Modelling Rating

```
> names(cereal)
 [1] "mfr"      "type"     "calories" "protein"  "fat"      "sodium"
 [7] "fiber"    "carbo"    "sugars"   "potass"   "vitamins" "shelf"
[13] "weigh"    "cups"     "rating"
> lm.all <- lm(rating ~ calories + protein + fat + sodium + fiber +
                        carbo + sugars + potass + vitamins,
               data=cereal)
> summary(lm.all)
Call:
lm(formula = rating ~ calories + protein + fat + sodium + fiber +
    carbo + sugars + potass + vitamins, data = cereal)
Residuals:
       Min         1Q     Median         3Q        Max
-5.343e-07 -2.537e-07  3.961e-08  2.424e-07  5.513e-07
Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  2.794e-07  196559702   <2e-16 ***
calories    -2.227e-01  7.501e-09  -29694282   <2e-16 ***
protein      3.273e+00  5.551e-08   58964906   <2e-16 ***
fat         -1.691e+00  8.101e-08  -20877762   <2e-16 ***
sodium      -5.449e-02  4.910e-10 -110974232   <2e-16 ***
fiber        3.443e+00  4.756e-08   72399805   <2e-16 ***
carbo        1.092e+00  3.492e-08   31287364   <2e-16 ***
sugars      -7.249e-01  3.311e-08  -21895192   <2e-16 ***
potass      -3.399e-02  1.601e-09  -21228850   <2e-16 ***
vitamins    -5.121e-02  1.779e-09  -28778552   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.069e-07 on 64 degrees of freedom
Multiple R-Squared:     1,Adjusted R-squared:     1
F-statistic: 1.696e+16 on 9 and 64 DF,  p-value: < 2.2e-16
>
```


Overfitting?

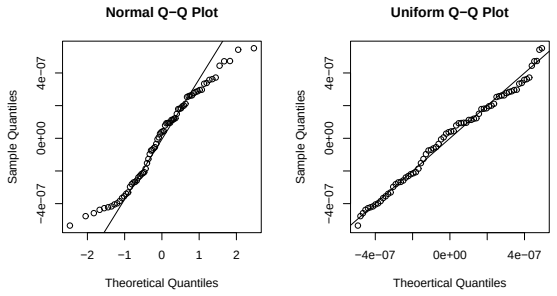
One way to rule this out is to fit the model on a small subset of the data:

```
> lm.20 <- lm(rating ~ calories + protein + fat + sodium + fiber +
                       carbo + sugars + potass + vitamins,
              data=cereal[1:20,])
> coef(lm.20)
(Intercept)    calories     protein         fat      sodium
54.92718341 -0.22272417  3.27317390 -1.69140791 -0.05449270
      fiber       carbo      sugars      potass    vitamins
 3.44347979  1.09245094 -0.72489510 -0.03399335 -0.05121197
> coef(lm.all)
(Intercept)    calories     protein         fat      sodium
54.92718413 -0.22272417  3.27317389 -1.69140795 -0.05449270
      fiber       carbo      sugars      potass    vitamins
 3.44347979  1.09245096 -0.72489512 -0.03399335 -0.05121197
> range(rating - predict(lm.20, newdata=cereal), na.rm=T)
[1] -6.156416e-07  1.046276e-06
>
```


The Residuals

The distribution is typical of error in rounding to 6 digits after the decimal point (i.e., the residuals have distribution $U[-5 \times 10^{-7}, +5 \times 10^{-7}]$).

```
> range(resid(lm.all))
[1] -5.342896e-07  5.513013e-07
> qqnorm(resid(lm.all))
> qqline(resid(lm.all))
> plot(qunif(ppoints(resid(lm.all)),-5e-7,5e-7),sort(resid(lm.all)),
+      main="Uniform Q-Q Plot",
+      xlab="Theoertical Quantiles",
+      ylab="Sample Quantiles")
> abline(0,1)
>
```



Why Such a Stupid Formula?

$$\begin{aligned} rating_i = \; & 54.92718413 - 0.22272417\, calories_i \\ & + 3.27317389\, protein_i - 1.69140795\, fat_i \\ & - 0.05449270\, sodium_i + 3.44347979\, fiber_i \\ & + 1.09245096\, carbo_i - 0.72489512\, sugars_i \\ & - 0.03399335\, potass_i - 0.05121197\, vitamins_i \end{aligned}$$

- Where do these ludicrous coefficients come from?
- Why is the coefficient on vitamins negative?
- Why are total calories *and* individual sources of calories both included in this formula?

A possible answer is given by a much smaller model:

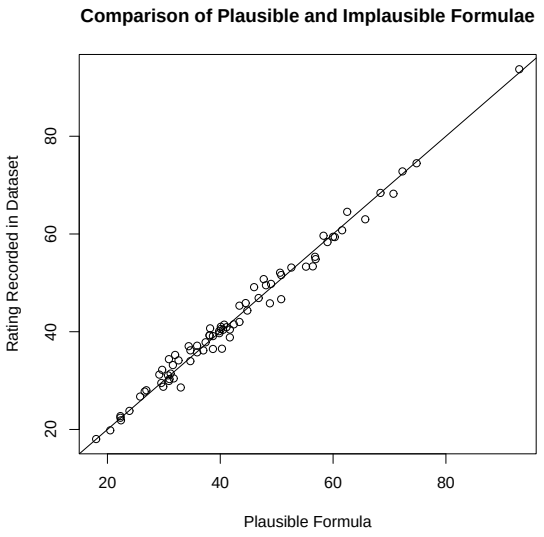
$$\begin{aligned} \widehat{rating}_i = \; & 58.15 + 1.978\, protein_i - 3.898\, fat_i \\ & - 1.803\, sugars_i + 2.422\, fiber_i - 0.0568\, sodium_i \end{aligned}$$

where the coefficients and standard errors are:

Factor	Coeff	Std.Err.
protein	1.978	0.232
fat	−3.898	0.214
sugars	−1.803	0.050
fiber	2.422	0.095
sodium	−0.057	0.002

A Plausible Formula

$$rating_i = 60 + 1.6\,protein_i - 3.8\,fat_i - 1.9\,sugars_i + 2.5\,fiber_i - 0.06 sodium_i$$



My Hypothesis

The ratings data was forged!

- *Consumer Reports'* editors used a simple formula like the one on the previous slide to calculate ratings;
- these ratings were displayed graphically in the magazine article;
- someone else measured the graphic with a ruler and recorded approximate ratings;
- to make them "more accurate", they fit them against all the nutritional information they had available to produce the implausible formula mentioned above;
- without paying any attention to whether or not the implausible formula made sense, they used it to regenerate new ratings for all cereals using an obscene amount of precision.

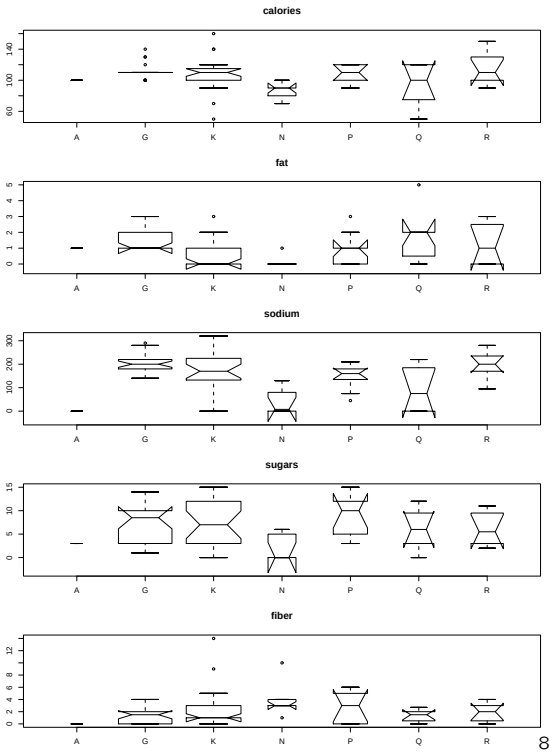
The nail in the coffin: the ratings for cereals with missing data (which was coded as -1 in the dataset) have been calculated using the implausible formula **and the value -1 for the missing values!**

Example of Bogus Rating

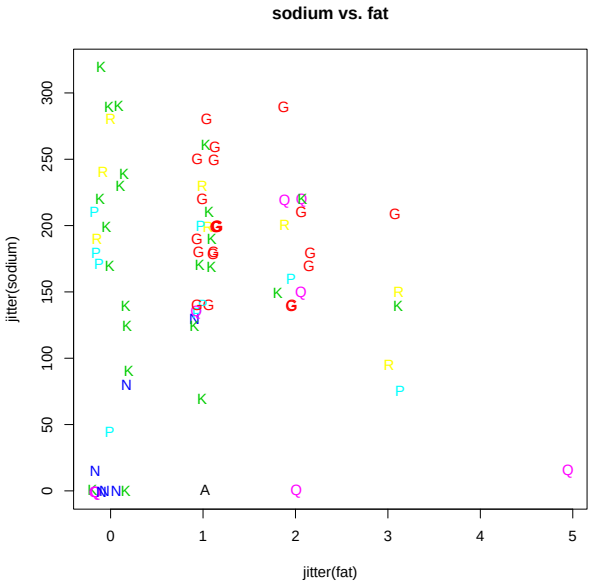
For example,

```
> cereal["Quaker_Oatmeal",]
              mfr type calories protein fat sodium fiber carbo
Quaker_Oatmeal   Q    H      100       5   2      0   2.7    NA
              sugars potass vitamins shelf weigh cups   rating
Quaker_Oatmeal     NA    110        0     1     1 0.67 50.82839
> coef(lm.all)
(Intercept)    calories     protein         fat      sodium
54.92718413 -0.22272417  3.27317389 -1.69140795 -0.05449270
      fiber       carbo      sugars      potass    vitamins
 3.44347979  1.09245096 -0.72489512 -0.03399335 -0.05121197
> sum(coef(lm.all)*c(1,100,5,2,0,2.7,-1,-1,110,0))
[1] 50.82839
>
```

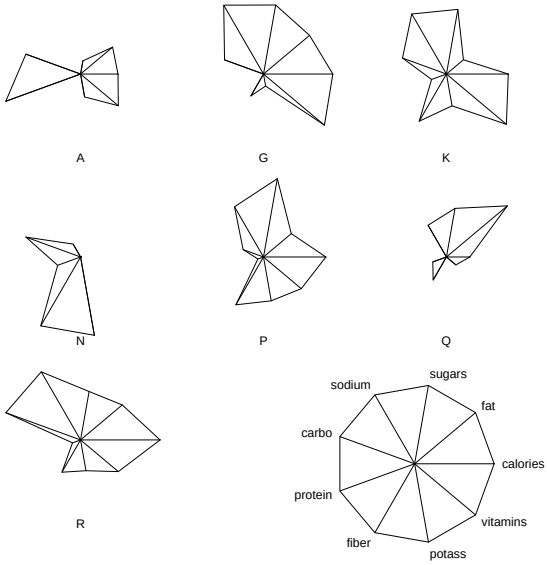

Nutrients by Manufacturer



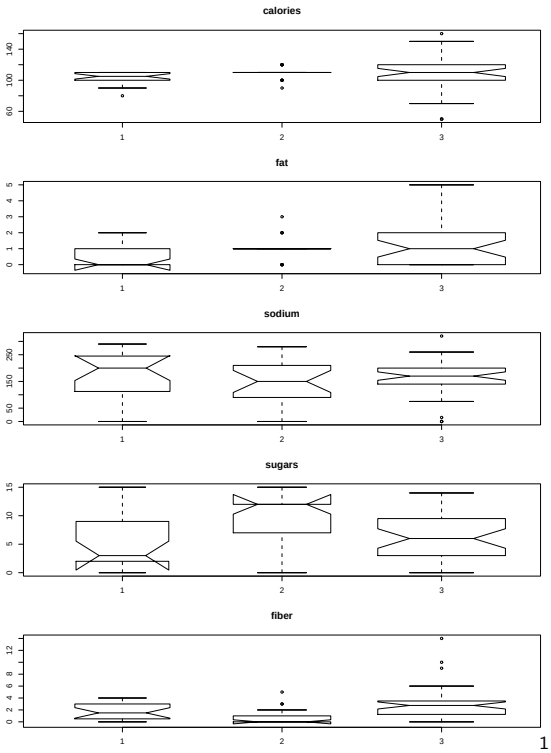
Two-Nutrient Manufacturer Profile



Manufacturer Profile



Nutrients by Shelf



Two-Nutrient Shelf Profile