

# **Module 4AB**

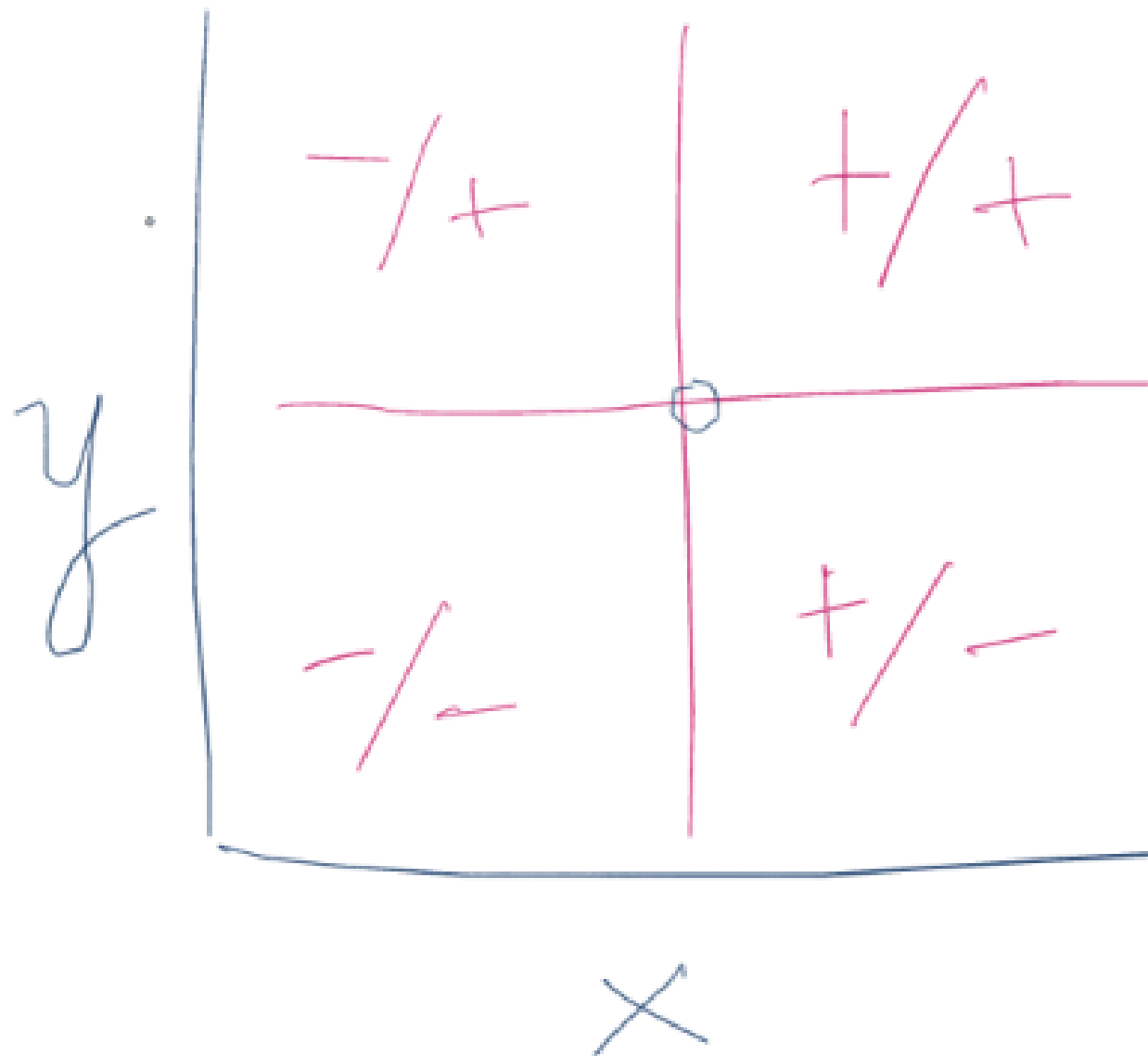
# **Supervised Machine Learning**

Different flavors of REGRESSION and General Linear Models

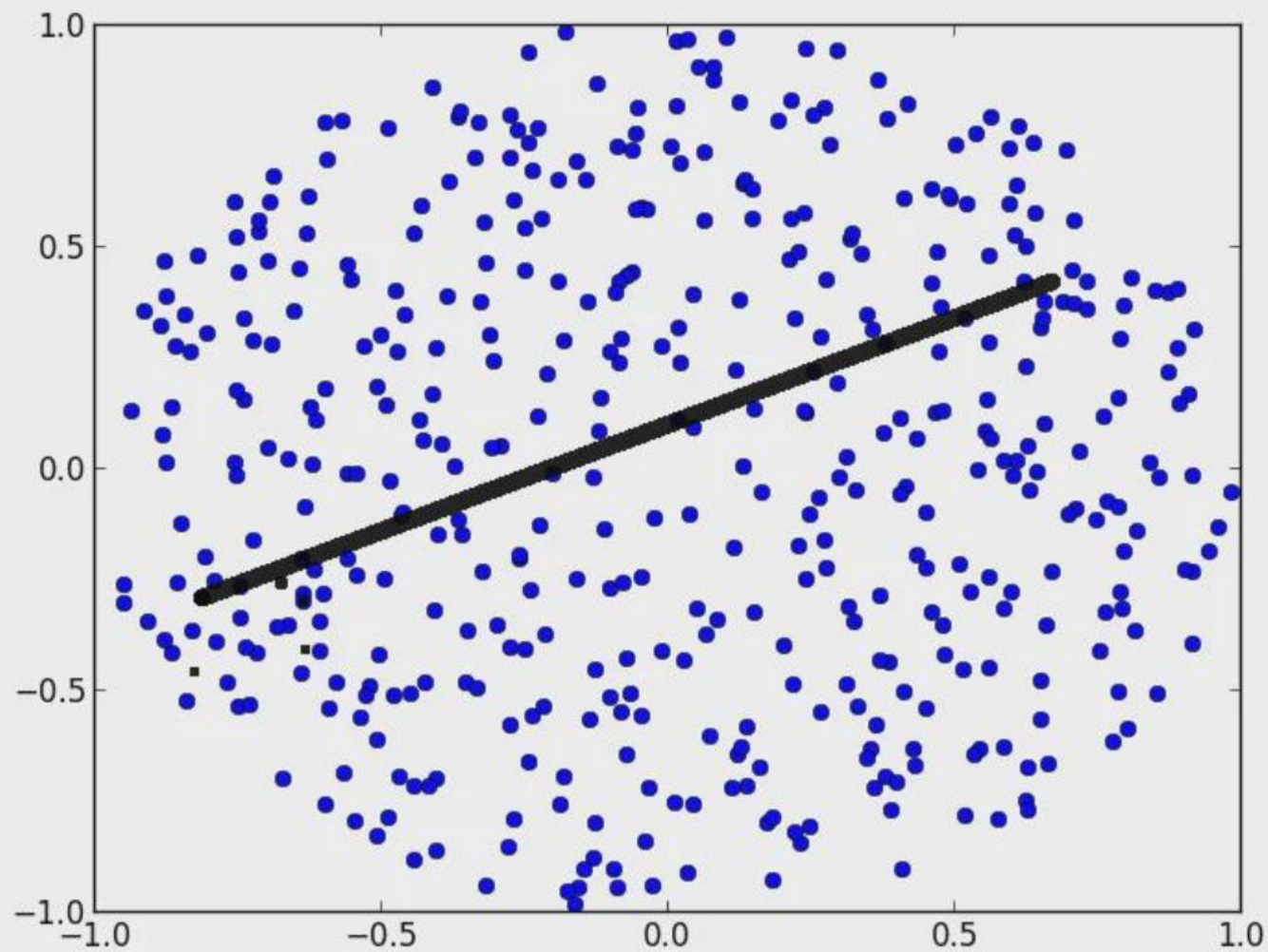
## Review: Correlation:

- Measures the amount/degree of linear association between two **numerical** variables
- Estimate the degree to which variables **covary**
  - With no attempt to interpret the causality of the association

Example: arm length and leg length covary together (individuals with longer arms often have longer legs) but they are influenced by other underlying variables **not** each other (longer legs do not cause longer arms)



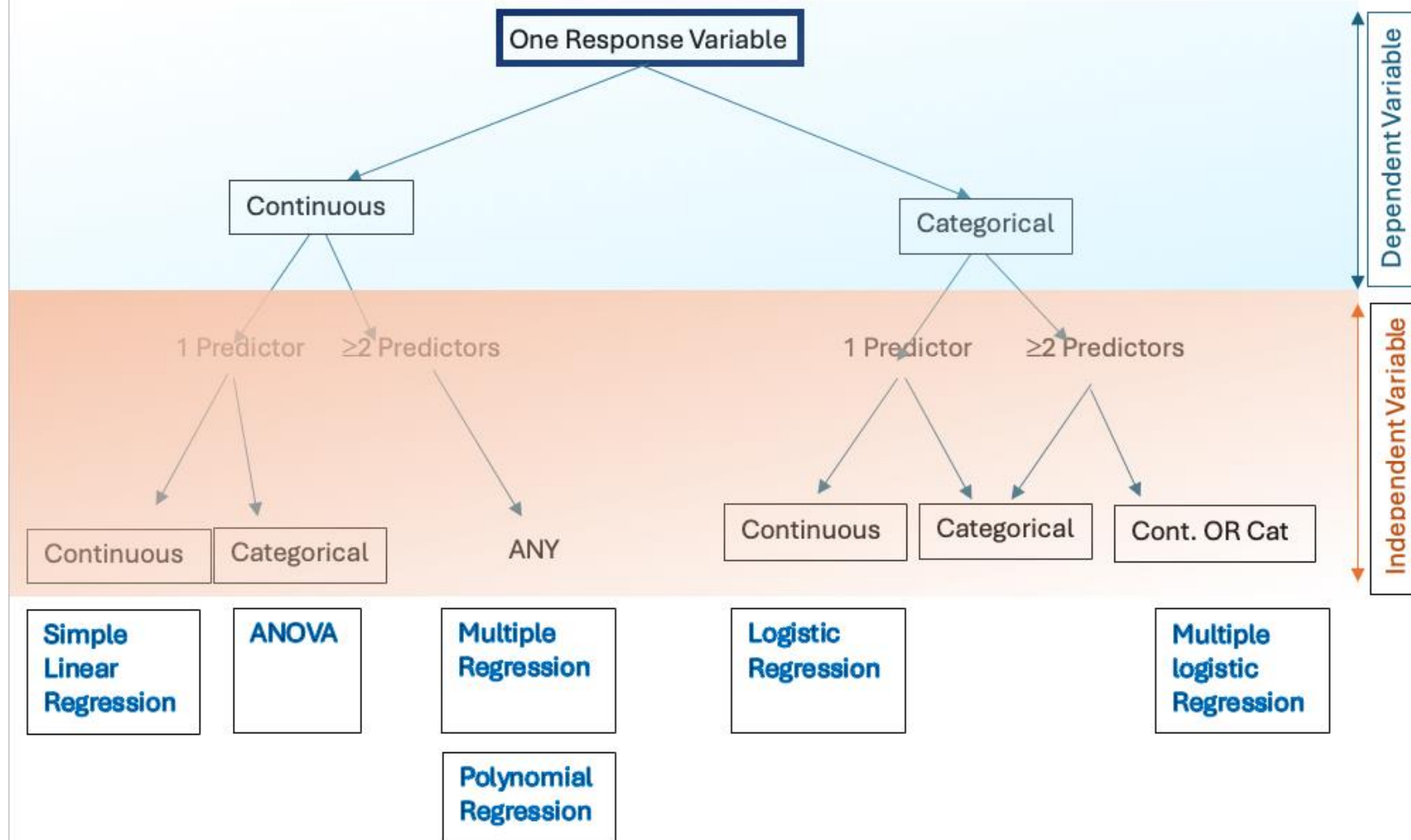
# Scientists be like



## Regression:

- Statistics is about prediction
- Used to **predict** value of one numerical variable from the value of another
  - predicting dependent/response variable, Y from independent/predictor X
- Linear regression assumes that the relationship between X and Y can be described by a line
  - Fits a straight line to a (messy) scatterplot

Example: ambient temperature may impact growth rate of a plant species, but the reverse is probably not true



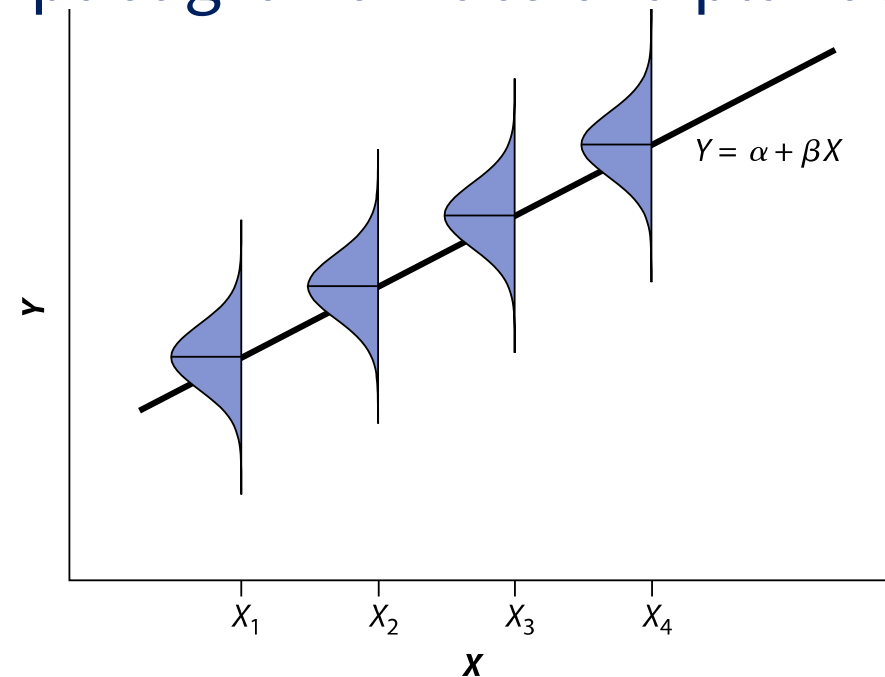
- Other kinds:

- Lasso
  - Variable selection (weighting a predictor variable by 0)
- Ridge
  - Allows analysis even in the face of **collinearity**

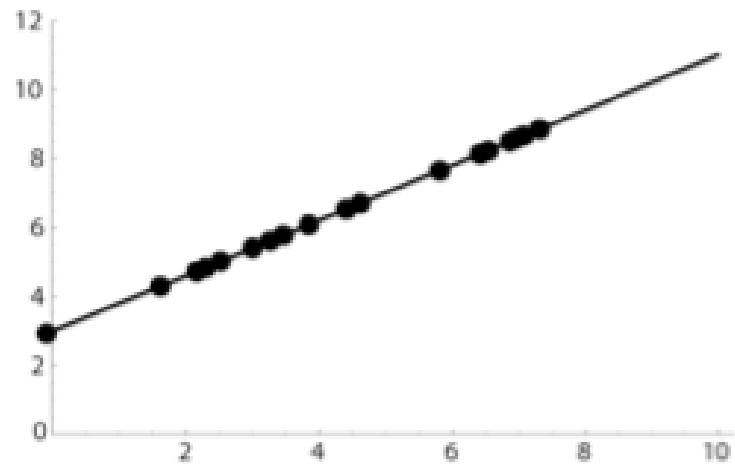
## Regression:

- Linear regression assumes that the relationship between  $X$  and  $Y$  can be described by a line
  - Fits a straight line to a (messy) scatterplot
- Homoscedasticity:  $Y$  is normally distributed with equal variance for all values of  $X$

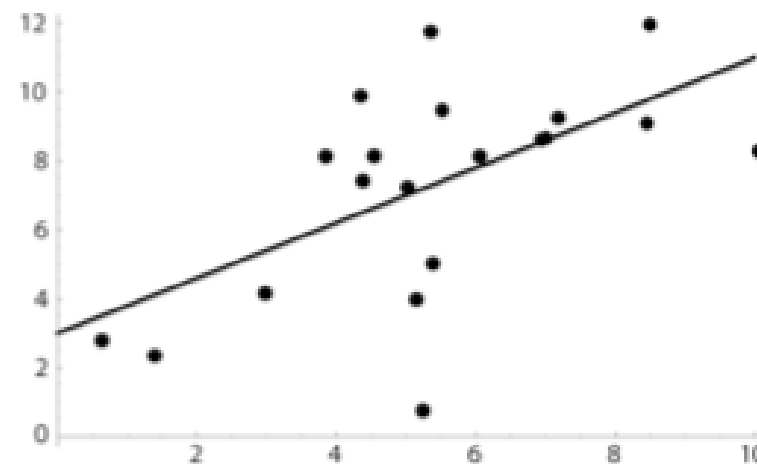
Example: ambient temperature may impact growth rate of a plant species, but the reverse is probably not true



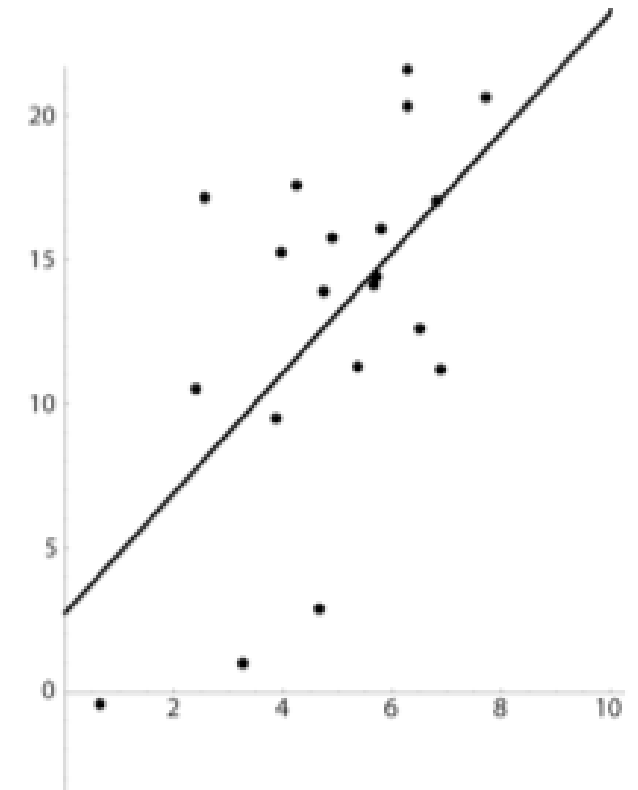
## correlation vs regression



$r = 1; Y = 3 + 0.8X$



$r = 0.6; Y = 3 + 0.8X$



$r = 0.6; Y = 3 + 2X$

←  
Different correlation;  
same slope

→  
Same correlation;  
different slope



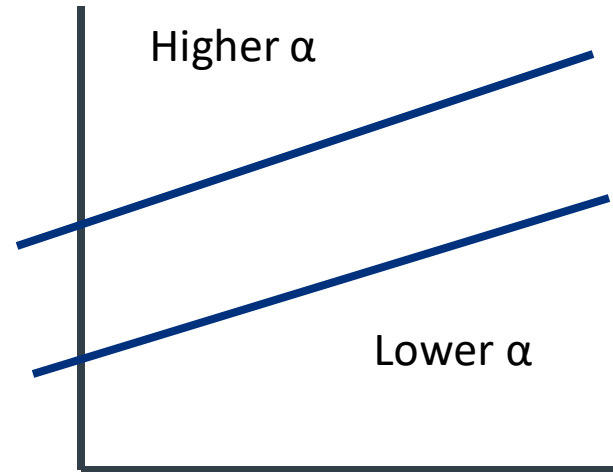
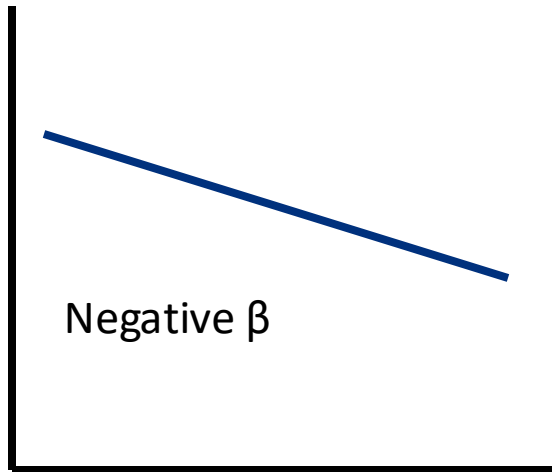
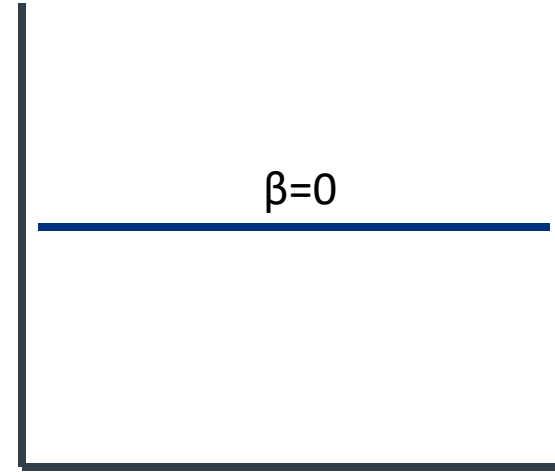
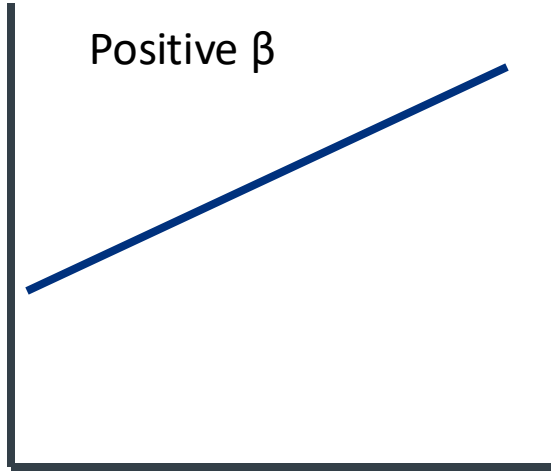
# The parameters of linear regression

$$Y = \alpha + \beta X + \varepsilon_1$$

intercept

slope

## Regression Overview

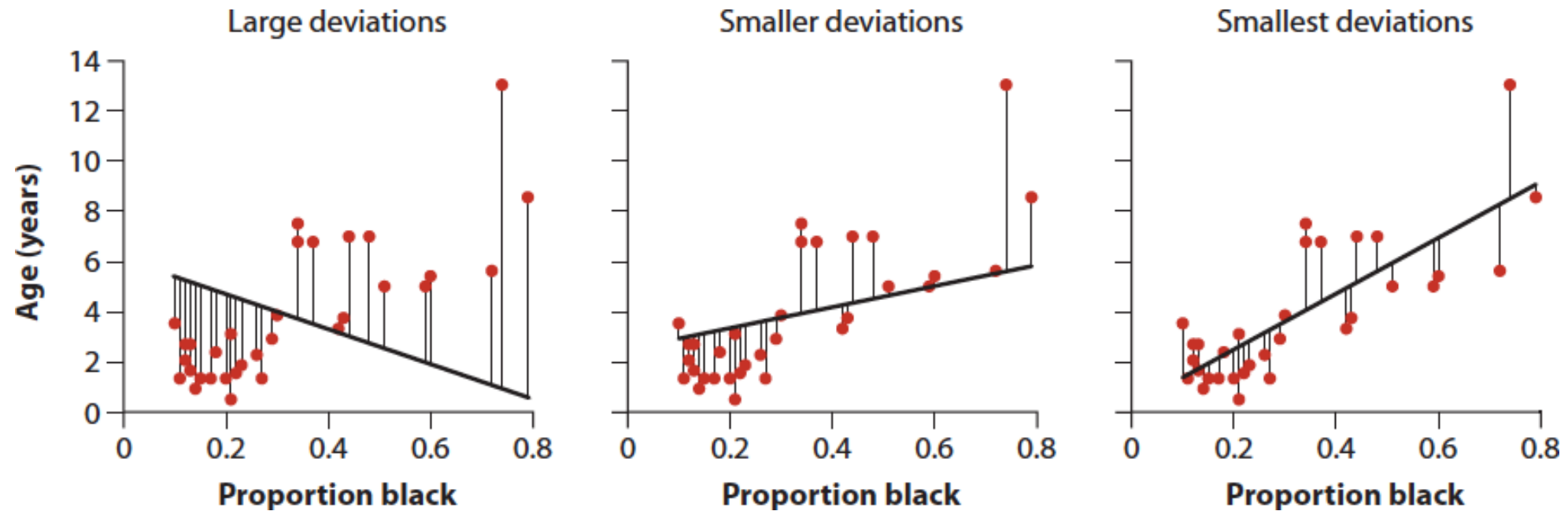


# Estimating a regression line

$$Y = a + bX + \varepsilon_1$$

# (Ordinary) Least Squares:

- Best fitting line through a scatterplot
  - Line that minimized spread of y values
- Minimize  $SS_{\text{residuals}}$ 
  - Measurement of how much the line's predicted  $y_i$  deviate from actual data values



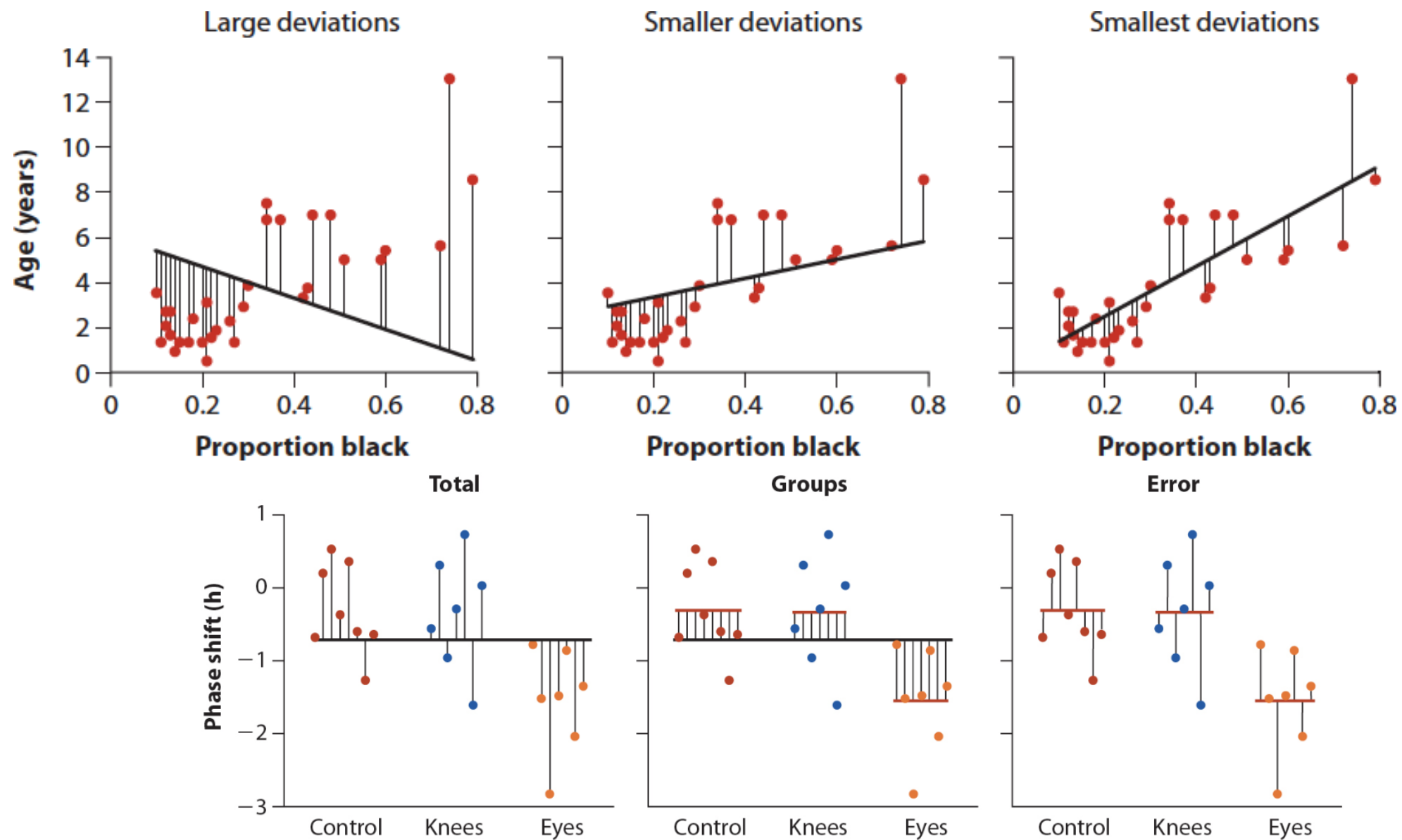


Figure 20.1: Whitlock and Schluter, Fig 15.1.2 – Illustrating the partitioning of sum of squares into  $MS_{group}$  and  $MS_{error}$  components.

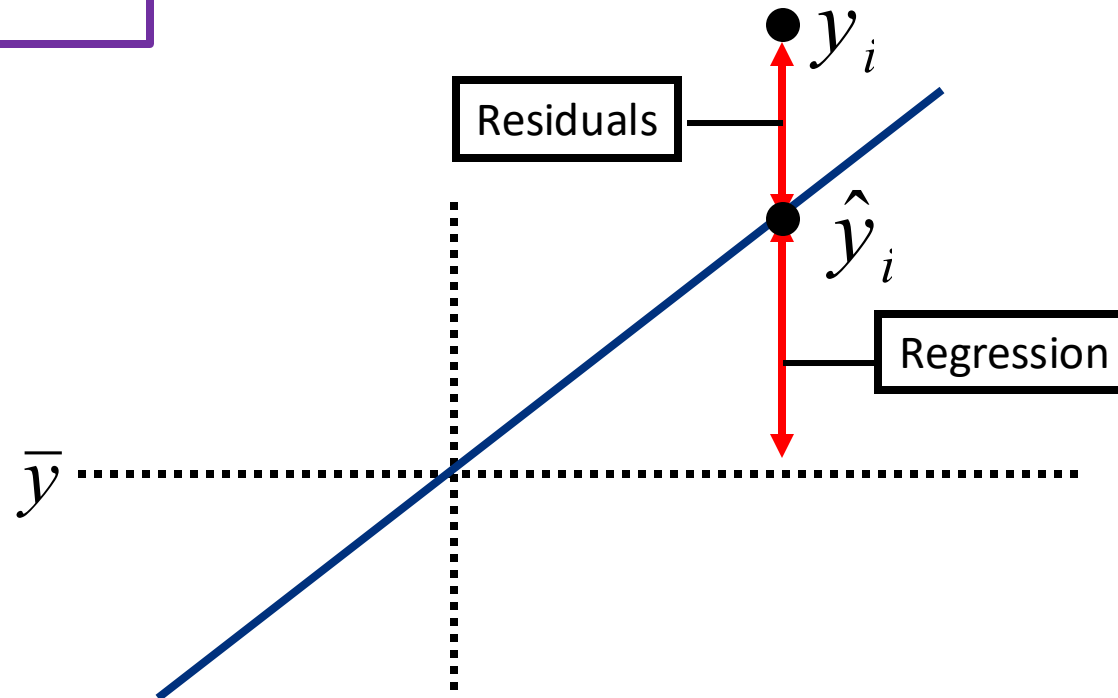
# Regression Overview

## Least Squares:

- What are the elements of this equation?

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

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



## • Residuals:

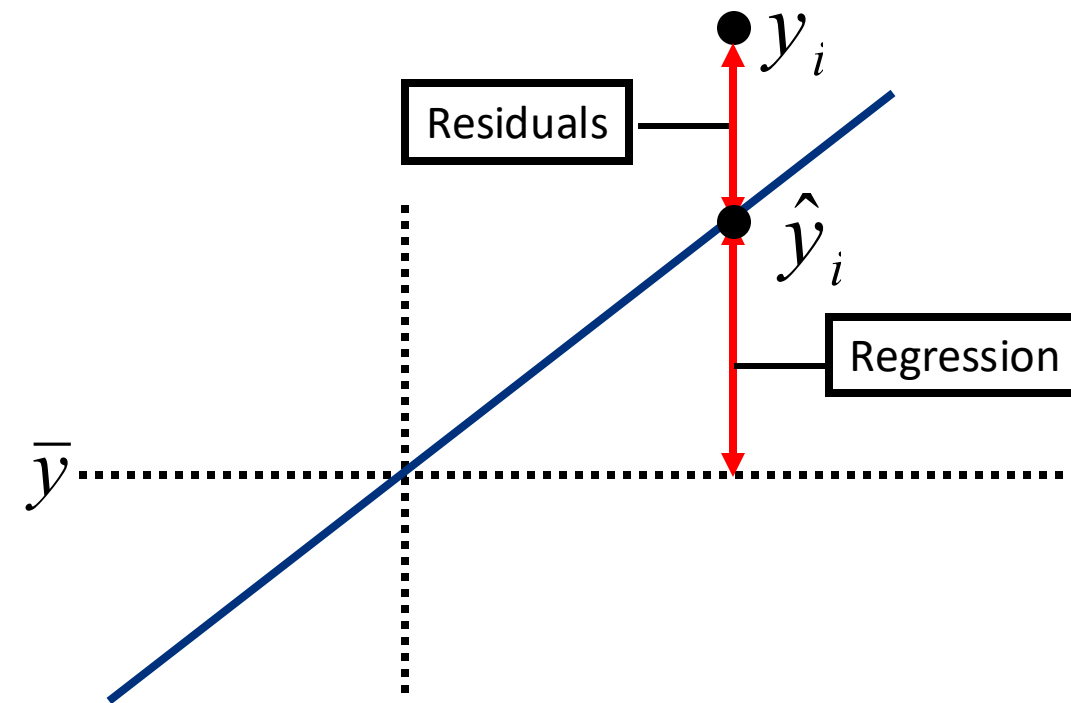
- Residuals measure the scatter of points above and below the least squares regression line
- $MS_{\text{residual}}$  is the variance of the residuals,  $\text{residual} = Y_i - \hat{Y}_i$

$$MS_{\text{residual}} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}$$

- $MS_{\text{regression}}$  is the variance of the regression

$$MS_{\text{regression}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{n - 2}$$

- Coefficient of determination ( $r^2$ ) = SSR/SST



$R^2$  predicts the amount of variance in Y explained by the regression line

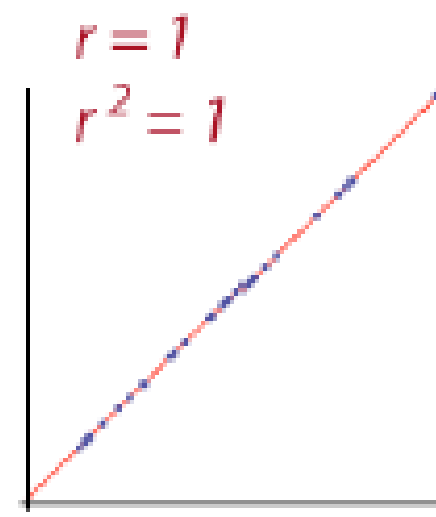
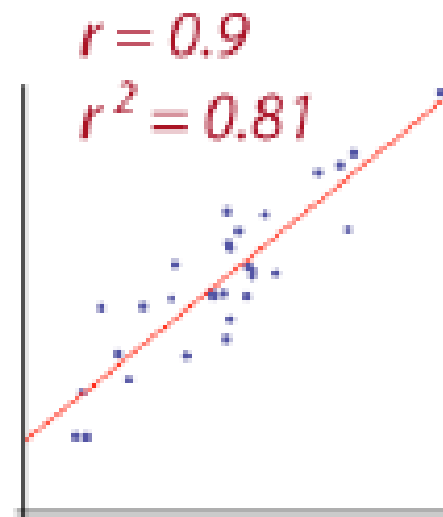
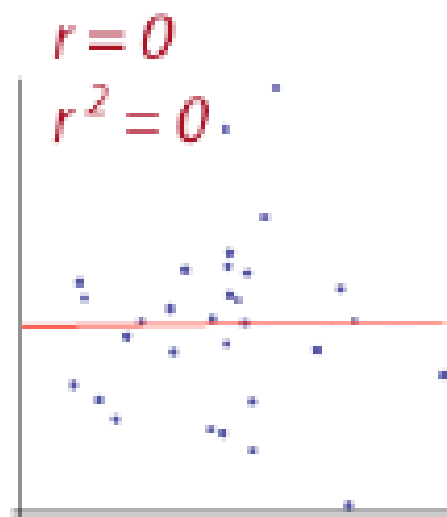
We saw this in ANOVA where  $R^2$  gave 'precision' of model (i.e. Ability of the model to explain variation)

- The coefficient of determination
- Sometimes written as  $r^2$
- Square of the correlation coefficient,  $r$

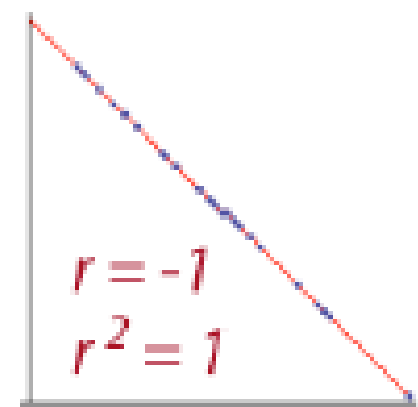
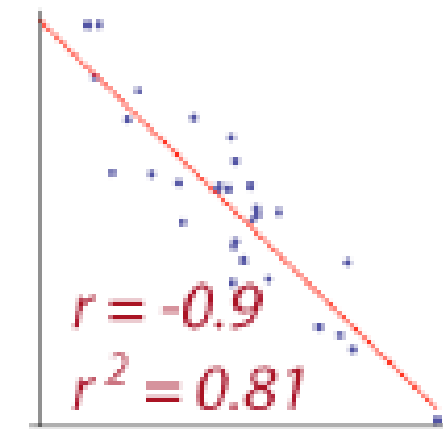
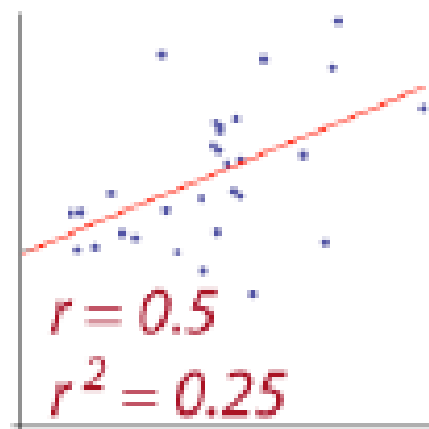
$$R^2 = \frac{SS_{regression}}{SS_{Total}}$$



$Y$



$Y$



$X$

$X$

$X$

## Best estimate of slope:

$b = \frac{\text{Sum of cross products}}{\text{Sum of squares of X}}$

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Finding **a**:

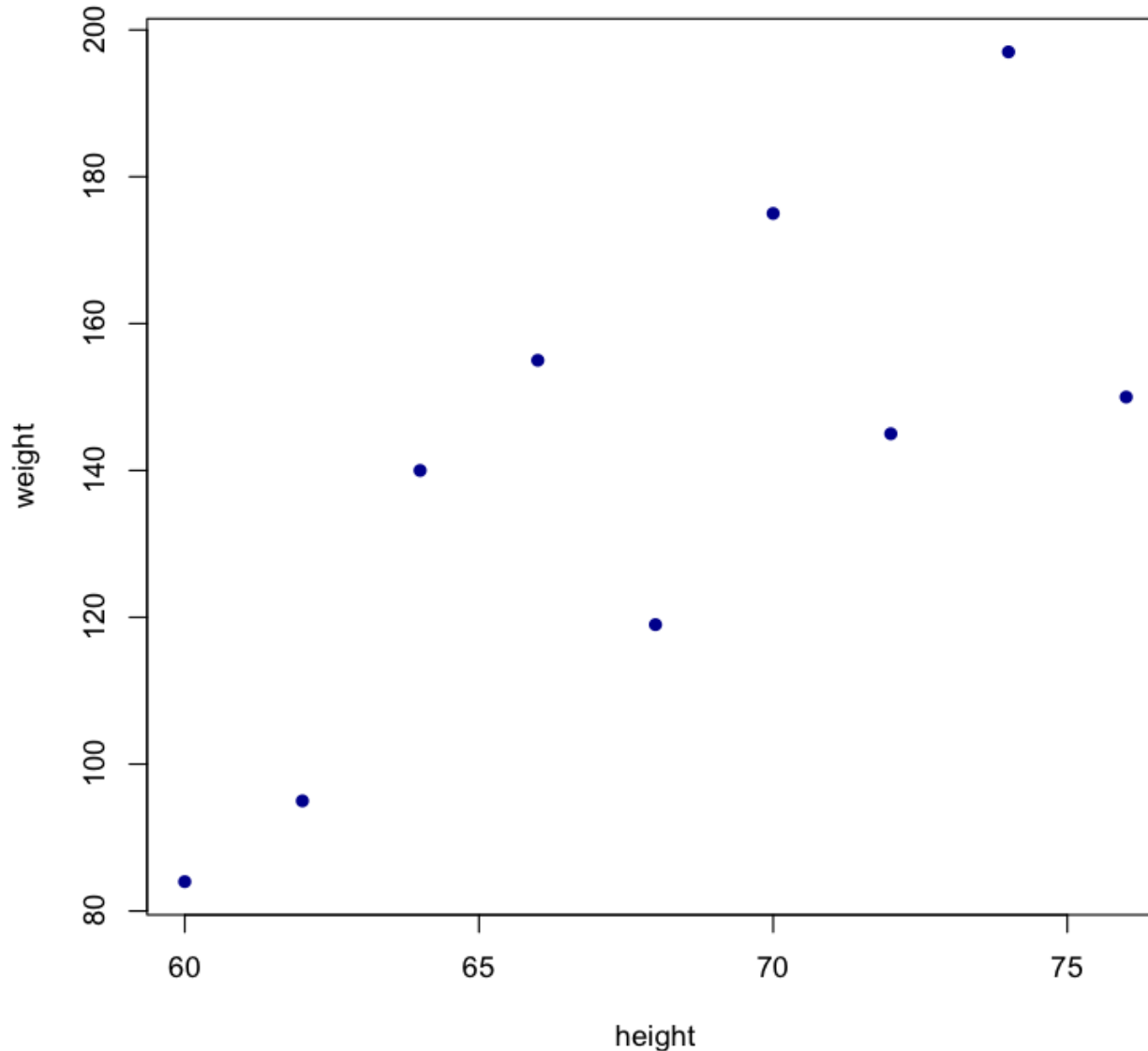
$$\bar{Y} = a + b\bar{X}$$

**OR**

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

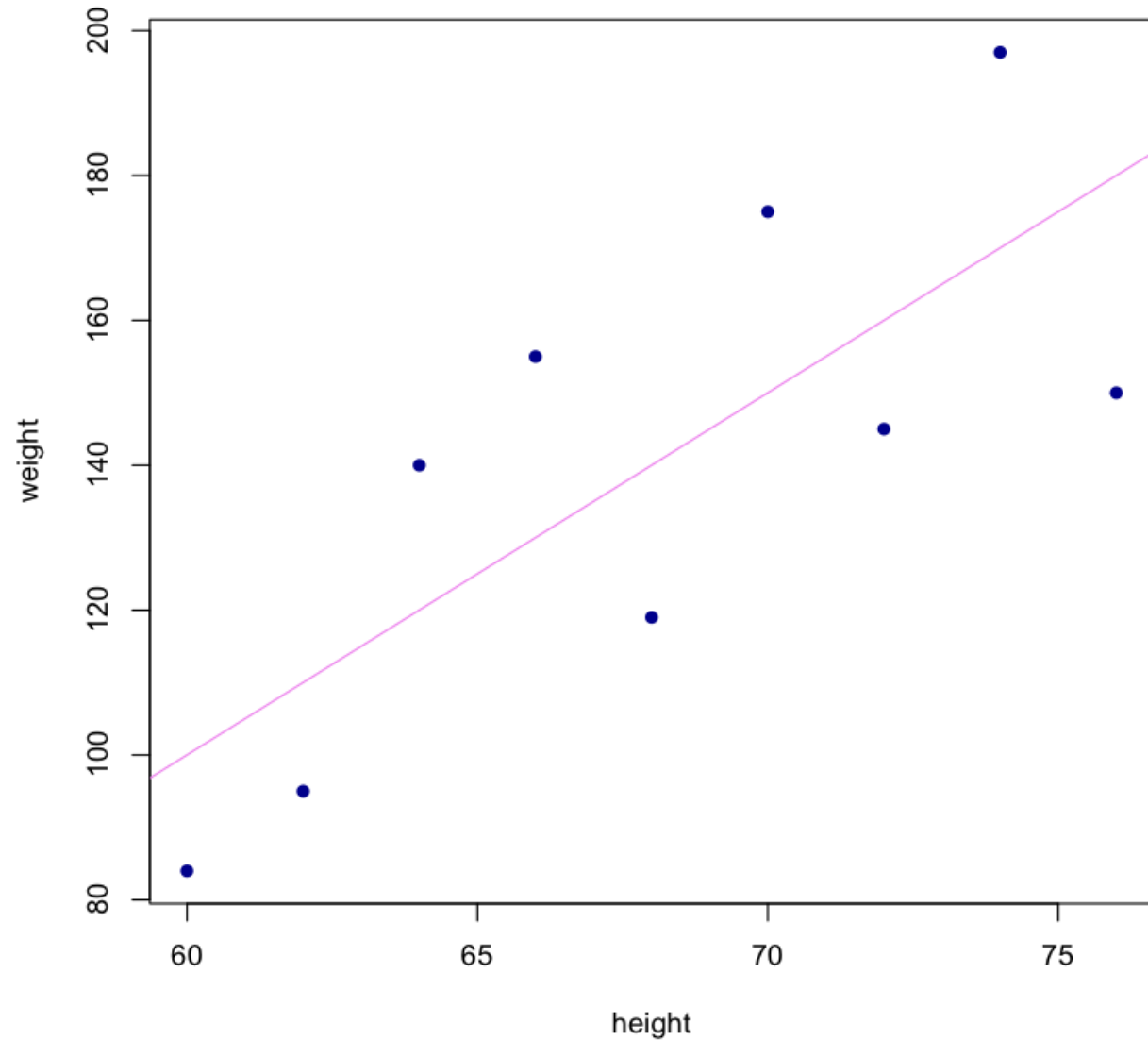
Example: Predicted weight for someone who is 65 inches tall?

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



Example: Predicted weight for someone who is 65 inches tall?

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



## Height Weight data:

$$\Sigma X = 612$$

$$\Sigma Y = 1260$$

$$n = 9$$

$$\Sigma X^2 = 41856$$

$$\Sigma Y^2 = 186826$$

$$\Sigma (XY) = 86880$$

$$\bar{Y} = 140$$

$$\bar{X} = 68$$

$$b = 5$$

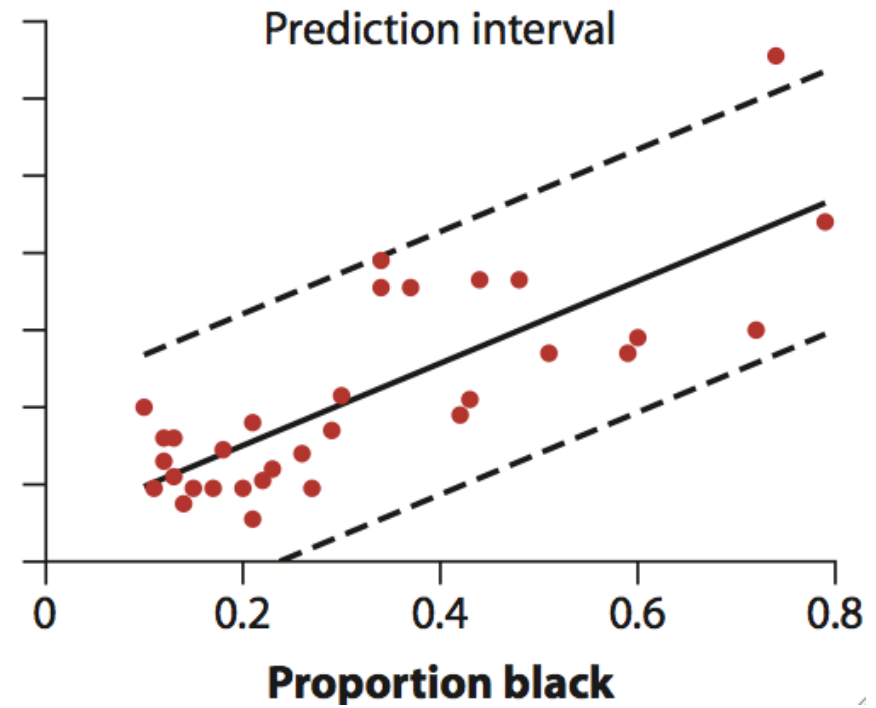
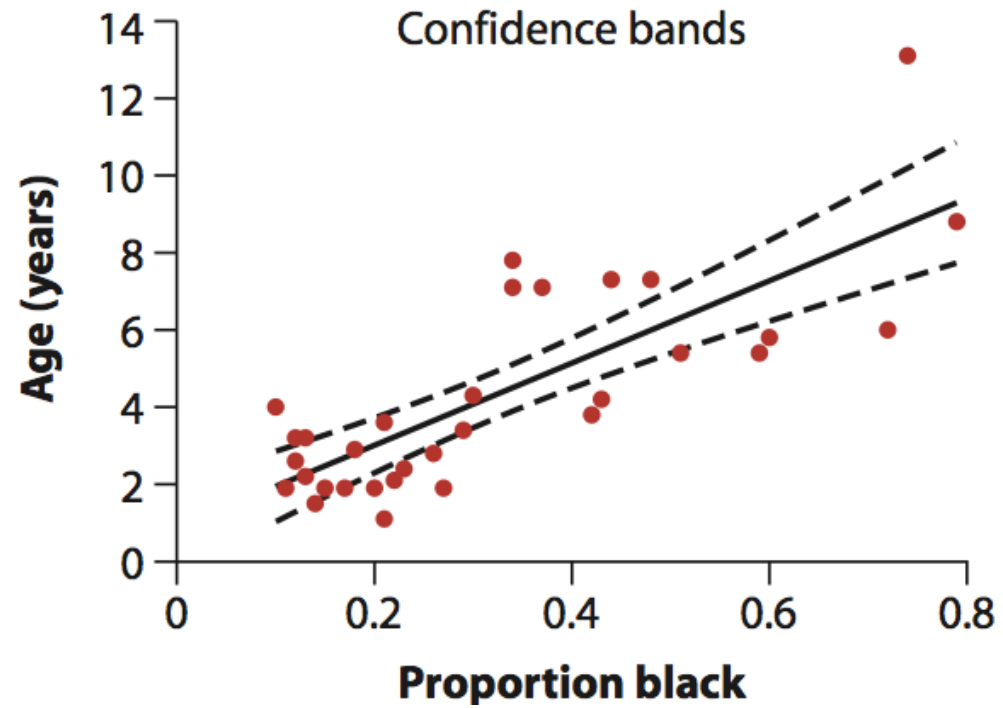
$$a = -200$$

$$\hat{Y} = -200 + 5X$$

An ice cream truck owner collects data on the number of sales made each day and the average temperature that day. He computes a regression line for predicting the number of sales based on how far the daily temperature is from freezing (32 degrees Fahrenheit) and finds  $\text{sales} = .22 + 1.8 (\text{degrees over } 32 \text{ Fahrenheit})$ . Identify the "y-intercept".

- A. 0.22
- B. 1.8
- C. 32.0
- D. Can't tell

# Prediction confidence:





## Prediction confidence:

The purpose of regression is to **predict**. There are two types of prediction:

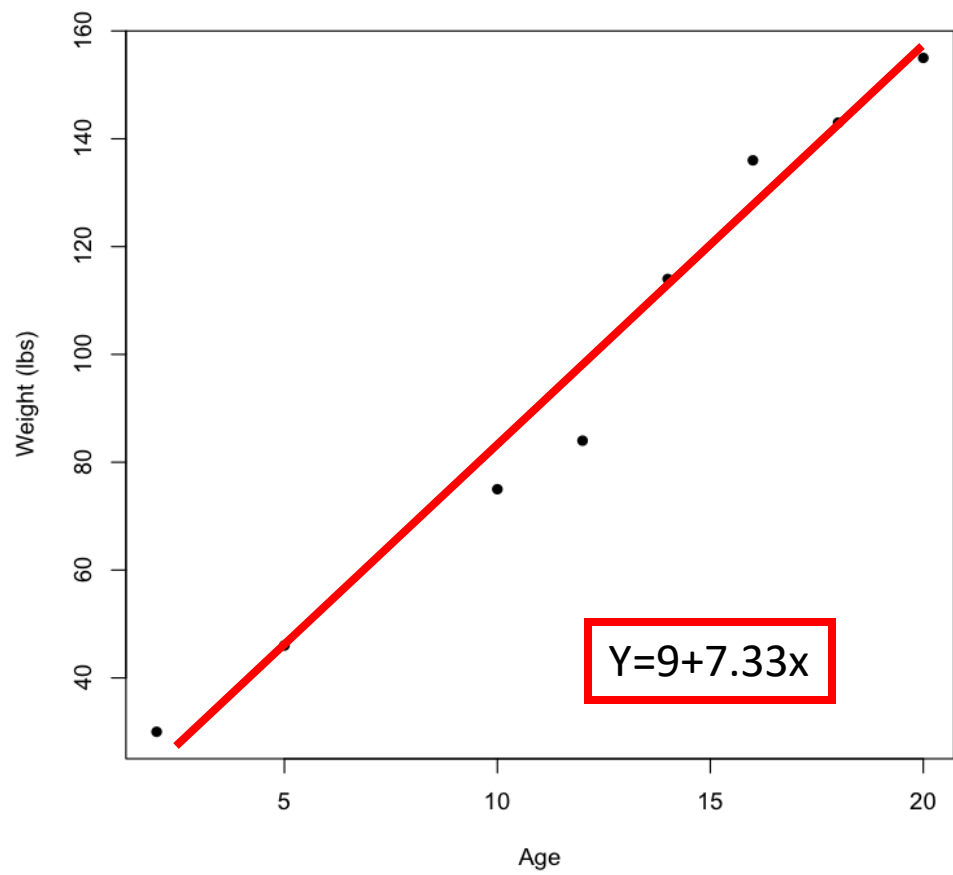
1.  $\bar{Y}$  for a given  $X$
2. Single  $Y$  for a given  $X$

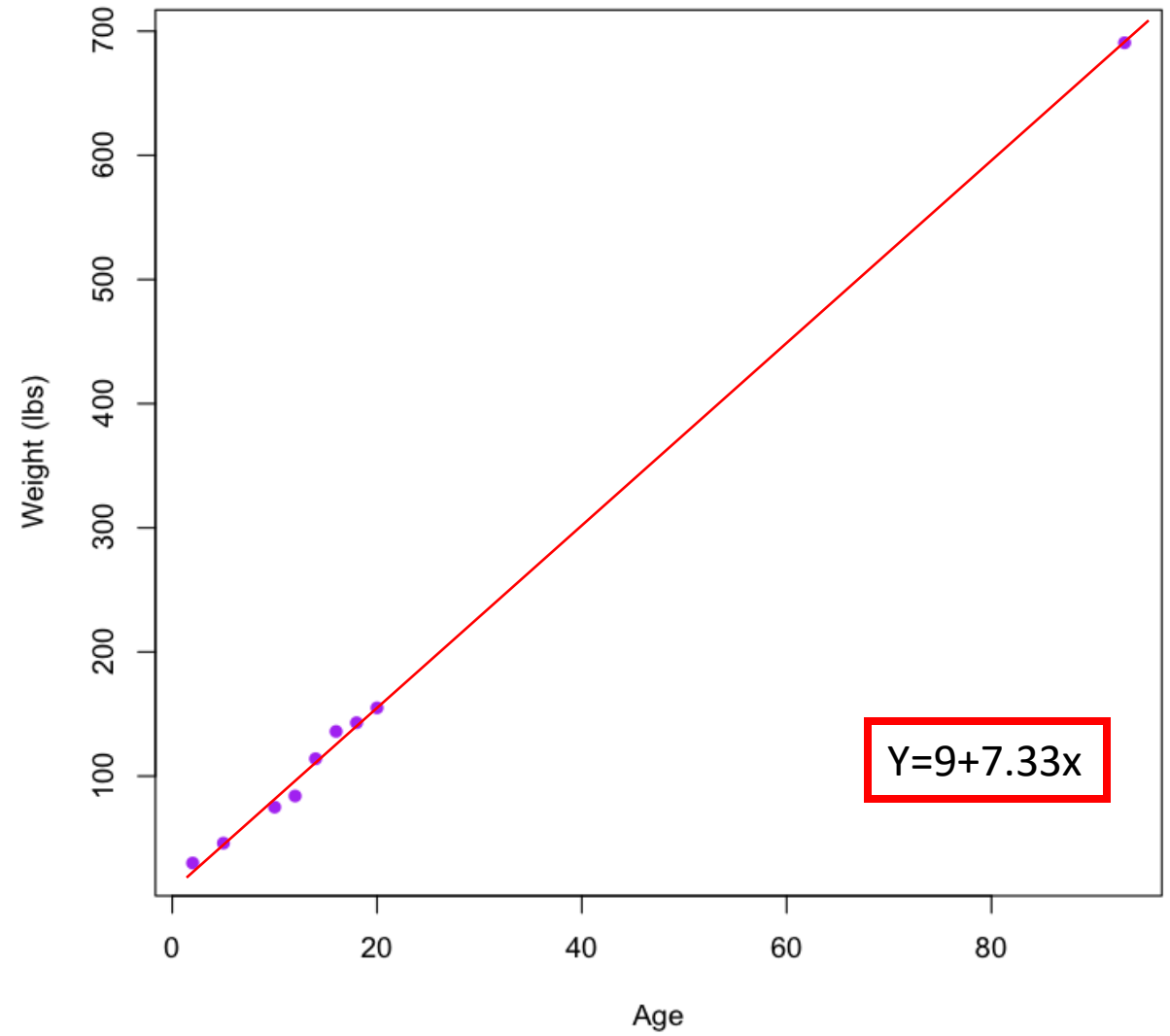
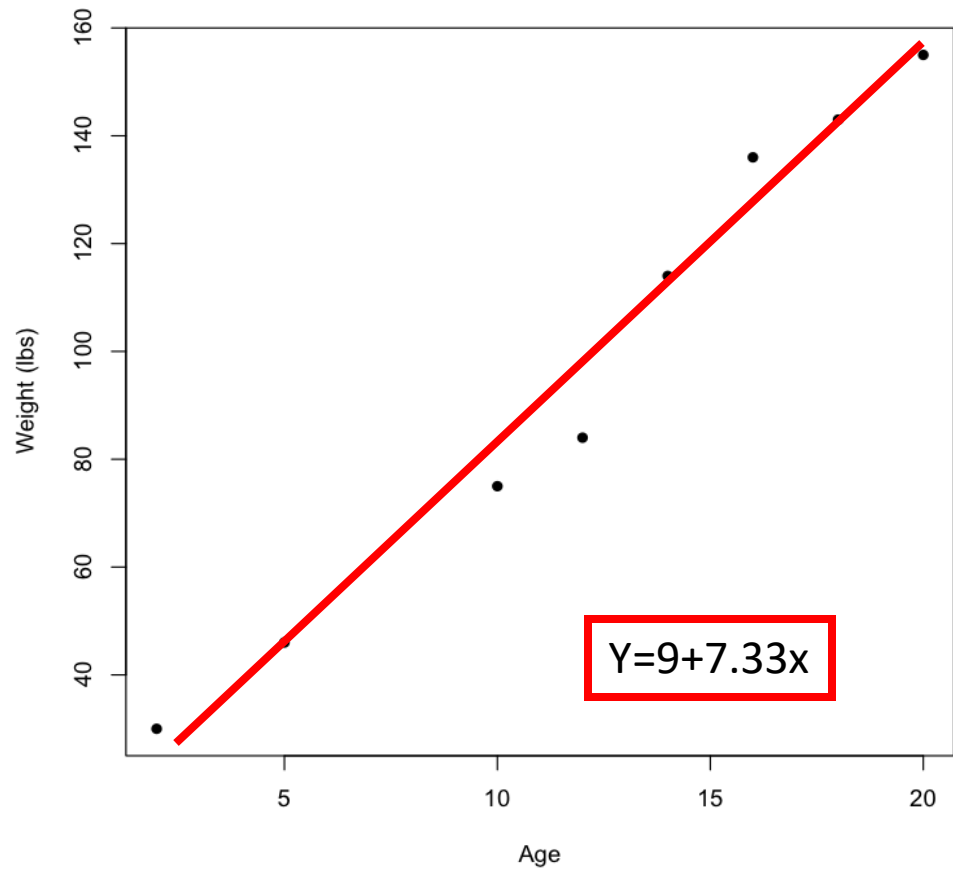
Both will generate  $\hat{Y}$  with the same value, but the prediction of a single  $Y$  point will have a lower precision.

**Caution!** Do not extrapolate beyond the range of the data

<b>Age</b>	<b>Weight (lbs)</b>	<b>Time to run one mile</b>	<b>Bench Press (lbs)</b>
<b>2</b>	30		
<b>5</b>	46		
<b>10</b>	75		
<b>12</b>	84	5:40	
<b>14</b>	114	5:05	
<b>16</b>	136	4:40	160
<b>18</b>	143	4:35	180
<b>20</b>	155	4:30	

Measurements taken over the course of an individual's life





## Regression towards the mean:

- Francis Galton invented the term to describe the observation that tall fathers had sons of average height
- He developed “regression analysis” to study this phenomenon of “regression towards mediocrity”
- results when two variables have correlation  $< 1$ 
  - Individuals who are far from the mean for one of the measurements will, on average, lie closer to the mean for the other measurement

## 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 than 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

## Regression fallacy: Rolling a die

Student	First	Second	Second roll lower?
1	4	5	no
2	4	3	yes
3	3	-	-
4	5	5	no
5	1	-	-
6	6	5	yes
7	5	2	yes
8	6	2	yes
9	3	-	-
10	2	-	-

Remaining students have a mean value of 5 (first roll) and 3.7 (second roll)

## 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:  $\mathbf{t = \frac{b - \beta_0}{Se_b}}$   $SE_b = \sqrt{\frac{MS_{residual}}{\sum (X_i - \bar{X})^2}}$

3. significance level; df=n-2

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



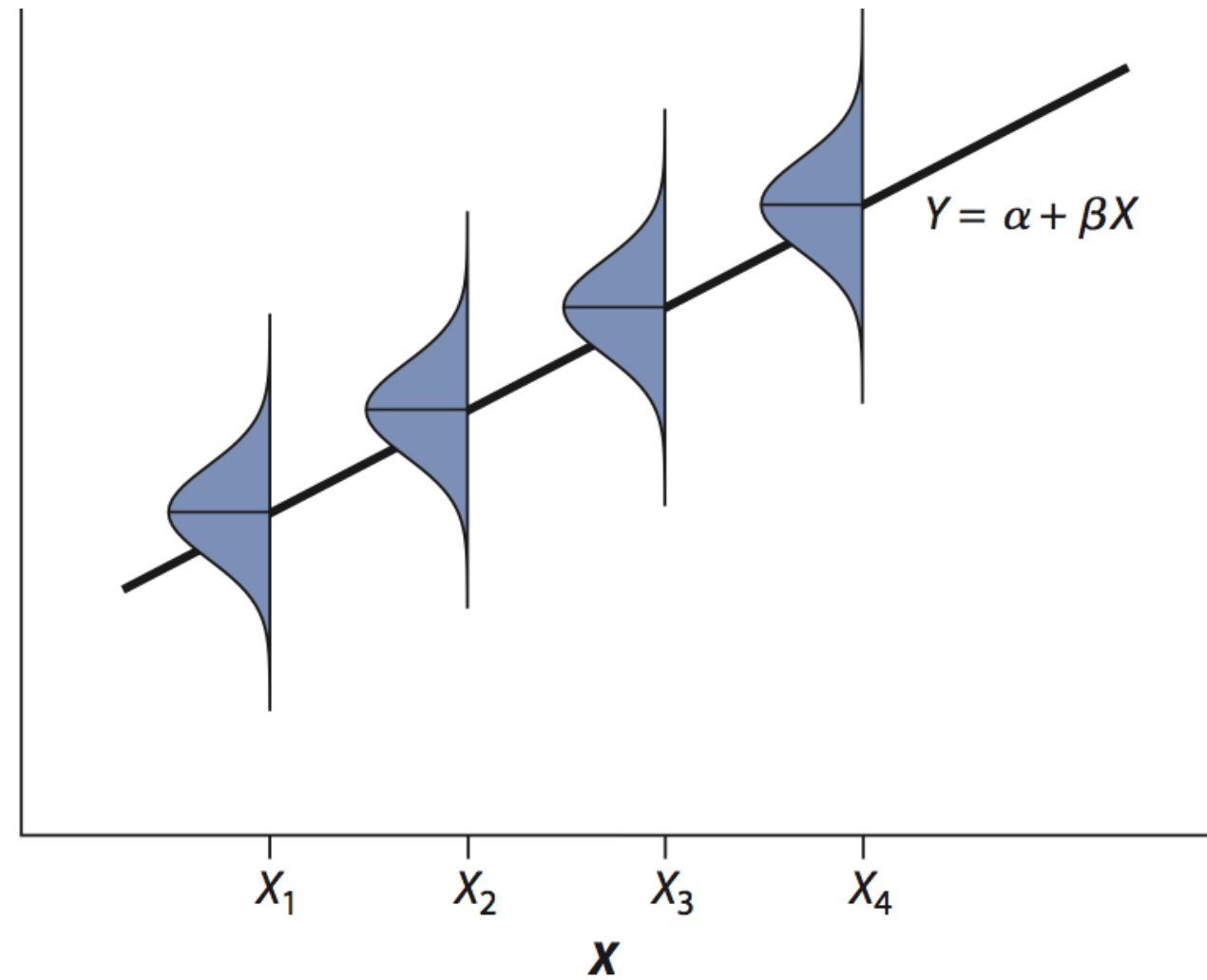
When test is two-tailed and  $H_0: \beta = 0$ , you can use ANOVA approach to testing regression slopes (for multiple models, too!)

- F-test versus t-test
- **If**  $H_0$  is true, then the mean squares corresponding to the two components should be equal

Source	DF	SS	MS	F
Regression (model)	1	$\sum (\hat{Y}_i - \bar{Y})^2$	$\Sigma(\hat{Y}_i - \bar{Y})^2 / 1$	$MS_{\text{regression}} / MS_{\text{residual}}$
Error (residual)	N-2	$\sum (Y_i - \hat{Y}_i)^2$	$\Sigma(\hat{Y}_i - \bar{Y})^2 / (n-2)$	
Total	N-1	$\sum (Y_i - \bar{Y})^2$	$\Sigma(Y_i - \bar{Y})^2 / (n-1)$	

## Assumptions of Regression Analysis:

- For each  $X_i$ , there is a population of  $Y$  values whose mean lies on the 'true' regression line
  - For each  $X_i$ , the  $Y$  are a random sample
  - For each  $X_i$ , the  $Y$  are normally distributed
- Homoscedasticity
  - For every  $X_i$ , the variance of  $Y$  is equal
- Nothing is assumed about the distribution of  $X$ 
  - It doesn't need to be normally distributed or randomly sampled - they might be fixed by the experimenter



## Major types of violation:

### 1. Outliers

- Violates homoscedasticity
- Violates normality of Y
- May make regression inappropriate; especially if they occur at the boundaries of X
- Compare results of regression with and without outlier
- Transformation of data ?

### 2. Non-linearity (we are dealing with linear regression)

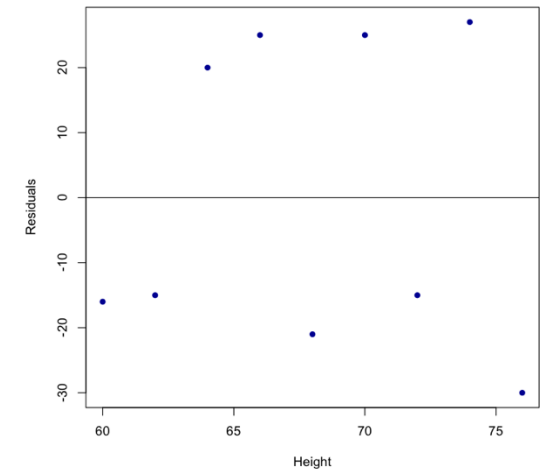
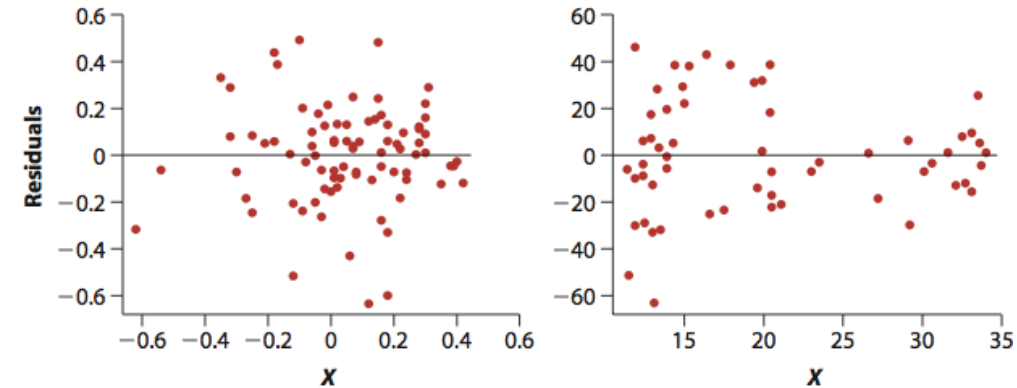
- Usually done by visual inspection of a scatterplot

### 3. Normality

- residual plot, where  $Y_i - \hat{Y}_i$  is plotted against  $X_i$
- cause a symmetric scatter of points above and below horizontal line

### 4. Measurement Errors

- Biological traits can be difficult to measure accurately
- Effects of measurement error depends on the variable

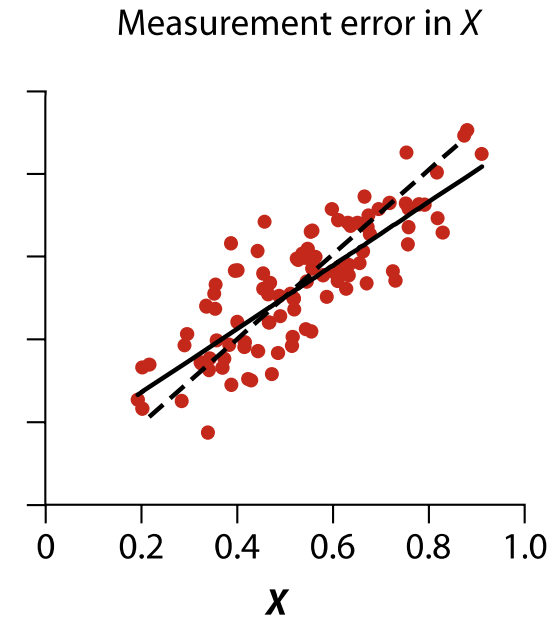
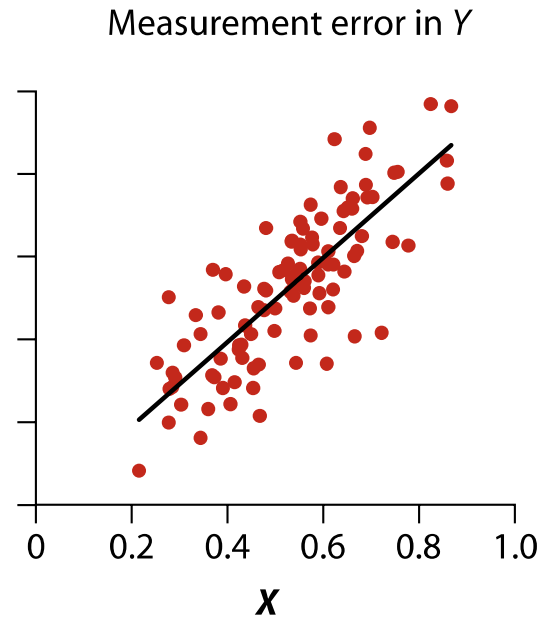
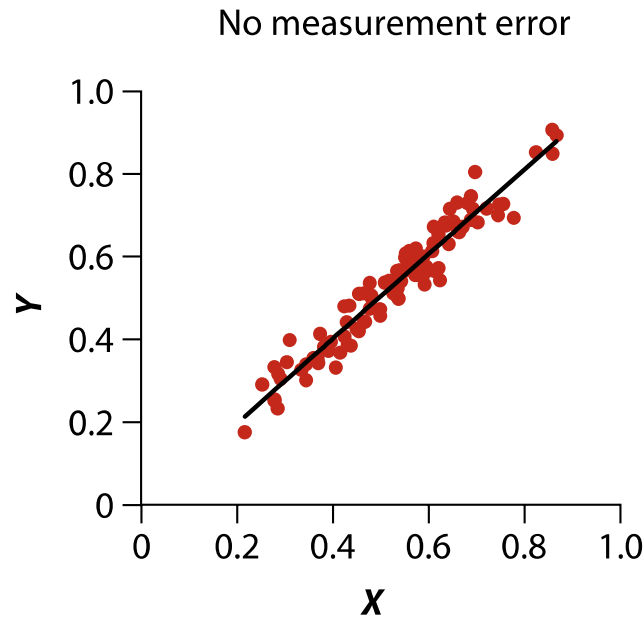


## If measurement error occurs on Y

- \* Increase variance of residuals
- \* Increases SE of slope

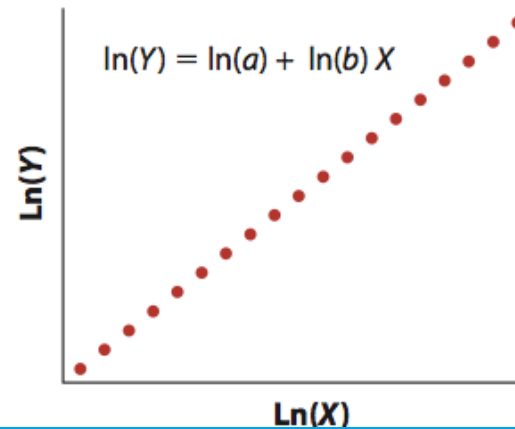
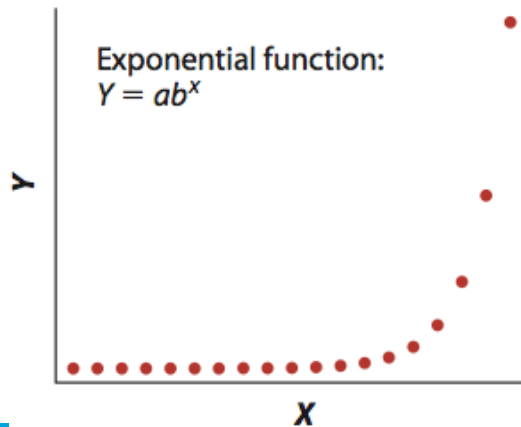
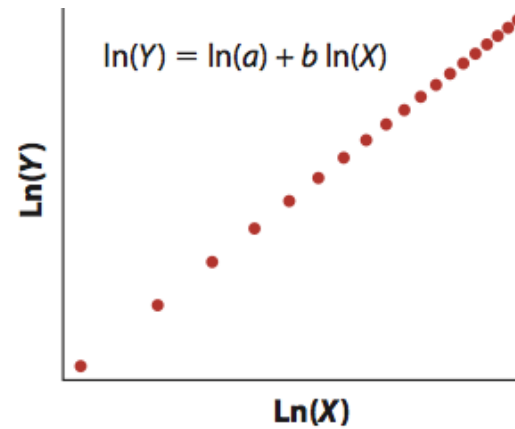
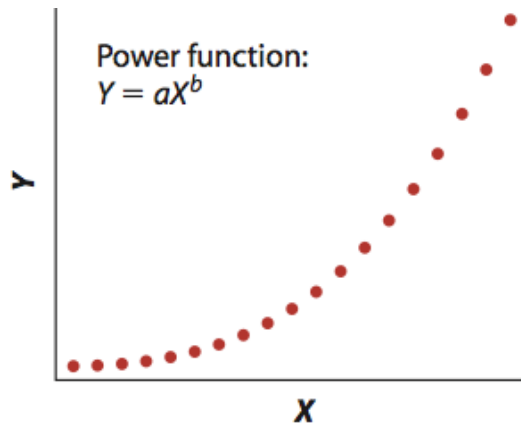
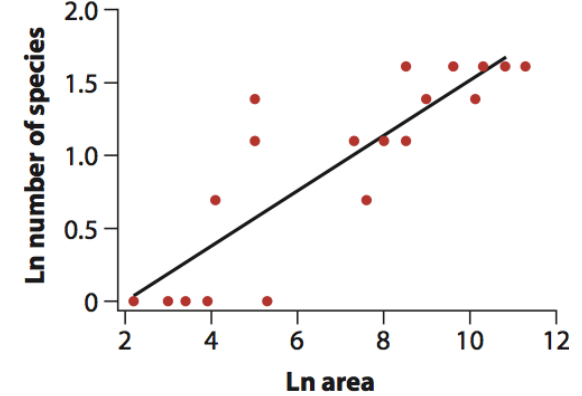
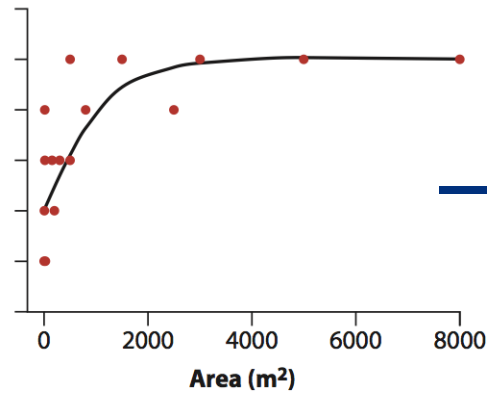
## If measurement error occurs on X

- \* Increases variance of residuals
- \* **Causes bias in estimate of  $b$**   
(underestimates slope)
  - $b$  will lie closer to 0 than  $\beta$
  - Remember: BIAS is really bad!



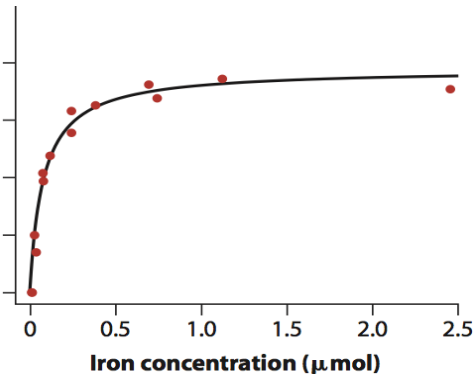
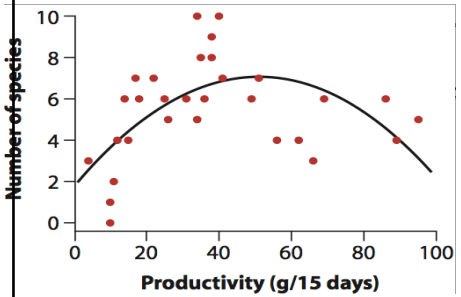
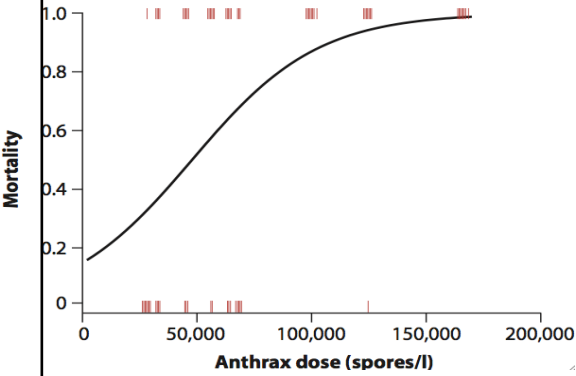
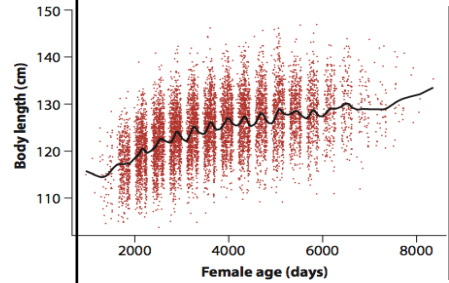
## Transformations:

- Non-linear relationships can sometimes be forced into linearity
- The usual suspects:
  - log transformation for power and exponential relationships



# Non-linear Regression:

- Same assumptions are linear regression but, obviously, doesn't assume a linear relationship
- Keep it simple: Don't **over fit**
  - It is possible to get a curve that fits each and every point ( $MS_{\text{residual}} = 0$ ) but it will not predict future points since the curve ***doesn't describe a general trend***

Curve with Asymptote	Quadratic curve	Binary response Variable	Smoothing
$Y = \frac{aX}{b + X}$	$Y = a + bX + cX^2$	$\text{Log-odds}(Y) = a + bX$	<ul style="list-style-type: none"> <li>depends on data</li> </ul>
Michaelis-menten eq <sup>n</sup>	Parabolic relationships	Dose response curve	Diagnosis of exclusion
			

## CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



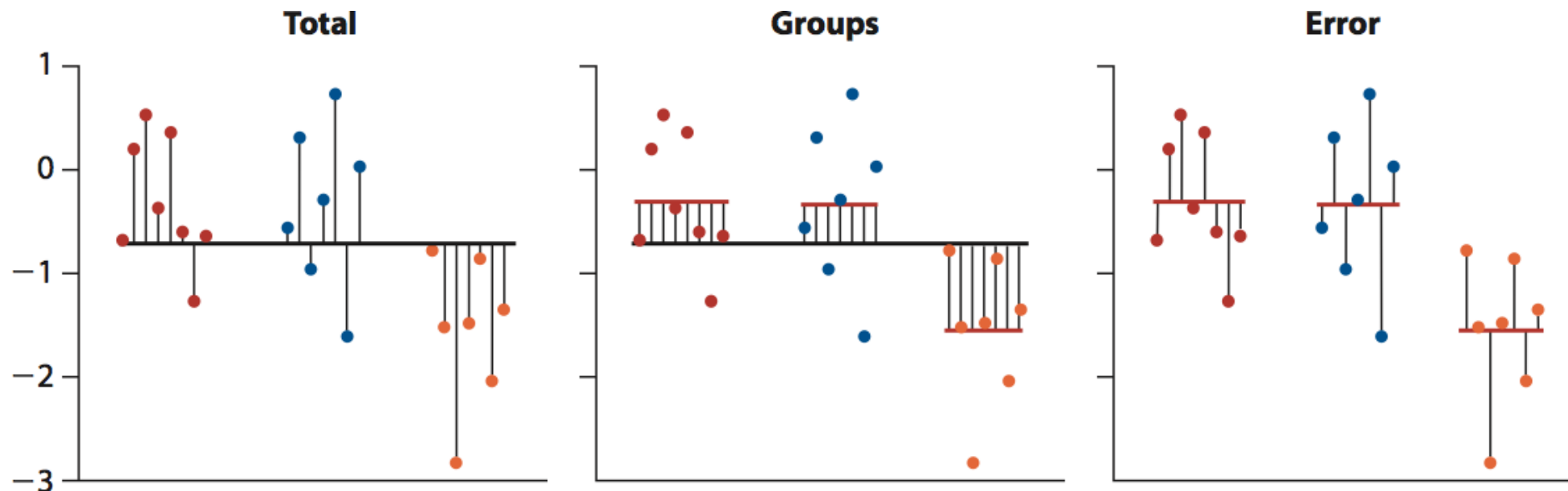


# Why do studies in biology often have more than one explanatory variable?

- a. Interactions; include blocking; cost efficiency
- b. Interactions; impress funding agencies; control for confounding variables
- c. Include blocking; cost efficiency; biologists need to justify their extra (field) work

# General Linear Models

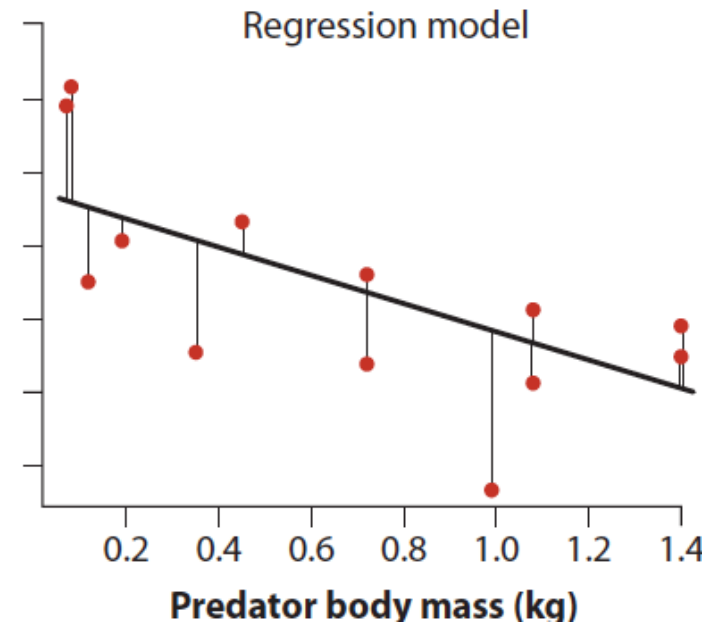
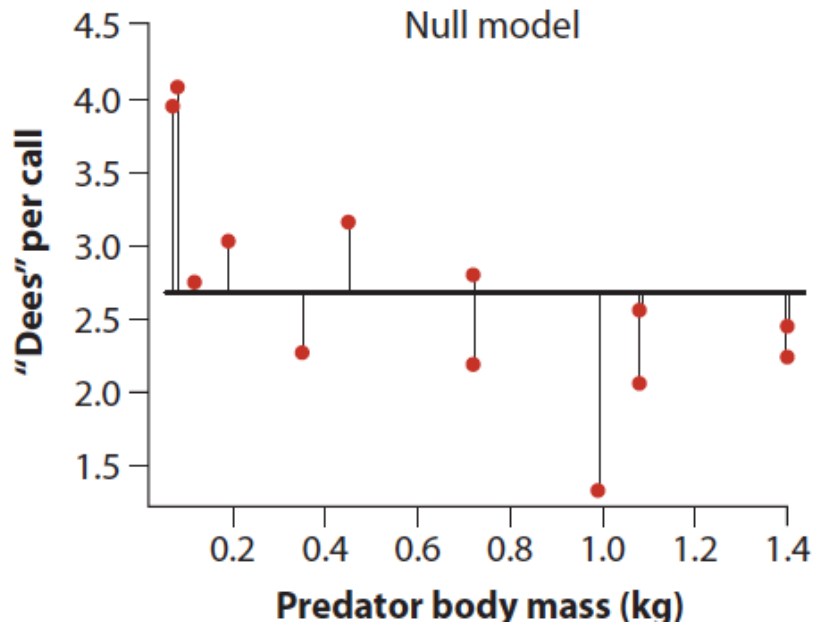
- Response variable,  $Y$ , can be represented by a linear model plus random error
  - Scatter of  $Y$  measurements around the model is random error
- So far, we have looked at (univariate) ANOVA, linear regression, and t-tests



# General linear model

We have also looked at the linear regression

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



## General linear model

- Extends the linear regression in two ways
  - More explanatory variables ( $>1$ )
  - Allows use of **categorical** explanatory variables

### Example:

Linear model for single-factor ANOVA

$$Y = \mu + A$$

Grand Mean

Treatment Effect

# 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

Linear Model	Other Name	Example-study Design
$Y = \mu + X$	Linear Regression	Dose-Response
$Y = \mu + A$	One-way ANOVA	Completely randomized
$Y = \mu + A + b$	Two-way ANOVA, no replication	Randomized block
$Y = \mu + A + B + A*B$	Two-way, fixed effects ANOVA	Factorial Experiment
$Y = \mu + A + b + A*b$	Two-way, mixed effects ANOVA	Factorial Experiment
$Y = \mu + X + A(+A*X)$	Analysis of Covariance (ANCOVA)	Observational Study
$Y = \mu + X_1 + X_2 + X_1*X_2$	Multiple Regression	Dose-Response

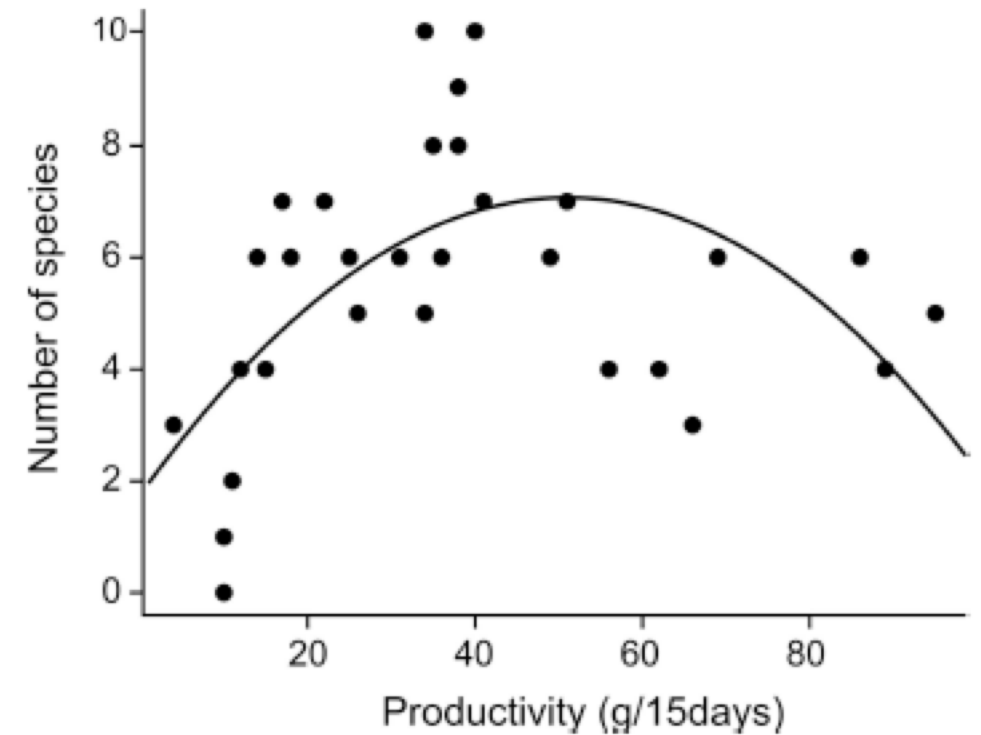
## Note: General linear model

In the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \text{error}$$

Doesn't have to be LINEAR relationship:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \text{ (Quadratic)}$$



## 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 effects of  $\geq 2$  treatment variables

Interactions

Covariates

Confounding variables

Nuisance variable is known but uncontrollable

**If the assumptions of general linear models are met, which of the following will NOT be true of the residual plot:**

- A) The model will have a roughly symmetric cloud of points above and below the horizontal line at 0.
- B) There will be noticeable curvature as we move left to right along the horizontal axis.
- C) Approximately equal variance of points above and below the horizontal line at 0.

[https://PollEv.com/multiple\\_choice\\_polls/GAoIY3qU6YczUUqB0NbYl/respond](https://PollEv.com/multiple_choice_polls/GAoIY3qU6YczUUqB0NbYl/respond)

## 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 ( $\geq 2$  groups)
- i.e., **Girls vs boys** in hours of tv watched

## Multi-Factor ANOVA:

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

# Multiple factor ANOVA:

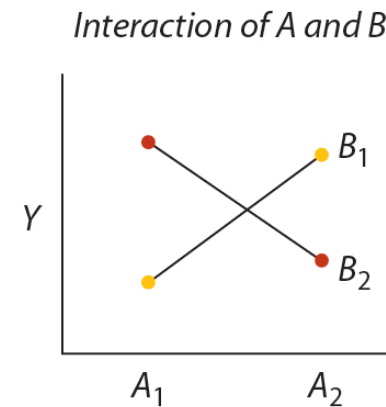
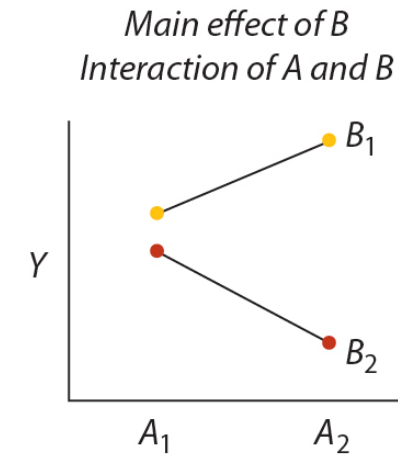
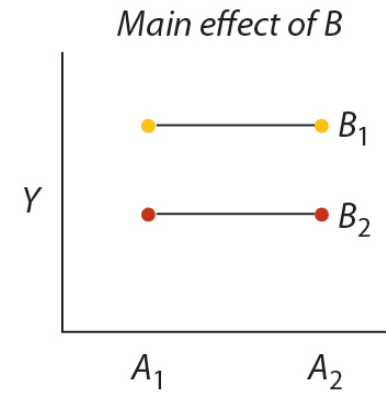
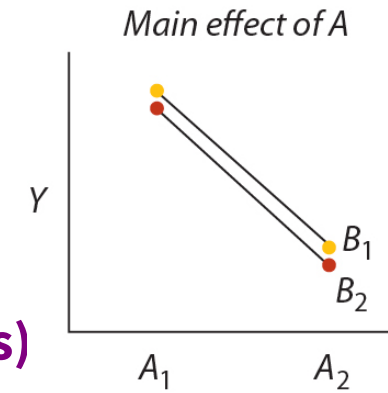
- 1 dependent variable and  $\geq 2$  (independent) categorical variables
- Produces 2 interesting results:

## 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**

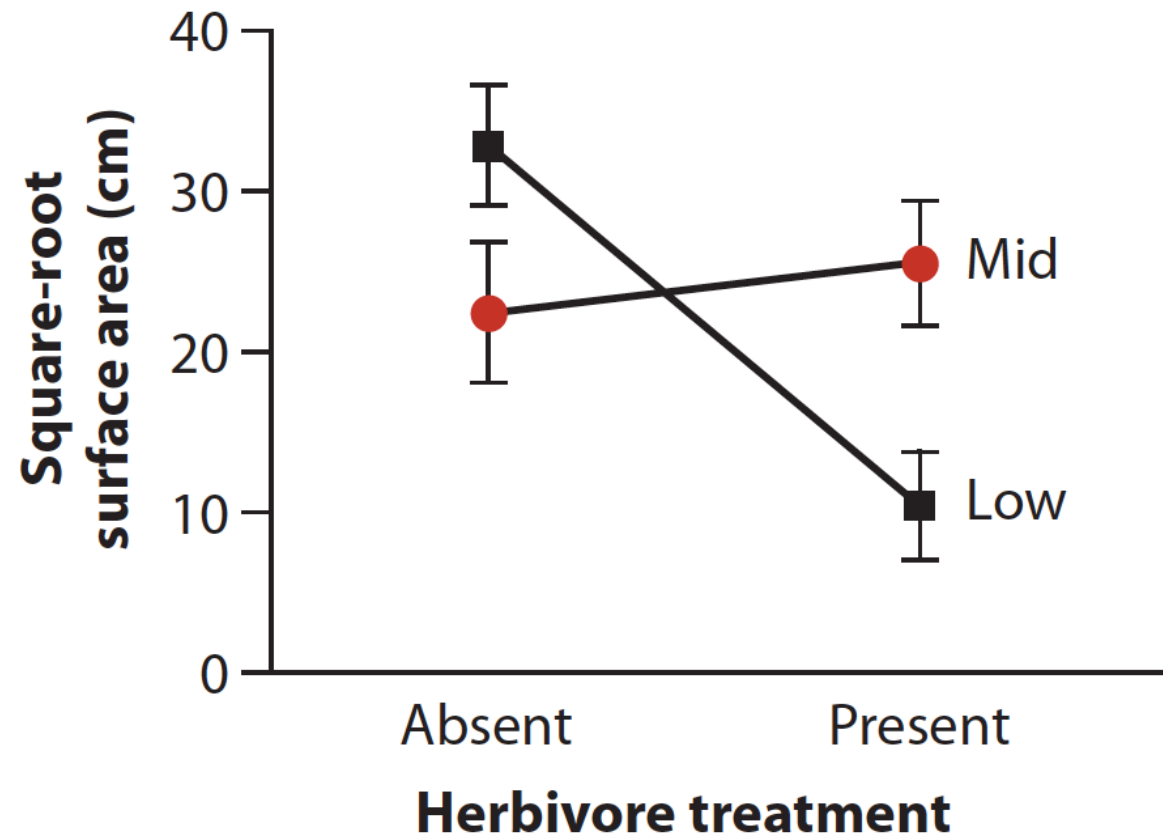
$$\text{F-test} = \frac{\text{Constant} + \text{Factor 1} + \text{Factor 2} + \text{Factor 1*Factor 2}}{\text{Constant} + \text{Factor 1} + \text{Factor 1*Factor 2}}$$

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:

**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	<u>237.842</u>		
Total	18488.63	63			

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

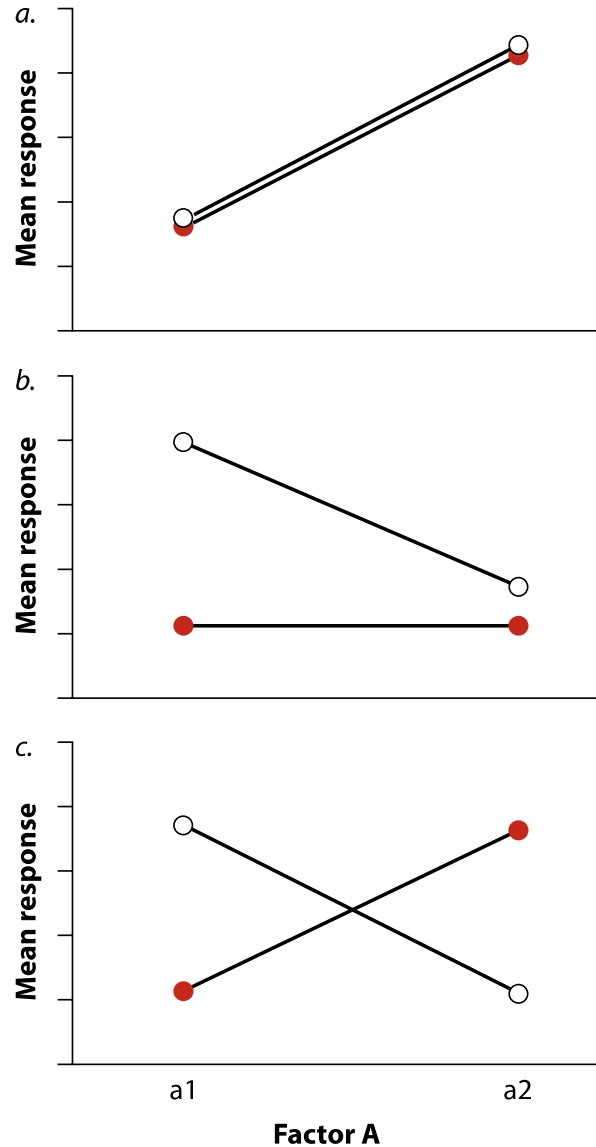
No interaction between height and herbivory is rejected

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		

# Multi-Factor ANOVA



1. No main effect of A or B but an interaction
  2. A main effect of A and B and an interaction
  3. A main effect of A, no main effect of B and no interaction
- \* It may be obvious, but the two different levels of B are indicated by the red dot (for one level of B) and the open circle (second level of B).

**A. (a,1), (b,2), (c,3)**

**B. (a,3), (b,1), (c,2)**

**C. (a,3), (b,2), (c,1)**

**D. (a,2), (b,1), (c,3)**

# 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?)

## Covariate effects:

Confounding variables bias estimates of treatment effects

**Experimental** - eliminate confounding variables by random assignment of treatment

**Observational** - include known confounding variables and correct for their distorting influence

## **ANCOVA: Analysis of Covariance**

### **Two rounds of model fitting:**

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

1. Interaction between covariate and treatment is tested  
Regression slopes differ among the 'groups' if interaction is present
2. If no interaction is detected, interaction term is dropped and treatment effect is tested

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

Two rounds of model fitting:

1. *Interaction between covariate and treatment is tested*

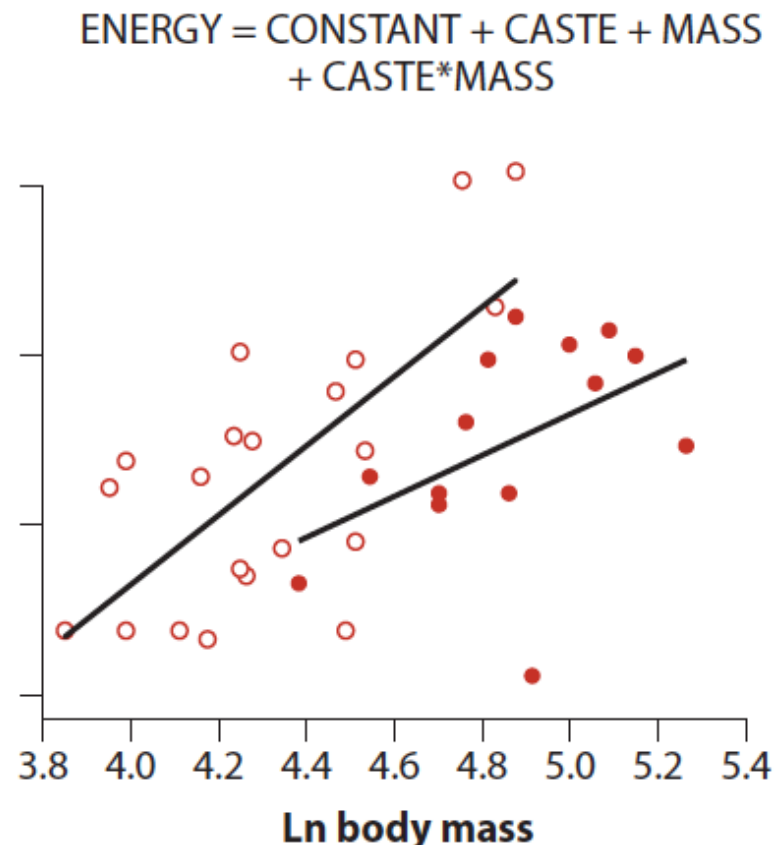
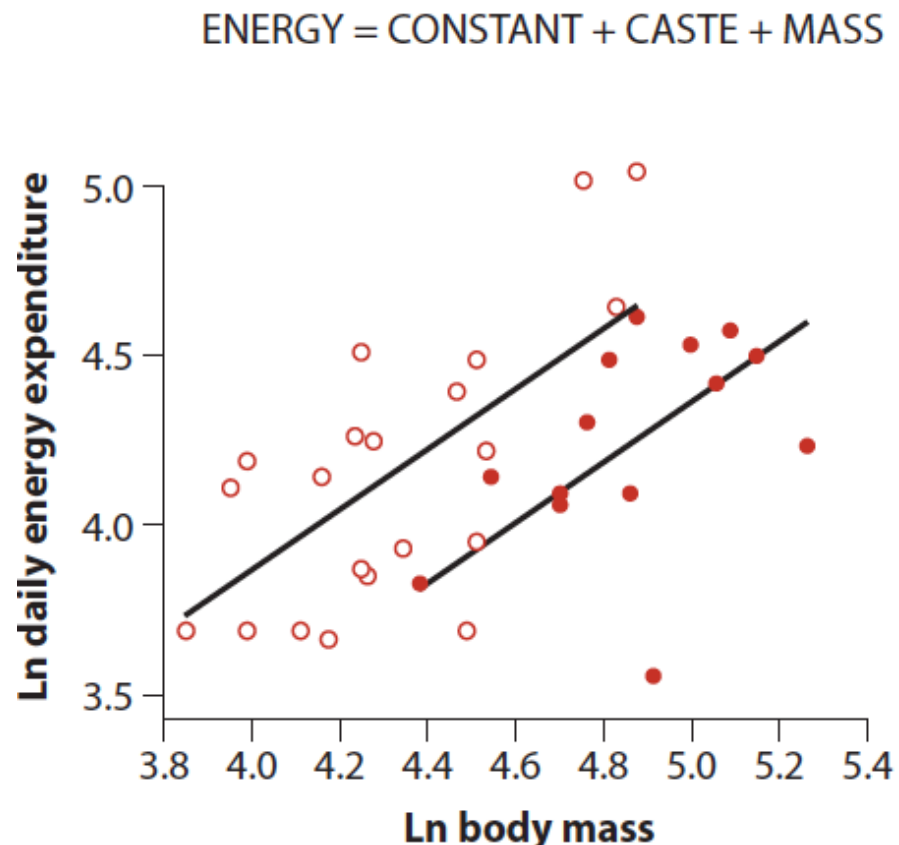
**Regression slopes differ among the ‘groups’ if interaction is present**

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

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

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

**Example:** Mole-rats are eusocial mammals with a queen, reproductive males and workers in a colony. It seems there might be two worker castes: “frequent workers”, who do most of the work of the colony, and “infrequent workers”, who do work after rains. Energy expenditure varies with body mass in both groups but infrequent workers are heavier than frequent workers. **How different is mean daily energy expenditure between two groups when adjusted for differences in body mass?**



## Covariate effects:

$$\text{Energy} = \text{Constant} + \text{Caste} + \text{Mass} + \text{Caste} * \text{mass}$$

### Two rounds of model fitting:

1. Interaction between covariate and treatment is tested

$H_0$ : There is no interaction between caste and mass

$H_A$ : There is interaction between caste and mass

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

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

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



# 1<sup>st</sup> Round Results:

**F-test =  $H_A$ : Constant + Caste + mass + Caste\*mass**  
 **$H_0$  Constant+Caste+Mass**

Source of Variation	SS	DF	MS	F	P
Caste	<b>0.0570</b>	1	<b>0.0570</b>		
Mass	1.3618	1	1.3618		
Caste*Mass	0.0896	1	0.0896	1.02	0.321
Residual	2.7249	32	0.0879		
Total	4.233	35			

Without the interaction term, the regression lines have slopes that are not significantly different – so we can drop interaction!

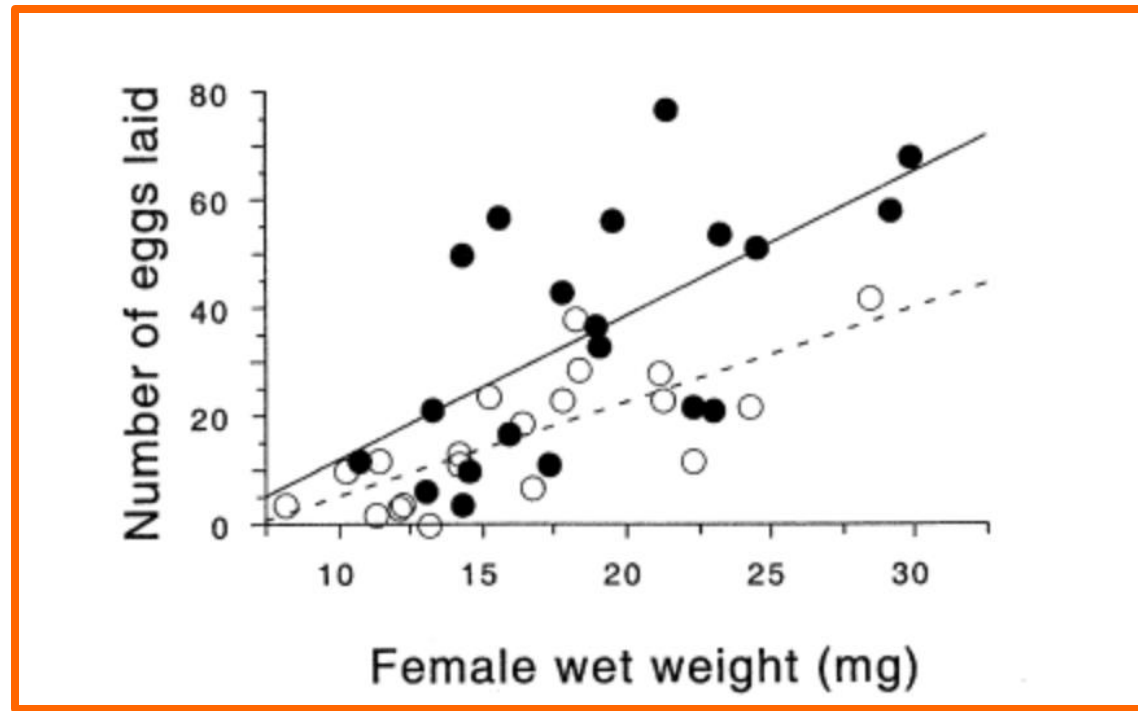
## Round 2: The updated model: **Energy = Constant + Caste+ Mass**

$H_0$ : There is no difference in energy expenditure between different castes

$H_A$ : There is a difference in energy expenditure between different castes

$$\text{F-test} = \frac{H_A: \text{Constant} + \text{Caste} + \text{Mass}}{H_0 \quad \text{Constant} + \text{Caste}}$$

Source of Variation	SS	DF	MS	F	P
Caste	<b>0.6375</b>	1	<b>0.6375</b>	7.25	0.011
Mass	1.8815	1	1.8815	21.39	<0.011
Residual	2.72.814	32	0.0880		
Total	5.3335	34			



Females were mated once (white circles) or three times (black filled circles).

Since larger females are known to produce more offspring, before mating, the female fireflies were weighed. The slopes between the once mated and thrice mated were not significantly different but the Y-intercepts are significantly different. Is there a difference in the number of eggs laid between once and thrice mated females?

- a. No
- b. Yes
- c. Can't tell

# Blocking

Results in an additional variable, a block, that must be included in analysis  
Can no longer use simple one-factor ANOVA

## Randomized block design

**Paired design** for > 2 treatments

Example:

Every treatment is replicated **once** within each block

Minimize “noise”

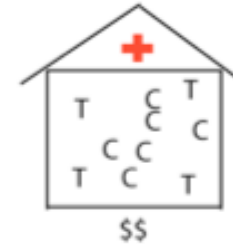
Accounting for any variation caused by blocking **can improve our treatment effect detection** (i.e., increase the power of our test)

Treatment effects are assessed by different treatments **within** each block so there is no interaction term

## Goals of experiments:

*determine how explanatory variable (treatment) affects response variable*

- Eliminate Bias
- Reduce Sampling Error
  - Blocking:



C = Control  
T = Treated

Variance among hospitals  
will not contribute to SE.

Only variance within hospitals  
will contribute to "noise"

# Main Principle of Blocking

$$\text{Response} = \text{Constant} + \text{Treatment} + \text{Block}$$

$$H_0: \text{Response} = \text{Constant} + \text{Block}$$

$$H_A: \text{Response} = \text{Constant} + \text{Block} + \text{Treatment}$$

- 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

# Example of Blocking:

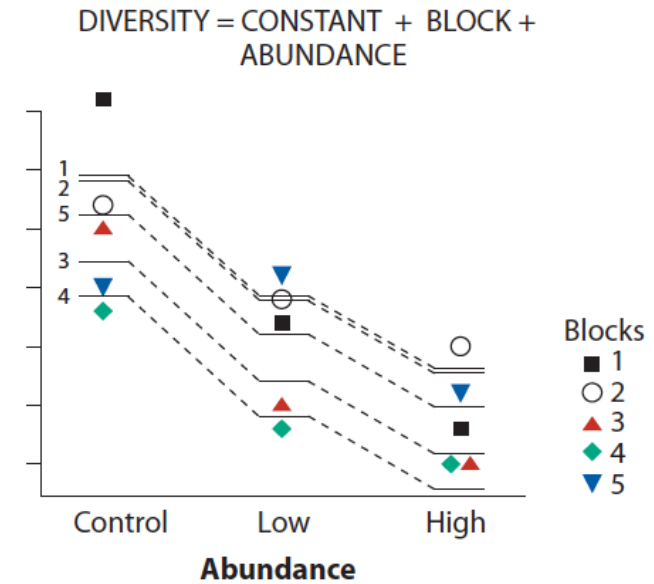
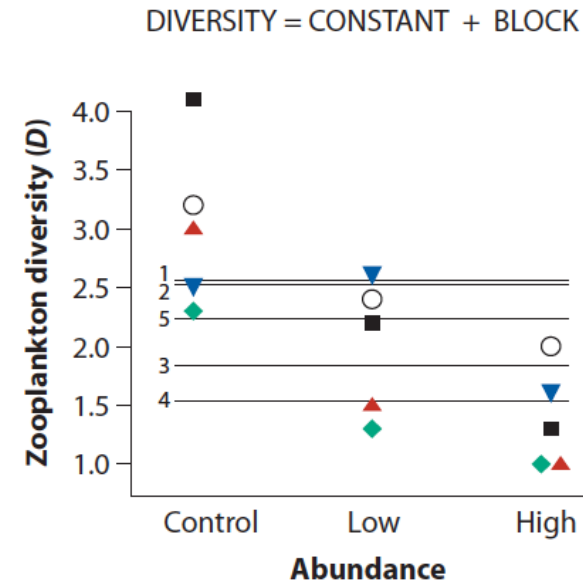
Response = Constant + Treatment + Block

$H_0$ : Response = Constant + Block

$H_A$ : Response = Constant + Block + Treatment

$$F = \frac{H_A}{H_0} = \frac{\text{Constant} + \text{Block} + \text{Treatment}}{\text{Constant} + \text{Block}}$$

$$= \frac{\text{residual} + \text{location} + \text{fish Abundance}}{\text{residual} + \text{location}}$$



Source of variation	Sum of Squares	df	Mean Square	F	P
<b>BLOCK</b>	2.340	4	0.5850		
<b>Treatment</b>	6.8573	2	3.4287	16.37	0.001
<b>Residual</b>	<u>1.6760</u>	<u>8</u>	<u>0.2095</u>		
<b>Total</b>	10.8733	14			

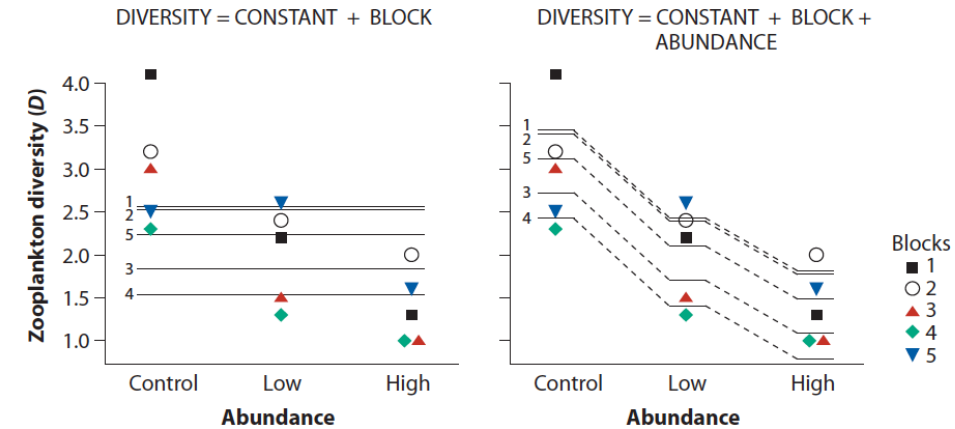
# Example of Blocking:

Response Full = Constant + Treatment + Block

Treatment:

$H_0$ : Response = Constant + Block

$H_A$ : Response = Constant + Block + Treatment



$$F = \frac{H_A}{H_0} = \frac{\text{Constant} + \text{Block} + \text{Treatment}}{\text{Constant} + \text{Block}} = \frac{\text{residual} + \text{location} + \text{fish Abundance}}{\text{residual} + \text{location}} = \frac{MS_{\text{treatment}}}{MS_{\text{block}}} = \frac{3.43}{0.59} = \mathbf{16.37}$$

$F_{0.05(1),2,14} = \mathbf{3.74}$  so we reject the  $H_0$

Source of variation	Sum of Squares	df	Mean Square	F	P
<b>BLOCK</b>	2.340	4	0.5850		
<b>Treatment</b>	6.8573	2	3.4287	16.37	0.001
<b>Residual</b>	<u>1.6760</u>	<u>8</u>	<u>0.2095</u>		
<b>Total</b>	10.8733	14			



# Example of Blocking:

Treatment:

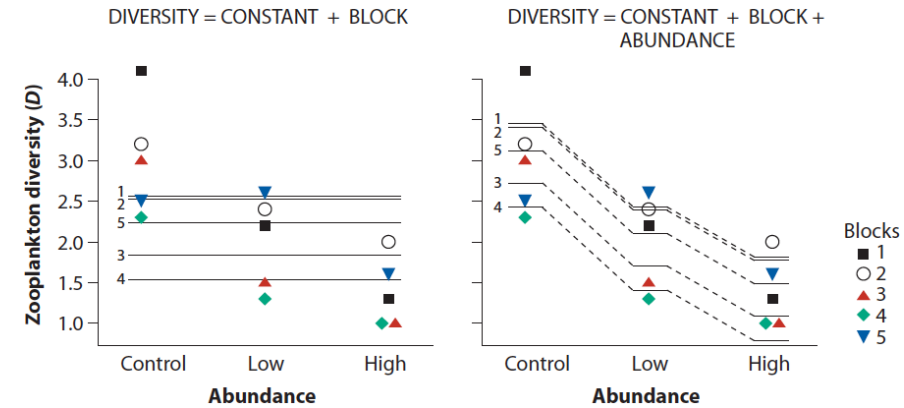
$H_0$ : Response = Constant + Block

$H_A$ : Response = Constant + Block + Treatment

$F = \frac{H_A}{H_0} = \frac{\text{Constant} + \text{Block} + \text{Treatment}}{\text{Constant} + \text{Block}}$

$$= \frac{\text{residual} + \text{location} + \text{fish Abundance}}{\text{residual} + \text{location}} = \frac{MS_{\text{treatment}}}{MS_{\text{block}}} = \frac{3.43}{0.59} = \mathbf{16.37}$$

$F_{0.05(1),2,14} = \mathbf{3.74}$  so we reject the  $H_0$



Block:

$$F_{\text{Block}} = \frac{H_A}{H_0} = \frac{\text{Residual} + \text{treatment} + \text{Block}}{\text{Residual} + \text{treatment}} = \frac{MS_{\text{block}}}{MS_{\text{residual}}} = \frac{0.5850}{0.2095} = 2.79$$

$F_{0.05(1),4,8} = \mathbf{3.84}$  so we fail to reject the  $H_0$

Source of variation	Sum of Squares	df	Mean Square	F	P
BLOCK	2.340	4	0.5850		
Treatment	6.8573	2	3.4287	16.37	0.001
Residual	1.6760	8	0.2095		
Total	10.8733	14			

## Blocking

Identify the blocking variable in the following example:

We have 3 different pastry recipes, and we are trying to determine which one is most delicious. Pastry is temperamental and can respond to many environmental features (temperature, moisture in the air etc.).

Day 1	Day 2	Day 3	Day 4	Day 5
1	3	1	2	1
3	1	2	3	2
2	2	3	1	3

**a. The particular recipe (#)**

**b. The particular day**

When I run this in RStudio, we get the following output:

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
block	4	6.9465	1.7366	3.6692	0.0555884
recipe	2	29.1472	14.5736	30.7918	0.0001747
Residuals	8	3.7864	0.4733		

- 
- A. Reject the null hypothesis and include block in further analysis**
- B. Reject the null hypothesis and DON'T include block in further analysis**
- C. Fail to reject the null hypothesis and include block**
- D. Fail to reject the null hypothesis and DON'T include block**