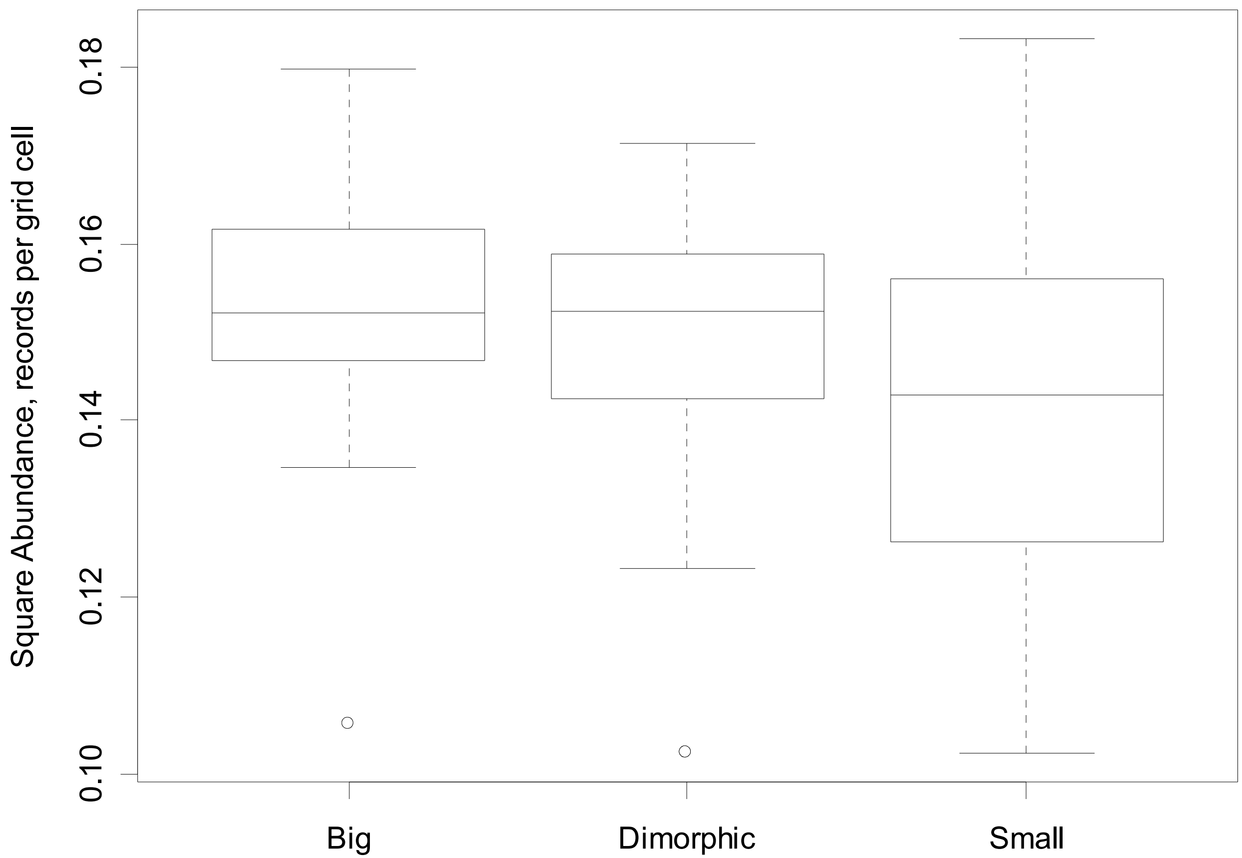
Lecture 3: ANOVA and Regression I

Bob O'Hara

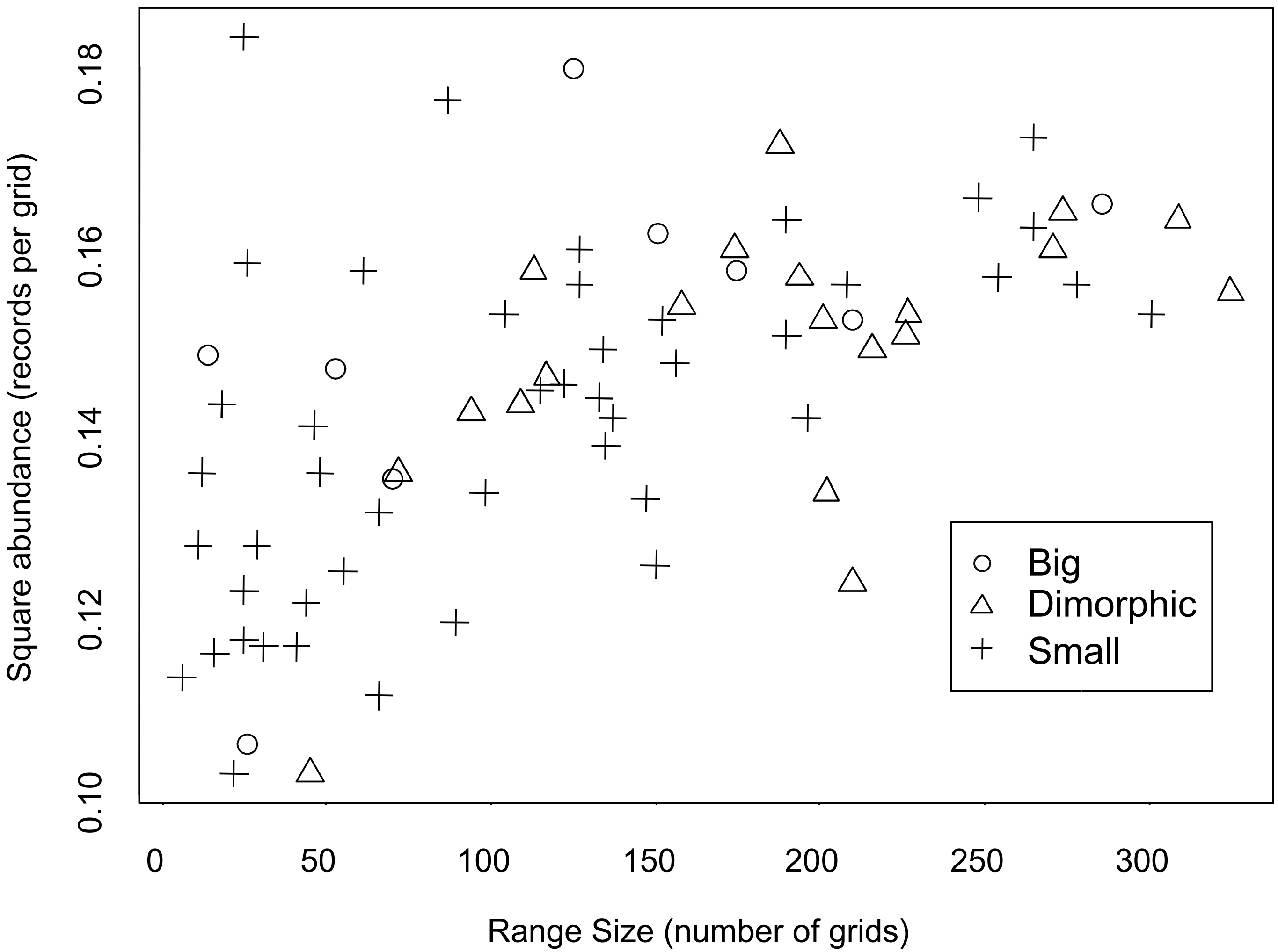
The Problem

- Is the abundance of a species related to
  - range size
  - wing type
- Data: from the ‘many’ habitat type
  - make it simpler
- Split the problem into small bits
  - easier to teach!

The Data I
Abundance of Wing Forms



The Data II
Abundance and Wing Form



Simple Problem

- Does wing form affect abundance?
  - are their mean abundances different?
- One way ANOVA, 3 groups
- Generalise - *m* groups, each with n_i observations

Model

- Start with a model
  - Each observation j, from group i has a value:
  - $y_{ij} = \mu_i + \varepsilon_{ij}$ (j=1…n_i for i=1, m)
  - μ_i is the group mean, **systematic** component
  - ε_{ij} is the error, the **random** component
- We can estimate the systematic part, but not perfectly because of the random part

Different Ways of Coding

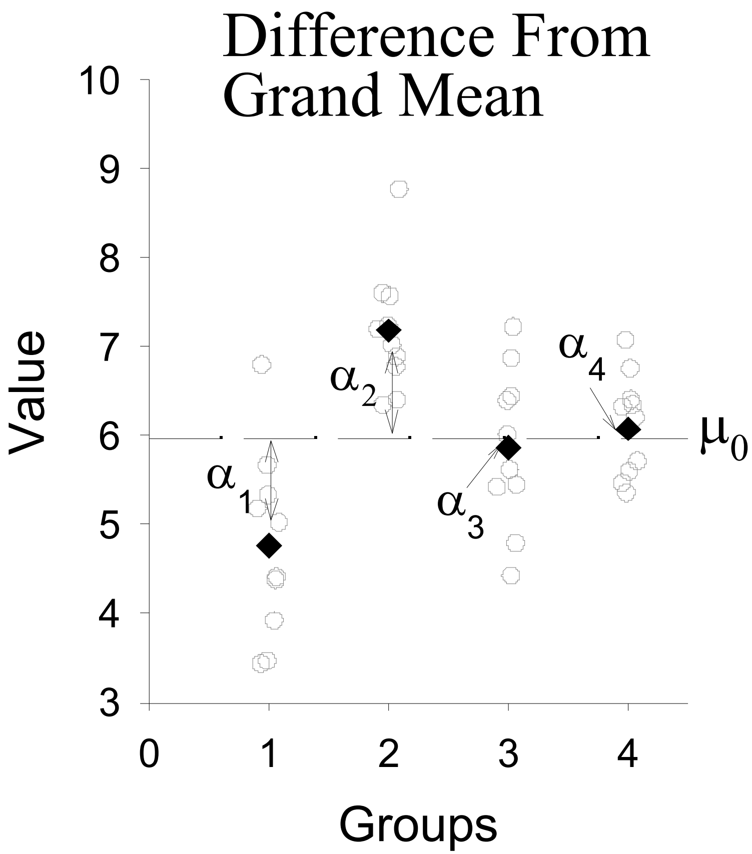
- At the moment, all μ_i's are different
  - no structure
- Can add it in several ways
  - Deviations from grand mean:

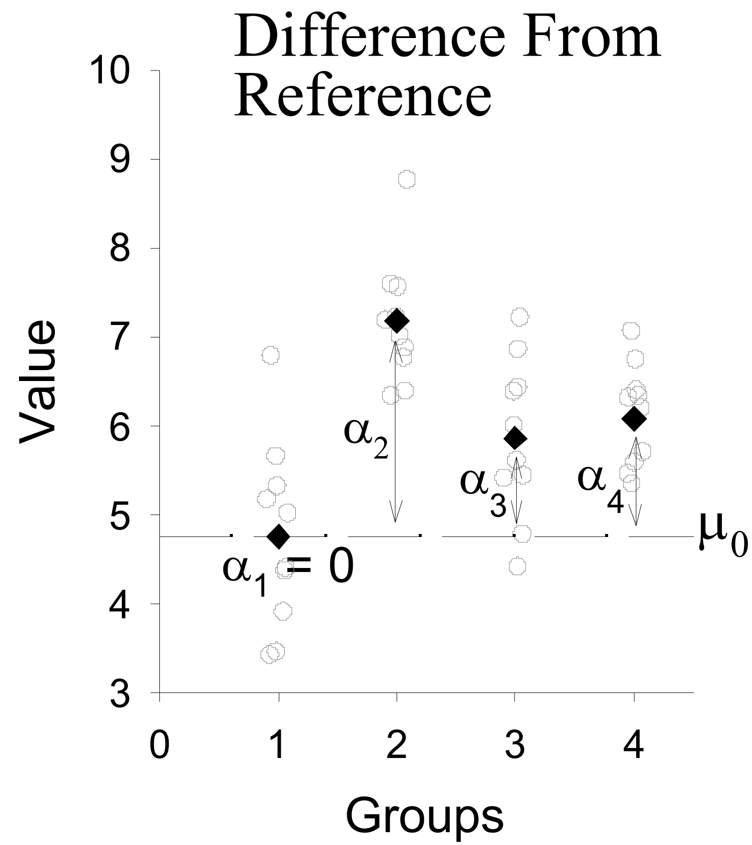
$$\mu_i = \mu_0 + \alpha_i \qquad \text{Constraint: } \sum_{i=1}^{m} \alpha_i = 0$$

  - Deviations from a reference

$$\mu_1 = \mu_0,\ \mu_i = \mu_0 + \alpha_i \qquad (i=2...m)$$

Coding in Pictures





Estimates

- Assume the ε_{ij}'s are normally distributed with mean 0 and variance σ^2.
  - notation: $\varepsilon_{ij} \sim N(0, \sigma^2)$
- Best estimates of μ_i's are their means $\bar{y}_i$
- Best estimate of the variance σ^2 is the sample variance, s^2

$$s^2 = \frac{1}{\sum_{j=1}^{m} n_j - m} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(y_{ij} - \bar{y}_j \right)^2$$

Tests

- Test whether μ_i's are the same
- Compare likelihoods
- Likelihoods are sums of squares

$$SS_E = \sum_{i=1}^{n} \sum_{j=1}^{m_j} y_{ij}^2 - \sum_{i=1}^{n} \bar{y}_{i\cdot}^2 \quad SS_B = \sum_{i=1}^{m} n_i \bar{y}_{i\cdot}^2 - m \bar{y}_{i\cdot\cdot}^2$$

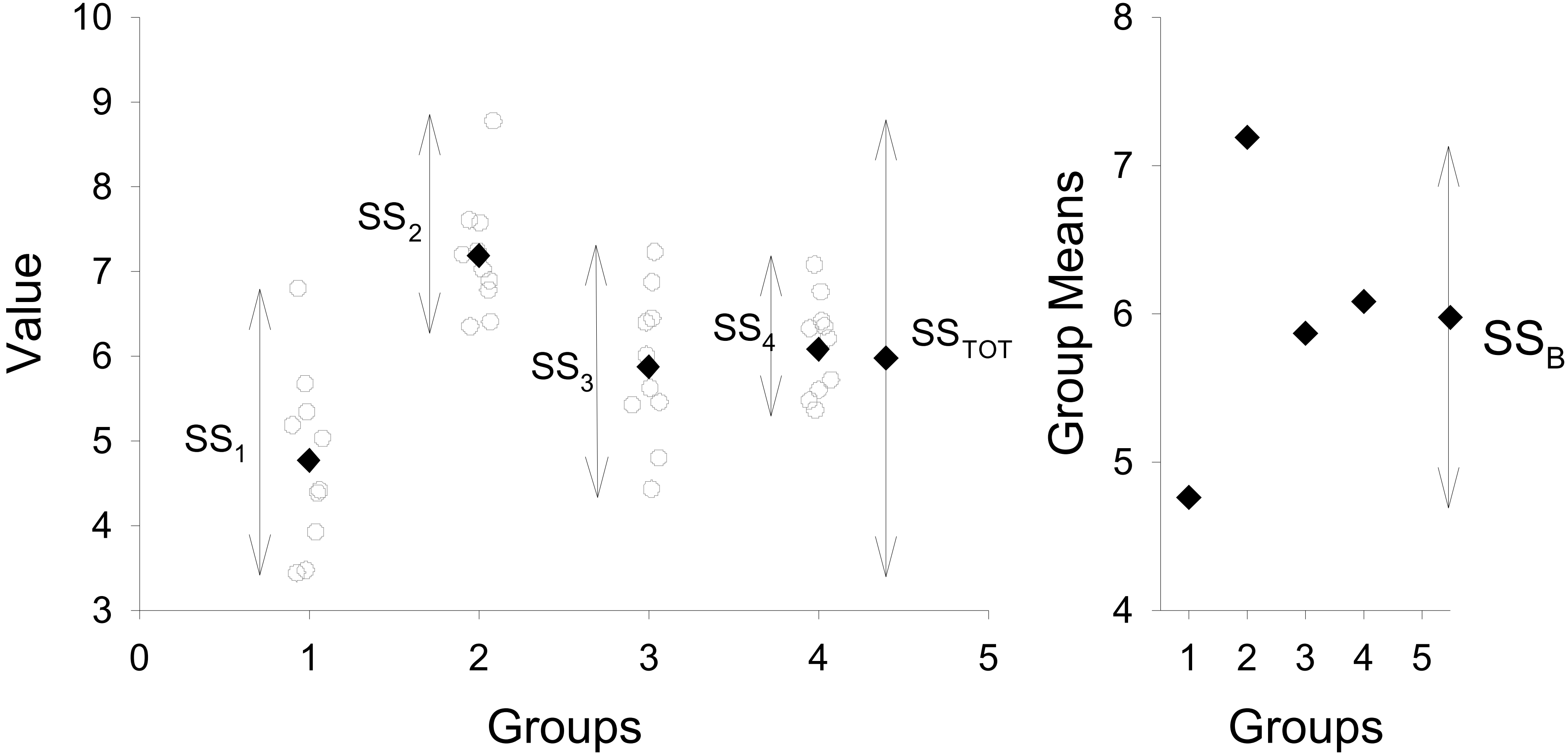
- Summarise in an ANOVA table

ANOVA Table

	df	SS	MS	F
Between	m-1	SS_B	SS_B/m	MS_B/MS_E
Within (error)	n-m	SS_E	$SS_E/(n-m)$	
Total	n-1	SS_T		

What does this mean?

$$SS_W = SS_1 + SS_2 + SS_3 + SS_4$$
$$SS_{TOT} = SS_B + SS_E$$



What it means II

- If the means are the same, then the differences in the means (as measured by SS_B) can all be explained by the error (as measured by SS_W)

Wing Size and Abundance

- 3 groups (Small, Dimorphic, Big)
- Are their abundances different?

ANOVA Table

	df	SS	MS	F	Pr(F)
Wings	2	1.12	0.56	1.69	0.19
Residuals (error)	73	24.4	0.33		
Total	75	25.5			

(Sums of Squares multiplied by 1000)

Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.150	0.0061	24.6	<2e-16
Dimorphic	-0.0015	0.00734	-0.21	0.83
Small	-0.0089	0.00665	-1.34	0.18

Residual standard error: 0.0183 on 73 degrees of freedom

- Level Big = Intercept, others give difference from this class
- *t*-tests are for difference between level and control level

(in this case Big)
Mean for Small is 0.150 - 0.0089 = 0.159
Variance for all is 0.0183

post hoc tests

- If there is a difference, where is it?
- Can use "orthogonal contrasts" for pre-planned tests
  - e.g. control vs treatments
- Most of the time: test everything
  - use *t*-tests

One Test

- Compare Dimorphic and Small
- Estimates:

	Mean	Standard Error
Dimorphic	-0.0015	0.00734
Small	-0.0089	0.00665

Residual Mean Square (σ^2): 0.33 (73 df)

- t-test:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma^2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)}}$$

Multiple Testing

- Do many tests at 5%
- If we do 20 tests where the null hypothesis is true, we would expect 1 to be rejected by chance
- We need to adjust the significance level
- Bonferroni: k tests at α%
  - change test level from α to α/k%
  - works OK if k not too large

Significance Tests are Evil

- A t statistic increases by $n^{1/2}$ as the sample size (n) increases
- For example: Correlation

  n=10, ρ=0.39, p=0.26

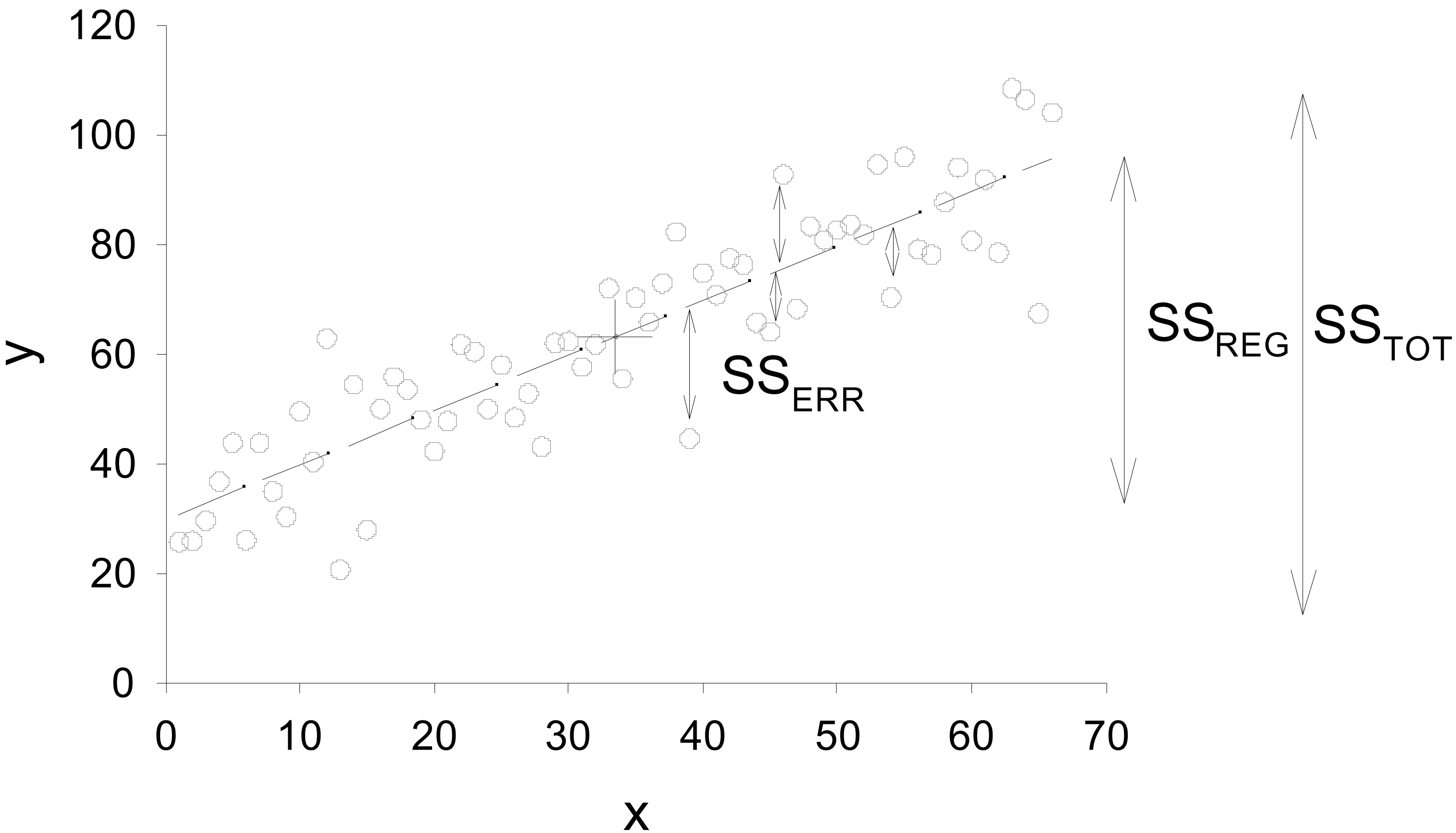
  n=30, ρ=0.39, p=0.033
- With very large sample sizes, almost p values will be <5%, but the effect sizes could be tiny
- You have been warned….

Simple Regression

- Similar to ANOVA
- Also use an ANOVA table to test hypotheses
- Estimate of slope and intercept:
  - intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - slope: $$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)}$$

Regression Picture

$$SS_{TOT} = SS_{REG} + SS_{ERR}$$



ANOVA Table

	df	SS	MS	F
Regression	1	SS_{REG}	$SS_{REG}/1$	MS_{REG}/MS_{ERR}
Error	n-2	SS_{ERR}	$SS_{ERR}/(n-2)$	
Total	n-1	SS_{TOT}		

Does Range Size Predict Abundance?

	df	SS	MS	F	Pr(>F)
Regression	1	8.20	8.20	35.09	$<10^{-7}$
Error	74	17.3	0.233		
Total	75	25.5			

(Sums of Squares multiplied by 1000)

Range Size - Abundance II

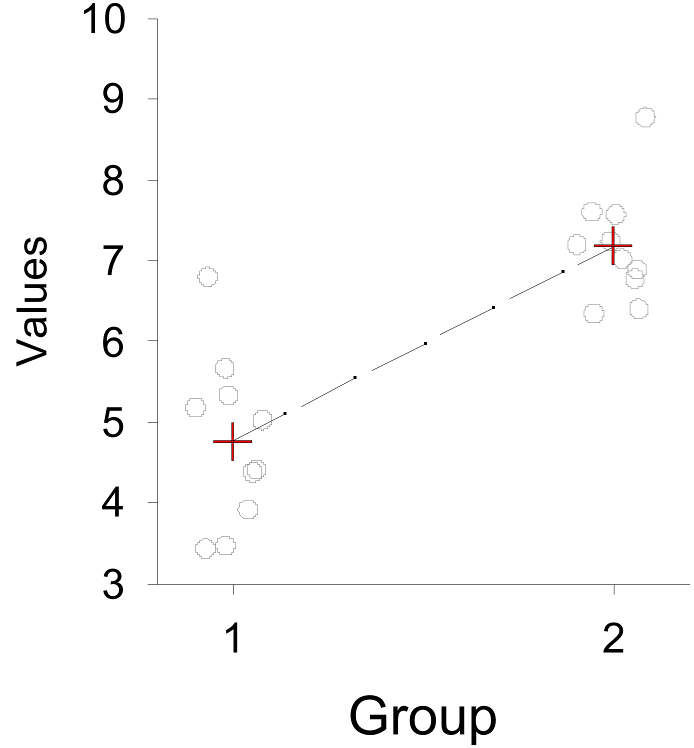
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.029	3.14×10^{-3}	41.0	$<2 \times 10^{-16}$
Range Size	1.18×10^{-4}	1.99×10^{-5}	5.92	$<10^{-7}$

Residual standard error: 0.0153 on 74 degrees of freedom
Multiple R-Squared: 32%, Adjusted R-squared: 31%

- R^2: Percent of variation explained by the regression
- Adjusted R^2: adjusts for the number of degrees of freedom
    - easier to compare different sized models

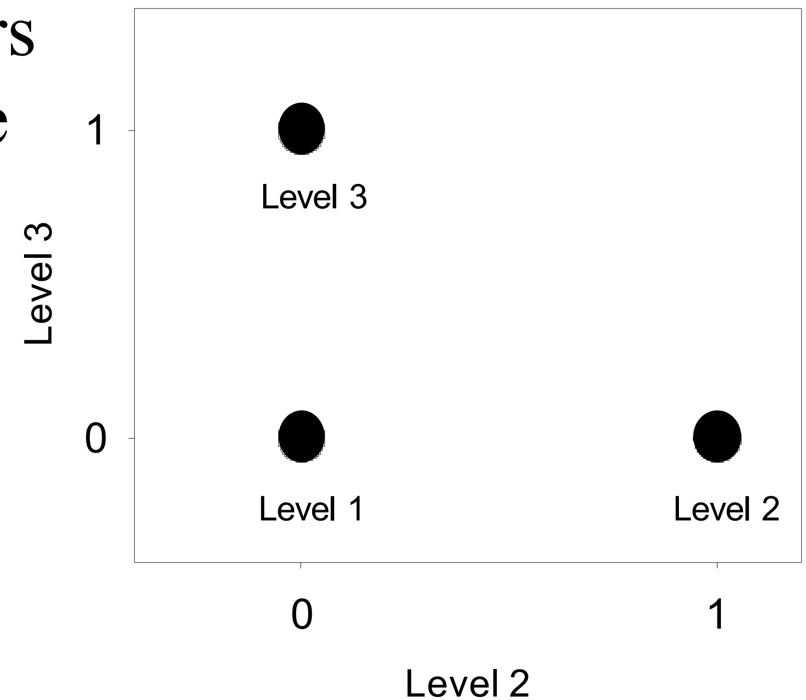
Regression and ANOVA

- Actually the same
- One-way ANOVA, 2 groups:



Regression = ANOVA

- One way ANOVA, 3 groups
  - Regression with 2 regressors
  - View Level 1 as a reference
  - Levels 2 and 3 change the mean from Level 1

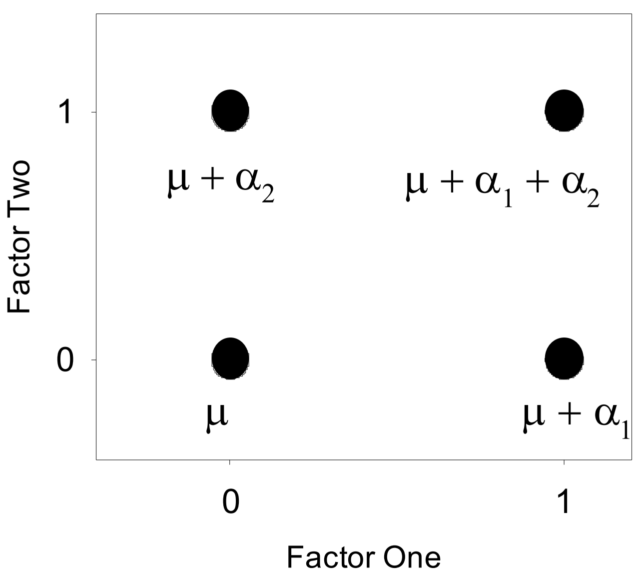


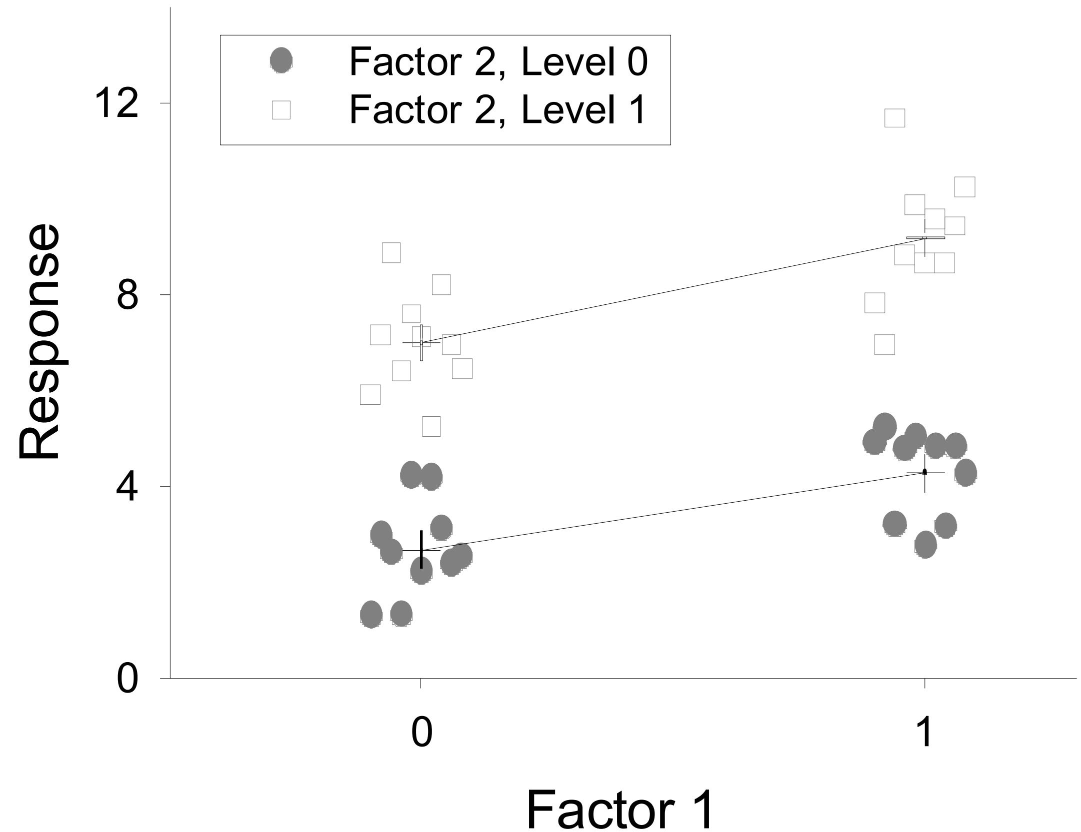
ANOVA - more parameters

- In 1 way ANOVA, adding each group adds another regression
  - number of parameters builds up
- This is why replication is such an issue
- Makes design of experiments important

Into something more difficult

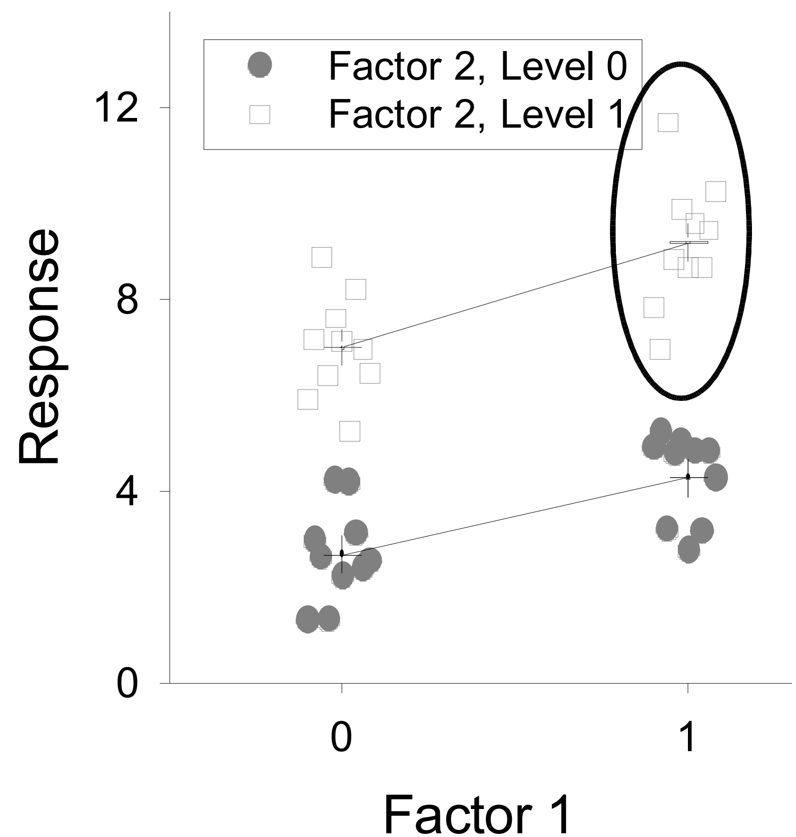
- 2 way ANOVA: no interactions





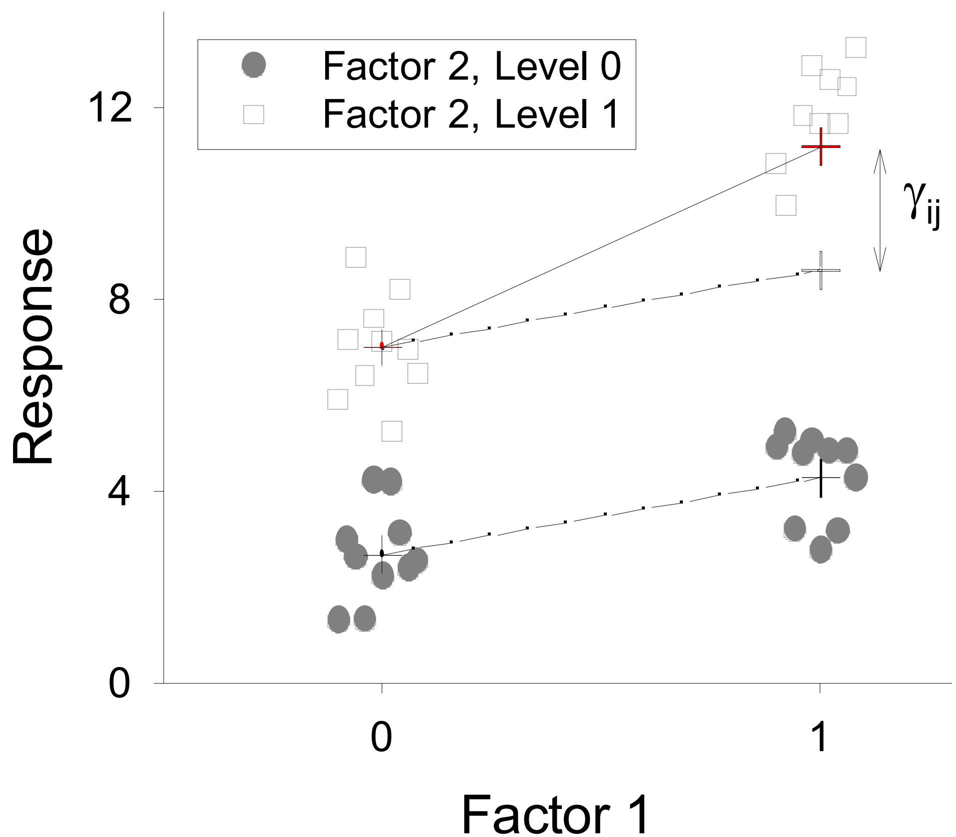
Interactions

- Here the mean of the ringed group is determined by the 3 other parameters
  - μ, α_1 and α_2
- What if it is different?



Interaction

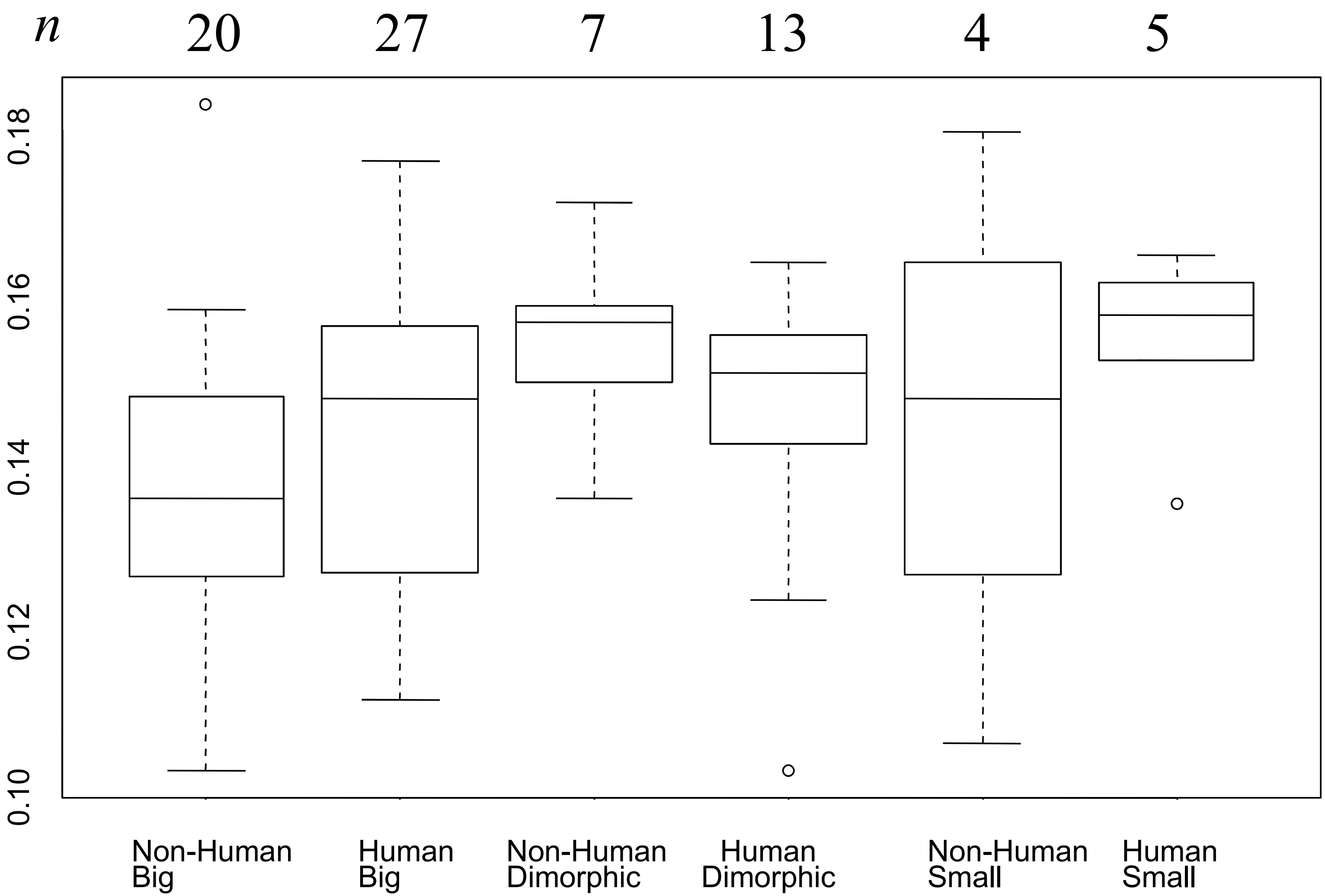
- $y_{ijk} = \mu_0 + \alpha_1 + \alpha_2 + \gamma_{12} + \boldsymbol{\varepsilon}_{\boldsymbol{12k}}$



Back to Beetles

- As well as wing form, we also know whether they like human habitats
- Does this have an effect?
- Does any effect vary with wing form?

Box plots



n
20
27
7
13
4
5
0.18
0.16
0.14
0.12
0.10
Non-Human
Big
Human
Big
Non-Human
Dimorphic
Human
Dimorphic
Non-Human
Small
Human
Small

The ANOVA

	df	SS	MS	F	Pr(F)
Wings	2	1.12	0.56	1.69	0.19
Human	1	0.14	0.14	0.42	0.52
Wings by Human	2	0.82	0.41	1.23	0.30
Residuals	70	23.4	0.33		
Total	75	25.5			

The ANOVA - change order

	df	SS	MS	F	Pr(F)
Human	1	0.17	0.17	0.51	0.48
Wings	2	1.09	0.54	1.63	0.20
Wings by Human	2	0.82	0.41	1.23	0.30
Residuals	70	23.4	0.33		
Total	75	25.5			

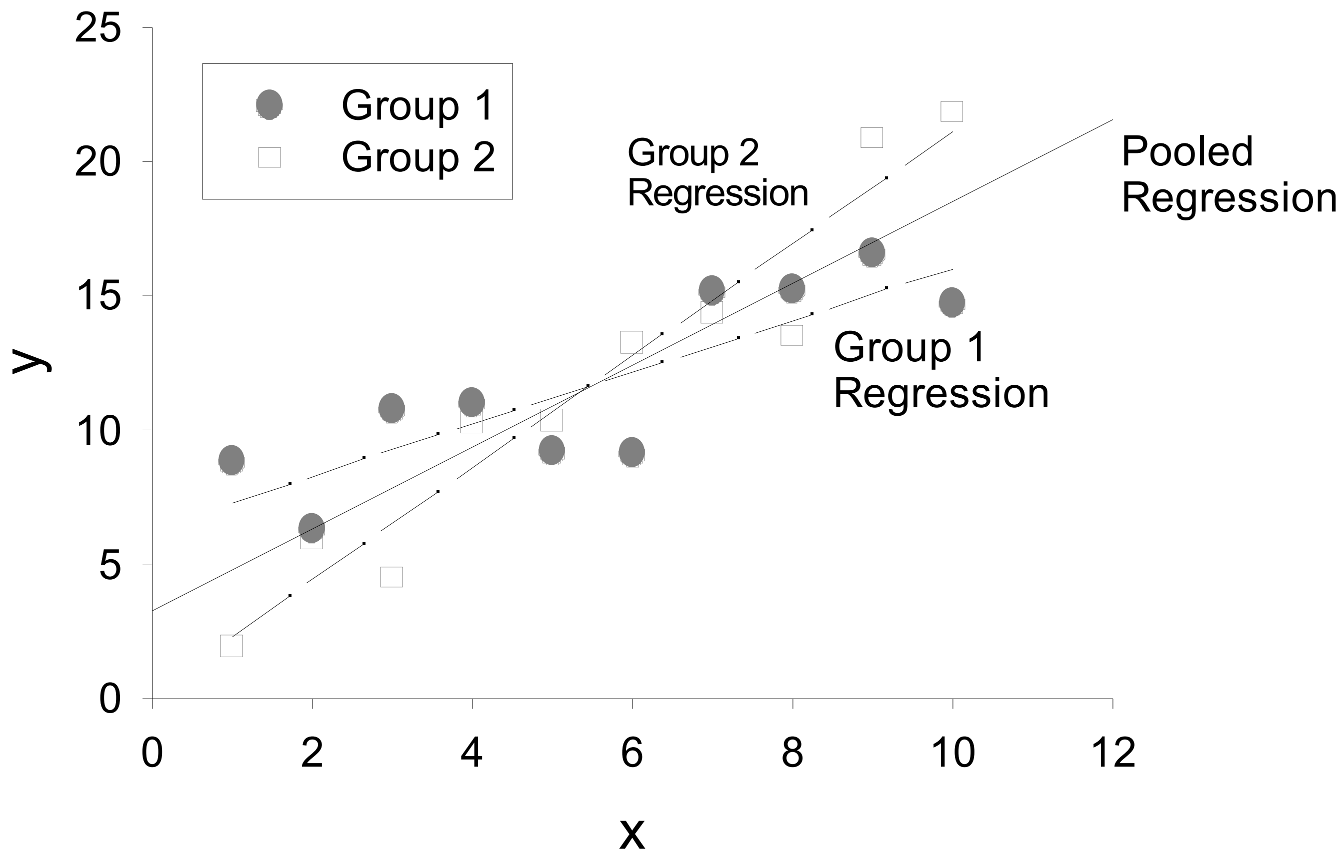
Comments on the ANOVA

- The total sums of squares and total df will be the same for different models
  - total amount of variation in the model is the same
- Terms are added sequentially
  - Terms are really (Wings), (Human | Wings), (Wings by Human | Human & Wings)
- Order in which they are added makes a slight difference

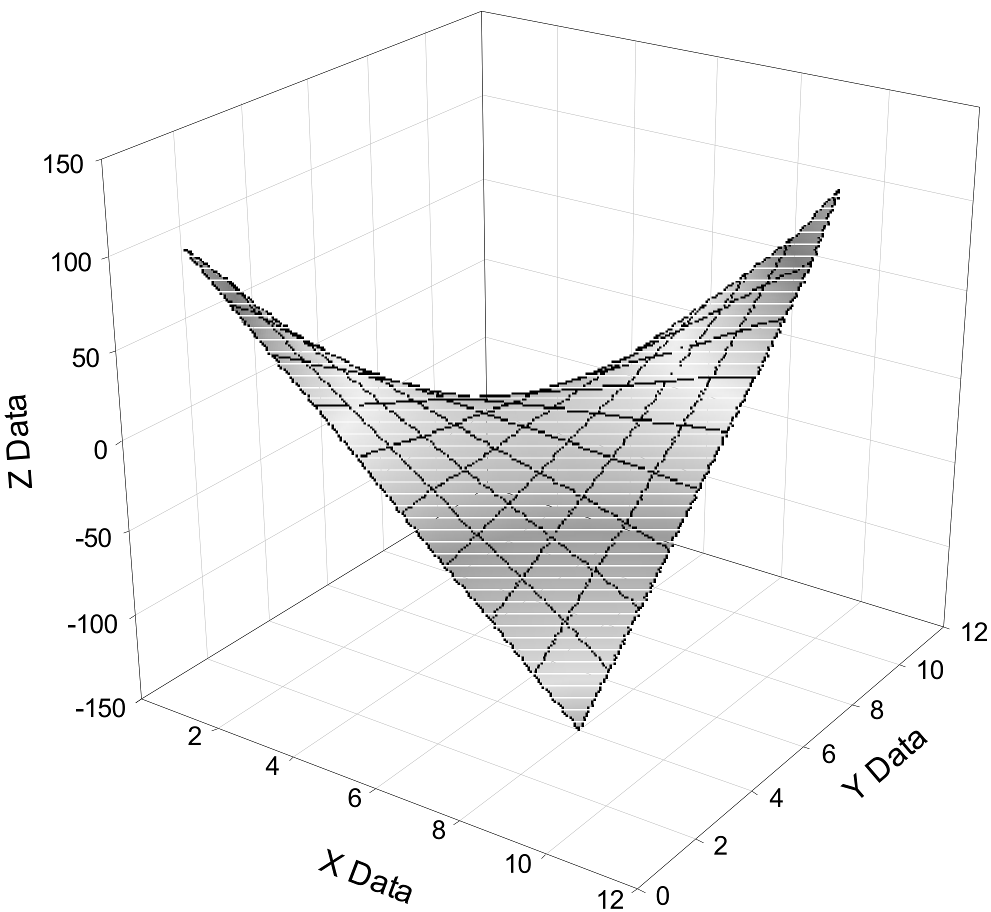
Regression + ANOVA

- As they are both the same, can do together
- And then can add lots of other factors...
- What does an interaction mean?
  - slopes are different for different levels of a factor
- What would an interaction between 2 slopes mean?
  - the model would include $\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$

Different Slopes



Regression Interaction



Getting much more complicated

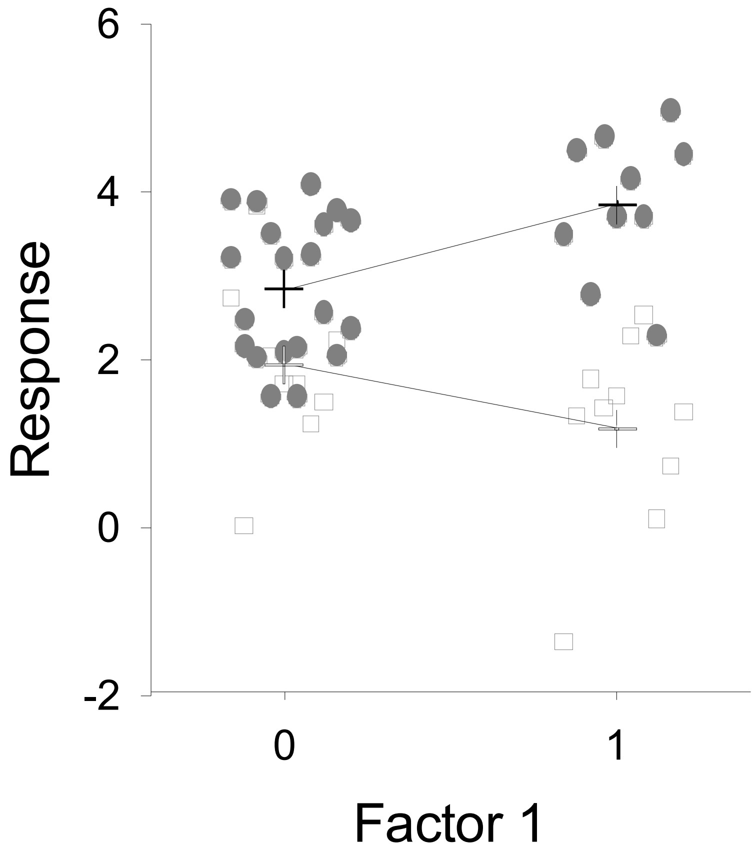
- As well as fitting linear terms for slopes, we can also fit polynomials
  - x^2, x^3, etc.
- Interactions are a natural extension
  - Model: $(1+x_1)\ (1+x_2) = 1+x_1+x_2+x_1x_2+x_1{}^2+x_2{}^2$
- When fitting polynomial terms, keep lower order terms
  - unless you've got a good reason to drop them!

Interactions

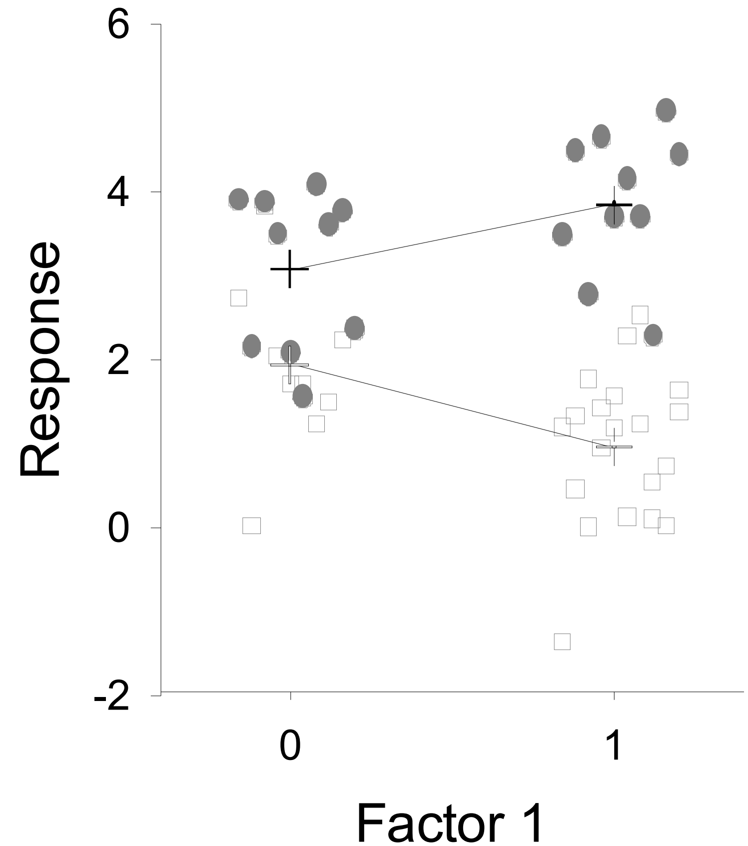
- The order they are added can be important for the ANOVA
  – Unless the experiment is balanced
- When adding interactions, keep main effects
  – unless you’ve got a good reason to drop them!
- In the presence of an interaction, the main effect is normally meaningless

Bad Main Effects...

Data Set 1



Data Set 2



ANOVA tables

	df	Data Set 1 SS	Data Set 1 F	Data Set 1 Pr(>F)	Data Set2 SS	Data Set2 F	Data Set2 Pr(>F)
Factor2	1	31.37	35.8	**$3x10^{-7}$**	11.9	8.41	**0.0057**
Factor1	1	0.75	0.85	0.36	7.17	5.05	**0.029**
Factor1 by 2	1	8.93	10.2	**0.0025**	20.8	14.66	**$3x10^{-4}$**
Error	46	40.26			65.30		

With a significant interactions

- Main effects can be informative when
  - All changes are in the same direction
  - from observational studies, the overall direction may be interesting
- If the interaction can be considered a random effect
  - then change the ANOVA to test against the correct error term