

Functional forms and model construction

Model specification

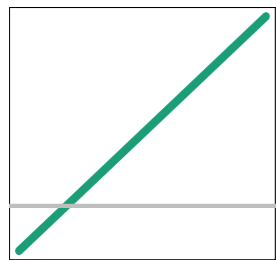
- Model formulation requires the selection of a functional form to formalize the relationship between the predictor variables and the process we are trying to understand.
- The functional form should clarify the verbal description of the mechanisms driving the process under study.
- Choosing among functional forms is a skill that needs to be developed over time.

How are predictor and response variables related?

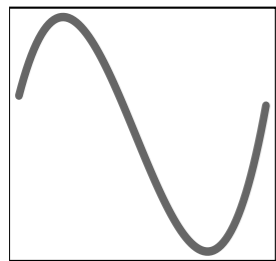
- Linear relationship?
- Saturating relationship?
- Accelerating relationship?
- Or more complicated?
- Is response bounded?
(eg, must be > 0 ?)

Two classes of functions

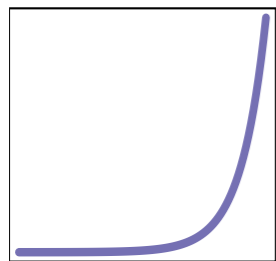
Non-asymptotic



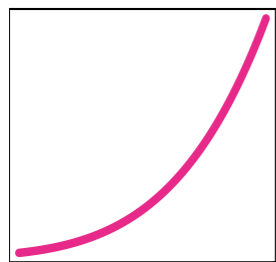
Linear



Polynomial

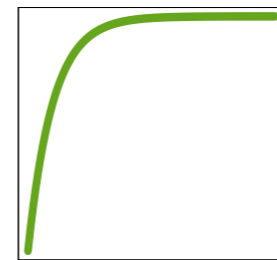


Exponential

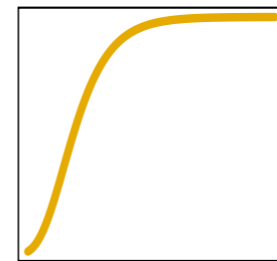


Power-law

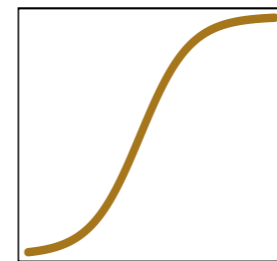
Asymptotic



Monomolecular



Gompertz



Logistic (3 & 4
param)

There are lots of other functions out there -
these are just some of the ones particularly appropriate for growth

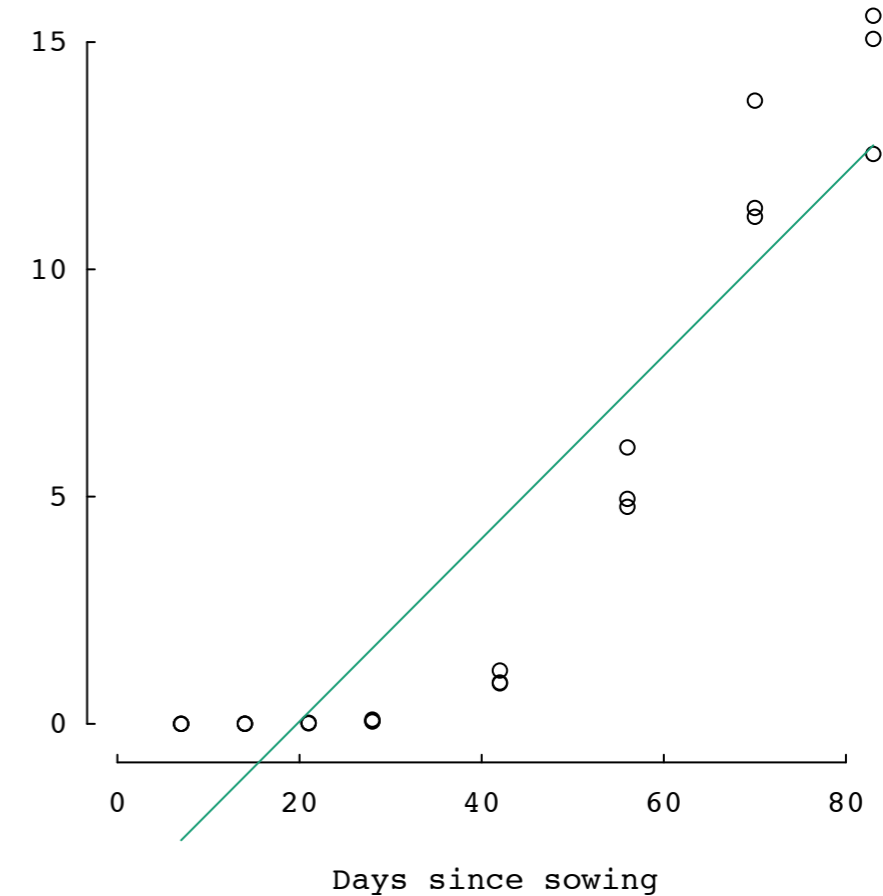
M: biomass

M₀: initial biomass

t: time

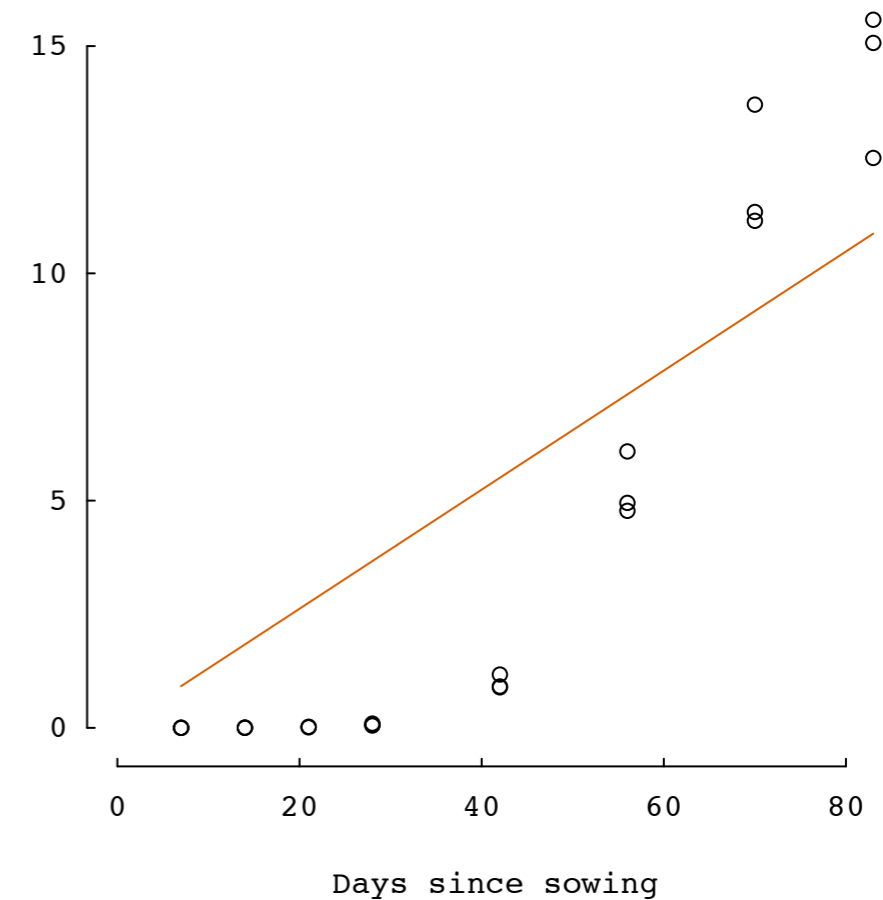
Linear

- $dM/dt = r$
- $M = M_0 + rt$
- 2 parameters
- Assumes constant amount of biomass added at every time point
- Unreasonable for growth



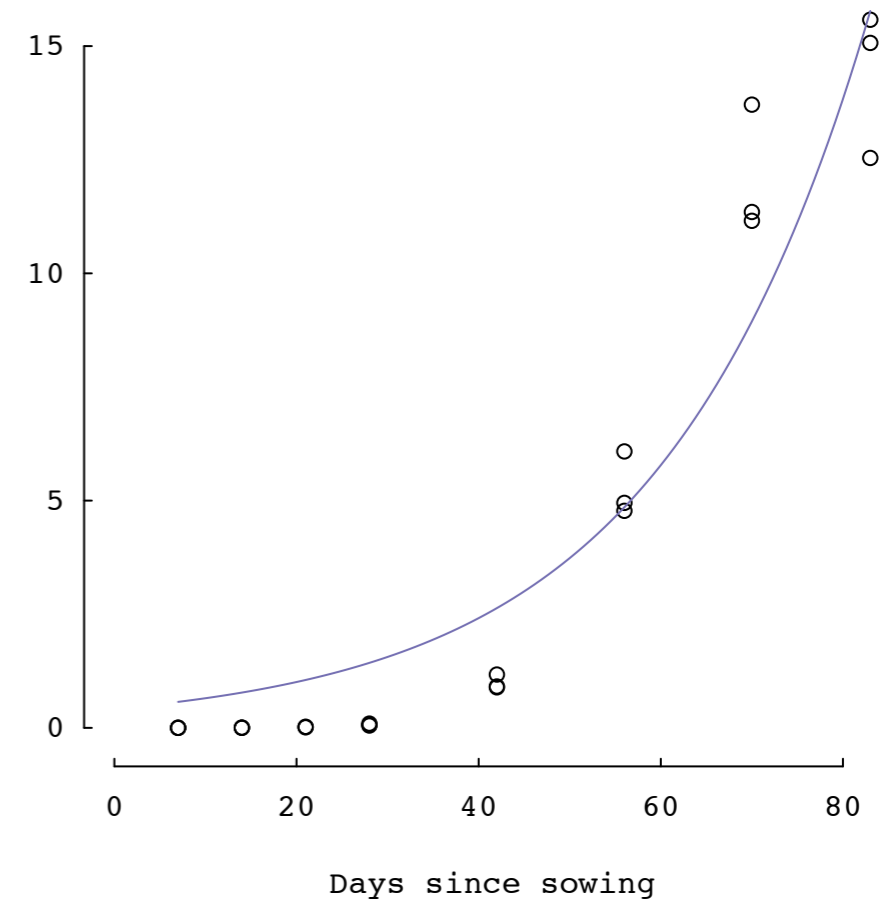
Linear (no-intercept)

- $dM/dt = r$
- $M = 0 + rt$
- 1 parameter
- Assumes constant amount of biomass added at every time point & that biomass starts at 0
- Also unreasonable for growth



Exponential

- $Y = M_0 * e^{rt}$
- 2 parameters
- Assumes constant **proportion** of biomass is added in every unit of time
- Reasonable for unlimited growth
- But, ecology gets in the way, and growth slows

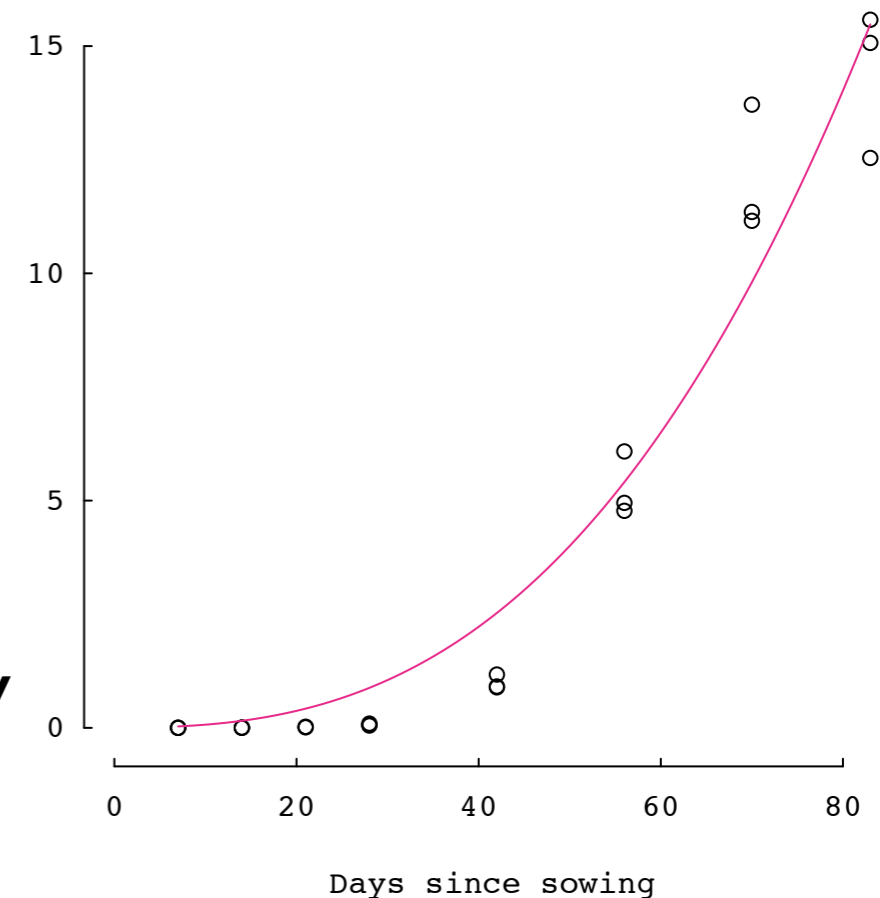




Brian Enquist

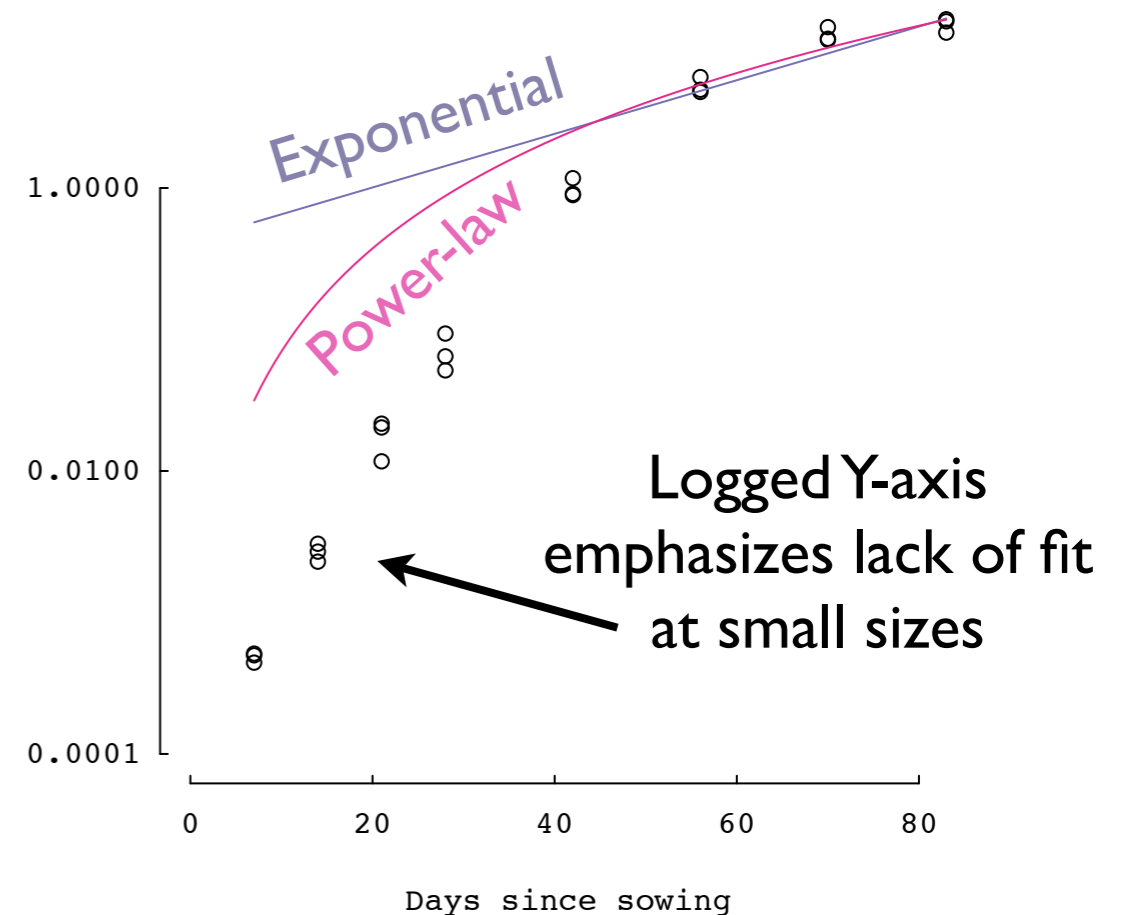
Power-law

- $dM/dt = rM^\beta$
- $M = \left(M_0^{1-\beta} + rt(1-\beta)\right)^{1/(1-\beta)}$
- 3 parameters
- Generally a good choice
- Hard to fit because of tradeoffs between parameters, and uncertainty in beta
- Metabolic theory suggests that beta = 3/4



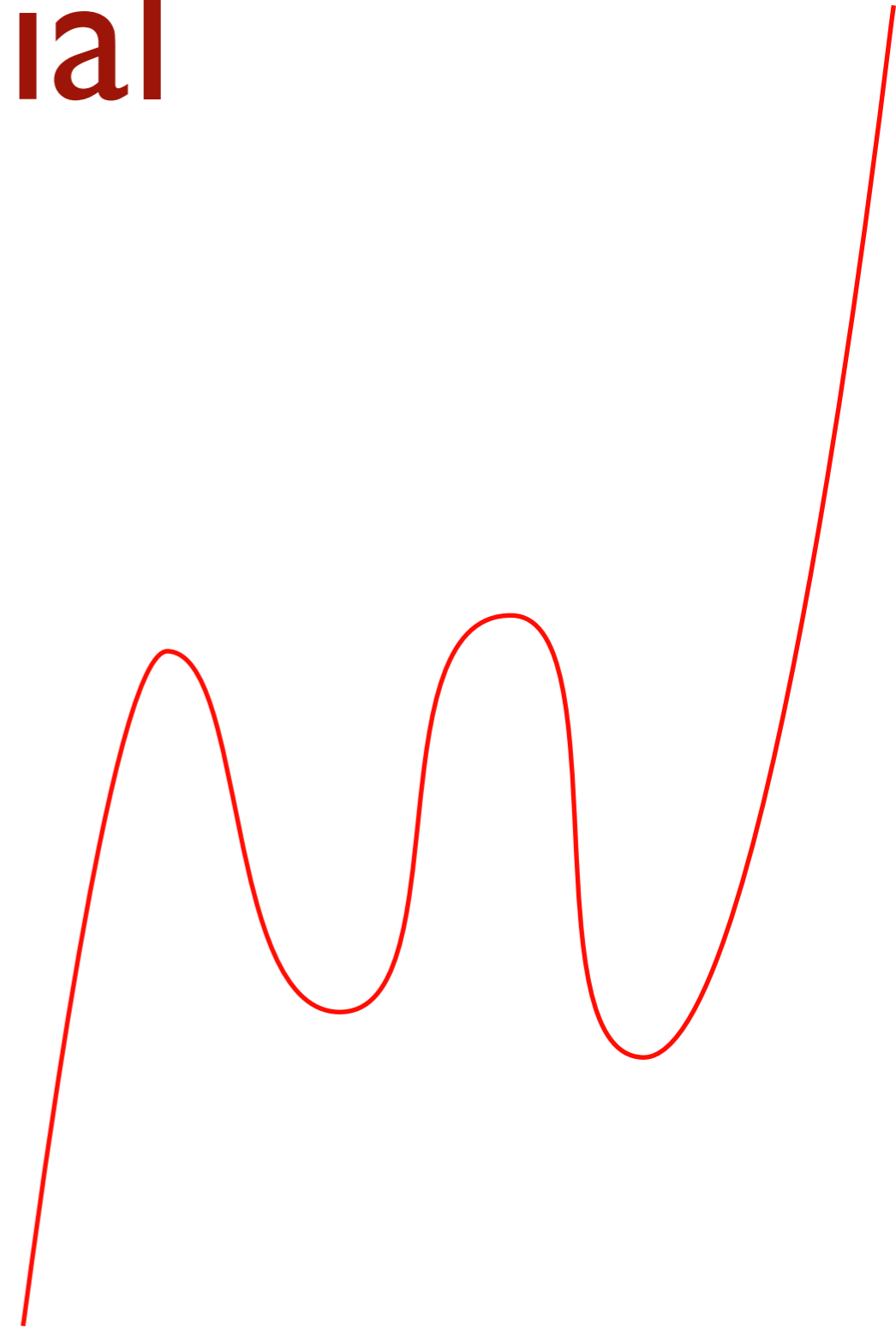
Compare exponential and power-law

- As plants grow, their growth slows because they build roots, trunks, etc... non-photosynthetic tissues
- So a flexible function that allows growth to slow is better
- Note: exponential fit appears linear on logged Y-axis



Polynomial

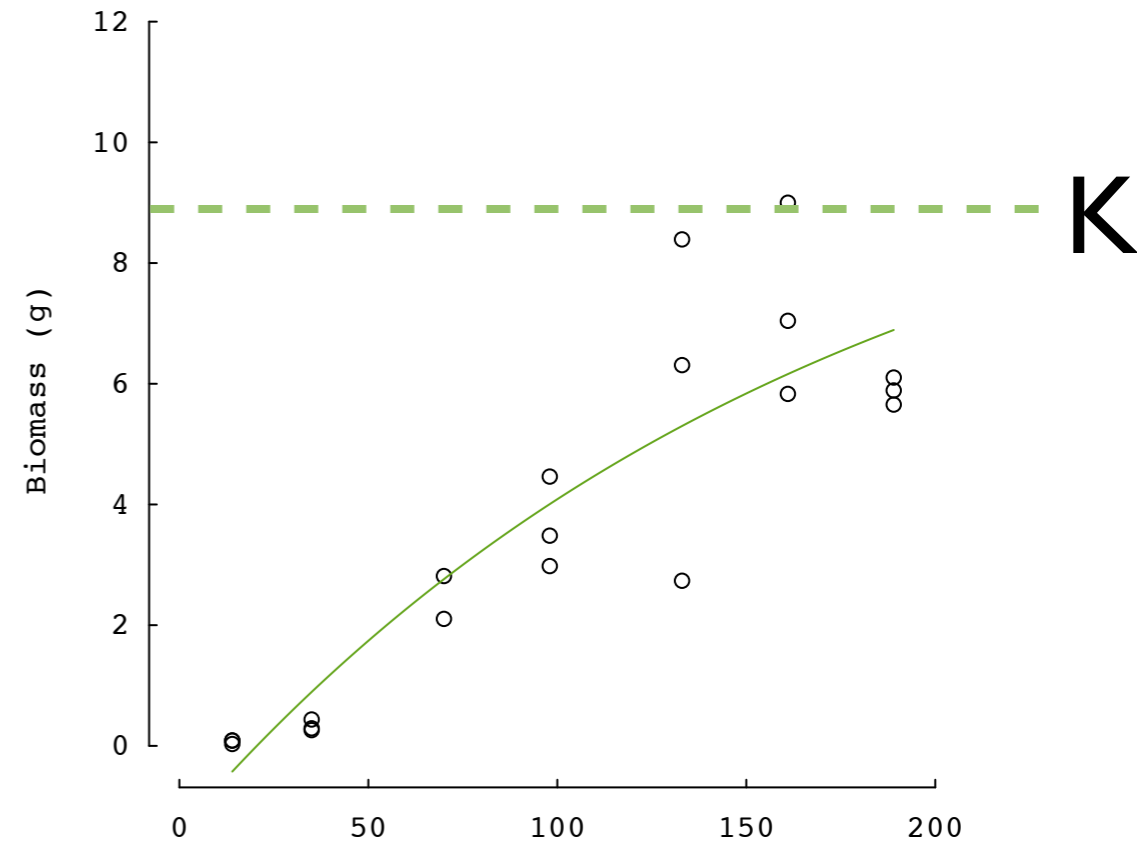
- $M = M_0 + r_1t + r_2t^2 + r_3t^3 + \dots + r_nt^n$
- Used to be popular, because easy to fit in linear framework
- Don't use it
- Easy to fit, hard to interpret
- Gives lousy predictions



Monomolecular

K: asymptotic biomass

- $dM/dt = r(K - M)$
- $M = K - e^{-rt}(K - M_0)$
- 3 parameters
- Ok for asymptotic growth, but assumes growth fastest initially, then continuous slowing

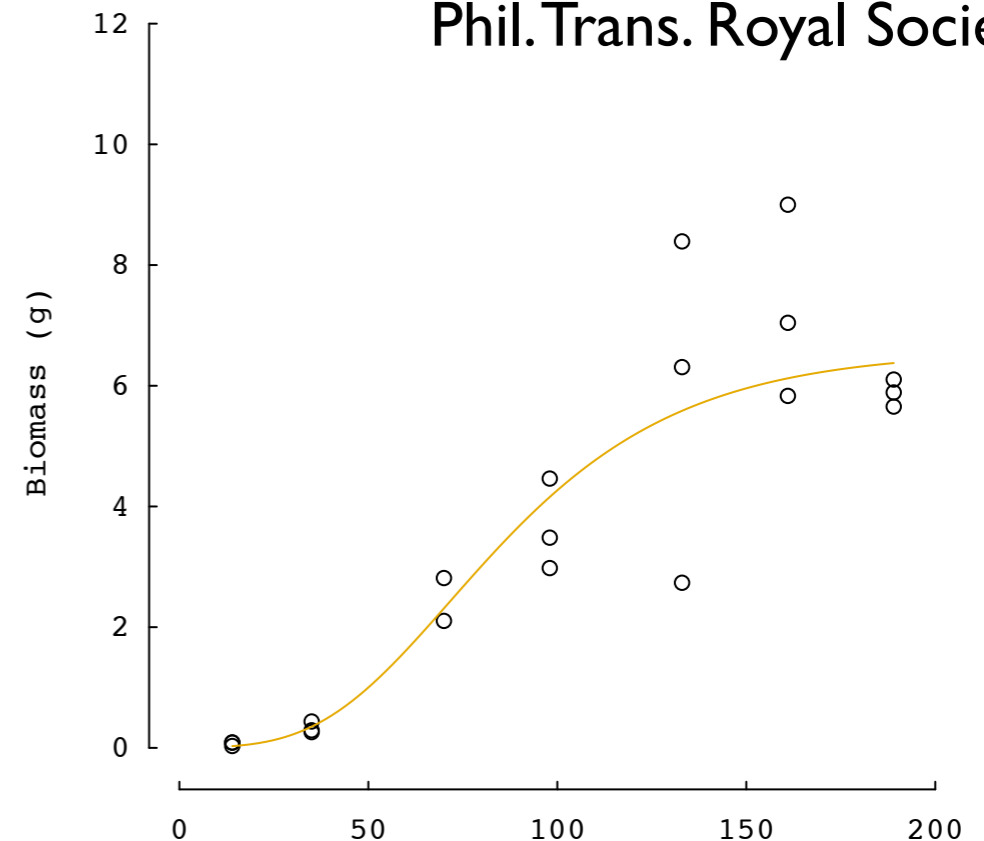


Gompertz



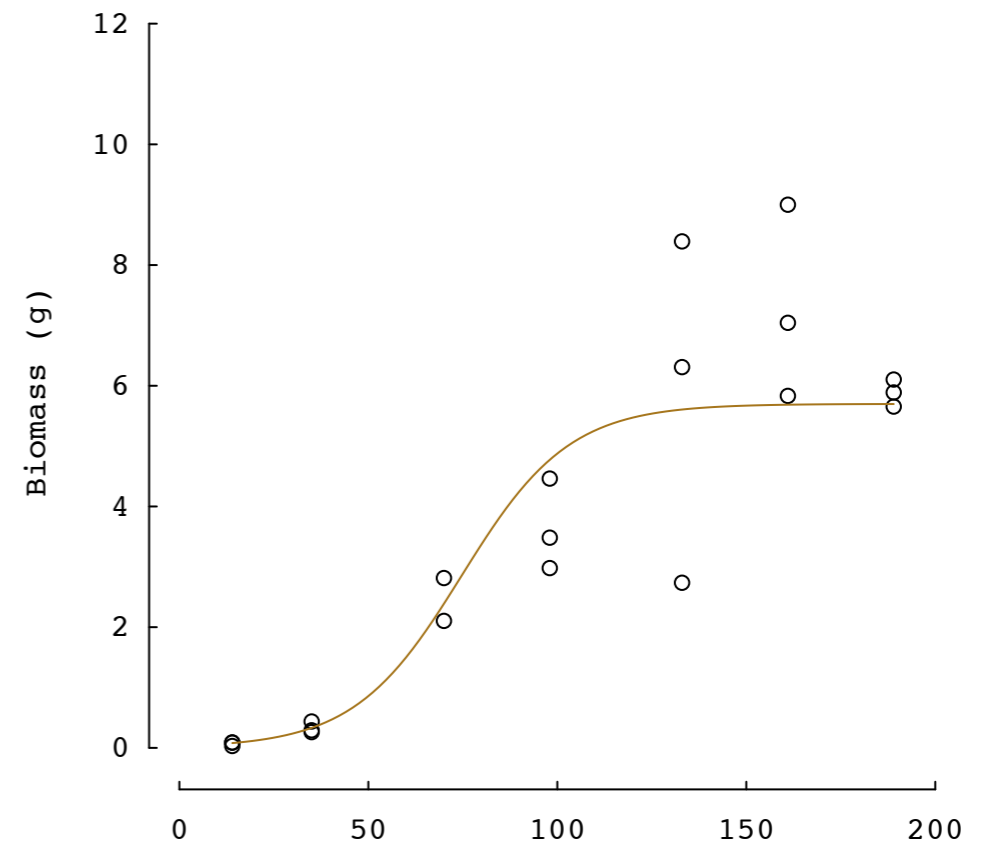
Gompertz 1825,
Phil. Trans. Royal Society

- $dM/dt = rM \left(\ln \frac{K}{M} \right)$
- $M = K \left(\frac{M_0}{K} \right)^{e^{-rt}}$
- 3 parameters
- Quite reasonable for asymptotic growth
- Inflection point (time of maximum growth) comes earlier than it does in logistic model



Logistic (three-parameter)

- $dM/dt = rM\left(1 - \frac{M}{K}\right)$
- $M = \frac{M_0 K}{M_0 + (K - M_0)e^{-rt}}$
- 3 parameters
- Reasonable, and widely used, for asymptotic growth
- Inflection point comes later than it does in Gompertz



Logistic (four-parameter)

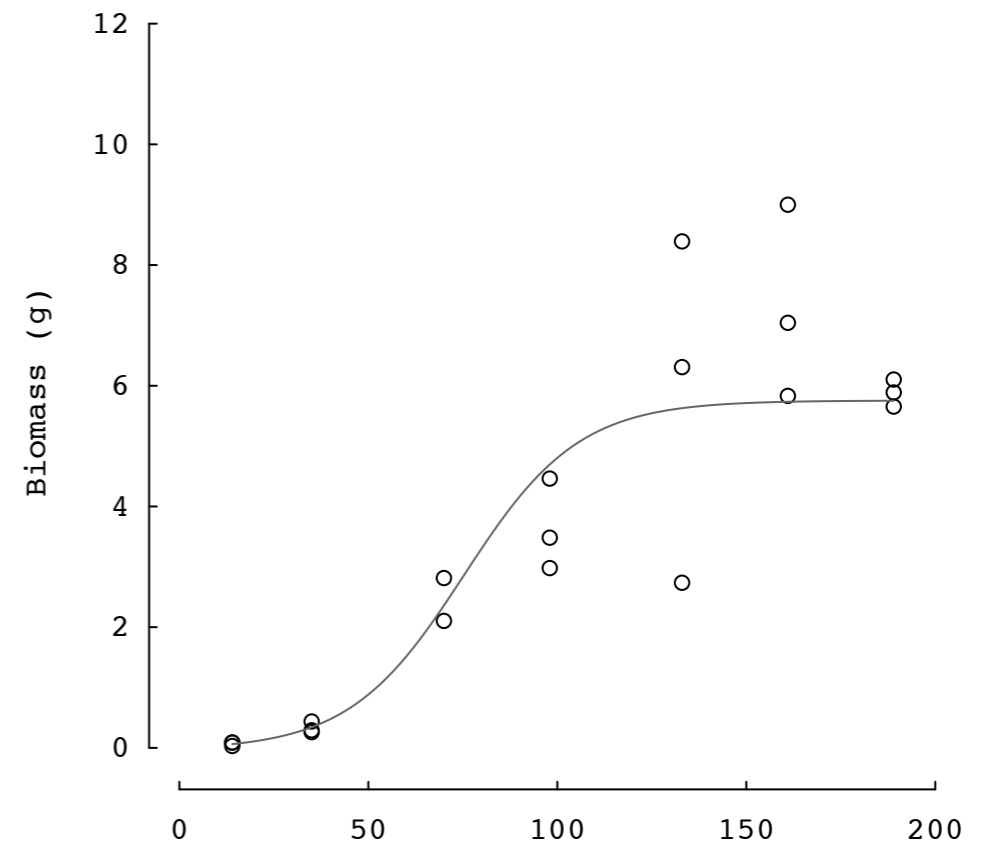
xmid: time of inflection point

- $$dM/dt = rM \left(1 - \left(\frac{M}{K} \right)^{1/xmid} \right)$$

- $$M = M_0 + \frac{K - M_0}{1 + e^{xmid - t/r}}$$

- 4 parameters

- Time of inflection point is a free parameter.
Sometimes superior fit to the 3-param version



Comparing models

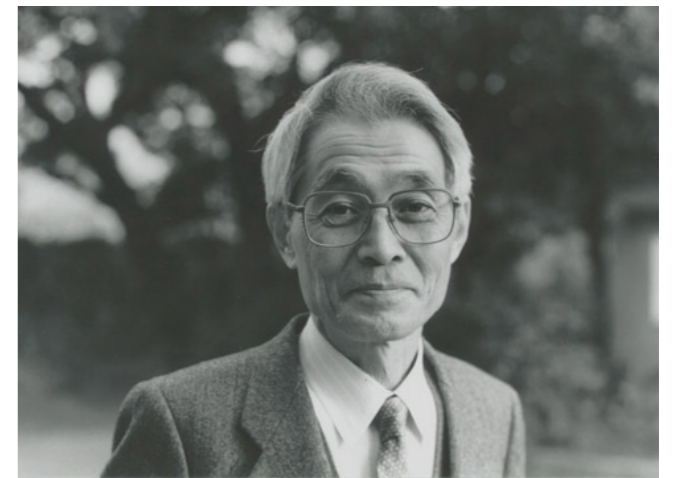
- Don't ask "Is the model right or wrong"? We ask "Do the data support a model more than a competing model"?
- Strength of evidence (support) for a model is relative:
 - to other candidate models: As models improve, support may change.
 - to data at hand: with different dataset, support may change.

Comparing models

- We penalize added complexity.
- A more complex model has to exceed a certain threshold of improvement over a simpler model.
- Added complexity usually makes a model less stable and less general
- Model selection is not about whether something is true or not but about whether we have enough information to characterize it properly.

Criteria for ranking models

- Rank by parsimony: use the simplest adequate model.
- Nested models: likelihood ratio test (χ^2 test)
- General: Akaike's Information Criterion (AIC)
 - $AIC = -2LL(\text{parameters}|\text{data}) + 2(\text{n. parameters})$
- Large datasets: Bayesian information criterion, which better balances likelihood and complexity
 - $BIC = -2LL(\text{parameters}|\text{data}) + \ln(\text{n. data}) * \text{n. parameters}$



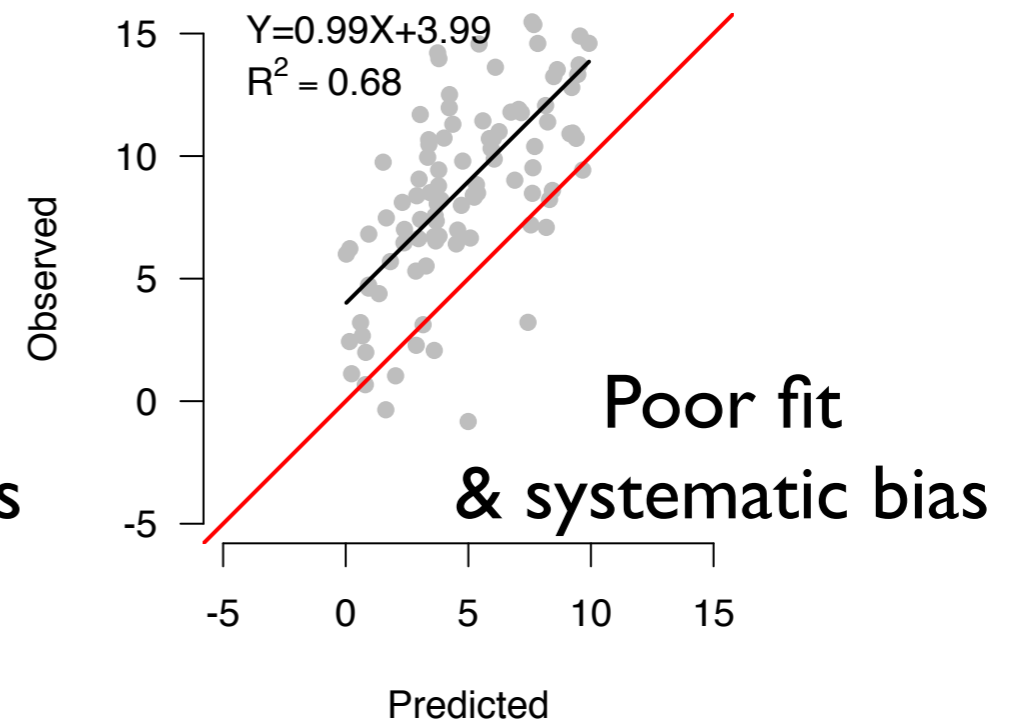
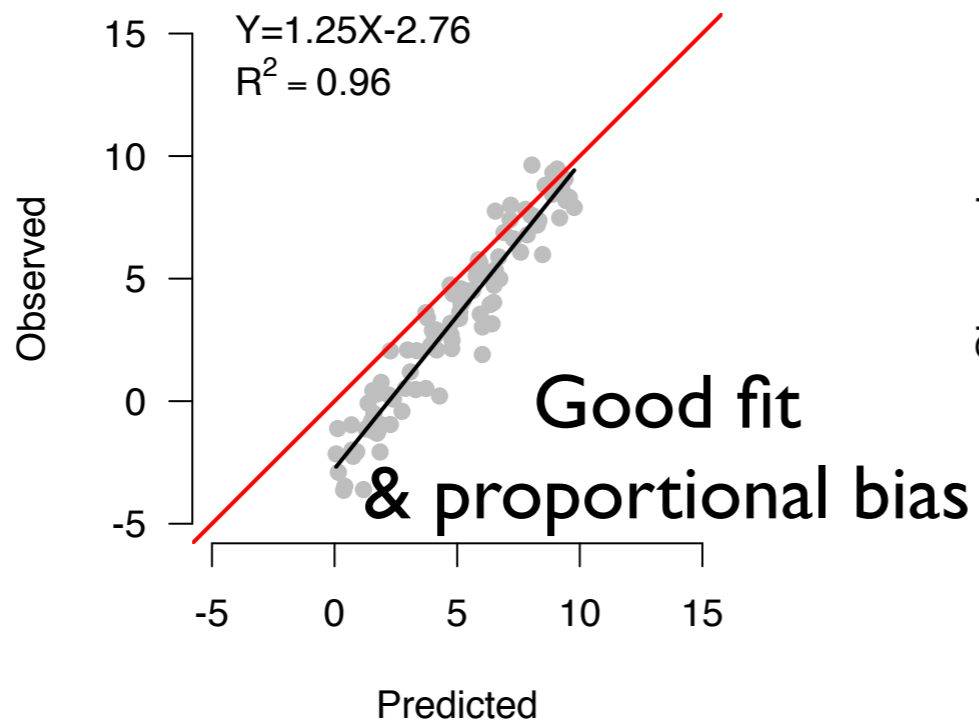
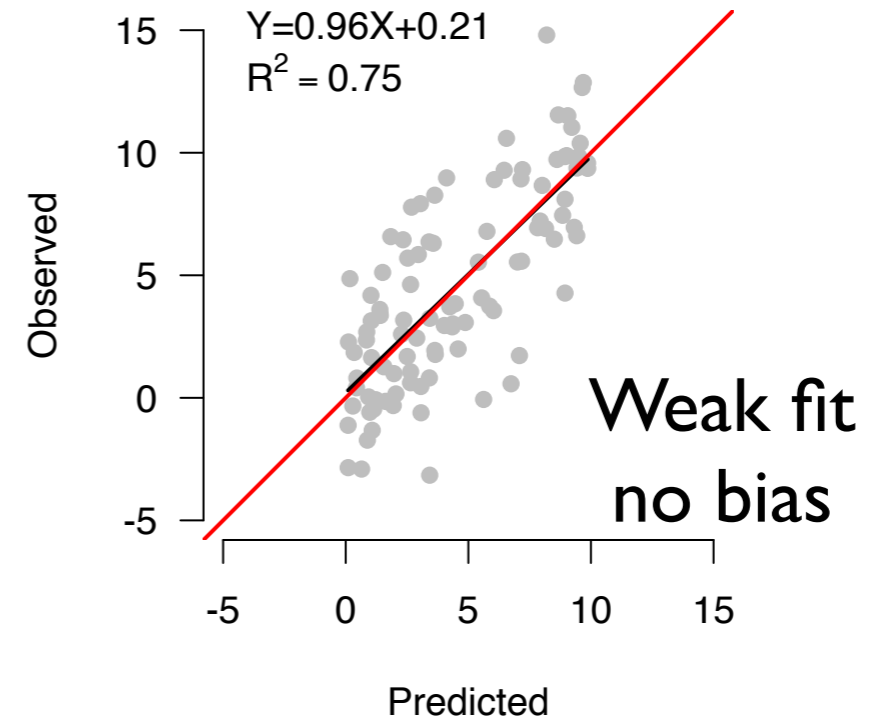
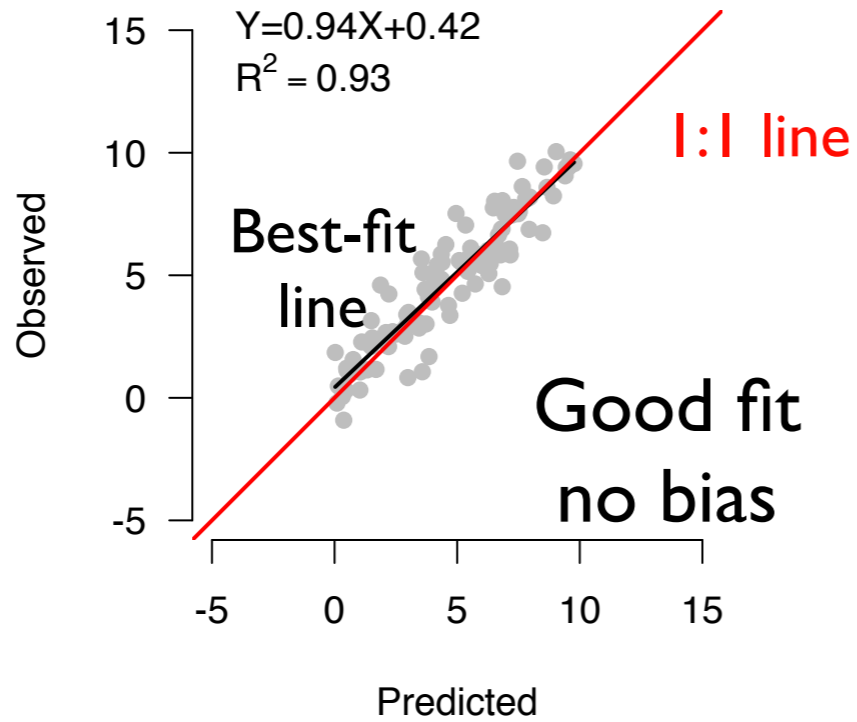
Criteria, continued

- AIC & BIC: more negative value is better
- Models within 2 AIC or BIC units are roughly equally supported by the data
- All criteria assume models are fit to **SAME** data set
 - Models not considered can't be compared.
 - Thus AIC will select the best model available, even if all the models are poor!
- It is your responsibility to assure that the set of candidate models includes well founded, realistic models.

Diagnostics

- r^2 = squared correlation (r) between observed (x) and predicted (y)
- Sensitive to data range
- Slope of relationship between observed and expected $\neq 1$ indicates proportional bias
- Intercept $\neq 0$ indicates systematic bias

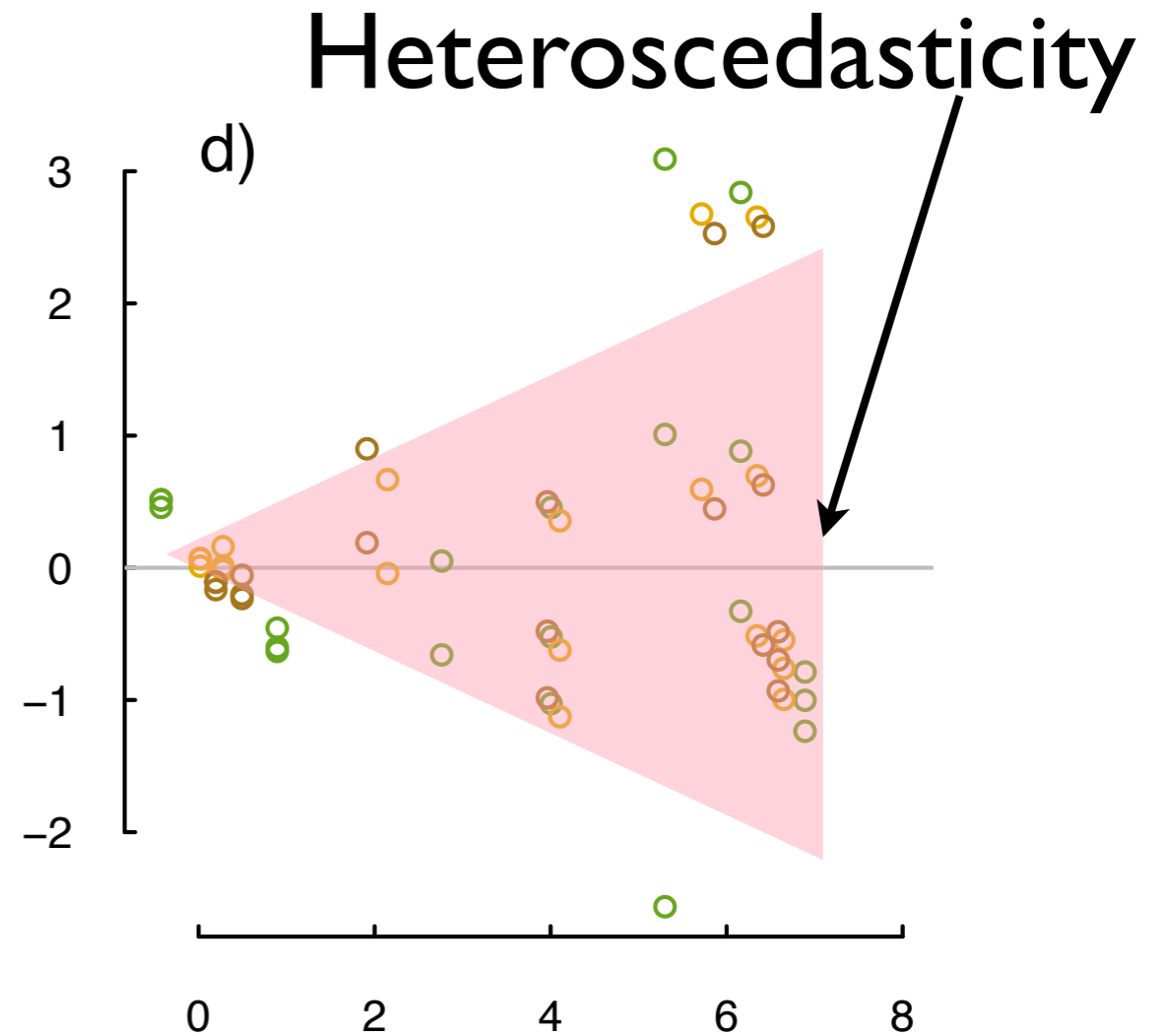
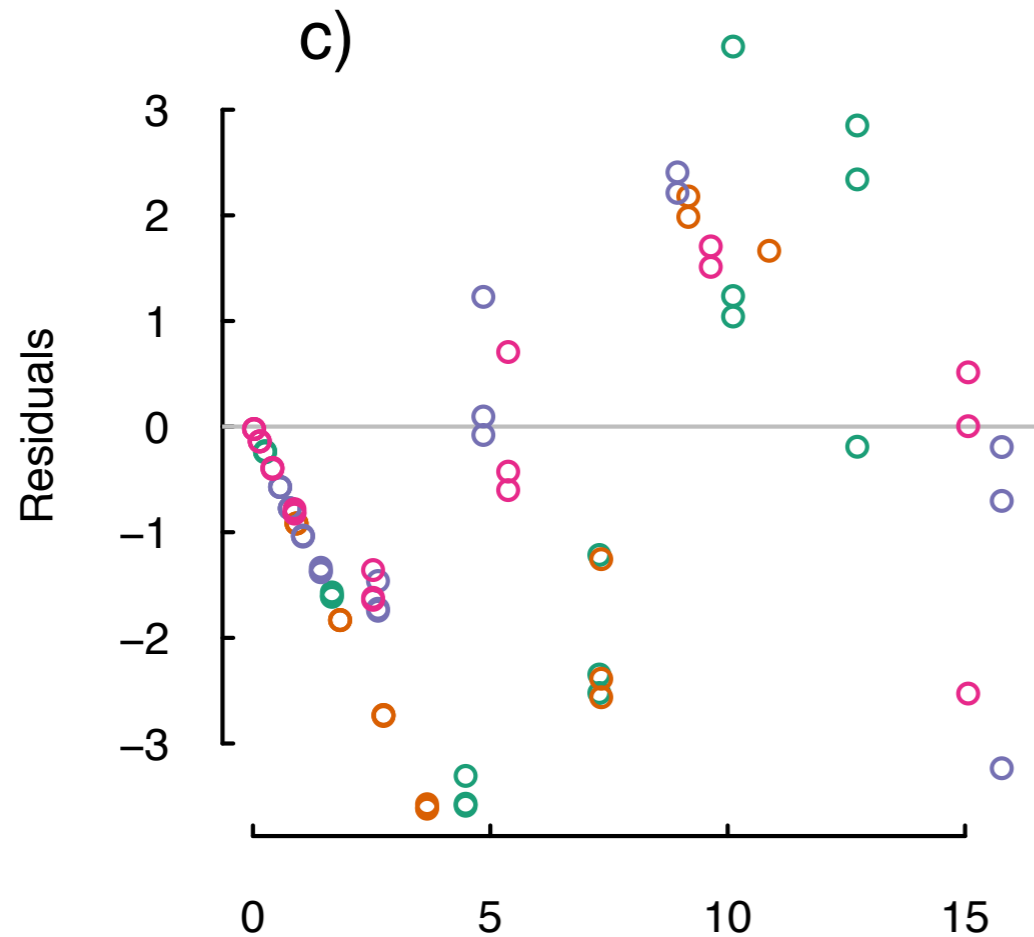
Diagnostics



Residuals

- Plot of fitted (predicted) values should show no relationship with residual (fitted - observed) values
- Particularly useful for detecting heteroscedasticity - a relationship between mean and variance of response (and a frequent issue in growth data)
- Deal with heteroscedasticity by log-transforming or variance modeling (later)

Residuals



	RMSE
Linear:	10.52
Linear no-int:	14.85
Exponential:	8.10
Power-law:	6.25

	RMSE
Monomolecular:	5.74
Logistic:	5.27
Gompertz:	5.32

Smaller root mean square error
(std dev of residuals; RMSE) is better

Considerations for model construction

- Scaling issues: Pay attention to units, scales, and conversions.
- Multiplicative functions and parameter tradeoff.
- parameter tradeoffs: covariance between estimated parameter values - frequently indicates overfitting