Factors and Model Matrices

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 22, 2007

The Big Picture

- A factor is a categorical variable.
- The possible values are called levels.
- In a regression framework, a single factor with $k$ levels is represented by $k - 1$ columns in the design matrix (or model matrix).
- There are several ways to represent the same factor with columns.
- The most usual parameterization uses 0s and 1s with a column corresponding to each level but one.
- The fitted values are the same for each parameterization, but the parameter estimates differ.
An experiment examines the sugar content in four plant lines including a control (wild type) and three treatments (genetically modified forms).

Data are sugar weights in mg after breaking down cellulose.

There are three individuals sampled in each group.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>82.0</td>
<td>58.3</td>
<td>68.1</td>
<td>50.7</td>
</tr>
<tr>
<td></td>
<td>97.8</td>
<td>67.9</td>
<td>70.8</td>
<td>47.1</td>
</tr>
<tr>
<td></td>
<td>69.9</td>
<td>59.3</td>
<td>63.6</td>
<td>48.9</td>
</tr>
</tbody>
</table>

Side-by-side boxplots are good for examining differences in center and spread as well as skewness.
Dot Plots

- For very small data sets, dot plots show the exact locations of points.
- Lattice graphics in R are useful here.
- We note that the control group is much more spread out than the genetically modified groups.

```r
> library(lattice)
> print(dotplot(weight ~ trt))
```

Transformations

- A log transformation would lessen the spread.
- According to Box-Cox, a more extreme transformation (such as $1/y^2$) would do better.
- I will show both transformations and their effects on the dot plots, but will simply analyze the untransformed data for simplicity.
Case Study: Sugar

Box-Cox Transformations

![Graph showing the relationship between Box-Cox transformations and log-likelihood.]

Log-transform

\[
\text{log-weight}
\]

Control A B C
● ● ● ● ● ● ● ● ● ● ● ●

Square-reciprocal

\[
\frac{1}{\text{weight}^2}
\]

Control A B C
● ● ● ● ● ● ● ● ● ● ● ●

Factors and Model Matrices

Standard Parameterization

- For this example, the single factor \( \text{trt} \) is parameterized with three columns of 0s and 1s.
- The intercept corresponds to the control group.
- The other variables correspond to differences between each other level and the control group.
- R will use the first level name in alphabetical order as the reference.

\[
\begin{align*}
\text{lm}(\text{weight} \sim \text{trt}) \\
\text{model.matrix}(\text{sugar1.lm})
\end{align*}
\]

For example:

\[
\begin{align*}
\text{(Intercept) trtA trtB trtC} \\
1 & 1 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 & 0 \\
3 & 1 & 0 & 0 & 0 \\
4 & 1 & 1 & 0 & 0 \\
5 & 1 & 1 & 0 & 0 \\
6 & 1 & 1 & 0 & 0 \\
7 & 0 & 1 & 1 & 0 \\
8 & 1 & 0 & 1 & 0 \\
9 & 1 & 0 & 1 & 0 \\
10 & 1 & 0 & 1 & 0 \\
11 & 1 & 0 & 0 & 1 \\
12 & 1 & 0 & 0 & 1
\end{align*}
\]

attr("assign")
[1] 0 1 1 1

attr("contrasts")$trt
\[
\begin{align*}
\text{contr.treatment}
\end{align*}
\]
**Factors and Model Matrices**

### Alternative Parameterization

- An alternative parameterization has a mean effect of 0 for the four groups.
- The model matrix is more complicated for this parameterization.
- You change the contrasts option to achieve this in R.

```r
> old = options(contrasts = c("contr.sum", "contr.poly"))
> sugar2.lm = lm(weight ~ trt)
> model.matrix(sugar2.lm)

(Intercept) trt1 trt2 trt3
1 1 1 0 0
2 1 1 0 0
3 1 1 0 0
4 1 0 1 0
5 1 0 1 0
6 1 0 1 0
7 1 0 0 1
8 1 0 0 1
9 1 0 0 1
10 1 -1 -1 -1
11 1 -1 -1 -1
12 1 -1 -1 -1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr("contrasts")$trt
[1] "contr.sum"
> options(old)
```

### Coefficients

- Different parameterizations lead to different estimates.
- The p-values correspond to different hypothesis tests.

```r
> summary(sugar1.lm)$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.23333 4.472757 18.608955 7.173441e-08
trtB -15.73333 6.325434 -2.487313 3.767970e-02
trtC -34.33333 6.325434 -5.427823 6.249550e-04

> summary(sugar2.lm)$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.366667 2.236379 29.2288028 2.033089e-09
trt1 17.866667 3.873521 4.6125129 1.726990e-03
trt2 -3.533333 3.873521 -0.9121761 3.883370e-01
trt3 2.133333 3.873521 0.5507478 5.968464e-01
```
**Factors and Model Matrices**

### ANOVA

> `anova(sugar1.lm)`

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>3</td>
<td>1822.21</td>
<td>607.40</td>
<td>10.121</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>480.13</td>
<td>60.02</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> `anova(sugar2.lm)`

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>3</td>
<td>1822.21</td>
<td>607.40</td>
<td>10.121</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>480.13</td>
<td>60.02</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

However, the ANOVA tables are identical.

> `cbind(sugar, fitted(sugar1.lm), fitted(sugar2.lm))`

<table>
<thead>
<tr>
<th>weight</th>
<th>trt</th>
<th>fitted(sugar1.lm)</th>
<th>fitted(sugar2.lm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.0</td>
<td>Control</td>
<td>83.23333</td>
<td>83.23333</td>
</tr>
<tr>
<td>97.8</td>
<td>Control</td>
<td>83.23333</td>
<td>83.23333</td>
</tr>
<tr>
<td>69.9</td>
<td>Control</td>
<td>83.23333</td>
<td>83.23333</td>
</tr>
<tr>
<td>58.3</td>
<td>A</td>
<td>61.83333</td>
<td>61.83333</td>
</tr>
<tr>
<td>67.9</td>
<td>A</td>
<td>61.83333</td>
<td>61.83333</td>
</tr>
<tr>
<td>59.3</td>
<td>A</td>
<td>61.83333</td>
<td>61.83333</td>
</tr>
<tr>
<td>68.1</td>
<td>B</td>
<td>67.50000</td>
<td>67.50000</td>
</tr>
<tr>
<td>70.8</td>
<td>B</td>
<td>67.50000</td>
<td>67.50000</td>
</tr>
<tr>
<td>63.6</td>
<td>B</td>
<td>67.50000</td>
<td>67.50000</td>
</tr>
<tr>
<td>50.7</td>
<td>C</td>
<td>48.90000</td>
<td>48.90000</td>
</tr>
<tr>
<td>47.1</td>
<td>C</td>
<td>48.90000</td>
<td>48.90000</td>
</tr>
<tr>
<td>48.9</td>
<td>C</td>
<td>48.90000</td>
<td>48.90000</td>
</tr>
</tbody>
</table>

The fitted values are also identical.