

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.

Sugar

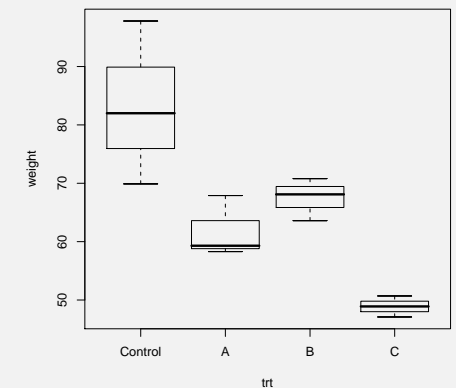
- An experiment examines the sugar content in four plant lines including a control (wild type) and three treatments (genetically modified forms).
- Data are sugar weights in mg after breaking down cellulose.
- There are three individuals sampled in each group.

Control	A	B	C
82.0	58.3	68.1	50.7
97.8	67.9	70.8	47.1
69.9	59.3	63.6	48.9

Side-by-side Box Plots

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

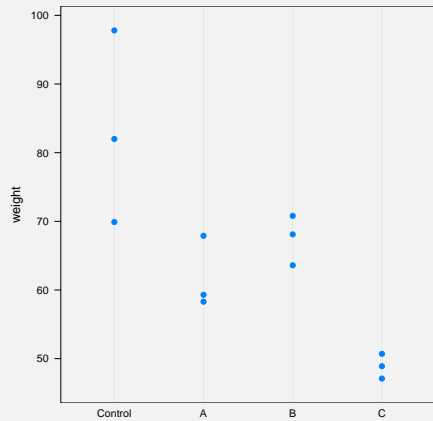
```
> library(DAAG)
> data(sugar)
> attach(sugar)
> plot(weight ~ trt)
```



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.

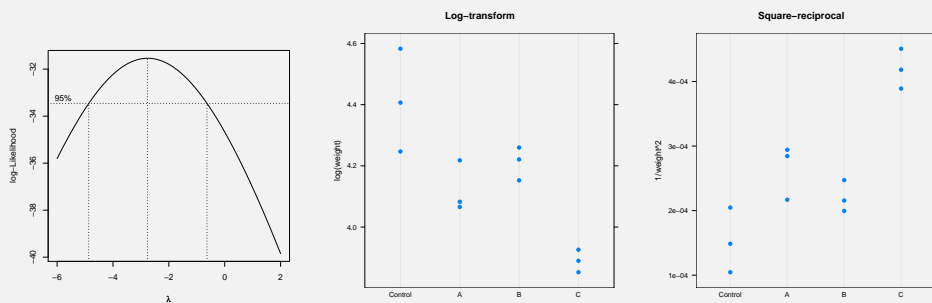
```
> library(lattice)
> print(dotplot(weight ~ trt))
```



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.

Box-Cox Transformations



Standard Parameterization

- For this example, the single factor `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.

```
> sugari.lm = lm(weight ~ trt)
> model.matrix(sugari.lm)

      (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                1    0    1    0
8                1    0    1    0
9                1    0    1    0
10               1    0    0    1
11               1    0    0    1
12               1    0    0    1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$trt
[1] "contr.treatment"
```

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.

```
> 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.

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

      (Intercept)      Estimate Std. Error  t value Pr(>|t|)
trtA            -21.40000     6.325434  -3.383167  9.596536e-03
trtB            -15.73333     6.325434  -2.487313  3.767970e-02
trtC            -34.33333     6.325434  -5.427823  6.249550e-04

> summary(sugar2.lm)$coefficients

      (Intercept)      Estimate Std. Error  t value Pr(>|t|)
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
```

ANOVA

- However, the ANOVA tables are identical.

```
> anova(sugar1.lm)

Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
trt         3  1822.21   607.40   10.121 0.004248 **
Residuals   8   480.13    60.02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(sugar2.lm)

Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
trt         3  1822.21   607.40   10.121 0.004248 **
Residuals   8   480.13    60.02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA

- The fitted values are also identical.

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

      weight      trt fitted(sugar1.lm) fitted(sugar2.lm)
1      82.0 Control             83.23333             83.23333
2      97.8 Control             83.23333             83.23333
3      69.9 Control             83.23333             83.23333
4       58.3           A             61.83333             61.83333
5       67.9           A             61.83333             61.83333
6       59.3           A             61.83333             61.83333
7       68.1           B             67.50000             67.50000
8       70.8           B             67.50000             67.50000
9       63.6           B             67.50000             67.50000
10      50.7           C             48.90000             48.90000
11      47.1           C             48.90000             48.90000
12      48.9           C             48.90000             48.90000
```