TONS OF USEFUL GLM STUFF ON REGRESSION!! DO IT ALL!!!

ANOVA & Regression, one big happy family: Review of General Linear models Part 1

SMEE
ZOL851
Thursday September 18th

$$Y \sim N(b_0 + b_1 X, \sigma^2)$$

Goals

- Review basic concepts for the general linear model.
- Discuss the implementation of design matrices.

Readings for today material

Review readings (if nesc):

Dalgaard: Chapters 6,7,11,12

Also see “LinearModelBasics” Folder on ANGEL

Main readings:

GelmanHill: Chapters 3, 4,Appendix A

R_intro_guide chapter 11 (a lot of useful advice for R syntax for using lm()).

Also helpful for people who want a more mathematical treatment:

Faraway, J. (Linear Models in R) Chapter 2

(His free online book (available in the books folder on ANGEL) also

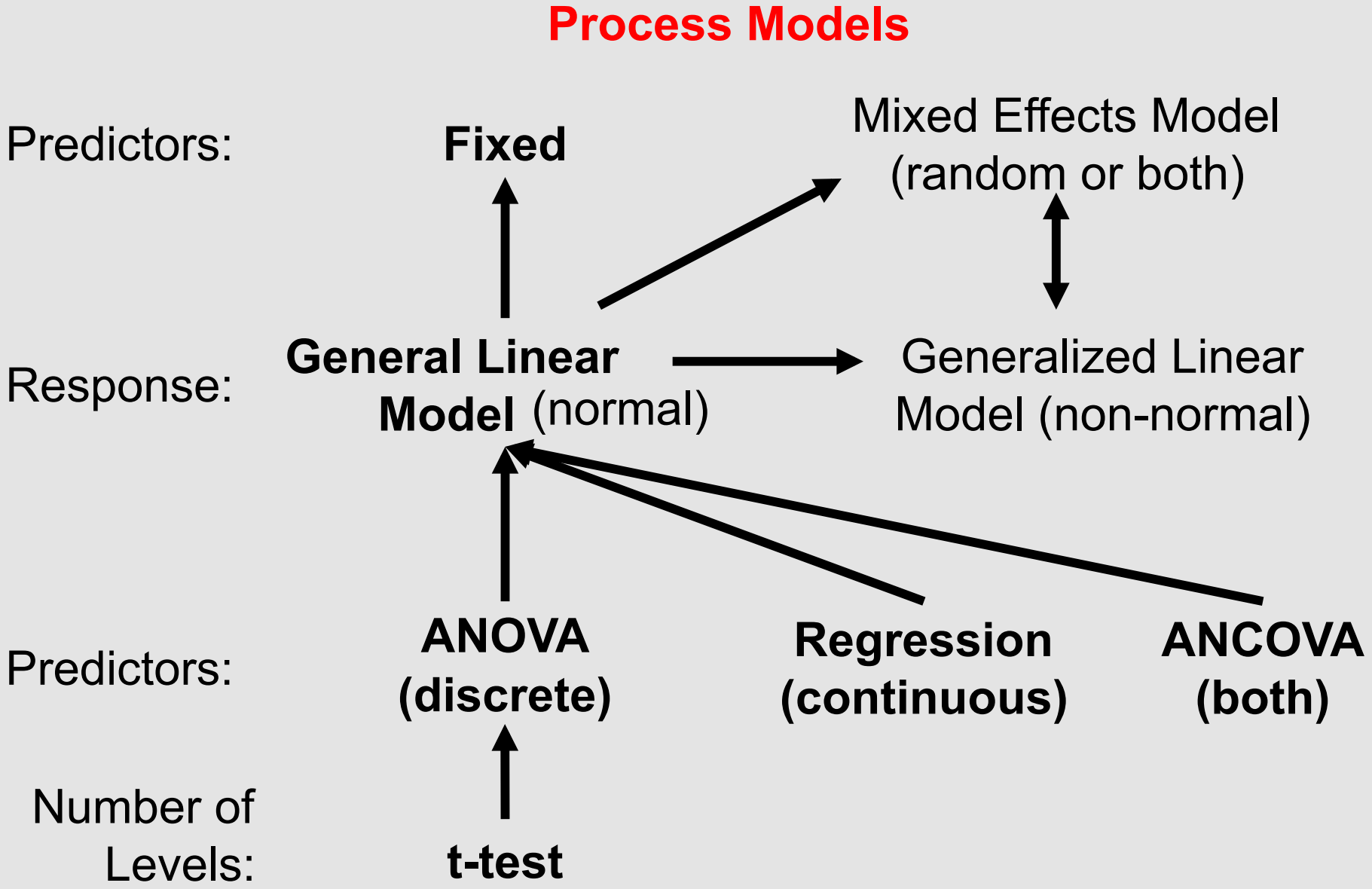
screencasts

https://github.com/idworkin/ZOL851/blob/master/LinksToScreencasts.md

Useful resource (in general)

http://www.statsoft.com/

Continuity of Statistical Approaches



What is the basic idea?

- Find the parameters that are “BEST” by some criteria (I.e. minimizing residual variation, maximizes the likelihood)
  - i.e., the line of “best fit”
  - But parameters are more than just slopes and intercepts
  - They can also represent the effects of one level of a factor
  - Right now we will use LSE to construct our objective function.
- Factors are coded as binary “is X?” questions
- Need to know how factors are coded (design matrix) in order to interpret the parameters that we get from the model.

Equivalent descriptions of general linear models

need more eqn's than unknowns

The following are all equivalent ways of stating the same model

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim \text{NID}(0,\sigma^2) \qquad i=1,\ldots,n$$

no bias in the error, normally identically distributed

$$y_i = \beta_0 + \left(\sum_{k=1}^{p} \beta_k x_{ik} \right) + e_i$$

notation for multiple predictors

Summation notation, useful for many covariates.

$$e_i \sim \text{NID}(0,\sigma^2) \qquad i=1,\ldots,n$$

For k=1,…,p covariates

$$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

N, normal. ß1x = mean. sd^2 = variance

Useful way of thinking about it for MLE or Bayesian inference.

Equivalent descriptions of general linear models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

bold is on purpose to reflect vectors **y**, β, & **e** and the design (incidence) matrix, **X**.

this = e from above

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$$

Where **I** is the identity matrix.

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}\right)$$

While we will rarely write it this way, this will become useful for mixed models where we model the random part of the model.

Vector and Matrix form for linear regression

Design matrix/incidence matrix
(latter has different meaning in quan gen)

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + e_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + e_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_n + e_n
\end{aligned}
$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
$$

general linear models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

vectors **y**, β, & **e**
design (incidence) matrix, **X**.

Computers software such as R (and SAS) use this matrix form (or a generalization of it). Thus it is worth getting familiar with it, and thinking a little about it.

If there are *n* observations and *p* observed (&/or indicator) variables, then..
y and **e** are vectors of length *n*.
β is a vector of length p.
X is matrix with *n* rows and *p* columns (written *n* x *p*).

Solving GLM: Least Squares

- How do we get our parameters (β)?

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{e}$$

$$\mathbf{e}=\mathbf{y}-\mathbf{X}\beta$$

$$\Sigma\mathbf{e}^2=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$$

- Minimize ε^2 wrt β (take derivative, set to zero, solve)

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$ (quadratic form)

(This is the OLS estimate).

Solving GLM

- How do we get our parameters (β)?
- We can also use Maximum Likelihood estimation (which we will do in a few weeks)
- Maximizes the likelihood of the model, given the data.

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

Thinking about models in statistics

Reminder: In statistics, most models have a “deterministic” & a “random” component .

The lowly regression model

$y_i = b_0 + b_1 x_i + e_i$

What is “fixed” what is random?

$Y \sim N(\mu, \sigma^2)$

$Y \sim N(b_0 + b_1 X , \sigma^2)$

General Linear Models (sometimes called LM or GLM*)

- A common statistical framework
- How are they coded?
- How are they implemented in R?

* Be careful as GLM sometimes refers to ***generalized linear models*** (*GLiM*), which are a further set of generalizations. In R general linear models use lm() and GLiMs use glm().

What is a General Linear Model?

- Response variable (y) is a linear function of a series of predictor variables (x_i's)

  e.g. $y = f(x_1, x_2, \ldots) + \varepsilon$

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$

- How is this different from regression and ANOVA?

  It isn't really, it is just a generalization.

  - x_i may be continuous or *discrete*
  - All x_i discrete is ANOVA
  - All x_i continuous is regression
  - Some of each is ANCOVA
- A unified framework for ANOVA, ANCOVA, regression
- A special case of ***generalized*** linear models

For our purposes, the various names..

- Y
- Dependent
- Response

- X
- Independent
- Explanatory
- Covariates
- predictor

general linear models

The following are all equivalent ways of stating the same model

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$y_i = \beta_0 x_0 + \beta_1 x_{i1} + e_i$$

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i$$

$$e_i \sim \text{NID}\left(0, \sigma^2\right) \qquad i=1,\ldots,n$$

What is x_0 (generally)?

general linear models

The following are all equivalent ways of stating the same model

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim \text{NID}\left(0, \sigma^2\right) \qquad i=1,\ldots,n$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

bold is on purpose to reflect vectors **y**, $\boldsymbol{\beta}$, & **e** and the design (incidence) matrix, **X**.

Computers use this final (matrix form for calculations)

general linear models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

vectors **y**, β, & **e**
design (incidence) matrix, **X**.

Computers software such as R (and SAS) use this matrix form (or a generalization of it). Thus it is worth getting familiar with it, and thinking a little about it.

If there are *n* observations and *p* observed (&/or indicator) variables, then..
y and **e** are vectors of length *n*.
β Is a vector of length p.
X is matrix with *n* rows and *p* columns (written *n* x *p*).

Y, X are observed/known. We are generally estimating β, and sometimes σ.

What makes a general linear model “linear”?

Which are linear models?

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

$$y \sim N\left(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2\right)$$

$$y \sim N\left(\beta_0 + \beta_1^x, \sigma^2\right)$$

What makes a general linear model "linear"?

Which are linear models?

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

$$y \sim N\left(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2\right)$$

$$y \sim N\left(\beta_0 + \beta_1^x, \sigma^2\right)$$

the first two are linear models:
while in math, the second eqn would be quadratic,
in statistics, X^2 could just be a transformation on X,
no different than a log transformation.

For statistics, we consider a model linear wrt to the parameters we are estimating, not wrt to the variables (the opposite of ecology and math).

wrt = w/ respect to

Assumptions of GLM

Structural Component

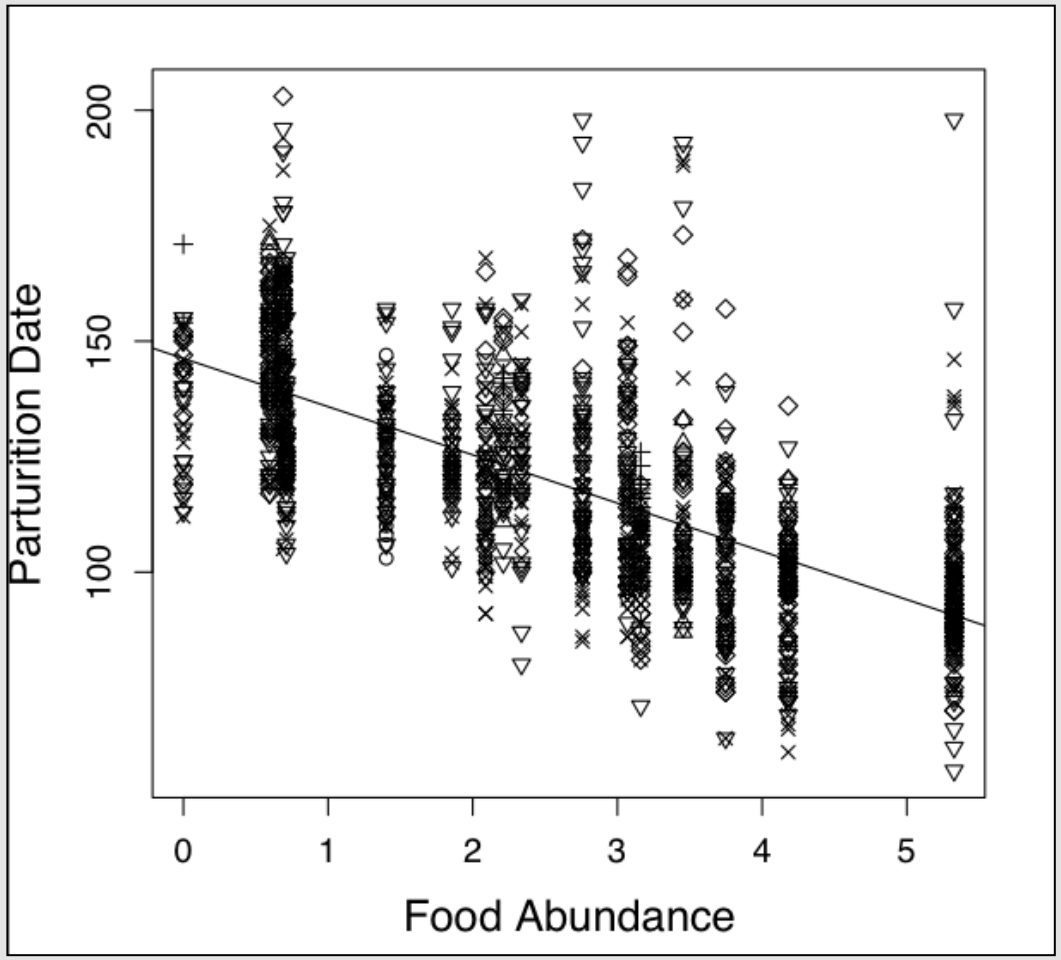
- Response (Y) is a continuous variable
- Y depends on 1 to many x_i's
- Y is **linearly** related to x_i's
  - Transformations ok
    - » e.g. $y_i = \beta_0 + e^{\beta 1\, x i}$
  - Quadratic terms ok
    - » e.g. $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$
  - Not ok
    - » Von Bertalanffy growth or Michaelis-Menton kinetics

Assumptions of GLM

- Normality of errors: $\varepsilon_i \sim N(0,\sigma_i)$
- Homoscedasticity: $\sigma_i=\sigma_j$
- Independence of errors: $cov(\varepsilon_i,\varepsilon_j)=0$
- Non-collinearity of predictors: $cov(x_i,x_j)=0$
- No errors in predictors

Plotting Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Thinking about the design matrix, **X**.

What makes general linear models so useful is how we go about setting up our Design matrix X.

For the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Where y and x are both continuous. We want to think of it like

$$y_i = \beta_0 x_0 + \beta_1 x_{i1} + e_i$$

The design matrix, **X**, for a simple regression

$$y_i = \beta_0 x_0 + \beta_1 x_{i1} + e_i$$

y_i is the observed response for the ith sample

x_0 is always 1 (for models with an intercept)

x_{i1} is the observed value for the explanatory variable for the ith sample.

β_0 is the intercept (estimated)
β_1 is the slope (estimated)

e_i is the residual (unexplained) variation for the ith sample.

$$\mathbf{X} = \begin{matrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{matrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

How would we generate a design matrix with multiple (say 4) covariates?

Work in groups

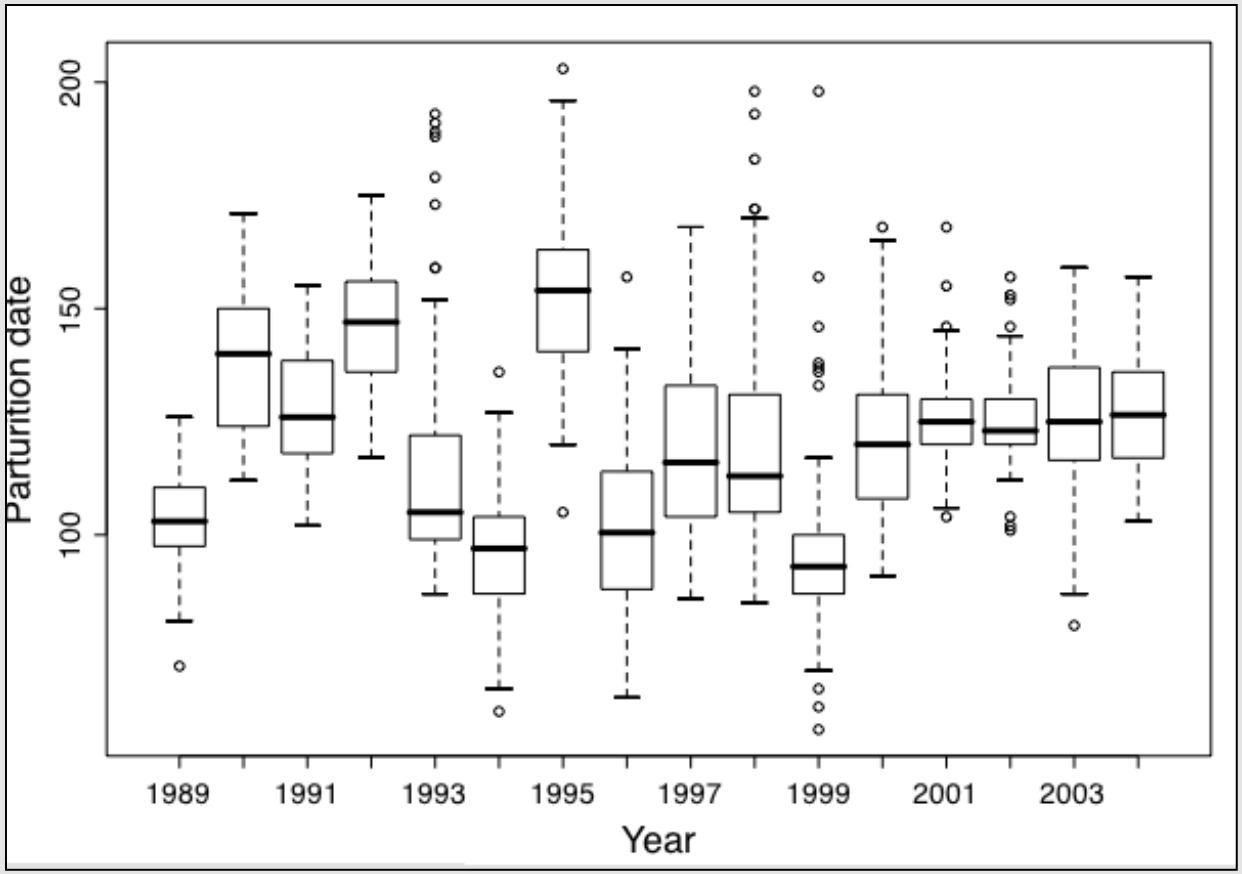
Design matrices with categorical variables (I.e. anova's).

- Once we get into using categorical variables we utilize "indicator" or dummy variables in the matrix.

- Say we have one factor with two treatment levels (lake, pond).
- How might we set up the design matrix?

Plotting Categories
(ANOVA Scenario)

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$



$i = 1, 2, \ldots, n$
$j = 1989, 1990, \ldots, 2004$

Work in groups to come up with one idea for describing categorical variables in a design matrix.

Design matrices for ANOVA's

- It turns out there are a number of options.
- Alas the most intuitive one can be the most problematic.

Design matrices for ANOVA's

- Your first thought may be to simply estimate the mean of pond and lake as a parameter each.

$$\mathbf{X} = \begin{matrix} \mathbf{1} & 0 \\ \mathbf{1} & 0 \\ \mathbf{0} & 1 \\ \mathbf{0} & 1 \\ \mathbf{0} & 1 \end{matrix}$$

Where the first column represents observations from the lake (2 observations) and the second column represents observations from the pond (3 observations).

$$\boldsymbol{\beta} = \begin{matrix} \mu_L \\ \mu_P \end{matrix}$$

This is called the "cell means" parametrized model

The problem with this form is that we have not estimated the overall mean of the population (no intercept). We also have redundancy in the matrix

So let’s just add a parameter

Design matrix for ANOVA with population mean

$$\mathbf{X} = \begin{matrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{matrix} \qquad \begin{matrix} \mu \\ \boldsymbol{\beta} = \tau_L \\ \tau_P \end{matrix}$$

Here τ_L and τ_P represent the estimated deviations for each factor level from the overall sample mean.

This is called an “Over-parameterized” or “factor-level” parametrized model.

Interpreting coefficients for the factor level parameterization of the model.

$$\beta = \begin{matrix} \mu \\ \tau_L \\ \tau_P \end{matrix}$$

$$\begin{matrix} \mu_L \\ \mu_P \end{matrix} = \begin{matrix} \mu + \tau_L \\ \mu + \tau_P \end{matrix}$$

Sadly, not the way to go..

- In an over-parameterized design matrix, our matrix is not of “FULL RANK”, since column 1 is a linear combination of columns 2 +3 (redundant).
- To solve this using the OLS approach you need to use something called a *generalized inverse*.
- It can be done via MLE, however.
- There is another way to parameterize the model…

So what should we do?

- We want to estimate the population mean, and not have redundant information, what should we do?

- We re-parametrize the model so that we can reduce number of parameters.

Design matrix for “ANOVA” with “treatment contrast” parametrization

$$\mathbf{X} = \begin{matrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{matrix}$$

$$\boldsymbol{\beta} = \begin{matrix} \mu + \tau_P \\ \tau_L - \tau_P \end{matrix}$$

Here τ_L represents a deviation of the the second treatment level (Lake), from the first level + population mean.

“”treatment contrast” parametrization of model

Interpreting the coefficients of a model set as a treatment contrast.

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \mu_P \\ \mu_L - \mu_P \end{bmatrix} = \begin{bmatrix} \mu + \tau_P \\ \mu + \tau_L - \mu - \tau_P \end{bmatrix} = \begin{bmatrix} \mu + \tau_P \\ \tau_L - \tau_P \end{bmatrix}$$

This is a useful way of parametrizing the model

- Each deviation for a level of a factor represents a test of the difference of that level from the first level.

What should we do if we have three levels for a factor?

Groups

What should we do if we have two factors each with two levels?

Groups

What should we do if we have one continuous covariate, and one categorical covariate with 3 levels?

This is a useful way of parametrizing the model

- This way each deviation for a level of a factor represents a test of the difference of that level from the first level.

What should we do if we have three levels?

What should we do if we have two factors each with two levels?

Groups

There is one other common parametrization of a model

- It is also a “sigma-restricted” parametrization of the model.

- However as you will see, we will interpret the parameters differently.

Design matrix for ANOVA with "sigma-restricted" parametrization

$$\mathbf{X} = \begin{matrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{matrix}$$

$$\boldsymbol{\beta} = \begin{matrix} \mu + \dfrac{(\tau_L + \tau_P)}{2} \\ \dfrac{(\tau_L - \tau_P)}{2} \end{matrix}$$

The first term represents the estimate half-way between the mean for Lake and Pond

"sigma-restriced", parametrization of model

Backup Slides

Handling of Discrete Variables

- If only two values (true/false, treatment/control) easy:
  - Use $x_i = 0$ or 1
- If n > 2 values, use n-1 binary values
  - Is category=2?, Is category=3?, Is category=4?, …
- In order to analyze effects of discrete variables the stats package (i.e. R) must code the data
  - You do not need to!
  - You need to understand this so that you can interpret the output.
  - Coding does not affect significance but it does affect the parameters

Example

- Three variables:
  - Body size (Mass in g)
  - Sex (M/F)
  - Breeding status (Juvenile/Firstyear/Repeat)
- Dependent variable (Y) is # of offspring

Example

How your data are setup

Mass	Sex	Breed
37.2	M	J
47.3	F	F
40.9	F	R
32.9	F	J

How stats package interprets

Mass	SEX isM?	Breed IsFY?	Breed IsR?
37.2	1	0	0
47.3	0	1	0
40.9	0	0	1
32.9	0	0	0

Coding

- Coding is the process of converting factors into numbers
- Need n-1 (not n) columns because otherwise introduce a redundancy which makes math break
- Different ways to code - “Contrast Matrix”
  - Treatment contrasts: Set control to be omitted, then β_0 is control mean and β_j is the treatment effect size of treatment *i* vs. control. This is the default in R.
  - Can remove intercept to get treatment means
  - Other more complicated coding used for mathematical reasons but tougher to interpret (e.g. Helmert contrasts in S-PLUS)
- Cannot interpret parameters without knowing coding!

One more step

- Add a constant column
- Now a linear algebra problem
- $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$
- β_0=intercept, β_1=slope for x_1,…

Y
12
14
20
11
…

=

Const	Mass	Sex	IsFY	IsR
1	37.2	1	0	0
1	47.3	0	1	0
1	40.9	0	0	1
1	32.9	0	0	0
…				

β
β_0
β_1
β_2
β_3
β_4

There is one other common parametrization of a model

- It is also a “sigma-restricted” parametrization of the model.

- However as you will see, we will interpret the parameters differently.

Design matrix for ANOVA with "sigma-restricted" parametrization

$$\mathbf{X} = \begin{matrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{matrix}$$

$$\boldsymbol{\beta} = \begin{matrix} \mu + \dfrac{(\tau_L + \tau_P)}{2} \\ \dfrac{(\tau_L - \tau_P)}{2} \end{matrix}$$

The first term represents the estimate half-way between the mean for Lake and Pond

"sigma-restriced", parametrization of model

Over to R….