Model Comparison and Selection

CSCS 530 - Marisa Eisenberg

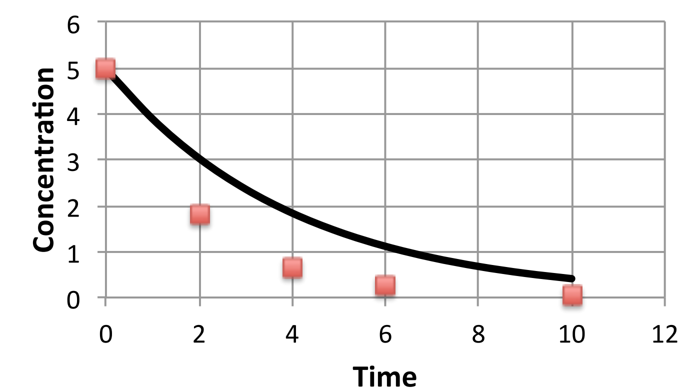
Examining the Model Fit to the Data

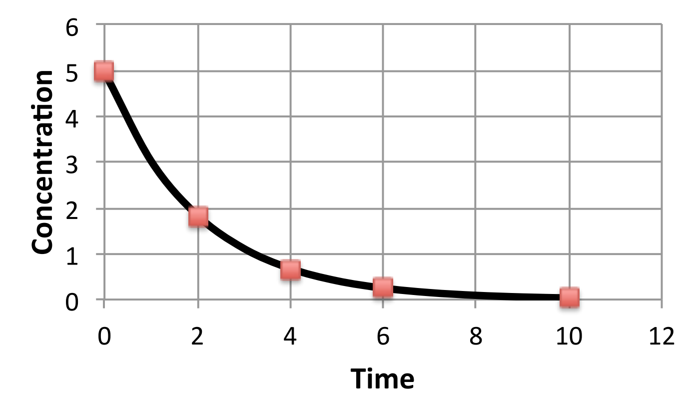
- The Eyeball Test!
- Negative log likelihood/RSS/posterior/etc.
- Parameter uncertainties & correlations - detect unidentifiability issues
- Distribution of residuals - should make sense based on data assumptions

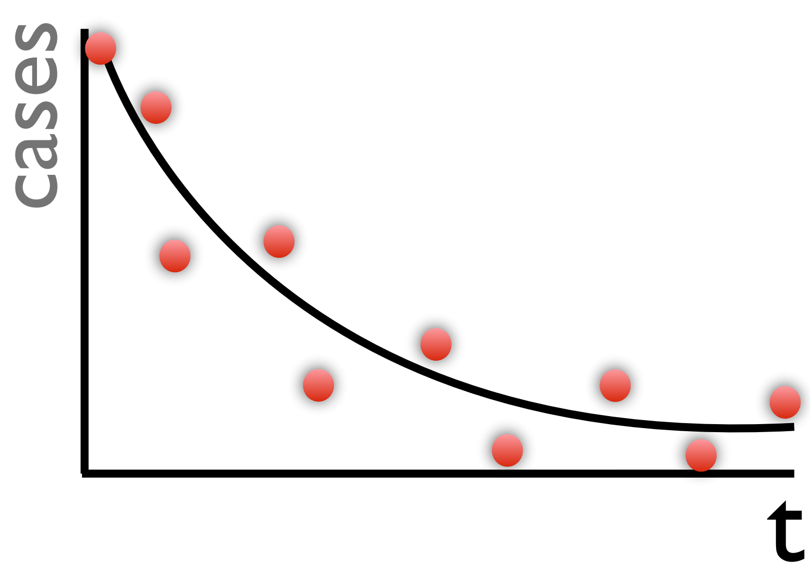
Examining the Model Fit to the Data

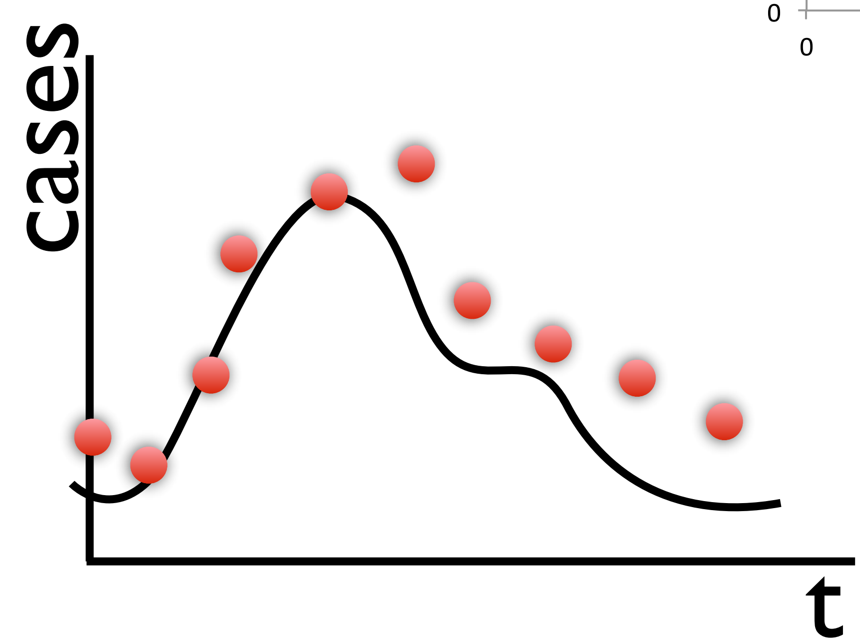
- Correlation of residuals
  (e.g serial correlation coefficient)

- Wald-Wolfowitz Runs Test

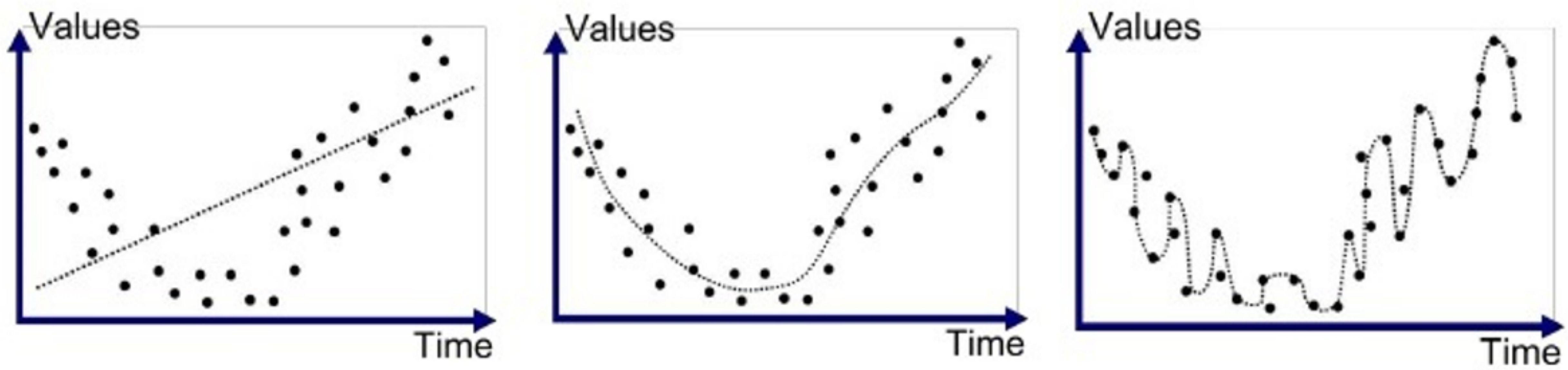








Overfitting



Underfitted   Good Fit/Robust   Overfitted

Image source: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Model misspecification

- All models are misspecified to some degree
- How to detect misspecification? (Can you always?)
- What to do about it?
- When testing models, often to compare a model vs. other candidate models
- So how to compare alternative models?

Model comparison & selection

- ‘As simple as possible, but not simpler’ (Einstein)
  - How simple to make the model? How to balance goodness-of-fit with parsimony?
- Often significant structural uncertainty—how to choose between candidate models or mechanisms?
- Think about what you are selecting for—do you want to figure out which model is more likely correct? Or do you want to pick the best predicting model?

Relationship between parameter uncertainty and model uncertainty

- Model selection/misspecification is related to structural and parameter uncertainty/sensitivity
- Uncertainty in model structure can be thought of as parameter unidentifiability in the “super-model” that includes both mechanisms
- Also related to feature selection and ideas of parameter subset selection

Model Comparison

- Many methods - F-test, likelihood ratio tests, simply comparing goodness of fit, Bayes factor, etc.
- Compare prediction accuracy with out-of-sample data
- One of the most common/popular—
- **Akaike Information Criterion** (AIC)

AIC

- More parameters - more degrees of freedom, more flexibility in the model
- So, we expect models with more parameters may be able to fit data better
- Danger of overfitting - need for parsimony
- **AIC** accounts for goodness of fit & overparameterization

AIC

$$AIC = -2\ln(\max(L)) + 2k$$
$$= 2\min(-LL) + 2k$$

- where *k* is the number of parameters, *L* is the likelihood, and *LL* is the log likelihood
- Smaller AIC is better (even if negative, i.e. more negative is better)
- AIC = -LL + penalty term for parameters

AIC

- AIC can be derived from information theory - “information loss” when using one model versus another (using the Kullback-Liebler divergence)
- One AIC has no real meaning by itself—generally need to compare AICs of competing models
- AIC comparisons also only make sense when using models fit to the same data set

AIC

- Some rough rules of thumb when comparing AICs
- Δ_i (difference in AIC) values less than 2 are often considered similarly good
- $\Delta_i \leq 6$ also may be considered
- "Δ_i values greater than 10 are sufficiently poorer than the best AIC model as to be considered implausible" (Symonds & Moussalli, Behav Ecol Sociobiol (2011) 65:13–21)

Other variations

- Many alternatives!

- BIC - stronger penalty on parameters

$$BIC = \ln(n)k - 2\ln(ML)$$

- cAIC - correction for small data sets

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

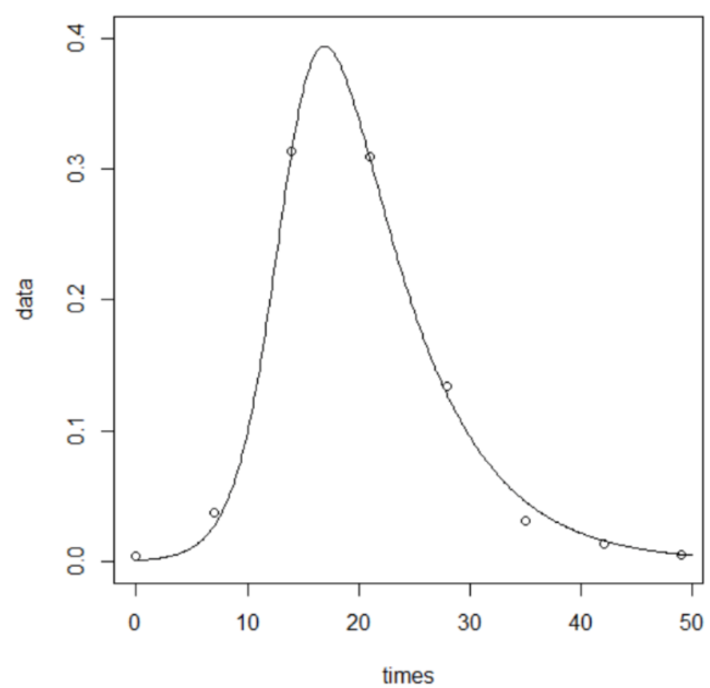
A word of caution about the AIC

- A lower AIC just that it gives you the simplest model among those tested that fits well—it may be 'more' misspecified in some ways

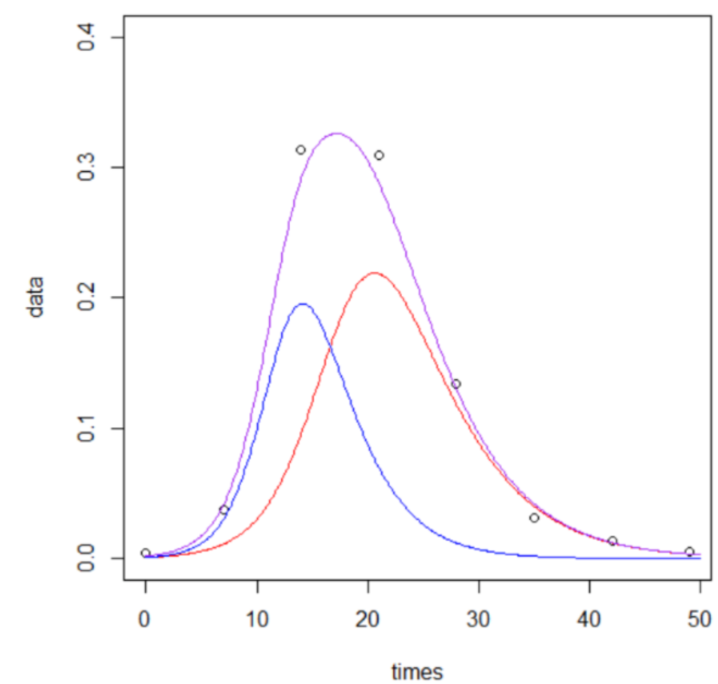
- Ex: an epidemic that spreads across 2 towns

Lower AIC because same fit but simpler

1 pop model



2 pop model



AIC and unidentifiability

- Unidentifiable models can complicate AIC evaluation
- Unidentifiability may make it harder to find optimal fit/ harder for estimation methods to converge
- What is the right degrees of freedom when unidentifiability is present?
- Cautionary tale: chronic wasting disease in deer

How to account for uncertainty in model selection?

- AIC and similar metrics just use the maximum likelihood value
- But we often have a distribution of parameters (e.g. the posterior or other uncertainty ranges)
- Goodness of fit can vary over this range, how to account for this?

Bayes factors

- Recall bayes theorem written for parameter estimation:

$$P(p|z) = \frac{P(z|p) \cdot P(p)}{P(z)}$$

- We can write this with the model explicitly included:

$$P(p|z, \mathcal{M}_1) = \frac{P(z|p, \mathcal{M}_1) \cdot P(p|\mathcal{M}_1)}{P(z|\mathcal{M}_1)}$$

Bayes factors

- Denominator: the marginal likelihood or the 'evidence' of the model—the overall probability that this model would generate the data z across all of parameter space:

$$P(z|\mathcal{M}_1) = \int_p P(z|p,\mathcal{M}_1) \cdot P(p|\mathcal{M}_1)dp$$

- For two models M_1 and M_2, the Bayes factor is given by:

$$\frac{P(z|\mathcal{M}_1)}{P(z|\mathcal{M}_2)} = \frac{\int_p P(z|p,\mathcal{M}_1) \cdot P(p|\mathcal{M}_1)dp}{\int_p P(z|p,\mathcal{M}_2) \cdot P(p|\mathcal{M}_2)dp}$$

- Reflects the balance of the evidence in favor of each candidate model vs. the other

Bayes factors

- Bayes factor accounts for uncertainty by integrating over the entire parameter space
- Note that if you use only the maximum likelihood instead of integrating the likelihood over parameter space, you get the likelihood ratio test instead of the Bayes factor

Bayes factors

TABLE 15.1: The Bayes factor scale as proposed by Jeffreys (1939). This scale should not be regarded as a hard and fast rule.

BF_{12}	**Interpretation**
> 100	Extreme evidence for $\mathcal{M}_1$.
$30 - 100$	Very strong evidence for $\mathcal{M}_1$.
$10 - 30$	Strong evidence for $\mathcal{M}_1$.
$3 - 10$	Moderate evidence for $\mathcal{M}_1$.
$1 - 3$	Anecdotal evidence for $\mathcal{M}_1$.
1	No evidence.
$\frac{1}{1} - \frac{1}{3}$	Anecdotal evidence for $\mathcal{M}_2$.
$\frac{1}{3} - \frac{1}{10}$	Moderate evidence for $\mathcal{M}_2$.
$\frac{1}{10} - \frac{1}{30}$	Strong evidence for $\mathcal{M}_2$.
$\frac{1}{30} - \frac{1}{100}$	Very strong evidence for $\mathcal{M}_2$.
$< \frac{1}{100}$	Extreme evidence for $\mathcal{M}_2$.

https://vasishth.github.io/bayescogsci/book/ch-bf.html#bayes-factor

Bayes factors

- You can actually take this whole idea one step further— we can use Bayes theorem to calculate the probability of a model (not the parameters!) given the data, if we can figure out a prior on our models:

$$P(\mathcal{M}_1|z) = \frac{P(z|\mathcal{M}_1)P(\mathcal{M}_1)}{P(z)}$$

- We can use this to calculate the posterior odds of the two models:

$$\frac{P(\mathcal{M}_1|z)}{P(\mathcal{M}_2|z)} = \frac{\frac{P(z|\mathcal{M}_1)P(\mathcal{M}_1)}{P(z)}}{\frac{P(z|\mathcal{M}_2)P(\mathcal{M}_2)}{P(z)}} = \boxed{\frac{P(z|\mathcal{M}_1)}{P(z|\mathcal{M}_2)}} \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}$$

Bayes factor

Ensemble models, comparative modeling, many-model approaches

- As we've seen, just because a model has the best AIC (or Bayes factor, etc), doesn't necessarily mean it's actually the true model

- Instead, often useful to use a multi-model approach (ensembles/comparative modeling/etc.)

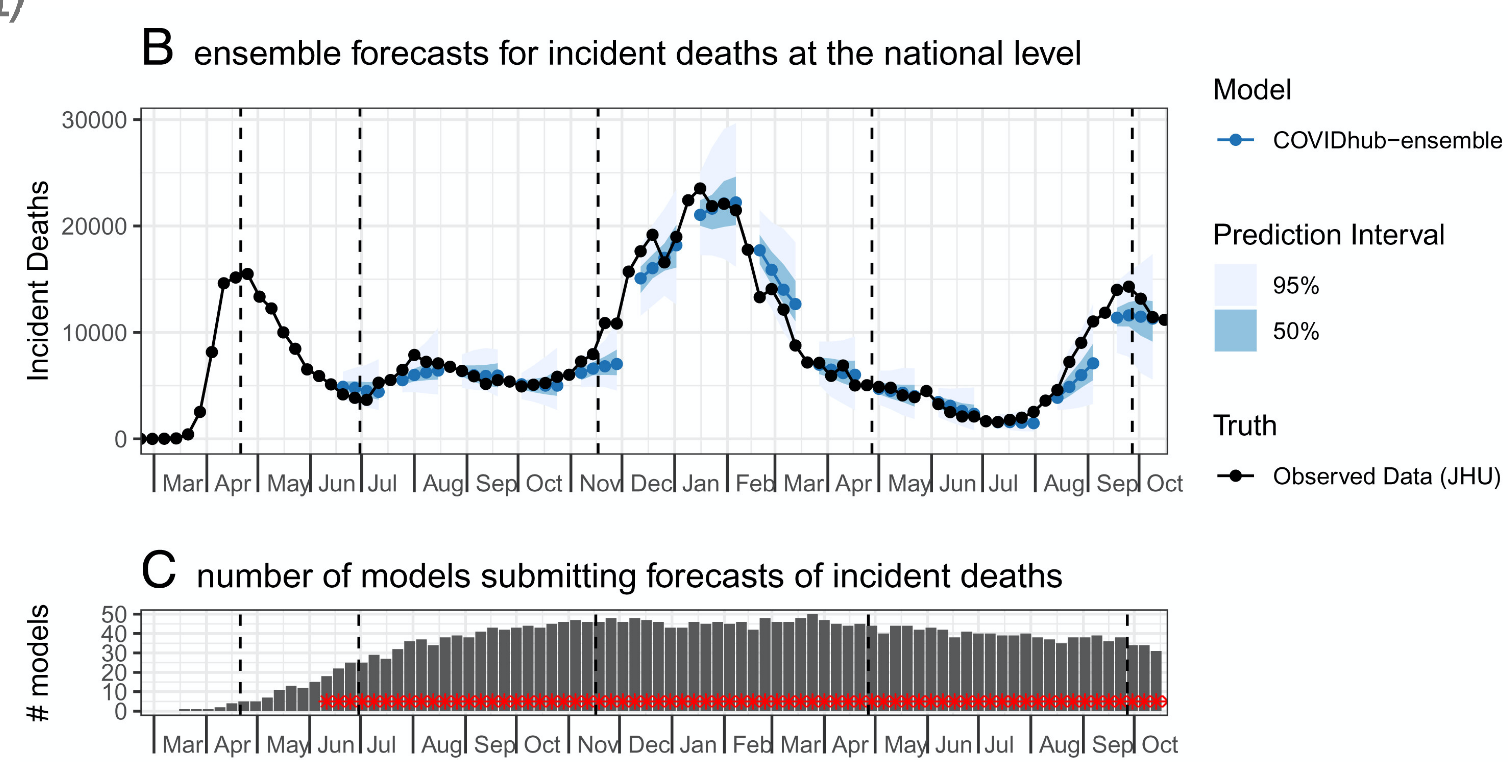
- Predictions from ensemble models often outperform individual models even when some individual models perform poorly

- Can evaluate whether inferences/predictions/conclusions are consistent across strongly-performing models and if not, figure out what data is needed to determine which model is correct

Ensemble models, comparative modeling, many-model approaches

- For example—model averaging can be done with weights based on AIC or model posterior (the probability of this being the correct model)

- Even simple averaging can be helpful, e.g. the CDC COVID-19 pandemic forecasting ensemble (https://covid19forecasthub.org)



https://www.pnas.org/doi/full/10.1073/pnas.2113561119

Ensemble models, comparative modeling, many-model approaches

- Lots of approaches to building ensembles—machine learning methods like bagging, voting methods, bucket of models (for when you have multiple problems/objectives) etc.