Module 4
Supervised Machine Learning

Different flavors of REGRESSION and General Linear Models

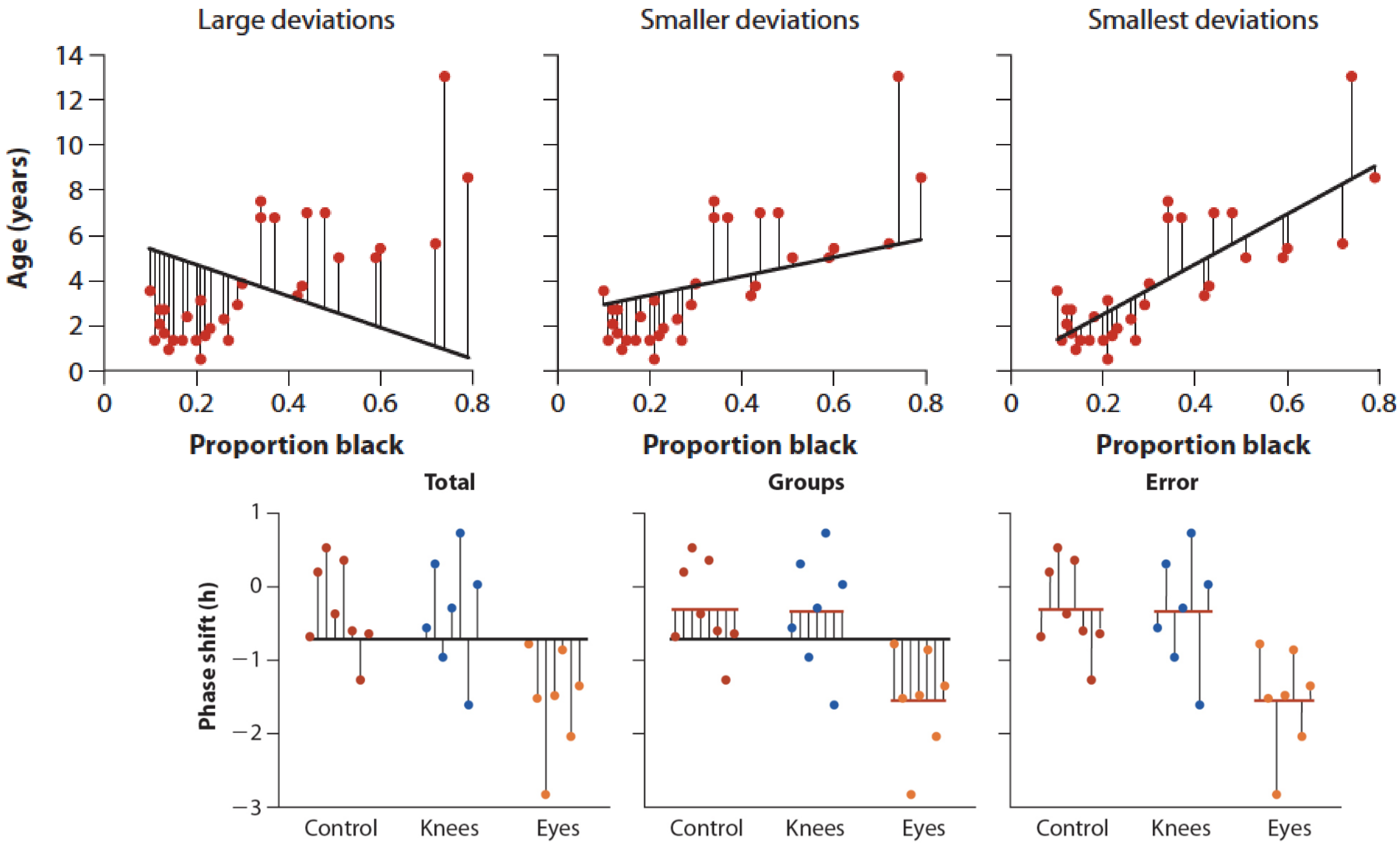


Figure 20.1: Whitock and Schluter, Fig 15.1.2 – Illustrating the partitioning of sum of squares into MS_{group} and MS_{error} components.

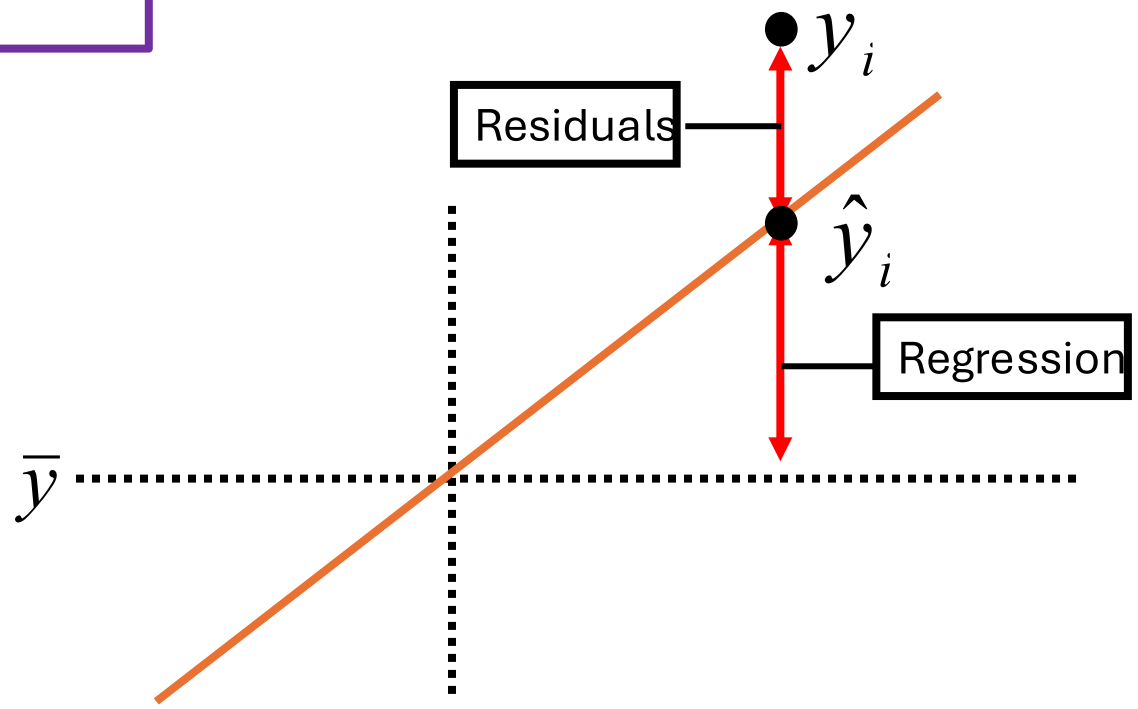
Regression Overview

Least Squares:

- What are the elements of this equation?

$$SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = a + bx_i$$



Finding **a**:

$$\boxed{\bar{Y} = a + b\bar{X}}$$

OR

$$\boxed{a = \bar{Y} - b\bar{X}}$$

Regression fallacy:

- Tricky concept:
  - Each individual has a **true** value, but the sampled value varies with time
    - the subset who scored highest on the first round included individuals who had higher values then their usual ‘true’ value
    - the second measurement captured these individuals when they happened to be closer to their own personal normal values

- failure to consider “regression towards the mean” when interpreting the results of **observational studies**

- can be a large problem when dealing with **sick** people - they are the tail of the distribution, and they might appear to improve even if the treatment applied has no real effect

Testing hypotheses about slope:

1. H_0: $\beta = \beta_0$ (N.B. The null hypothesis is that Y cannot be predicted from X)

   H_A: $\beta \neq \beta_0$

2. Test statistic: $\mathbf{t = \dfrac{b - \beta_0}{SE_b}}$ $$SE_b = \sqrt{\frac{MS_{residual}}{\square\,(X_i - \bar{X})^2}}$$

3. significance level; df=n-2

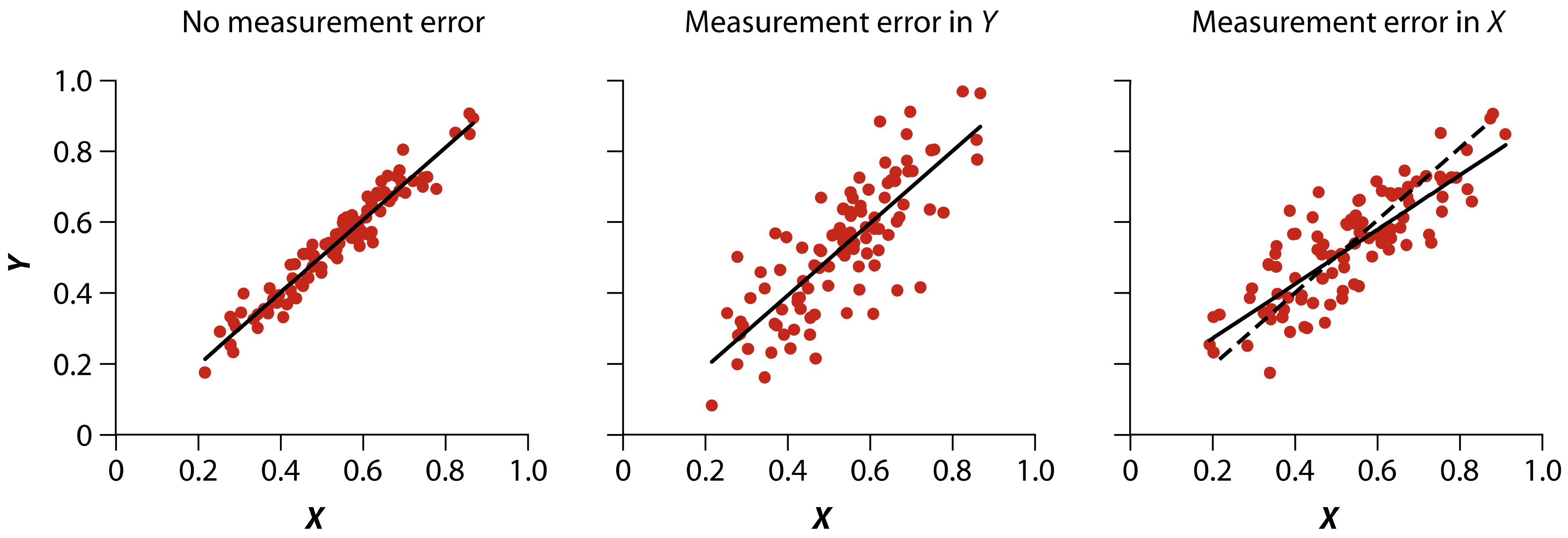
4. Reject or FTR and: $$\boxed{b - t_{\alpha(2),n-2}SE_b < \beta < b + t_{\alpha(2),n-2}SE_b}$$

If measurement error occurs on Y

* Increase variance of residuals
* Increases SE of slope

If measurement error occurs on X

* Increases variance of residuals
* **Causes attenuation bias in estimate of b** (underestimates slope)
  - b will lie closer to 0 than β
  - Remember: BIAS is <u>really bad</u>!



General linear model

- Linear Model for single-factor ANOVA
- Linear Regression

$$Y = \mu + A_i \qquad\qquad A_i = \text{group mean} - \mu$$

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

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**

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

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

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**

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

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

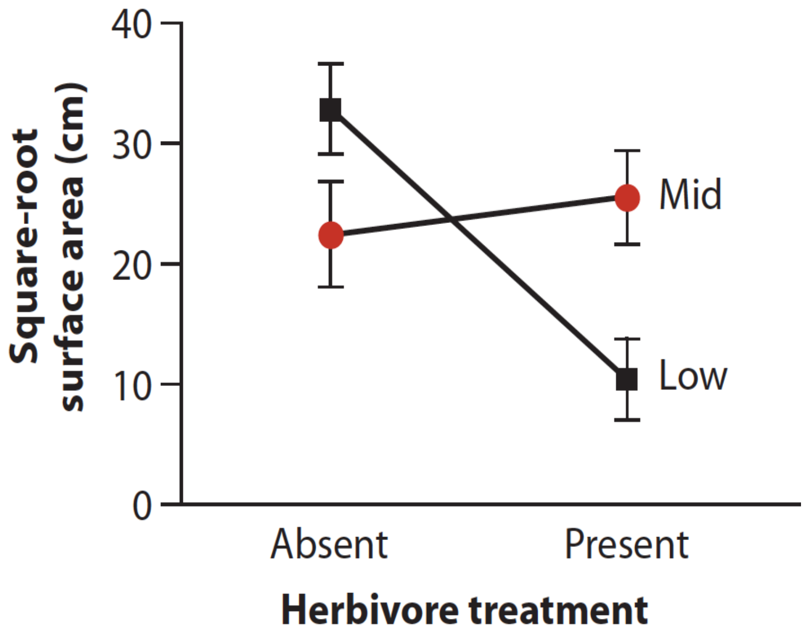
$$\text{F-test} = \frac{\text{Constant + Factor 1 + Factor 2 + Factor 1*Factor 2}}{\text{Constant + Factor 1 + Factor 1*Factor 2}}$$

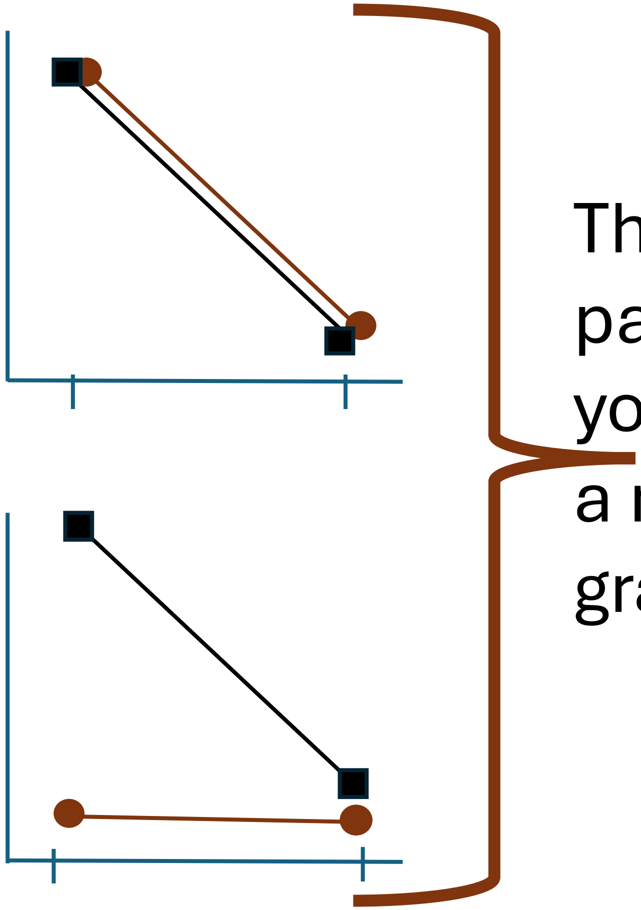
3. H_0:**Interaction effect: Factor 1*Factor 2**

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

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

<u>Multi- factor ANOVA **Example**</u>**:** 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

ANCOVA

- Increases precision
- Attempts to adjust for bias
- Often will include “pre” and “post” treatment to try to account for **confounding** individual differences
- Common: SES, age

Covariate effects:

- Confounding variables bias estimates of treatment effects

Covariate: a variable (or group of variables) that accounts for a portion of the variance in the dependent variable

- Allows researchers to test for group differences while controlling for effects of the covariate

ANCOVA VS Multi-factor ANOVA?

- ANCOVA doesn’t require that the variable we are trying to control for is ***necessarily an independent categorical variable***

Ex. Amount of tv watching for girls versus boys (**independent variable** 1 = gender), in four different geographic locations (**independent variable** 2 = North, South, West, East), the interaction (gender*geography) ***and S.E.S.*** (what type of variable could this be?)

Response = Constant + Factor 1 + Covariate + Factor 1*Covariate

Two rounds of model fitting:

1. *Interaction between covariate and treatment is tested*

   Regression slopes differ among the 'groups' if interaction is present

   $$\text{F-test} = \frac{H_A\text{: Constant + Factor 1 + Covariate + Factor 1*Covariate}}{H_0\text{: Constant + Factor 1 + Covariate}}$$

2. *If no interaction is detected, interaction term is dropped and treatment effect is tested*

   $$\text{F-test} = \frac{H_A\text{: Constant + Factor 1 + Covariate}}{H_0\text{: Constant + Covariate}}$$

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

ANCOVA:

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

Main Principle of Blocking

Response = Constant + Treatment + Block

H_0: Response = Constant + Block
H_A: Response = Constant + Block + Treatment

- Determine significance via ANOVA table which includes a row for the **block**
- Calculates a F value for block - examines how much better fit is with the block versus without the block

Continue to add to your flowchart!