Conditional Inference Trees and ANOVA comparison

Iain Proctor, BASF SE, 29th May 2017
ANOVA is a standard procedure, but unfortunately not always straightforward for everyone.

Practice for exploratory models follows protocol for ANOVA for registration assessments:

- Choice of transformation
- Normally distributed residuals
- Understanding of significance level
- Choice of post hoc test
- Interpretation of the result
Needs

**Problem**
- Choice of transformation
- Normally distributed residuals
- Understanding of significance level
- Choice of post hoc test
- Interpretation of the result

**Requirement**
- No transformation necessary
- No residual check required
- Obvious answer to significance of factor
- Defined test
- Easily understandable result
Conditional inference trees

- *partykit* package in R
- Iris data: are species different in terms of petal length?
  - Petal length ~ Species
- Local optimum chosen for split in data
- Where there is more than one covariate, the most significant covariate in model is chosen
- Minimum adequate model returned

```
     setosa versicolor, virginica

     Species
     p < 0.001

     virginica  versicolor

     Species
     p < 0.001

     Node 2 (n = 50)  Node 4 (n = 50)  Node 5 (n = 50)
```

7 6 5 4 3 2 1
7 6 5 4 3 2 1
7 6 5 4 3 2 1
Example data
Comparison of output

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>F-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>2</td>
<td>253111</td>
<td>126556</td>
<td>47.046</td>
<td>0.03101</td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>322807</td>
<td>26901</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Simulation comparison

- Data simulated for two treatment groups with varying power
- Linear model, Ctree and Wilcoxon Rank Sum Test conducted on data
Real data

- Measurements from the BASF field trial database tested using ANOVA and Ctree, using response as dependent variable and treatment as independent covariate

- 8000 models tested:

<table>
<thead>
<tr>
<th>Model</th>
<th>Sig. Difference</th>
<th>No sig. difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>2456</td>
<td>5544</td>
</tr>
<tr>
<td>Ctree</td>
<td>2376</td>
<td>5624</td>
</tr>
</tbody>
</table>

- 82 models where ANOVA returned a significant difference and Ctree did not

- 2 models where Ctree returned a significant difference and ANOVA did not
Potential drawbacks

- Complexity of post hoc test result can be lost in conditional inference tree result
- Change of result based on inclusion of third parties
- In agricultural field trials, such a result can be justified (or avoided) by only including those treatments in the model which were part of the original trial design.
- Reappearance of covariate in tree (Covariate A, followed by B followed by A again) could be considered as confusing as an interaction effect to explain
Extension to 2 factors – trial series
Extension to model fit at each node
Conclusion

- Conditional inference trees are useful for exploratory analysis, prior to registration.
- Requires little statistical knowledge for interpretation of the resultant plot.
- Presence of more than one end node is closely correlated with significance in an ANOVA with treatment; improved power relative to Wilcoxon Rank Sum.
- Minimum adequate model results in simplified effects and interactions.
- Can be combined for more complex assessment with regression model at the end of each node.