

Statistical Modeling

Models in Science

- A conceptual construct intended to represent a phenomenon of interest



Models in R

- R is built on the notion that statistical analysis can be viewed as an exercise in statistical modeling, an exercise that *is tightly linked* to the original scientific question.
- This view provides a coherent framework for
 - conducting standard hypothesis tests, *and*
 - dealing with data that contain complexities that restrict the use of standard hypothesis tests
 - estimating effect sizes
 - prediction

What is Statistics?

- "I like to think of statistics as the science of learning from data..."

Jon Kettenring, ASA President, 1997

Example model

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb (α).

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Example model

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb (α).

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim N(0, \sigma^2)$

Grand mean of all Y_{ij}

Effect of concentration i

Random variability in Y after accounting for Pb concentration

Example model

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb (α).

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{i.} \sim N(0, \sigma^2)$$

Errors within each level of α are normally distributed with mean=0 and variance = σ^2

Example model

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb (α).

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Analysis of Variance (ANOVA)

An alternative model

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb. ***Let's consider Pb as a continuous variable (X).***

$$Y_i = \mu + \beta_1 X + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

An alternative model

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb. ***Let's consider Pb as a continuous variable (X).***

$$Y_i = \mu + \beta_1 X + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$



Rename μ as β_0

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

Simple Linear Regression

Dummy Variables

- We could rewrite the ANOVA model using the regression “terminology” via dummy variables. For example, assume 3 concentrations.
- Strategy
 - Recode the independent variables (X_i) using 0 or 1 to represent treatment levels.

Analysis of Variance (ANOVA)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

	X_1	X_2
α_1	0	0
α_2	1	0
α_3	0	1

Contrast Matrix:

The way we perform the coding of dummy variables determines how to interpret model parameters. This coding scheme is called “Treatment Contrasts” - the default in R

A further complication

- We think that the concentration of a blood enzyme (Y) is the result of exposure to Pb. We design an experiment and expose organisms to a series of concentrations of Pb (α). **Assume we also want to get rid of the possibly confounding effects of body size (S).**

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \& \quad Y_i = \beta_0 + \beta_1 S + \varepsilon_i$$

The general linear model

- Forms the basis for most classical statistics

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon_i \\ &= \beta \mathbf{X} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2 \mathbf{I}) \end{aligned}$$

Example Data Set

age	sbp
30	108
30	110
30	106
40	125
40	120
40	118
40	119
50	132
50	137
50	134
60	148
60	151
60	146
60	147
60	144
70	162
70	156
70	164
70	158
70	159

- Demo & Handout

Example 17.8 from
Zar, J. 1999. Biostatistical Analysis. 4th
Ed. Prentice Hall. ISBN 0-13-081542-X

Ancova

- Demo and Handout