

# Module 4C

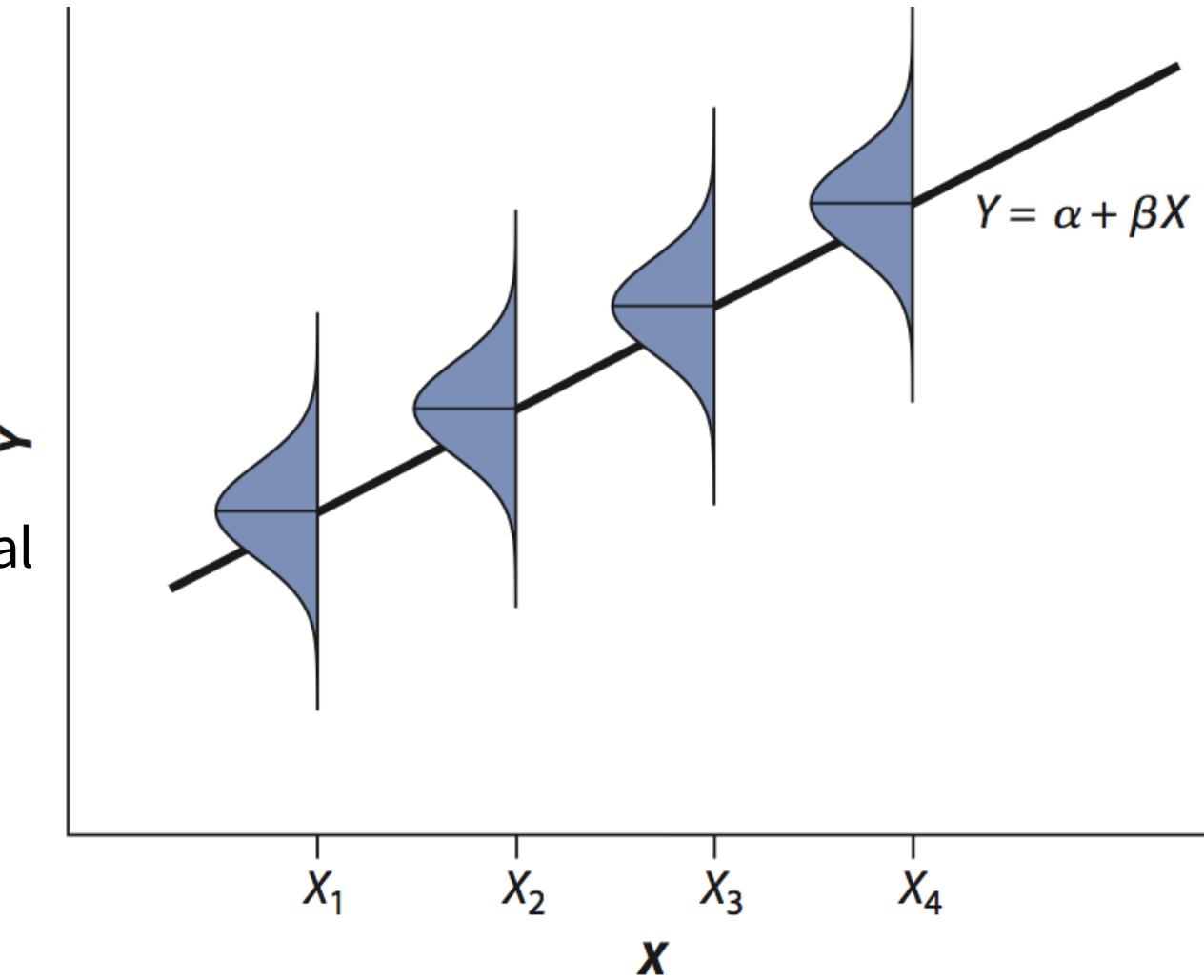
# Supervised Machine

# Learning

Different flavors of REGRESSION and General Linear Models

## 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)

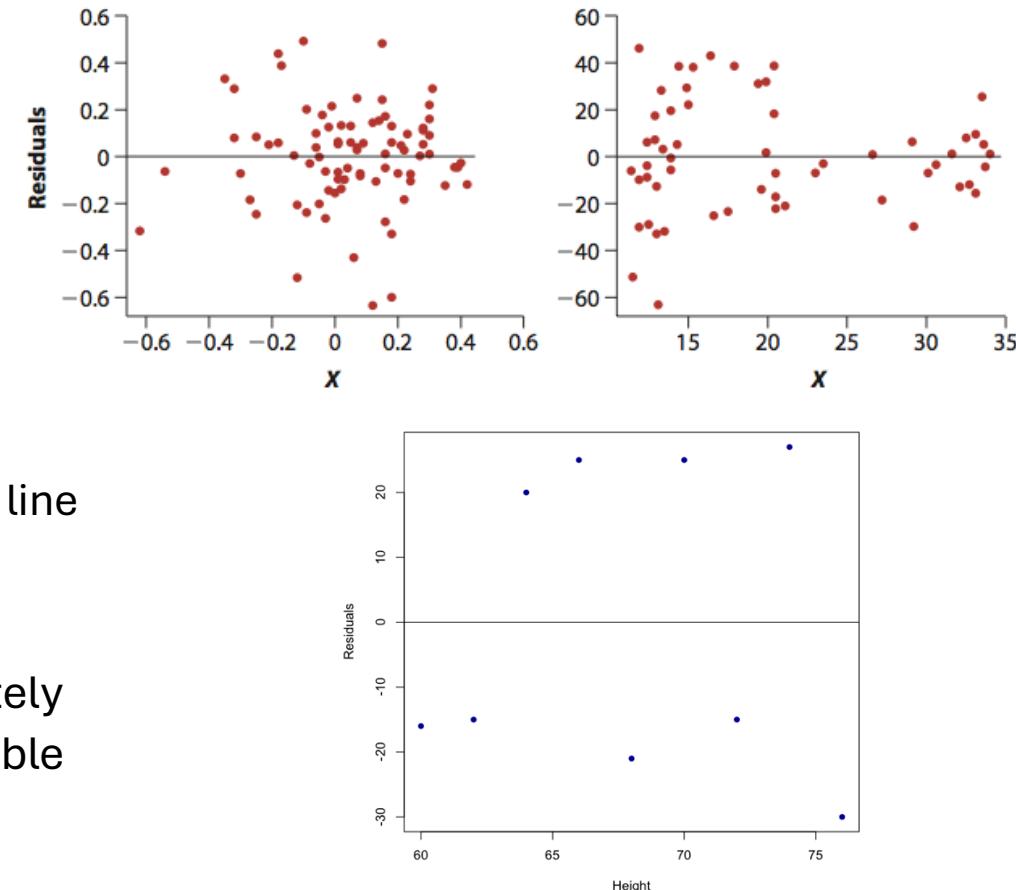
- Usually done by visual inspection of a scatterplot

### 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



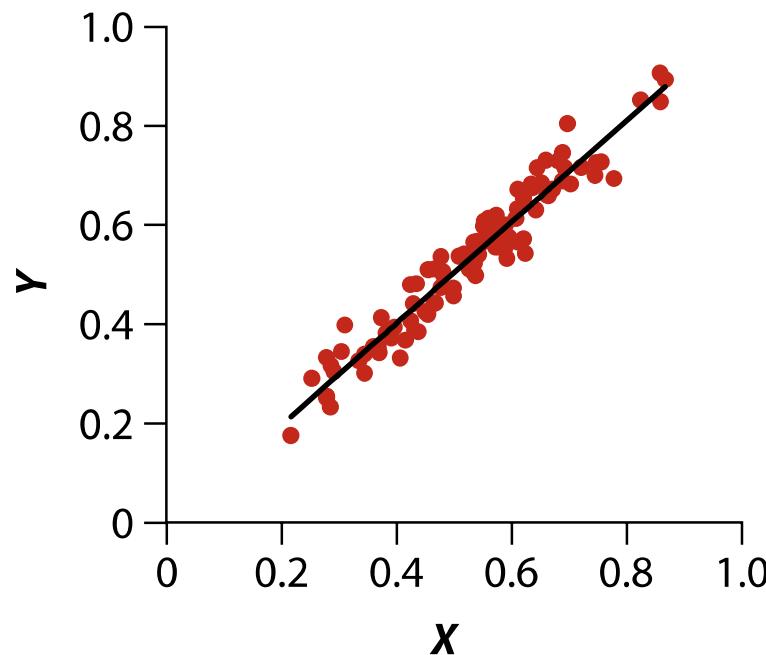
## If measurement error occurs on Y

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

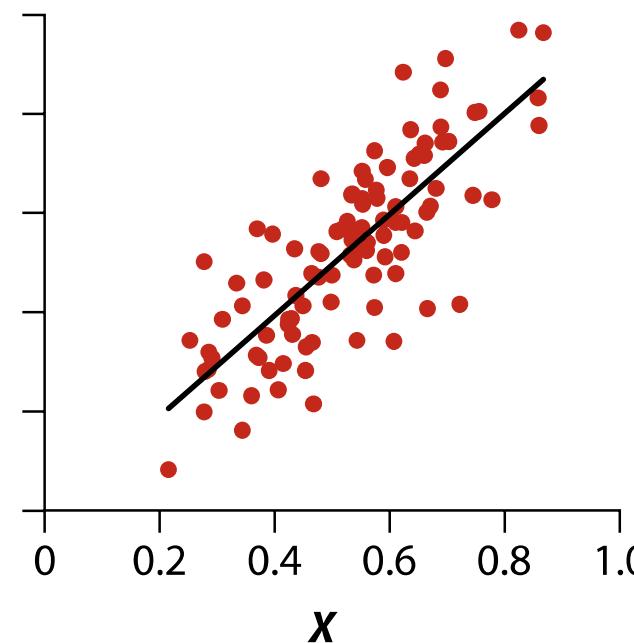
## If measurement error occurs on X

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

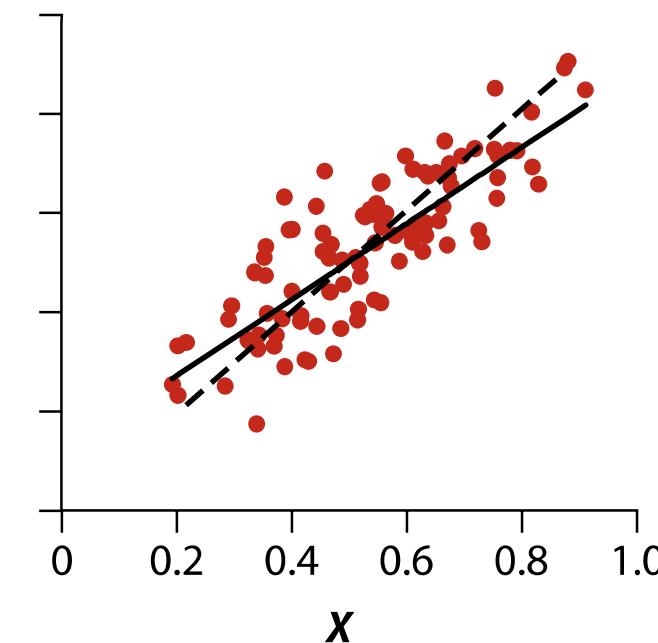
No measurement error



Measurement error in  $Y$

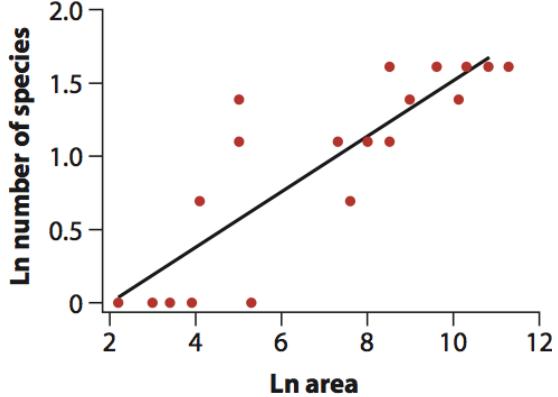
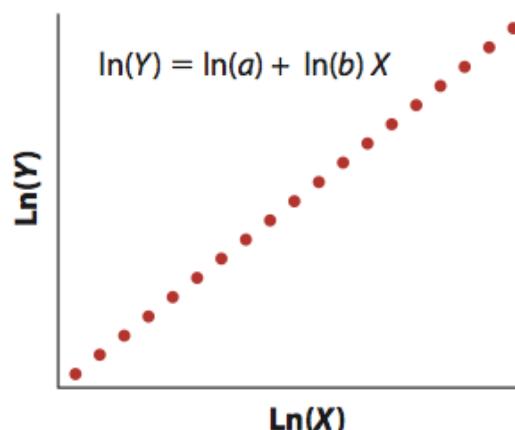
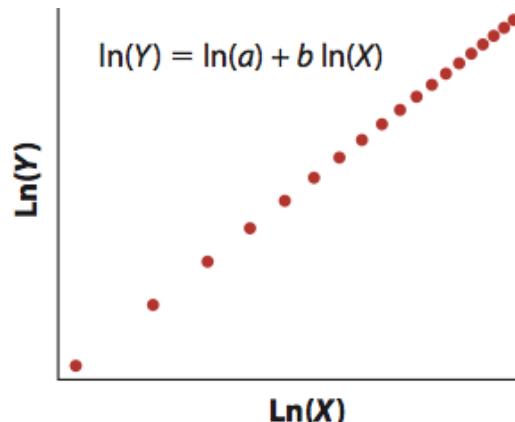
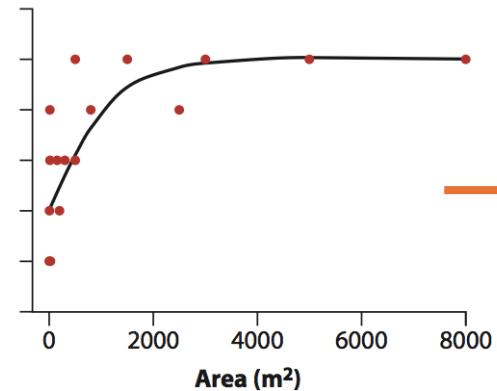
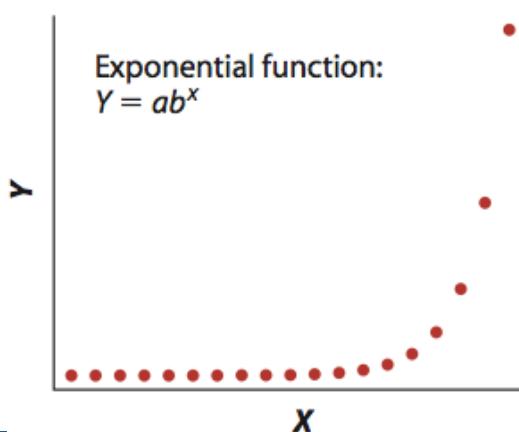
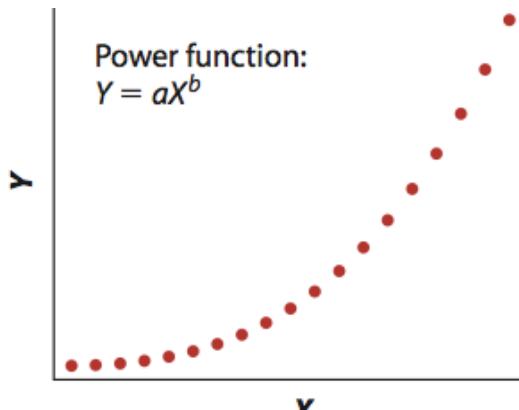


Measurement error in  $X$



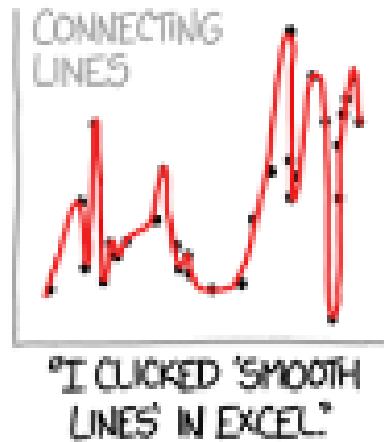
## Transformations:

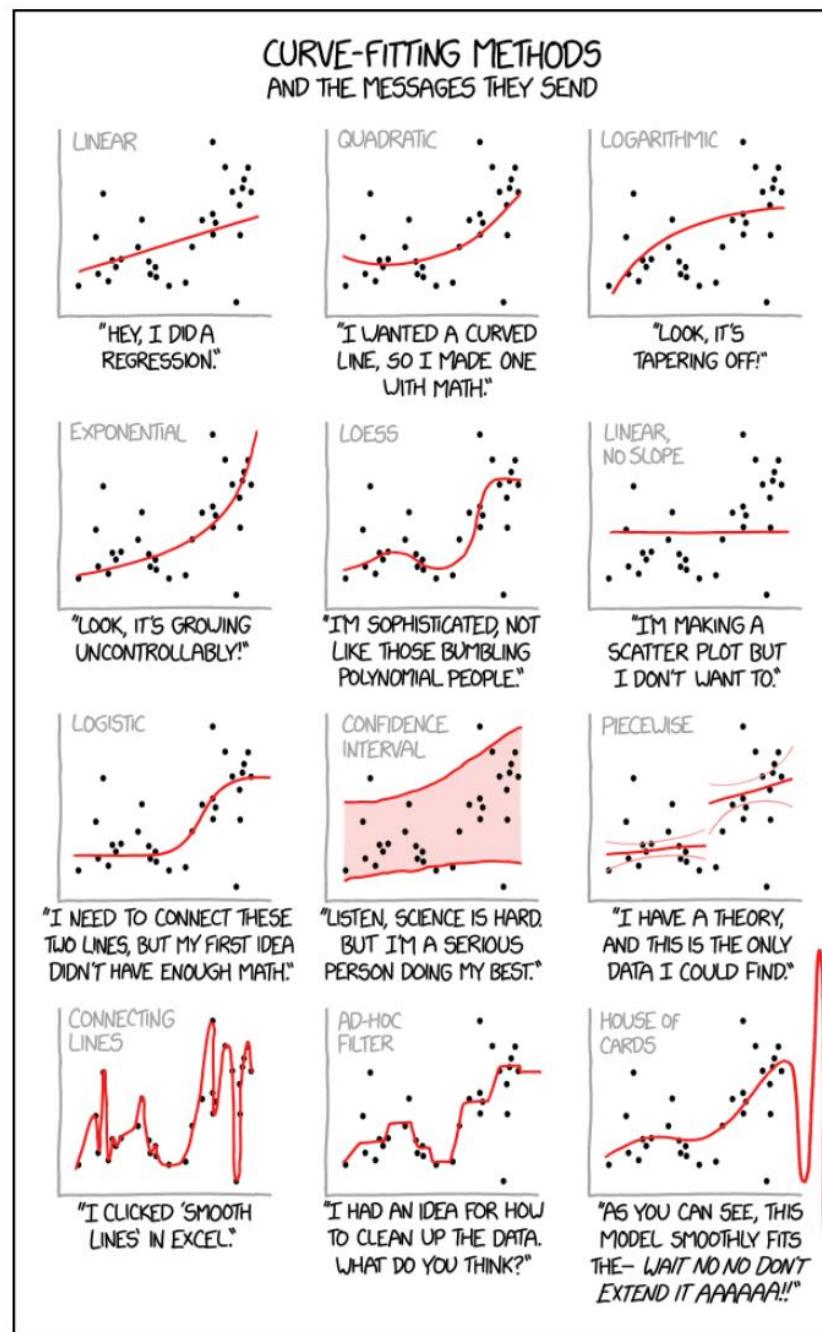
- Non-linear relationships can sometimes be forced into linearity
- The usual suspects:
  - log transformation for power and exponential relationships

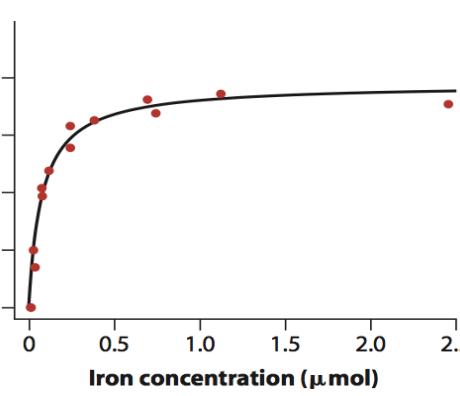
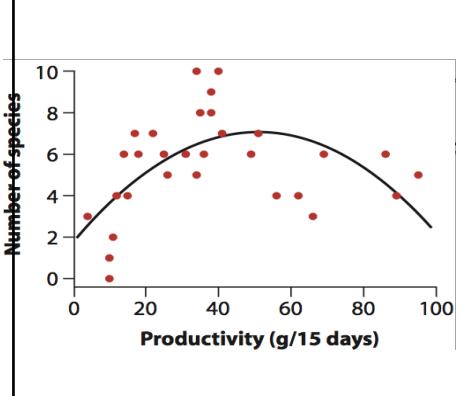
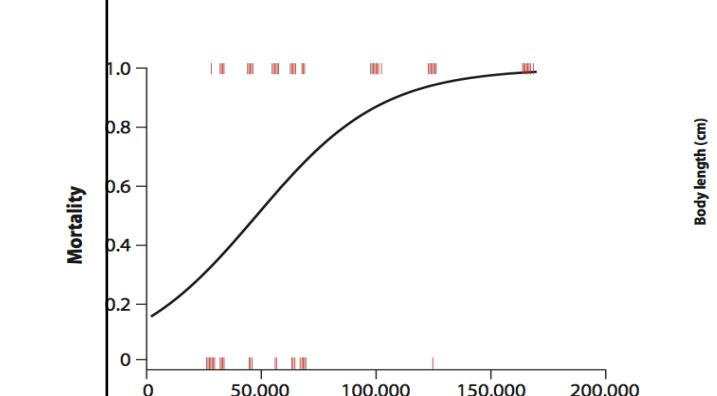
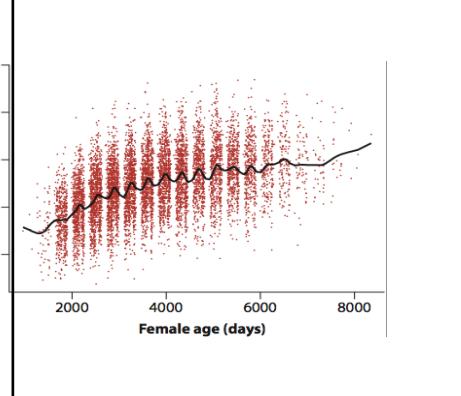


## Non-linear Regression:

- Same assumptions as 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**

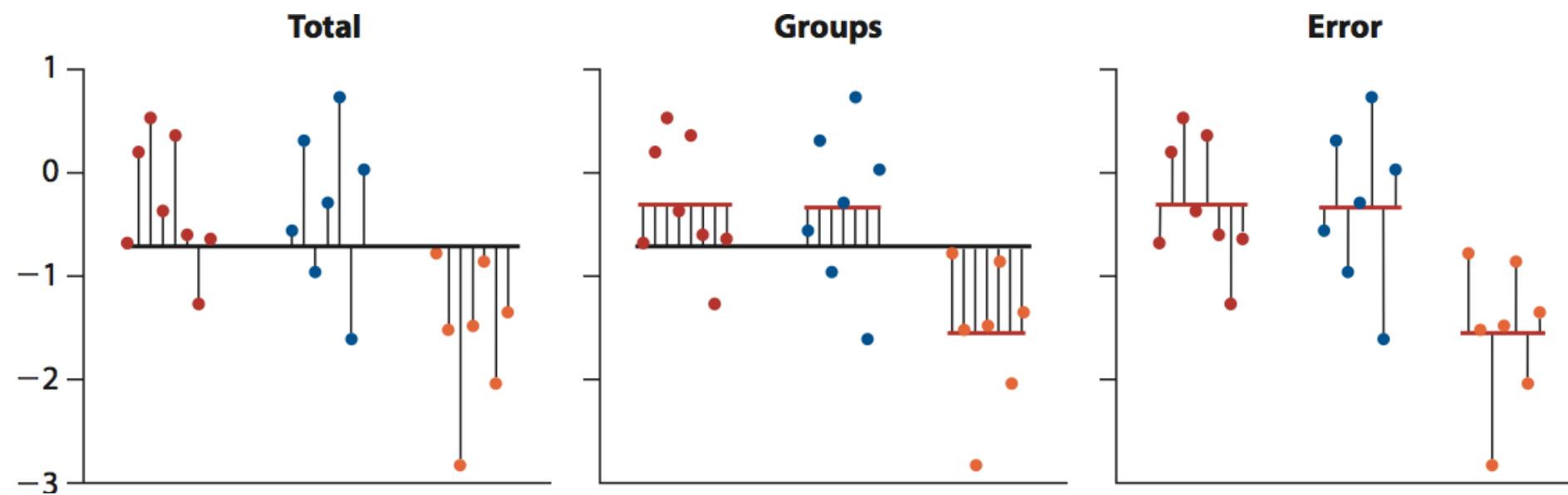




Curve with Asymptote	Quadratic curve	Binary response Variable	Smoothing
$Y = \frac{aX}{b + X}$	$Y=a+bX+cX^2$	$\text{Log-odds}(Y)=a+bX$	<ul style="list-style-type: none"> <li>depends on data</li> </ul>
Michaelis-Menten eq <sup>n</sup>	Parabolic relationships	Dose response curve	Diagnosis of exclusion
 <p>A scatter plot showing the relationship between Iron concentration (<math>\mu\text{mol}</math>) on the x-axis (ranging from 0 to 2.5) and Number of species on the y-axis (ranging from 0 to 15). A smooth black curve is fitted to the data points, which show a rapid increase in species number at low concentrations that levels off as it approaches a maximum value.</p>	 <p>A scatter plot showing Productivity (g/15 days) on the x-axis (ranging from 0 to 100) and Number of species on the y-axis (ranging from 0 to 10). Red dots represent data points, and a black parabolic curve is fitted to them, showing a peak in productivity around 40 units before decreasing.</p>	 <p>A scatter plot showing Mortality on the y-axis (ranging from 0 to 1.0) and Anthrax dose (spores/l) on the x-axis (ranging from 0 to 200,000). Red dots indicate mortality events, and a black S-shaped curve shows the probability of mortality increasing from near zero to near one as the dose increases.</p>	 <p>A scatter plot showing Body length (cm) on the y-axis (ranging from 110 to 150) and Female age (days) on the x-axis (ranging from 2000 to 8000). Red dots represent individual data points, and a black line shows a smoothed trend that remains relatively flat around 125 cm until about 6000 days, after which it rises slightly to approximately 130 cm.</p>

# General Linear Models

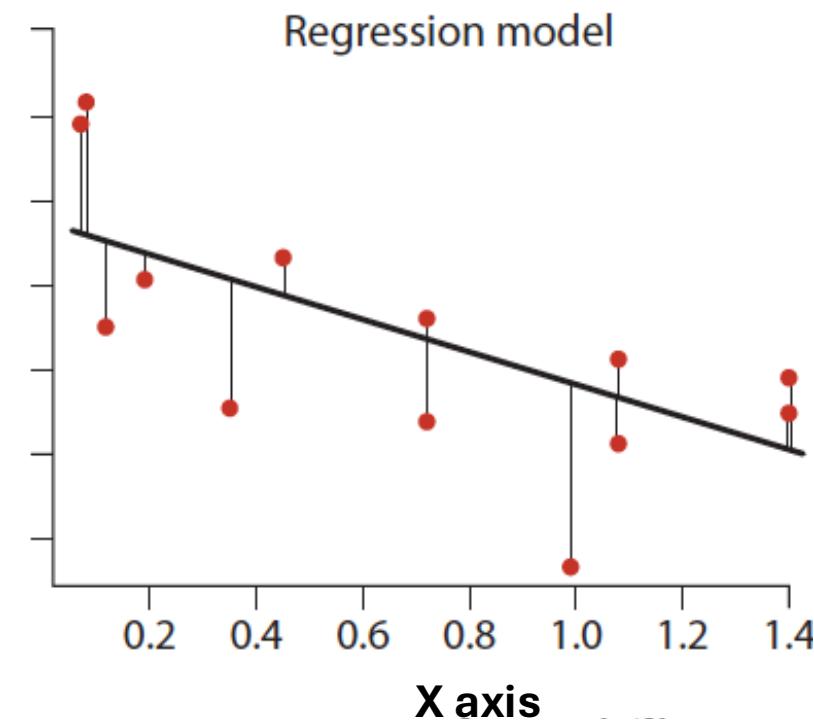
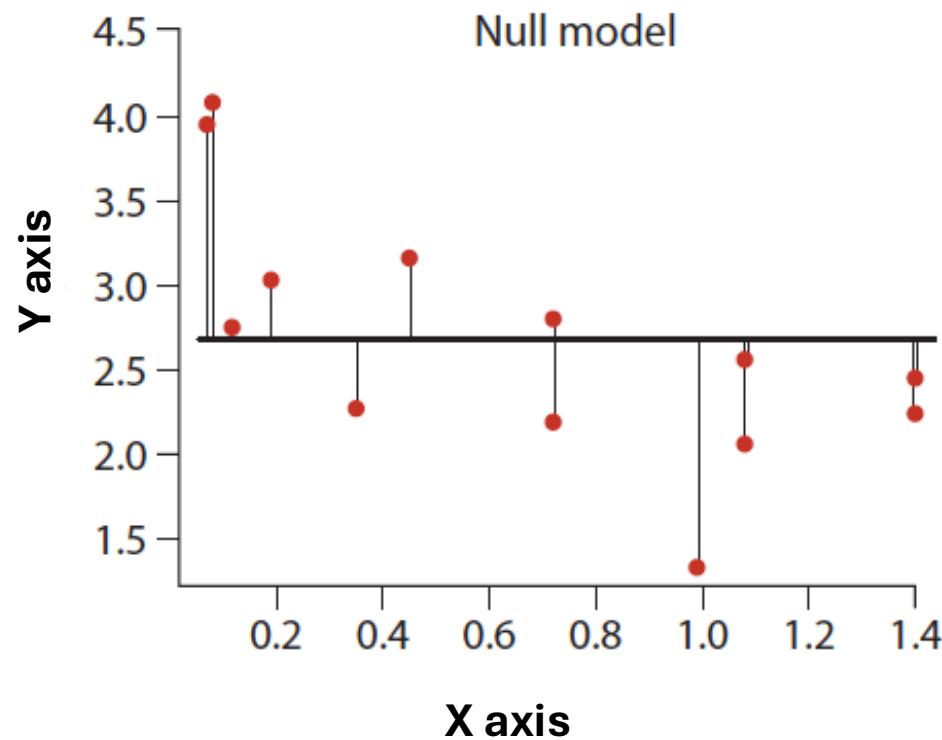
- Response variable,  $Y$ , can be represented by a linear model plus random error
  - Scatter of  $Y$  measurements around the model is random error
- So far, we have looked at (univariate) ANOVA, linear regression, and t-tests



## General linear model

We have also looked at the linear regression

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

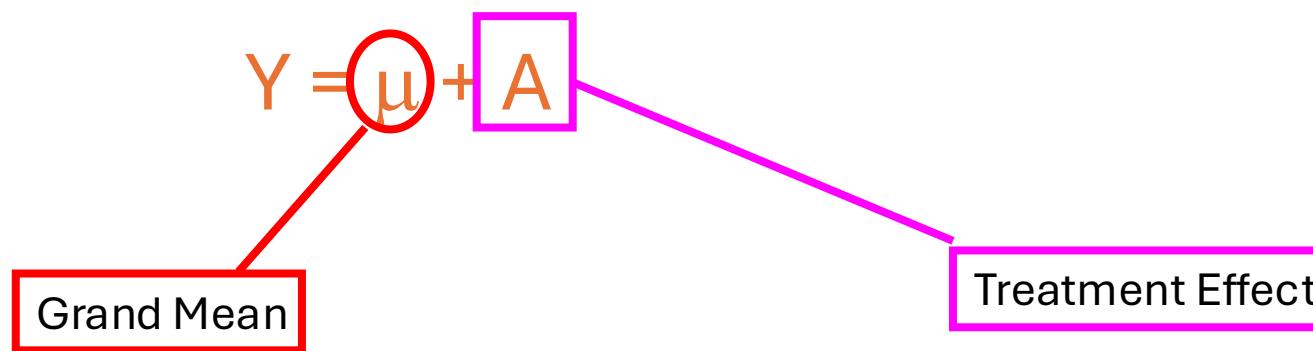


## General linear model

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

### Example:

Linear model for single-factor ANOVA



## General linear model

- Linear Model for single-factor ANOVA
- Linear Regression

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

$A_i$  = group mean -  $\mu$

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

Purple = we've already seen; Orange = We will see next



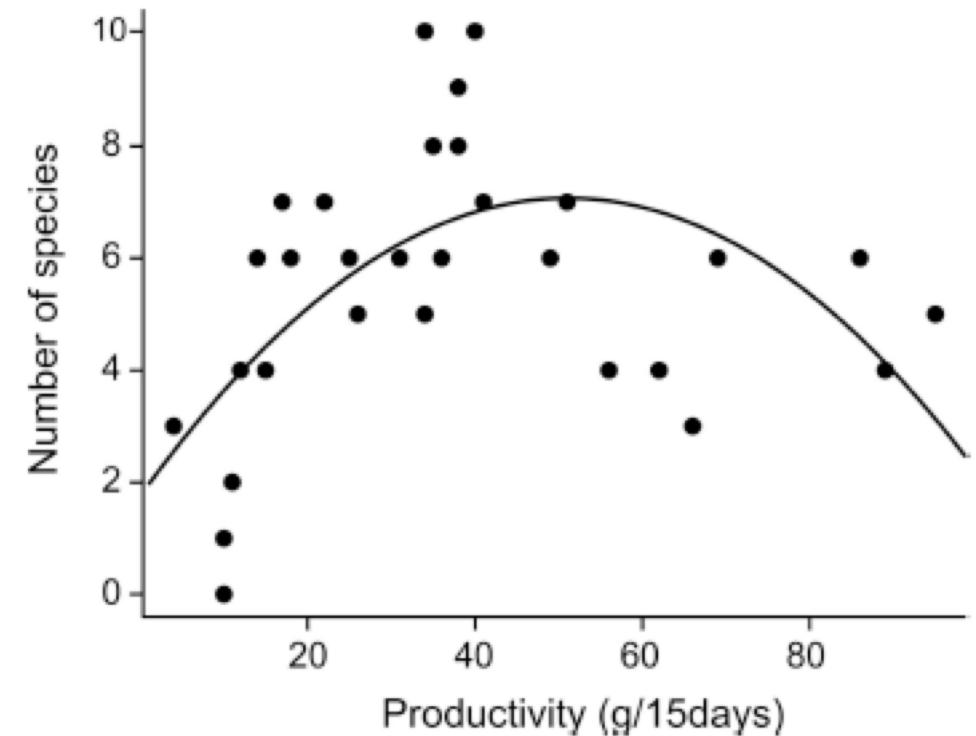
## Note: General linear model

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

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

***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

**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 → paired t-test is an example of blocking design

#### **Factorial experiment**

Investigate main effects of  $\geq 2$  treatment variables

Interactions

#### **Covariates**

Confounding variables

Nuisance variable is known but uncontrollable