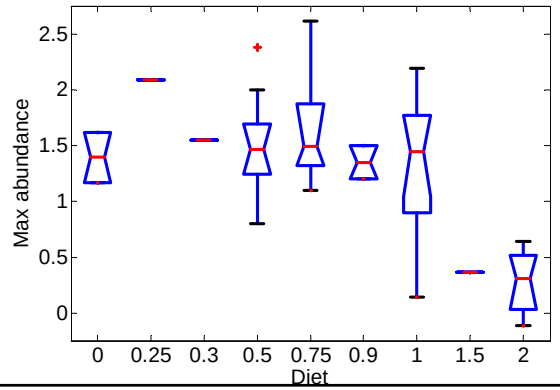
Basic GLM

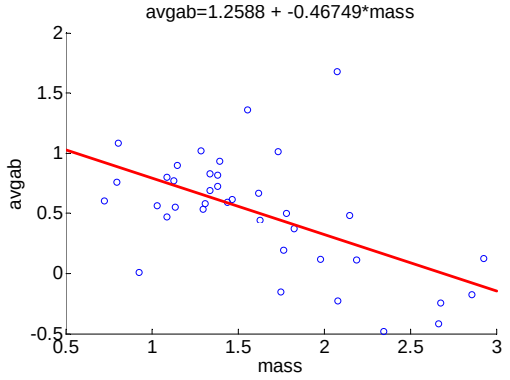
What is GLM

- $y=f(x_1,x_2,...)+\varepsilon$ where f linear
- $y=\beta_0+\beta_1x_1+\beta_2x_2+...+\varepsilon$
- How different from regression?
  - 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

Plotting ANOVA ($y_i=\mu+\tau_i+\varepsilon_i$)

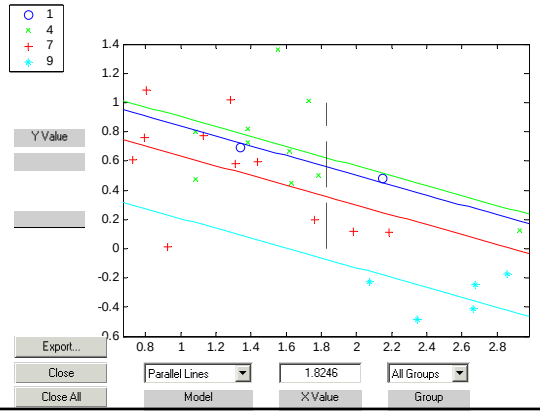


Plotting regression ($y_i=a+bx_i+\varepsilon_i$)



Plotting ANCOVA

$(y_{ij}=\mu+\tau_i+\beta_i(x_j-\underline{x})+\varepsilon_{ij})$



Handling of discrete variables

- If only two values (true/false, treatment/control) easy:
  - Use x_i=0/1
- If n>2 values, use n-1 binary values
  - Is2, Is3, Is4…

Example

- Three variables:
  - Body size (M)
  - Gender (G=M/F)
  - Breeding status (Juvenile/Firstyear/Repeat)
- Dependent variable (Y) is # of offspring

Example

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

Mass	Gender	IsF	IsR
37.2	1	0	0
47.3	0	1	0
40.9	0	0	1
32.9	0	0	0

Coding

- Coding=process of converting discrete into numbers
- Need n-1 (not n) because otherwise introduce a redundancy which makes math break
- Various tricks/variations in coding
  - Set control to be omitted, then β_i is the treatment effect size of treatment i vs. control
  - Using –1/1 for two variables (or for three variables –1/0/1) gives deviation of each factor from grand mean rather than from treatment
  - Other more complicated codings are used for mathematical reasons but impossible to interpret
- Cannot interpret without knowing coding!

One more step

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

Y
12
14
20
11

Const	Mass	Gend	IsF	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

Solving GLM

- $Y=X\beta+\varepsilon$
- $\varepsilon = Y-X\beta$
- $\varepsilon^2 = (Y-X\beta)^2 = Y^2-2YX\beta+(X\beta)^2$
- Minimize ε^2 wrt β (take derivative, set to zero, solve)
- $d\varepsilon^2/d\beta=-2X'Y+2X^2\beta=0$
- $\beta=(X'X)^{-1}X'Y$
- Least squares method and known X,Y gives us β

Three things come out of GLM

- P (alpha)
- Effect size
  - 3 types
- Explanatory power
  - Usually r^2=(variance explained)/(total variance)=% explained variance
  - Physics often get 90% or more
  - Ecology 50% and sometimes even 25% is good

Three types of effect sizes

- Absolute (μ_{treat}-$\mu_{control}$)
  - E.g population goes up by 20 individuals, yield goes up by 10 kg/ha
  - Has units
- Relative (μ_{treat}-$\mu_{control}$)/ $\mu_{control}$*100
  - Units of % - easy to detect importance
  - Independent of units (g, kg)
- Normalized d=(μ_{treat}-$\mu_{control}$)/σ
  - Normalized by variance (std deviation really)
  - Allows for comparison between studies
  - Useful in "metaanalysis"

Which advances science the most?

	p	r^2	Effect
1	0.0001	0.10	1%
2	0.06	0.9	10%
3	0.06	0.2	150%
4	0.50	0.9	100%
5	0.05	0.9	150%

Ethics

- I claim it is unethical to:
  - Report only p values without effect size and coefficient of determination (r^2)
  - Argue that the null hypothesis is true when fail to reject

Interaction terms

- Non-additive
  - +N=+2
  - +H_2O=+10
  - Interaction:
    - 0N/0H_2O=0, +N/0H_2O=0, +H_2O/0N=0, +N/+H_2O=+4

	0N	+N
0H_2O	0	+2
+H_2O	+10	+16

As a GLM

	0N	+N
$0H_2O$	0	+2
$+H_2O$	+10	+16

0
2
10
16

1	0	0
1	1	0
1	0	1
1	1	1

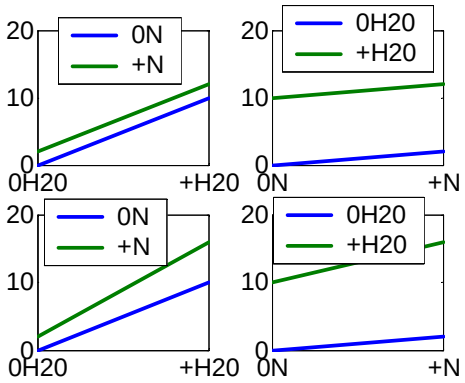
ANOVA interactions

- Result is an estimate of non-additive interactions between factors
- In practice, interaction goes in as multiplication $x_i * x_j$
- Do a significance test for

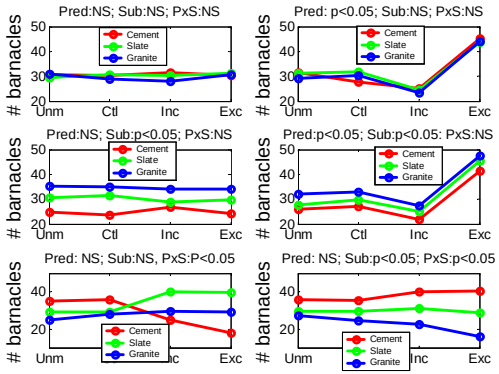
Y
0
2
10
16

c	N	W	I
1	0	0	0
1	1	0	0
1	0	1	0
1	1	1	1

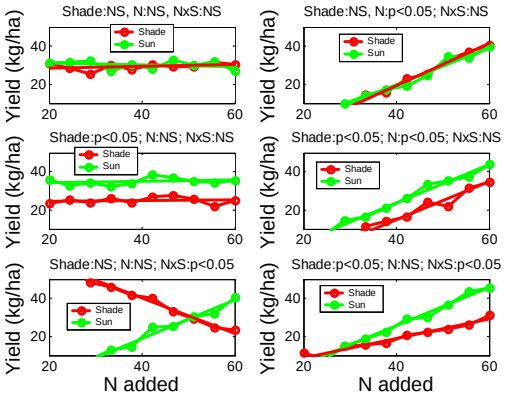
ANOVA Interaction plots



Complex interactions



ANCOVA interactions



Sum squares

- Recall ANOVA hypothesis tests are based on variance explained by a factor vs. unexplained variance
- How to calculate
  - Type I – build up sequentially A+B+A*B
    - Error vs A, A vs A+B
    - SS add up (good for variance partitioning
    - But depends on order of factors!
  - Type III – subtract out from full model
    - A+B+A*B vs B+A*B
    - Free of order
- In R must use Anova to get type II, III

Significance

- N*W p>0.05 remove N*W & rerun
- N p<0.05, N*W p<0.05 ->N has effect
- N p>0.05 N*W p<0.05 -> interaction significant, can't say anything about N
- N p>0.05, N*W p>0.05 -> N has no effect

	0N	+N
$0H_2O$	0	0
$+H_2O$	+10	+16

GLM in R

- Key concept in R is the formula
- Written as "dependent~independents"
- "+" includes additional terms
- Constant term automatically included
  - -1 omits constant (e.g. y~x-1) (or +0)
- Interactions via : & *
  - ":" interaction only Y~x:z
  - "*" includes subterms y~x*z = y~x+z+x:z
- Polynomial:
  - (var1+var2)^2 or even (a+b)*(c+d)
  - poly(var1,var2,2)
- Arithmetic
  - log(var) etc. allowed
  - I(x+y) does actual arithmetic
- Usually can specify dataframe after formula

Plotting formulae

plot(Mass~Passerine,data=birds)

boxplot(log(Mass)~Passerine,data=birds)

plot(TotalAbund~Mass,data=birds)

interaction.plot(birds$Invasive,birds$Aquatic,log(birds$Mass))

Solving formulae

- Use the "lm" function
- m<-lm(y~diet+mass)
- Usually work with a dataframe, so
  - m<-lm(y~diet+mass,data=mydata)
- What can you do with m
  - print(m)
  - plot(m)
  - summary(m)
  - predict(m,newdata)

R examples

```
#
#fitting models
#
ma<-lm(Mass~Aquatic,data=birds) #run a model
summary(ma) #summarize it
plot(ma) #plot it
ma<-lm(log(Mass)~Aquatic,data=birds) #oops - need log transform
summary(ma) #despite enormous variance, is a significant difference
plot(ma)
ma<-lm(log(Mass)~Invasive,data=birds) #no difference
summary(ma)
ra<-lm(Rangesize~TotalAbund,data=birds)
summary(ra)
plot(ra)
fitted(ra)
```

On a formula

- plot(y~x1+x2)
- interaction.plot()
- replications()

R continued

- On a model object
  - Use resid(m) or m$resid
  - resid()
  - coef()
  - fitted() # predicted
  - model.matrix()
  - abline(m) #after plot of formula)

R continued

- summary.aov(m) # factor p
- summary.lm(m) #regression & overall model p, r2, effect sizes
- anova(m) #same as summary.aov(m)
- anova(m1,m2)
- Anova(m1,m2) #library CAR allows type II, III Sum-Squares

R – contrasts (p. 370-381 Crawley)

- model.matrix(m)
- var=relevel(var,level1,..)
- contrasts(var)
- contrasts(var)=
- options(contrasts=c("contr.treatment"))