

Module 4D

Supervised Machine Learning

Different flavors of REGRESSION and General Linear Models

General linear model

- Linear Model for single-factor ANOVA
- Linear Regression

$$Y = \mu + A_i$$

$$A_i = \text{group mean} - \mu$$

$$Y = \alpha + \beta X$$

You are fundamentally fitting two models in both cases

RESPONSE = CONSTANT + VARIABLE

- Analysis of covariance
- Multiple regression

General linear models:

H_0 : Treatment means are same

H_A : Treatment means are not all the same

Significance of a treatment variable is tested by comparing the fit of two models, H_0 and H_A , to the data by using **F-test**

$$\text{F-test} = \frac{H_A}{H_0} = \frac{\text{Constant} + \text{Variable}}{\text{Constant}}$$

Does the additional parameter, the variable, improve the fit of the data significantly?

- ANOVA table
- P-value leads to rejection or FTR H_0
- Assumptions are same (residual plots): random sample, normal distribution,
Variance of response variable is the same for all combinations of the explanatory variables

GLM: just a curated taste (there are many more)!

Often appropriate/useful to investigate >1 explanatory variable simultaneously

Efficiency

Interactions

Three major approaches:

Blocking

Improve detection of treatment effects

If nuisance variable is known and controllable

Factorial experiment

Investigate main effects of ≥ 2 treatment variables

Interactions

Covariates

Confounding variables

Nuisance variable is known but uncontrollable

Multiple factor ANOVA:

- A factor is a categorical variable
- ANOVAs can be generalized to look > 1 categorical variable at a time
 - Same principles as one-way ANOVA
 - partitioning of variance
 - Same assumptions as one-way ANOVA
 - Equal variances
 - Equal sizes
- *Not only can we ask whether each categorical variable affects a numerical variable, but also do they **interact** in affecting the numerical variable*
 - The most important aspect of multi-factor ANOVA is that we can determine whether groups differ on some dependent variable while controlling for the effects of the other independent variables
- Similar to ANCOVA but ANCOVA is more general

One-way ANOVA:

- 1 continuous dependent variable
- 1 categorical independent variable (≥ 2 groups)
- i.e., **Girls vs boys** in hours of tv watched

Multi-Factor ANOVA:

- 1 continuous dependent variable
- ≥ 2 categorical independent variables
- i.e., **Girls vs boys** in hours of tv watched in **four regions** of the United States

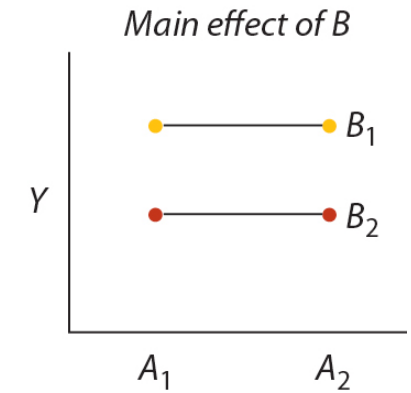
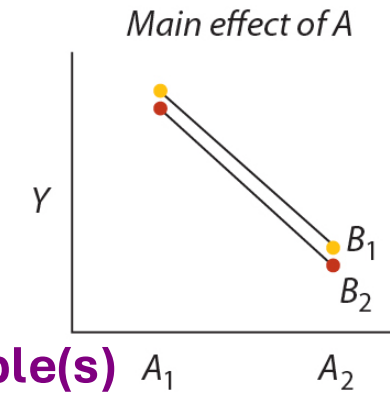
Multiple factor ANOVA:

- 1 dependent variable and ≥ 2 (independent) categorical variables
- Produces 2 interesting results:

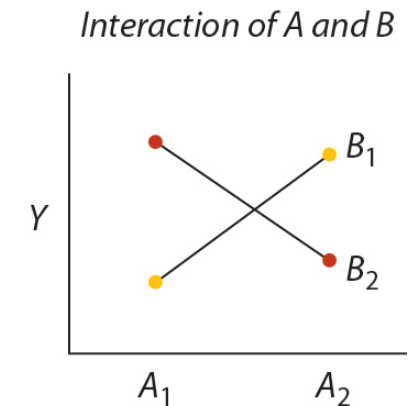
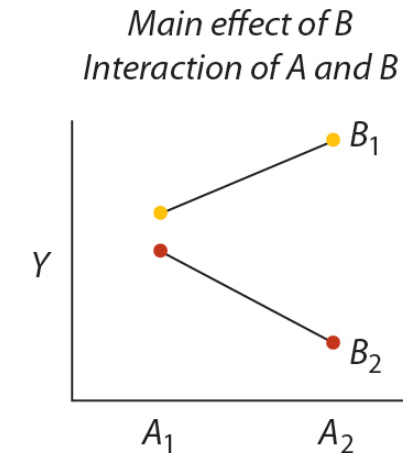
1. Main effects

- 1 F-value for each category
- Like one-way ANOVA but with **one glorious difference:**

Control for (partial out the effects of) the other independent variable(s)



2. Interaction effects



Fixed Factorial Designs:

- Effects of factors (treatments) and their interactions on a response variable
 - All combinations of the two (or more) explanatory variables are investigated
- Fixed, repeatable factors
 - Interaction term:** if it equals 0, there is no interaction
 - Main effects:**
 - Factor 1 and Factor 2 since they represent the effects of that factor alone when averaged over the other factor, ie. Marginal values

$$\text{Response} = \text{Constant} + \text{Factor 1} + \text{Factor 2} + \text{Factor 1} * \text{Factor 2}$$

- **F-test**
 - Contribution of each main effect and their interaction to the fit of the model to the data

Fixed Factorial Designs: **Response = Constant + Factor 1 + Factor 2 + Factor 1* Factor 2**

Three sets of null/alternate hypotheses to test:

1. H_0 : Main effect: **Factor 1**

$$\text{F-test} = \frac{\text{Constant} + \text{Factor 1} + \text{Factor 2} + \text{Factor 1*Factor 2}}{\text{Constant} + \text{Factor 2} + \text{Factor 1*Factor 2}}$$

2. H_0 : Main effect: **Factor 2**

$$\text{F-test} = \frac{\text{Constant} + \text{Factor 1} + \text{Factor 2} + \text{Factor 1*Factor 2}}{\text{Constant} + \text{Factor 1} + \text{Factor 1*Factor 2}}$$

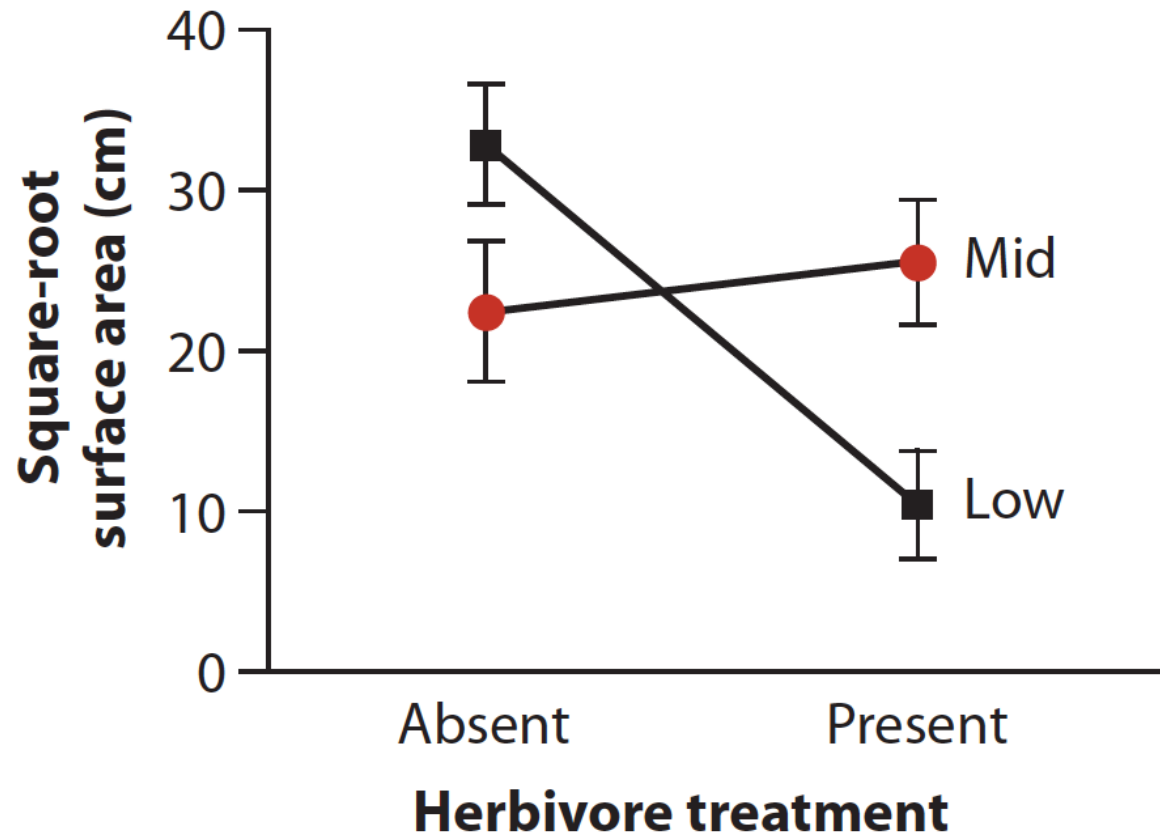
3. H_0 : Interaction effect: **Factor 1*Factor 2**

$$\text{F-test} = \frac{\text{Constant} + \text{Factor 1} + \text{Factor 2} + \text{Factor 1*Factor 2}}{\text{Constant} + \text{Factor 1} + \text{Factor 2}}$$

Source of Variation	Sum of Squares	df	Mean Square	F	P
Factor 1					
Factor 2					
Interaction					
<u>Residual</u>					
Total					

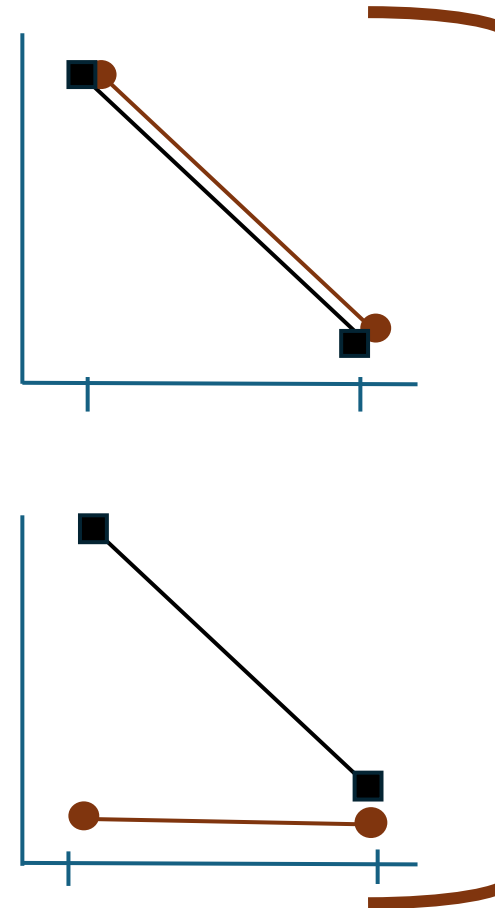
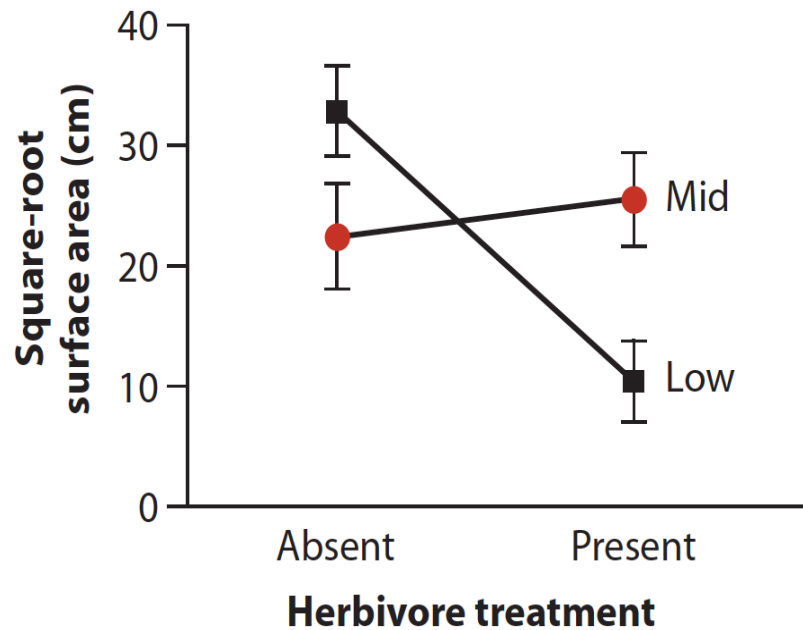
Multi- factor ANOVA

Example: Herbivores affect on red algae in an intertidal zone: exclusion and presence. Two locations variables, low tide mark and middle mark.



Multi- factor ANOVA

Example: Herbivores affect on red algae in an intertidal zone: exclusion and presence. Two locations variables, low tide mark and middle mark.



The other types of patterns that you might see on a multi-factor graph

Multi- factor ANOVA:

Testing three hypothesis pairs:

Herbivory (main effect):

H_0 : **No difference between** herbivory treatments in mean algal cover

H_A : There is a difference between herbivory treatments in mean algal cover

Height (main effect):

H_0 : **No difference** between height treatments in mean algal cover

H_A : There is a difference between height treatments in mean algal cover

Herbivory*Height (interaction effect):

H_0 : The effect of herbivory on algal cover **does not** depend on height in the intertidal region

H_A : The effect of herbivory on algal cover **does** depend on height in the intertidal region

Source of Variation	SS	DF	MS	F	P
Herbivory	1512.18	1	1512.18	6.36	0.014
Height	88.97	1	88.97	0.37	0.543
Herbivory*Height	2616.96	1	2616.96	11.00	0.002
Residual	14270.52	60	237.842		
Total	18488.63	63			

Three F ratios in the table; two of them are significant.

No interaction between height and herbivory is rejected

No effect of herbivory is rejected

Source of Variation	SS	DF	MS	F	P
herbivores	1512.18	1	1512.18	5.5227	0.02197
Residuals	16976.5	62	273.81		

Source of Variation	SS	DF	MS	F	P
height	88.973	1	88.973	0.2998	0.586
Residuals	18400	62	296.769		