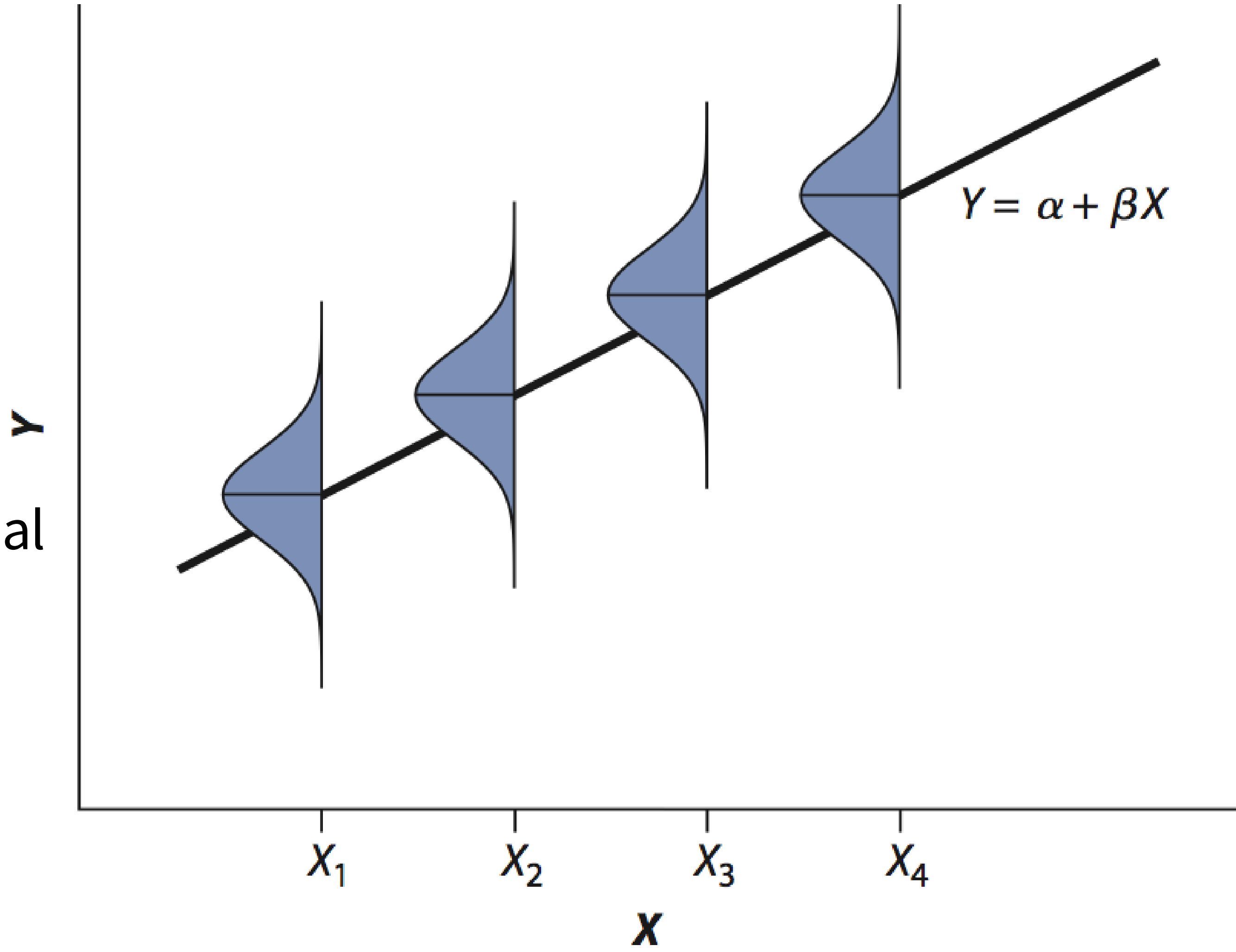
Module 4C
Supervised Machine Learning

Different flavors of REGRESSION and General Linear Models

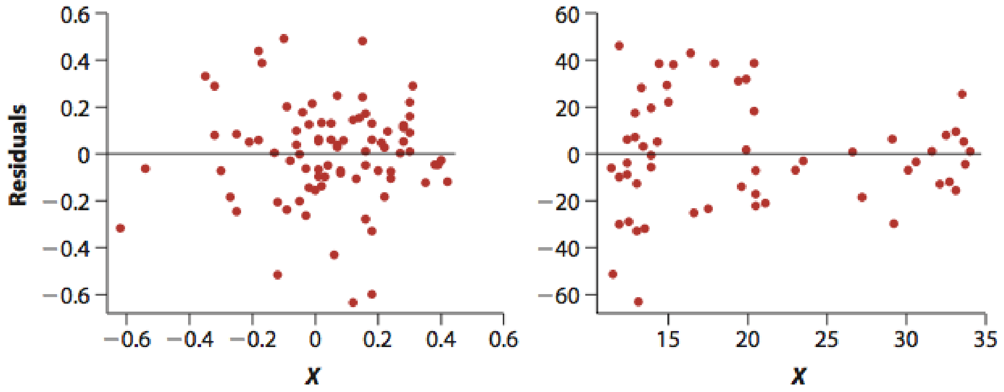
Assumptions of Regression Analysis:

- 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



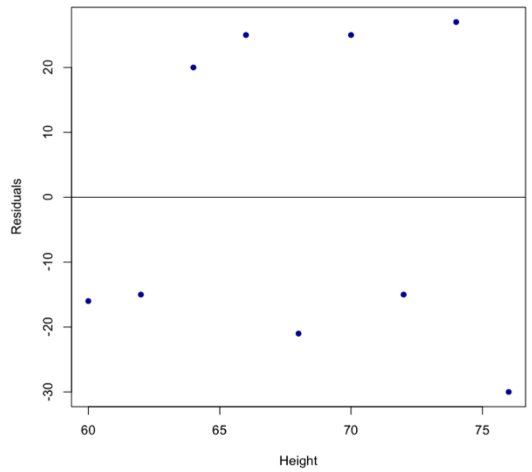
Major types of violation:

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



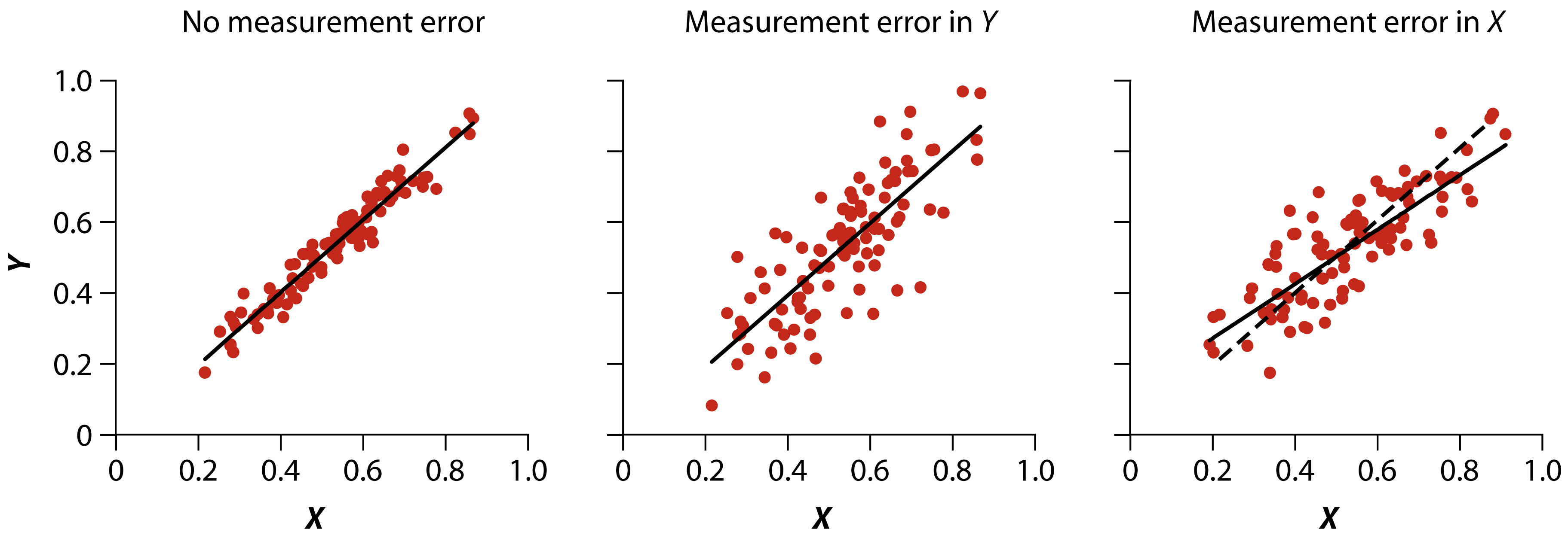
4. Measurement Errors
   - Biological traits can be difficult to measure accurately
   - Effects of measurement error depends on the variable

If measurement error occurs on Y

* Increase variance of residuals
* Increases SE of slope

If measurement error occurs on X

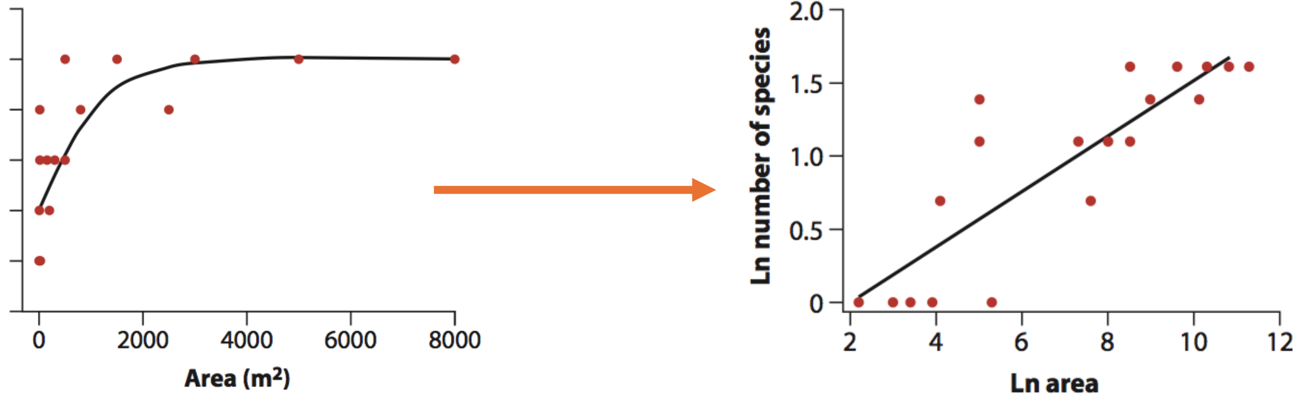
* Increases variance of residuals
* **Causes attenuation bias in estimate of b** (underestimates slope)
  - b will lie closer to 0 than β
  - Remember: BIAS is <u>really bad!</u>

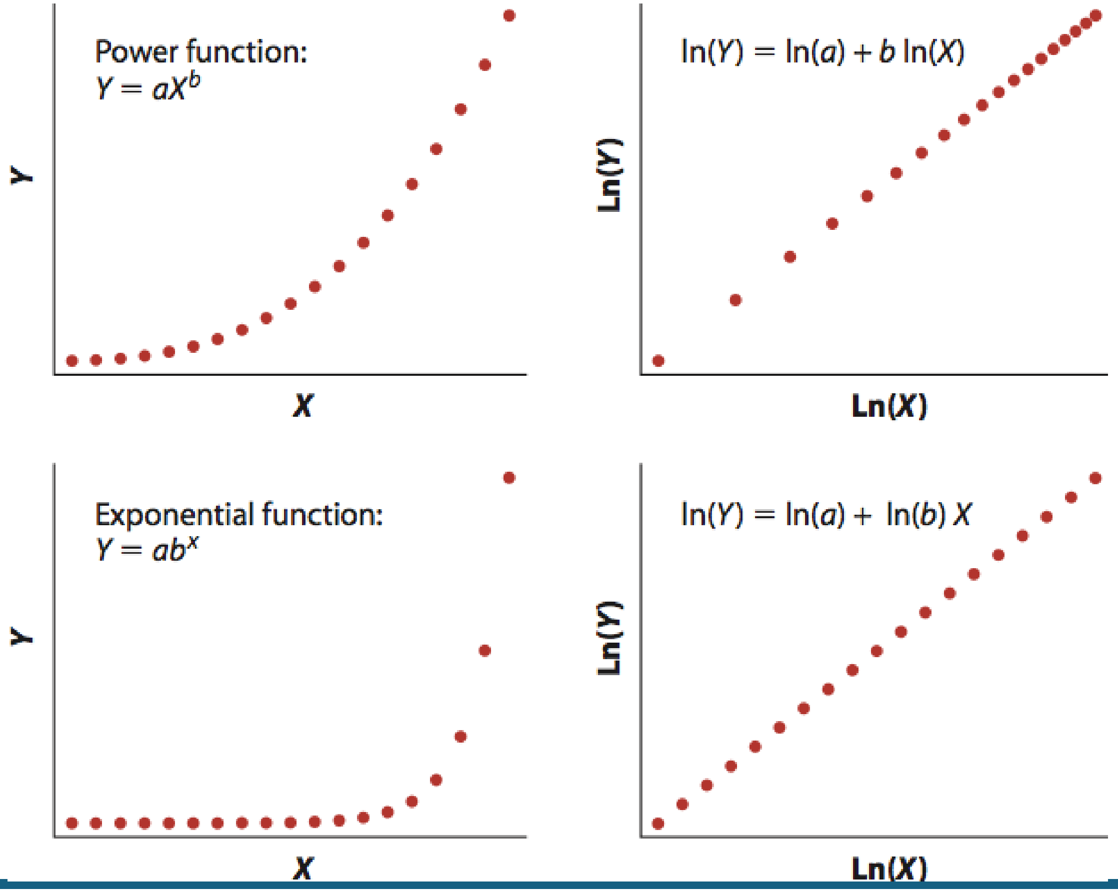


Transformations:

- Non-linear relationships can sometimes be forced into linearity

- The usual suspects:
  - log transformation for power and exponential relationships



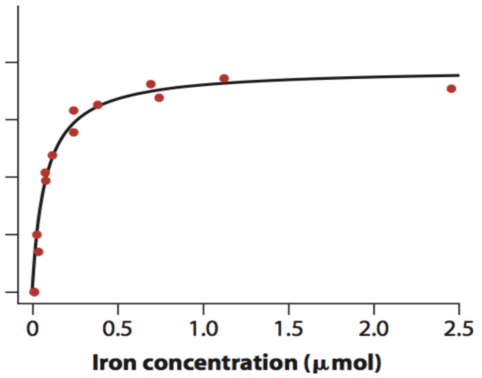


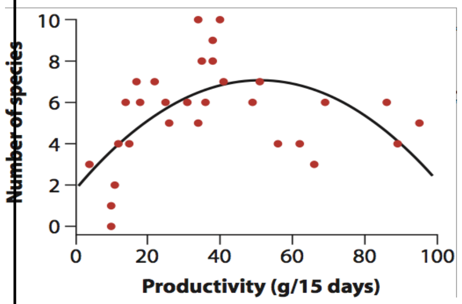
Non-linear Regression:

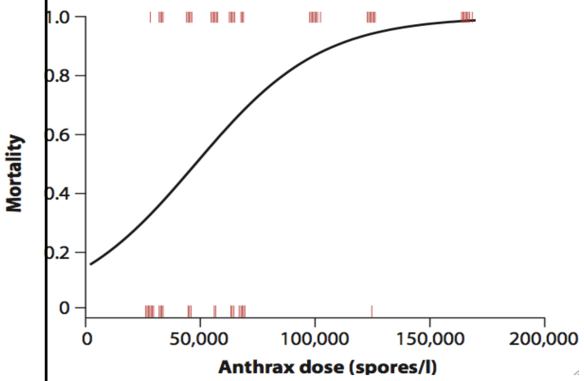
- Same assumptions are linear regression but, obviously, doesn't assume a linear relationship

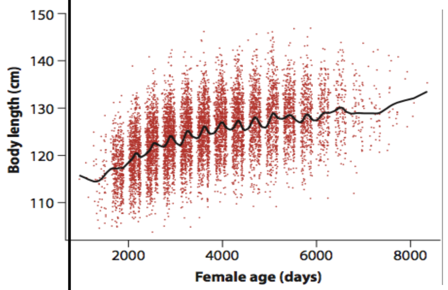
- Keep it simple: Don't **over fit**
  - It is possible to get a curve that fits each and every point ($MS_{residual} = 0$) but it will not predict future points since the curve ***doesn't describe a general trend***

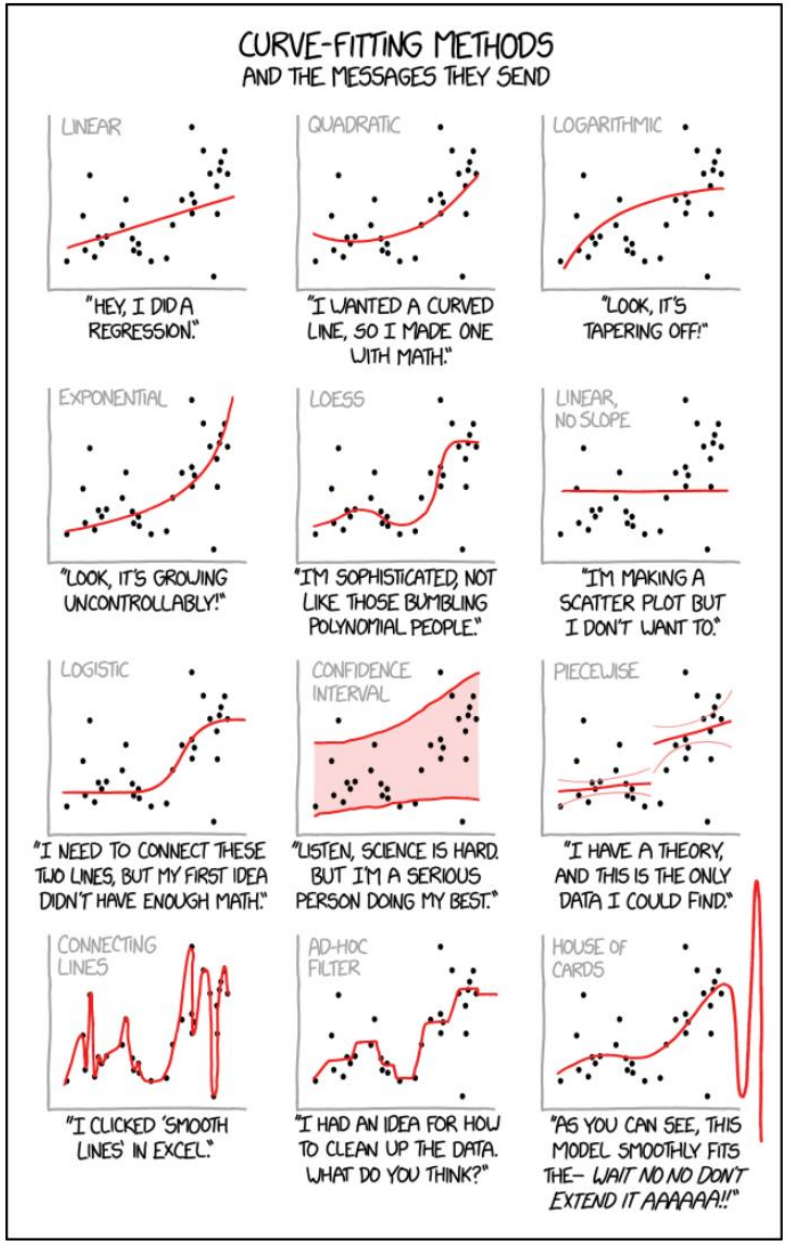
Curve with Asymptote	Quadratic curve	Binary response Variable	Smoothing
$Y = \dfrac{aX}{b+X}$	$Y=a+bX+cX^2$	Log-odds(Y)=a+bX	• depends on data
Michaelis-menten eq$^{\underline{n}}$	Parabolic relationships	Dose response curve	Diagnosis of exclusion







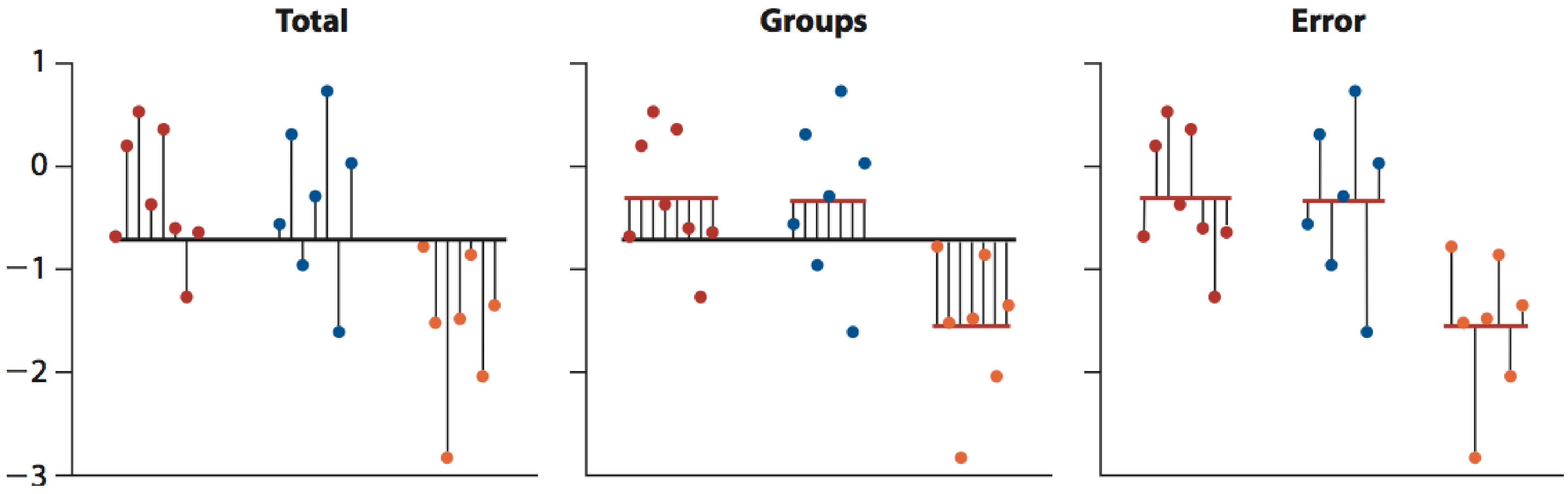




https://xkcd.com/2048/

General Linear Models

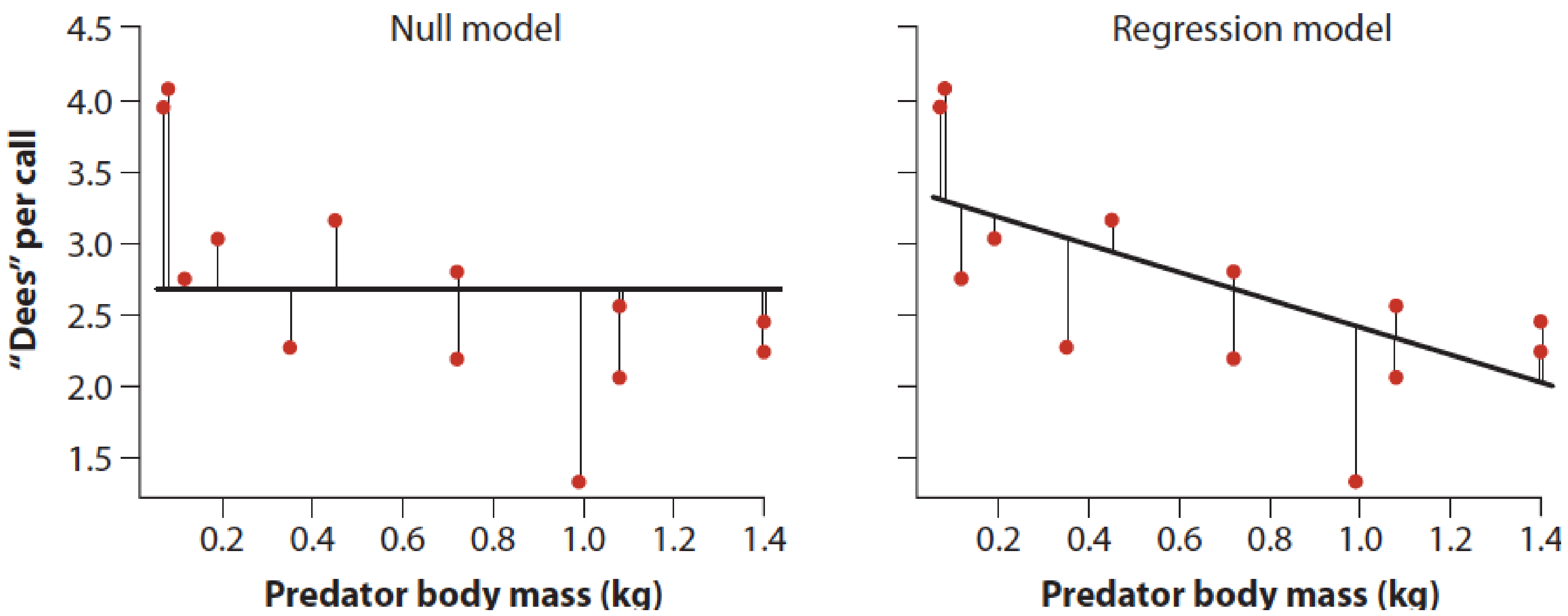
- Response variable, **Y**, can be represented by a linear <u>model</u> plus random error
  - Scatter of Y measurements around the model is random error
- So far, we have looked at (univariate) ANOVA, linear regression, and t-tests



General linear model

We have also looked at the linear regression

$$\boxed{Y = \alpha + \beta X + \varepsilon}$$



General linear model

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

Example:

Linear model for single-factor ANOVA

$$Y = \underbrace{\mu}_{\text{Grand Mean}} + \underbrace{\boxed{A}}_{\text{Treatment Effect}}$$

General linear model

- Linear Model for single-factor ANOVA
- Linear Regression

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

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

You are fundamentally fitting two models in both cases

RESPONSE = CONSTANT + VARIABLE

- Analysis of covariance
- Multiple regression

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

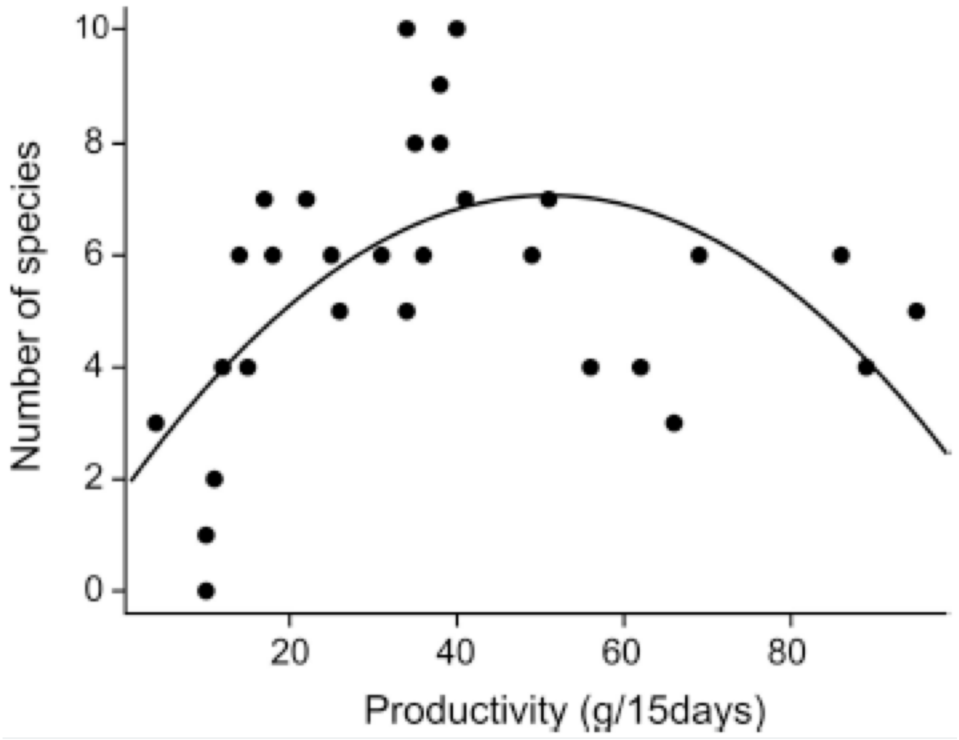
Note: General linear model

In the form:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\ldots+\text{error}$$

Doesn't have to be LINEAR relationship:

$$Y = \beta_0+\beta_1X+\beta_2X^2 \text{ (Quadratic)}$$



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**

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

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 ≥ 2 treatment variables
- Interactions

Covariates

- Confounding variables
- Nuisance variable is known but uncontrollable