ozgur, and K.A. Dunning, *Journal of Data Science*, 5, 199-215) points out how unusually small $F$-statistics can occur if a needed interaction effect is left out of a model, a needed predictor is not included, or there is lack of fit of the linear regression model.

The situation where $n_{ij}$ is constant for all $i$ and $j$ (that is, a balanced design) is worth special comment, for four reasons.

1. In a balanced design, the effects are orthogonal (uncorrelated) with each other. That is, the significance of the row effect (for example) does not change if the column effect is included or omitted from the model. This corresponds to predictors in a regression model that have variance inflation factors equal to one (that is, a balanced design leads to a complete lack of collinearity). This is useful, since it means that we can examine each effect separately, without worrying whether other effects are significant or not. If the design is very unbalanced, it can happen that a row effect (for example) is highly significant when the model including an insignificant column effect is fit, but becomes insignificant if the apparently unneeded column effect is omitted from the model, which is counterintuitive and undesirable (but not surprising in the presence of collinearity). Similarly, in the presence of an unbalanced design, the inclusion of an interaction effect (even if it is not close to statistical significance) can affect the statistical significance of main effects; for that reason, in that situation the interaction effect should be removed from the model to see what happens.
(2) If the model is balanced, all observations have equal leverage. In very unbalanced designs, on the other hand, observations in cells with relatively few observations can have high leverage, with the usual problems that come with that. In the most extreme situation, if a cell has only one observation \(n_{ij} = 1\) and you fit a model that includes an interaction effect, the leverage value for that observation will equal 1 (since one of the constructed effect codings will effectively be an indicator for that one observation). As we saw before when discussing handling outliers or leverage points in time series data with an indicator variable, in that case the standardized residual for that observation given in Minitab will be undefined (*), but you can set it to 0 manually.

(3) If the design is balanced, and a model with only the two main effects is fit (no interaction), the fitted values correspond to \(\bar{y}_i + \bar{y}_j - \bar{y}_{..}\) for the observations in the \((i, j)\)th cell. That is, the regression approach to ANOVA and the cell means approach coincide. This is useful since it ties together the intuition of cell means with the powerful machinery of regression modeling. If the design is very unbalanced the fitted values can be very different from this cell means form.

(4) If in your (unbalanced) design you have a cell with \(n_{ij} = 1\) and you are fitting a model that includes the interaction effect the leverage value for that observation will be 1, and the standardized residual will not be defined. Since the residual must be 0 in that case you should set the standardized residual to 0 also. Note that if you do not do this a Levene’s test that includes the interaction cannot be fit, since from the point of view of the test the cell is empty.

(5) If \(n_{ij} = 1\) for all cells (for example, if the numbers in the table on page 1 were the mileages for only one auto of each paired type, rather than the averages of five such autos), the interaction effect cannot be fit. The reason for this is that if the interaction effect was included, the observed and fitted values would be equal, yielding identically zero residuals for all observations.

Note, by the way, that all of these comments related to unbalanced designs also hold if a model includes one or more numerical predictors; that is, the analysis of covariance situation.