Lecture 21
Factor Effects Model

STAT 512
Spring 2011

Background Reading
KNNL: 16.7, 16.10-16.11
Topic Overview

- Factor Effects Model for Single Factor ANOVA
- Cash Offers example
Review: Cell Means Model

\[ Y_{ij} = \mu_i + \varepsilon_{ij} \]

- \( Y_{ij} \) is the value of the response variable in the \( j^{th} \) trial for the \( i^{th} \) factor level.
- \( \mu_i \) is the (unknown) theoretical mean for all of the observations at level \( i \).
- \( \varepsilon_{ij} \) are independent normal errors with means 0 and variances \( \sigma^2 \).
- Since \( \varepsilon_{ij} \) are normal RV, \( Y_{ij} \) also are normal RV with means \( \mu_i \) and variances \( \sigma^2 \).
Parameters in ANOVA

- Need to estimate all of the cell means $\mu_1, \mu_2, \ldots, \mu_r$ and also $\sigma^2$

- F-test answers the question of whether $\mu_i$ depends on $i$. That is we test the null hypothesis $H_0: \mu_1 = \mu_2 = \ldots = \mu_r$ against the alternative that not all the means are the same.
Notation

• “DOT” indicates to sum over that index, “BAR” indicates to take the average.

• **Overall or grand mean** is

\[
\bar{Y}_{..} = \frac{1}{n_T} \sum_{i} \sum_{j} Y_{ij}
\]

• Mean for factor level \( i \) is

\[
\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j} Y_{ij}
\]
Estimates

- Each group mean is estimated by the mean of the observations within that group:
  \[
  \hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i} \sum_j Y_{ij}
  \]

- Variance estimated by
  \[
  MSE = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i (n_i - 1)} = \frac{\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2}{n_T - r}
  \]
Factor Effects Model

• Re-parameterize cell means model by taking
  \[ \mu_i = \mu + \tau_i \]

• The “factor effect” \( \tau_i \) represents the difference between the grand/overall mean and the factor level mean.

• Model becomes:
  \[ Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \]
  where the usual assumptions apply
Parameters / Constraints

- Parameters are $\mu$ (or $\mu_\cdot$), $\tau_1, \tau_2, \ldots, \tau_r, \sigma^2$
- Note that there is an “extra” parameter (for cell means model, had $\mu_1, \mu_2, \ldots, \mu_r, \sigma^2$
- One of the $\tau_i$'s is redundant (if you know the grand mean, and $r - 1$ of them, you can compute the rest).
- To avoid redundancy and make the models equivalent, we assume $\sum_i \tau_i = 0$
- See pages 702-703 for further info.
Example

- Suppose $r = 3$ and we have means
  \[ \mu_1 = 10, \mu_2 = 20, \mu_3 = 30 \]

- With no constraint, any of the following would be valid sets of parameters for the factor effects model:
  \[ \mu = 0, \tau_1 = 10, \tau_2 = 20, \tau_3 = 30 \]
  \[ \mu = 20, \tau_1 = -10, \tau_2 = 0, \tau_3 = 10 \]
  \[ \mu = 500, \tau_1 = -490, \tau_2 = -480, \tau_3 = -470 \]
  and infinitely many others

- Constraint needed for parameters to be well defined (i.e. have a unique solution).
Factor Effects Model

\[ Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \]

where \( \varepsilon_{ij} \sim N(0, \sigma^2) \)
and \( \sum \tau_i = 0 \)

- Null hypothesis for F-test becomes
  \[ H_0 : \tau_1 = \tau_2 = \ldots = \tau_r = 0 \]

- Parameter Estimates become
  \[ \hat{\mu} = \bar{Y}_. \quad \text{and} \quad \hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_. \]
Constraints

• As long as we make a constraint that brings us back to the correct number of parameters, we have a valid model.

• $\sum \tau_i = 0$ is a convenient constraint because it means that the $\tau_i$'s represent differences from the grand mean.

• **Important:** SAS uses instead $\tau_r = 0$, which means that $\mu$ will be the mean for the $r^{th}$ level instead of the grand mean. So in SAS treatments are all compared to the $r^{th}$ level.
Cash Offers Example

- **Goal:** Estimate the parameters for the cell means model and for the factor effects model using the constraint $\sum \tau_i = 0$.
- **Easiest way** to get the cell-means estimates is to use PROC MEANS. Alternatively, one can use the MEANS statement and put things together from PROC GLM.
- The cell-means estimates are then used to produce the factor effects estimates.
- Still using code: cashoffers.sas
Cash Offers (Cell Means)

```sas
proc sort data=cash; by age;
proc means data=cash noprint;
  class age;
  var offer;
  output out=means mean=average;
proc print; run;
```

- Using CLASS statement causes means to be produced for each level of the class variable(s)
## Output

<table>
<thead>
<tr>
<th>Obs</th>
<th>age</th>
<th><em>TYPE</em></th>
<th><em>FREQ</em></th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>36</td>
<td>23.5556</td>
</tr>
<tr>
<td>2</td>
<td>Elderly</td>
<td>1</td>
<td>12</td>
<td>21.4167</td>
</tr>
<tr>
<td>3</td>
<td>Middle</td>
<td>1</td>
<td>12</td>
<td>27.7500</td>
</tr>
<tr>
<td>4</td>
<td>Young</td>
<td>1</td>
<td>12</td>
<td>21.5000</td>
</tr>
</tbody>
</table>

- Type = 0 is the Grand Mean
  \[ \hat{\mu} = 23.56 \]

- Type = 1 are Cell Means
  \[ \hat{\mu}_{eld} = 21.42 \quad \hat{\mu}_{mid} = 27.75 \quad \hat{\mu}_{yng} = 21.50 \]
Cash Offers (Factor Effects)

- Grand mean: \( \hat{\mu} = 23.56 \)
- Factor Effects:
  \[ \hat{\tau}_{eld} = 21.42 - 23.56 = -2.14 \]
  \[ \hat{\tau}_{mid} = 27.75 - 23.56 = 4.19 \]
  \[ \hat{\tau}_{eld} = 21.50 - 23.56 = -2.08 \]
Cash Offers (Model using SAS)

```
proc glm data=cash;
  class age;
  model offer=age /solution;
```

Solution option produces estimates:

| Param   | Estimate | SE   | t Value | Pr>|t| |
|---------|----------|------|---------|-----|
| Int     | 21.500   | 0.456| 47.20   | <.0001 |
| age     | Elderly  | -0.083| 0.644  | -0.13 | 0.8979 |
| age     | Middle   | 6.250 | 0.644  | 9.70  | <.0001 |
| age     | Young    | 0.000 | .      |       |       |

SAS Estimates

- Are “biased” by the choice of constraint; in this case the intercept represents the cell mean for YOUNG (since it is alphabetically last)
- ELDERLY and MIDDLE levels are compared to YOUNG
SAS Estimates (2)

- We could reproduce estimates for the textbook parameterizations from the SAS estimates:

\[ \hat{\mu} = 21.5 + \left( -0.083 + 6.250 + 0 \right) \frac{3}{3} = 23.56 \]

\[ \hat{\tau}_{middle} = \left( 21.5 + 6.25 \right) - 23.56 = 4.19 \]
Big Picture

Whatever parameterization is used, we will still be looking to determine answers to the questions:

- Is there a difference among the levels of the factor? (F-test)

- Where do the differences lie? How big are the differences? (multiple comparisons – Chapter 17)
Power / Sample Size Issues

• In ANOVA, the power is the probability that we will find a difference in treatment means, given that one exists.

• Power depends on:
  o The size of difference in trt means the researcher believes is practically significant
  o Variance
  o Significance level (alpha)
  o Sample size
Power / Sample Size Issues

- Sections 16.10 and 16.11 discuss how to use Table B.12 in order to find the appropriate sample size for a given power level.

- Example: Suppose I want to detect treatment differences (4 groups) greater than $\Delta = 0.5$, and I believe that $\sigma = 0.2$. Then $\Delta / \sigma = 2.5$ and $n = 6$ or 7 per group is needed.
TABLE B.12 (concluded) Table for Determining Sample Size for Analysis of Variance (fixed factor levels model).

Power 1 − \( \beta = .90 \)

<table>
<thead>
<tr>
<th>( \Delta / \sigma = 1.0 )</th>
<th>( \Delta / \sigma = 1.25 )</th>
<th>( \Delta / \sigma = 1.50 )</th>
<th>( \Delta / \sigma = 1.75 )</th>
<th>( \Delta / \sigma = 2.0 )</th>
<th>( \Delta / \sigma = 2.5 )</th>
<th>( \Delta / \sigma = 3.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
<td>( .2 )</td>
<td>( .1 )</td>
<td>( .05 )</td>
<td>( .01 )</td>
<td>( .2 )</td>
<td>( .1 )</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>18</td>
<td>23</td>
<td>32</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>22</td>
<td>27</td>
<td>37</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>40</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>27</td>
<td>32</td>
<td>43</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>29</td>
<td>34</td>
<td>46</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>31</td>
<td>36</td>
<td>48</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>32</td>
<td>38</td>
<td>50</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>33</td>
<td>40</td>
<td>52</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
<td>35</td>
<td>41</td>
<td>54</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>

Power 1 − \( \beta = .95 \)

<table>
<thead>
<tr>
<th>( \Delta / \sigma = 1.0 )</th>
<th>( \Delta / \sigma = 1.25 )</th>
<th>( \Delta / \sigma = 1.50 )</th>
<th>( \Delta / \sigma = 1.75 )</th>
<th>( \Delta / \sigma = 2.0 )</th>
<th>( \Delta / \sigma = 2.5 )</th>
<th>( \Delta / \sigma = 3.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
<td>( .2 )</td>
<td>( .1 )</td>
<td>( .05 )</td>
<td>( .01 )</td>
<td>( .2 )</td>
<td>( .1 )</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>23</td>
<td>27</td>
<td>38</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>27</td>
<td>32</td>
<td>43</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>30</td>
<td>36</td>
<td>47</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>33</td>
<td>39</td>
<td>51</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>35</td>
<td>41</td>
<td>53</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>37</td>
<td>43</td>
<td>56</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
<td>39</td>
<td>45</td>
<td>58</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>33</td>
<td>40</td>
<td>47</td>
<td>60</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>42</td>
<td>48</td>
<td>62</td>
<td>22</td>
<td>27</td>
</tr>
</tbody>
</table>

Upcoming in Lecture 22...

- Multiple Comparisons (Chapter 17)