Parsimony and Collinearity

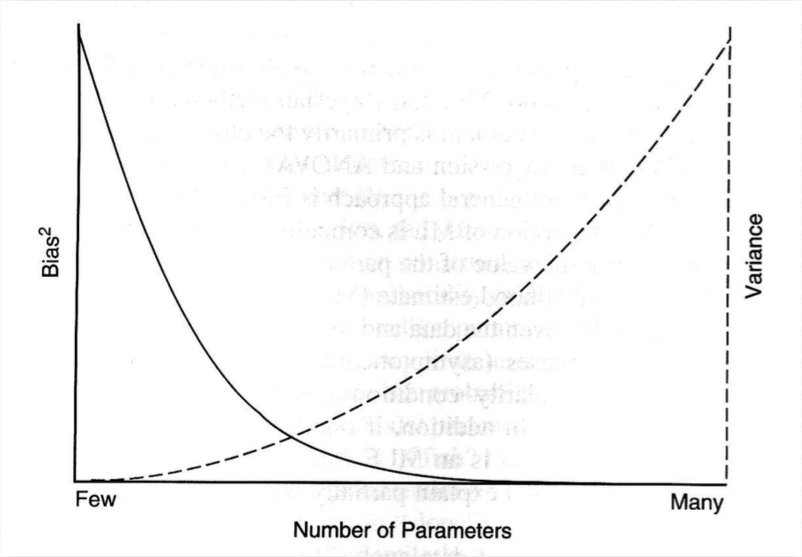
M. Drew LaMar
September 21, 2016

Introduction to Quantitative Biology, Fall 2016

Class announcements

- R packages installed on one campus computer will work on all campus computers!

The Principal of Parsimony



"…too few parameters and the model will be so unrealistic as to make prediction unreliable, but too many parameters and the model will be so specific to the particular data set so to make prediction unreliable."

- Edwards

The Principal of Parsimony

Quote: “Each time a parameter is estimated, some information is "taken out” of the data, leaving less information available for the estimation of still more parameters.“

Quote: "In model selection, we are really asking which is the best model ***for a given sample size.***”

In other words, what's the best model given the amount of information that we have?

Quote: “We are really asking - how much ***model structure*** will the data support?”

Tapering Effect Sizes

Large effects -> Medium effects -> Small effects -> …

We achieve the ability to detect ever smaller effects in a system through:

- larger sample sizes,
- better study designs, and
- better models based on
- better hypotheses.

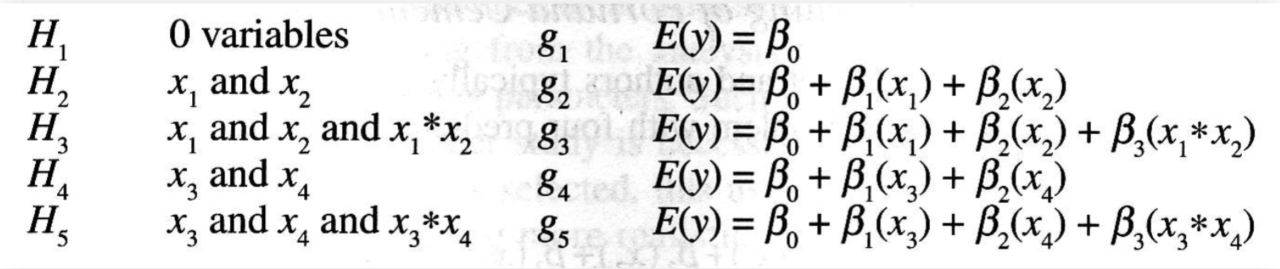
Example: Hardening of Portland Cement

Variables

- x_1: calcium aluminate
- x_2: tricalcium silicate
- x_3: tetracalcium alumino ferrite
- x_4: dicalcium silicate
- y: calories of heat per gram of cement following 180 days of hardening

Example: Hardening of Portland Cement

Hypotheses/Models



H_1	0 variables	g_1	$E(y) = \beta_0$
H_2	x_1 and x_2	g_2	$E(y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2)$
H_3	x_1 and x_2 and x_1*x_2	g_3	$E(y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_1*x_2)$
H_4	x_3 and x_4	g_4	$E(y) = \beta_0 + \beta_1(x_3) + \beta_2(x_4)$
H_5	x_3 and x_4 and x_3*x_4	g_5	$E(y) = \beta_0 + \beta_1(x_3) + \beta_2(x_4) + \beta_3(x_3*x_4)$

Example: Hardening of Portland Cement

Data

```
library(AICcmodavg)
data(cement)
str(cement)
```

```
'data.frame':   13 obs. of  5 variables:
 $ x1: int  7 1 11 11 7 11 3 1 2 21 ...
 $ x2: int  26 29 56 31 52 55 71 31 54 47 ...
 $ x3: int  6 15 8 8 6 9 17 22 18 4 ...
 $ x4: int  60 52 20 47 33 22 6 44 22 26 ...
 $ y : num  78.5 74.3 104.3 87.6 95.9 ...
```

Discuss: How many parameters can we reasonably estimate with this amount of data?

Answer: Rule-of-thumb: Number of estimable parameters = $n/10$.

Example: Hardening of Portland Cement

Data

Collinearity between variables!!!

```
cor(cement %>% select(starts_with("x")))
```

```
           x1         x2         x3         x4
x1  1.0000000  0.2285795 -0.8241338 -0.2454451
x2  0.2285795  1.0000000 -0.1392424 -0.9729550
x3 -0.8241338 -0.1392424  1.0000000  0.0295370
x4 -0.2454451 -0.9729550  0.0295370  1.0000000
```

Example: Hardening of Portland Cement

Data

Quote: “Rigorous experimental methods were just being developed during the time these data were taken (about 1930). Had such design methods been widely available and the importance of replication understood, then it would have been possible to break the unwanted correlations among the x variables and establish cause and effect if that was a goal.”

Quote: “Orthogonality arises in controlled experiments where the factors and levels are designed to be orthogonal. In observational studies, there is often a high probability that some of the regressor variables will be mutually quite dependent.”