Unbalanced Data

Part 1: The basic problem with unbalanced data

Unbalanced one way classification - no problem.

A two way classification by job category and gender:

<table>
<thead>
<tr>
<th></th>
<th>Workers</th>
<th>Executives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>30, 50, 30, 35, 55</td>
<td>100</td>
</tr>
<tr>
<td>Females</td>
<td>20</td>
<td>75, 85, 80</td>
</tr>
</tbody>
</table>

Males are complaining because

* There are more female than male executives

* Females make more $ on average
\[(30+50+30+35+55+100)/6 = 50,000 \text{ M} \]
\[(20+75+85+80)/4 = 260/4 = 65,000 \text{ F} \]

This is very interesting in light of the fact that every female makes at least $10,000 less than the worst paid of her male counterparts. The comparison of the male to female average salary clearly is unfair. Why is this?

Maybe it is because of some sort of interaction in the salaries which would mean that the difference between male and female salaries is a function of the job level. To check this out, we look at a table of means:

Table of Mean Salaries (thousand $)

<table>
<thead>
<tr>
<th></th>
<th>Workers</th>
<th>Executives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>Females</td>
<td>20</td>
<td>80</td>
</tr>
</tbody>
</table>
The interaction is 0!! In both columns males make an average of $20,000 more than females. Therefore the difference in overall male and female salaries has nothing to do with interaction. It is simply a result of the imbalance in the data.

**Part 2: Contrasts and sums of squares:**

Model:

\[ Y_{ijk} = \mu + G_i + L_j + e_{ijk} \]

\[ Y = \text{Overall Mean} + \text{Gender Effect} + \text{Level of Job Effect} + e_{ijk} \]

Summing all the table entries (using \( e_{ij} \) to denote a sum over \( k \)):

<table>
<thead>
<tr>
<th></th>
<th>Workers</th>
<th>Executives</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>5( \mu + 5G_1 + 5L_1 + 5e_{11} )</td>
<td>( \mu + G_1 + L_2 + e_{121} )</td>
</tr>
<tr>
<td>F</td>
<td>( \mu + G_2 + L_1 + e_{211} )</td>
<td>3( \mu + 3G_2 + 3L_2 + e_{221} )</td>
</tr>
</tbody>
</table>

so the original data table has rows with means
M: \(( \mu + G_1 + \frac{5}{6}L_1 + \frac{1}{6}L_2 + \text{error term} )\)

F: \(( \mu + G_2 + \frac{1}{4}L_1 + \frac{3}{4}L_2 + \text{error term} )\)

The difference of these two means, row 1 mean minus row 2 mean, is thus an estimate of

\[ G_1 - G_2 + \frac{7}{12}(L_1 - L_2), \]

not an estimate of just \( G_1 - G_2 \).

We have four means, \( \bar{Y}_{ij} \) in the cells of our table. The mean of the 6 male salaries minus the mean of the 4 female salaries is seen to be

\[ \frac{5}{6}\bar{Y}_{11} + \frac{1}{6}\bar{Y}_{12} - \frac{1}{4}\bar{Y}_{21} - \frac{3}{4}\bar{Y}_{22}. \]

Now with 4 cell means we can compute any linear combination

\[ c_{11}\bar{Y}_{11} + c_{12}\bar{Y}_{12} + c_{21}\bar{Y}_{21} + c_{22}\bar{Y}_{22}. \]

Clearly, this is an estimate of
(c_{11} + c_{12} + c_{21} + c_{22})\mu +
(c_{11} + c_{12})G_1 + (c_{21} + c_{22})G_2 +
(c_{11} + c_{21})L_1 + (c_{12} + c_{22})L_2

to which we would add
\begin{align*}
c_{11}(GL)_{11} + c_{12}(GL)_{12} + \\
c_{21}(GL)_{21} + c_{22}(GL)_{22}
\end{align*}

if our model included interaction. It has standard error

\[\sqrt{\frac{c_{11}^2 \sigma^2}{n_{11}} + \frac{c_{12}^2 \sigma^2}{n_{12}} + \frac{c_{21}^2 \sigma^2}{n_{21}} + \frac{c_{22}^2 \sigma^2}{n_{22}}}\]

and sum of squares (where \(\bar{Y}_{ij.}\) is the mean and \(n_{ij}\) is the number of observations in the \(ij^{th}\) cell of the table):

\[
\frac{Q^2}{den.} = \frac{(c_{11}\bar{Y}_{11.} + c_{12}\bar{Y}_{12.} + c_{21}\bar{Y}_{21.} + c_{22}\bar{Y}_{22.})^2}{\frac{c_{11}^2}{n_{11}} + \frac{c_{12}^2}{n_{12}} + \frac{c_{21}^2}{n_{21}} + \frac{c_{22}^2}{n_{22}}}
\]

**TYPE I:**
I want to estimate $G_1 - G_2$ so, if I do not worry about contamination from other parts of my model I want

$$(c_{11} + c_{12})G_1 + (c_{21} + c_{22})G_2 = G_1 - G_2$$

so that any c's with $c_{12} = 1-c_{11}$ and $c_{22} = -1-c_{21}$ will work. Since there are many ways to do this, I will pick the one with the smallest standard error. Using calculus to find the $c_{11}$ and $c_{21}$ that do the job, we find

$$c_{11} = \frac{5}{6} \quad \text{and} \quad c_{21} = -\frac{1}{4}.$$ 

Thus

$$\frac{5}{6}\bar{Y}_{11\cdot} + \frac{1}{6}\bar{Y}_{12\cdot} - \frac{1}{4}\bar{Y}_{21\cdot} - \frac{3}{4}\bar{Y}_{22\cdot}$$

is the estimate of $G_1 - G_2$ (plus other contaminating effects) that has minimum variance. Its sum of squares is the Type I sum of squares for GENDER in a model containing GENDER, LEVEL, and possibly
GENDER*LEVEL, in that order. It is also sometimes called the sum of squares for GENDER ignoring LEVEL.

Using the cell means we have our Type I sum of squares:

\[
\frac{(c_{11}\bar{Y}_{11.}+c_{12}\bar{Y}_{12.}+c_{21}\bar{Y}_{21.}+c_{22}\bar{Y}_{22.})^2}{c_{11}^2/n_{11}+c_{12}^2/n_{12}+c_{21}^2/n_{21}+c_{22}^2/n_{22}}
\]

\[
\frac{(\frac{5}{6}40+\frac{1}{6}100-\frac{1}{4}20-\frac{3}{4}80)^2}{(\frac{5}{6})^2/5+(\frac{1}{6})^2/1+(\frac{-1}{4})^2/1+(\frac{-3}{4})^2/3} = 540
\]

**Type II:**

Suppose I decide that I want to estimate \(G_1 - G_2 + 0L_1 + 0L_2\).

I now need to have:

\[
\begin{align*}
(c_{11} + c_{12}) &= 1, \\
(c_{21} + c_{22}) &= -1 \\
(c_{11} + c_{21}) &= 0 \\
\text{and } (c_{12} + c_{22}) &= 0
\end{align*}
\]

or, in a table,
\[
\begin{array}{|c|c|c|}
\hline
\text{c}_{11} & 1-\text{c}_{11} & \Rightarrow (1G_1) \\
\hline
-\text{c}_{11} & \text{c}_{11}-1 & \Rightarrow (-1G_2) \\
\hline
\Rightarrow (0L_1) & \Rightarrow (0L_2) & \Rightarrow (0\mu) \\
\hline
\end{array}
\]

Now we minimize the standard error of such a linear combination by minimizing

\[
\sigma \sqrt{\frac{c_{11}^2 \frac{1}{5} + (1-c_{11})^2 \frac{1}{1} + c_{11}^2 \frac{1}{1} + (c_{11}-1)^2 \frac{1}{3}}}
\]

so we set

\[
2c_{11} \frac{1}{5} - 2(1-c_{11}) \frac{1}{1} + 2c_{11} \frac{1}{1} + 2(c_{11}-1) \frac{1}{3} = 0
\]

\[
\Rightarrow c_{11} = \frac{10}{19}. 
\]

Using the cell totals we have our Type II sum of squares:

\[
\frac{(c_{11}\bar{Y}_{11\bullet} + c_{12}\bar{Y}_{12\bullet} + c_{21}\bar{Y}_{21\bullet} + c_{22}\bar{Y}_{22\bullet})^2}{c_{11}^2n_{11} + c_{12}^2n_{12} + c_{21}^2n_{21} + c_{22}^2n_{22}} = 
\]

\[
\frac{(\frac{10}{19}40 + \frac{9}{19}100 - \frac{10}{19}20 - \frac{9}{19}80)^2}{(\frac{10}{19})^2/5 + (\frac{9}{19})^2/1 + (-\frac{10}{19})^2/1 + (-\frac{9}{19})^2/3} = 633.33
\]
This would also be referred to as the sum of squares for GENDER adjusted for LEVEL.

**Type III:**

The Type II sum of squares above is fine for the model without interaction. If we had an interaction, our Type II linear combination of means would be

\[
\frac{10}{19} \bar{Y}_{11} + \frac{9}{19} \bar{Y}_{12} - \frac{10}{19} \bar{Y}_{21} - \frac{9}{19} \bar{Y}_{22}
\]

and thus would estimate

\[
0\mu + 1G_1 + (-1)G_2 + 0L_1 + 0L_2 + \frac{10}{19}(GL)_{11} + \frac{9}{19}(GL)_{12} - \frac{10}{19}(GL)_{21} - \frac{9}{19}(GL)_{22}
\]

which still seems like a bizarre quantity in which to be interested. Looking at the four cells it is clear that we cannot get the coefficients of these interactions to be 0 unless we set all the c's to 0. Perhaps we
can get a linear combination that is zeroed out by the "standard assumptions" that 
\[ \sum_i (GL)_{ij} = \sum_j (GL)_{ij} = 0. \] This is accomplished by this table of c's:

| \( .5 \) | \( .5 \) | \( \Rightarrow (1G_1) \) |
| \( -.5 \) | \( -.5 \) | \( \Rightarrow (-1G_2) \) |
| \( \Rightarrow (0L_1) \) | \( \Rightarrow (0L_2) \) | \( \Rightarrow (0\mu) \) |

The margin restrictions and the restriction that the coefficients in each row be equal (so the sum of coefficients times interactions will be the coefficient times the sum of the interactions \( \Rightarrow 0 \) in each row) are enough to completely specify the c's. Notice that the linear combination we are discussing compares the average of the two row 1 cell means to the average of the two cell means in row 2. This is 
\[ .5(40 + 100 - 20 - 80) = 20 \text{ thousand dollars} \] and represents what we would say, after a little careful
thought, is the correct salary comparison for males versus females if we adjust for job level.

**Type IV:**

The Type IV and Type III sums of squares are the same unless a cell in your table has no entries. In that case, the Type IV SS has the unfortunate property of sometimes having different values depending on the alphabetical order of the levels of your factors. Thus if, instead of "workers" and "executives" we had used the terms "blue-collar" and "VIP" the actual Type IV sums of squares might change. For this reason I recommend you never use Type IV and I do not discuss this topic further here.

**Part 3: LSMEANS**
PROC GLM can produce LSMEANS or least squares adjusted means. The \(i^{th}\) GENDER LSMEAN for example would be an estimate of

\[
\mu + \frac{1}{2} \left[ (GL)_{i1} + (GL)_{i2} \right]
\]

so these are estimates of what we would have had if the data had been balanced.

**Part 4: Some SAS Examples:**

```
Data unbal;
  input gender $ 1-7 level $ 9-19  n @;
  do worker = 1 to n; input salary @; output; end;
cards;
  male      worker      5    30 50 30 35 55
  male      executive   1    100
  female    worker      1    20
  female    executive  3    75 85 80
;
proc glm;       class gender level;
model salary = gender level gender*level / ss1 ss2 ss3 ss4;
lsmeans gender / pdiff;
run;
```
General Linear Models Procedure
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>2</td>
<td>female male</td>
</tr>
<tr>
<td>LEVEL</td>
<td>2</td>
<td>executive worker</td>
</tr>
</tbody>
</table>

Number of observations in data set = 10

Dependent Variable: SALARY

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>6240.0000</td>
<td>2080.0000</td>
<td>20.80</td>
<td>0.0014</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>600.0000</td>
<td>100.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cor Tot</td>
<td>9</td>
<td>6840.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square    C.V.    Root MSE    SALARY Mean
0.912281    17.85714 10.0000     56.000

Source    DF    Type I SS   Mean Square   F Value   Pr > F
GENDER    1    540.0000   540.0000       5.40       0.0591
LEVEL     1    5700.0000  5700.0000     57.00      0.0003
GEN*LEV   1    0.0000     0.0000        0.00       1.0000

Source    DF    Type II SS  Mean Square  F Value   Pr > F
GENDER    1    633.3333   633.3333     6.33       0.0455
LEVEL     1    5700.0000  5700.0000    57.00      0.0003
GEN*LEV   1    0.0000     0.0000       0.00       1.0000

Source    DF    Type III SS Mean Square  F Value   Pr > F
GENDER    1    631.5789   631.5789     6.32       0.0457
LEVEL     1    5684.2105  5684.2105    56.84      0.0003
Example 2: Paint example from the book.

In this example I point out that no amount of statistical gymnastics will make up for a poor experiment. Here there are 2 additives, A and B, each at two levels in some paint. Drying time for the paint is the response. The data are badly unbalanced.

<table>
<thead>
<tr>
<th>Paint Drying Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>no A</td>
</tr>
<tr>
<td>no B 20,33,24,23,26</td>
</tr>
<tr>
<td>B 26</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>20,31</td>
</tr>
<tr>
<td>14,17,18,22,16,17,8,7</td>
</tr>
</tbody>
</table>
Data paint; input A B n @;
   do board = 1 to n; input dry_t @; output; end;
cards;
0 0 5 20 33 24 23 26
0 1 1 26
1 0 2 20 31
1 1 8 14 17 18 22 16 17 8 7
;
proc glm; class a b;
model dry_t= a b a*b/ss1 ss2 ss3 ss4;
lsmeans a b/pdiff; run;

General Linear Models Procedure
Class Level Information
Class   Levels   Values
A       2       0 1
B       2       0 1
Number of observations in data set = 16

Dependent Variable: DRY_T

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>441.57500</td>
<td>147.19167</td>
<td>5.25</td>
<td>0.0152</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>336.17500</td>
<td>28.01458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr Total</td>
<td>15</td>
<td>777.75000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>260.41667</td>
<td>260.41667</td>
<td>9.30</td>
<td>0.0101</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>109.63470</td>
<td>109.63470</td>
<td>3.91</td>
<td>0.0713</td>
</tr>
<tr>
<td>A*B</td>
<td>1</td>
<td>71.52363</td>
<td>71.52363</td>
<td>2.55</td>
<td>0.1361</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type II SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A  1  38.61883  38.61883  1.38  0.2631
B  1  109.63470  109.63470  3.91  0.0713
A*B  1  71.52363  71.52363  2.55  0.1361

Source  DF  Type III SS  Mean Square  F Value  Pr > F
A  1  64.208562  64.208562  2.29  0.1559
B  1  52.893493  52.893493  1.89  0.1945
A*B  1  71.523630  71.523630  2.55  0.1361

Source  DF  Type IV SS  Mean Square  F Value  Pr > F
A  1  64.208562  64.208562  2.29  0.1559
B  1  52.893493  52.893493  1.89  0.1945
A*B  1  71.523630  71.523630  2.55  0.1361

Least Squares Means

A  DRY_T  Pr > |T|  H0:
   LSMEAN  LSMEAN1=LSMEAN2
  0  25.6000000  0.1559
  1  20.1875000

B  DRY_T  Pr > |T|  H0:
   LSMEAN  LSMEAN1=LSMEAN2
  0  25.3500000  0.1945
  1  20.4375000

Points: Overall F is significant (P=.0152) but nothing is significant individually except for the Type I for A. If you put B in there first, it would be significant but not A in the
Type I list $SS(A|B)$ would be 38.62 as we can tell from the current Type II list. In other words, there is clearly some effect of these additives but it is virtually impossible to sort out the nature of the effect. The cell means suggest the interesting hypothesis that both additives need to be present. The comparison of the lower right cell to the rest uses up almost all the treatment sum of squares (check it out).

*No amount of statistical calculation can save a poorly designed experiment like this!* There is no magic - the only inference we can make is based on arbitrary uncheckable assumptions about the treatment effects (e.g. assuming no B or AB we have significant A)