Model diagnostics and information criteria

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 4, 2021

Outline

Last week:

- MCMC!

Today:

- Information theory and predictive accuracy
- Scoring models to avoid overfitting

Model selection and information theory

Comparing models

One of the major problems in applied statistics is the problem of choosing between different models, or different sets of predictors.

Two distinct goals:

- Predictive accuracy: a model should produce predictions that agree with observed data
- Causal explanation: a model should inform causal relationships between observed variables

Today's tools focus on the first case. So what makes a prediction "good"? What makes models perform well or poorly?

The problem with parameters

With causal DAGs:

- We saw that adding a variable can hurt our causal estimates – if there are colliders or unobserved variables, or if we lack a trusted DAG
- What if we only care about predicting an outcome variable, and we aren't concerned with causal effects – do we then just want to add everything to the model?

Problem with parameters:

- adding more parameters to the model nearly always improves the fit to the sample (as measured by, e.g. R^2)

$$R^2 = 1 - \frac{\text{var}(\text{residuals})}{\text{var}(\text{outcomes})}$$

- however, more complex models often fit new data *worse*

Overfitting and underfitting

Two ways a model can fit badly:

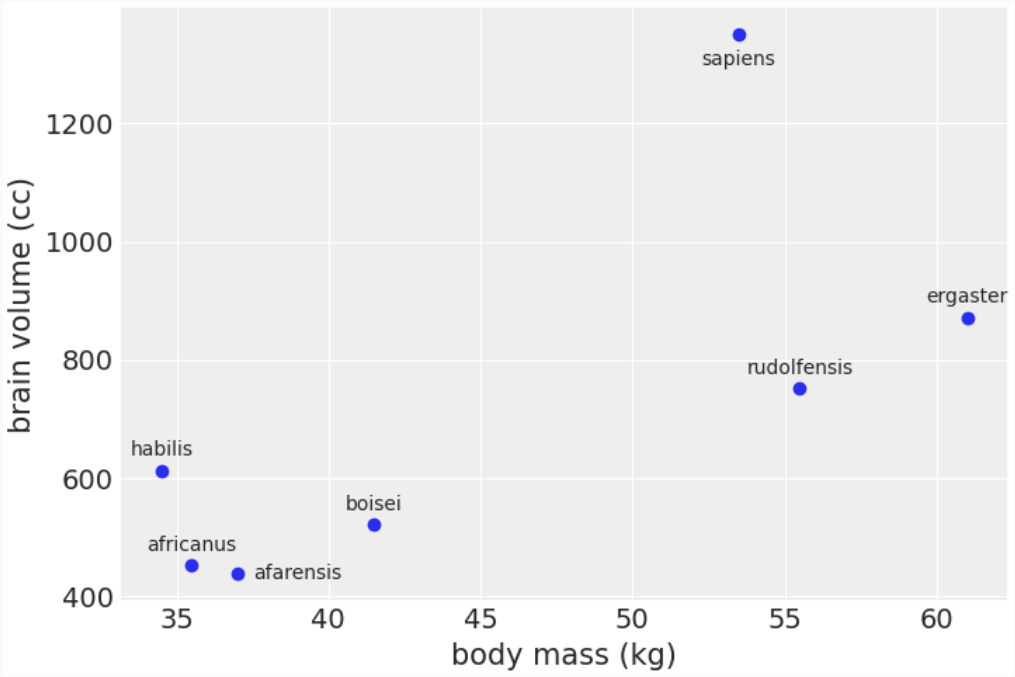
- *overfitting*: a model learned too much from the data; model fit random noise that doesn't generalize to new observations; usually, because of too many parameters
- *underfitting*: a model learned too little from the data; model is unable to fit the true relationships in the data; usually, because of insufficient parameters

Caused by too much / too little sensitivity to the sample.

Example

Example from Rethinking Ch. 7:

- Relationship between body mass and brain volume for hominid species



Example

Compare several models:

- Linear:

$$\widehat{\text{brain}} = b_0 + b_1(\text{body})$$

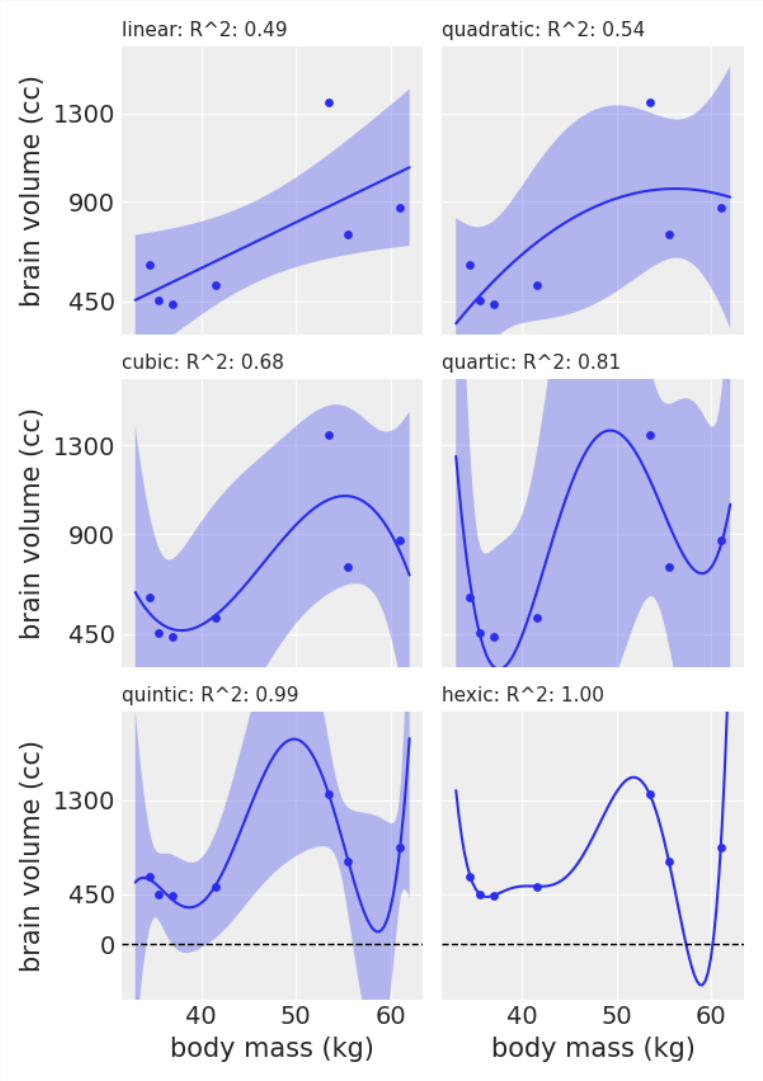
- Quadratic:

$$\widehat{\text{brain}} = b_0 + b_1(\text{body}) + b_2(\text{body})^2$$

- Cubic, etc.

Example

- Polynomial linear models of increasing degree, 1-6. Posterior mean is the dark line, 89% interval of the mean shaded.
- Increasing number of parameters always increases R^2



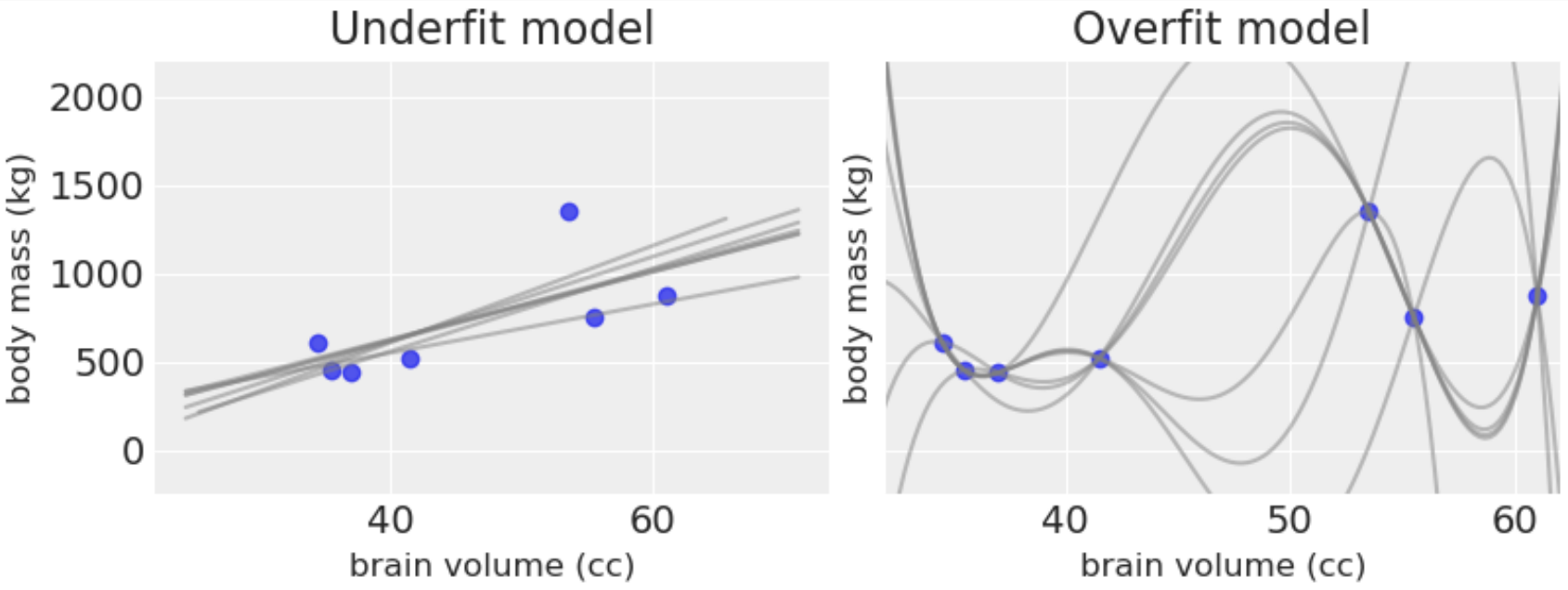
The extreme end

In the extreme end, the degree 6 polynomial has $R^2 = 1$

- Curve passes through all the points
- 7 points determine a unique degree 6 polynomial (7 parameters, 7 points)
- All our model does is "encode" the sample in a different form – doesn't learn any relationships

Sensitivity to sample

- High degree models: sensitive to specific samples
- Low degree models: insensitive to specific samples
- Below: linear and degree 4 models fit by dropping each data point once



A little information theory

Uncertainty

The main contribution of information theory to statistics is a measurable notion of uncertainty.

What is uncertainty?

- We don't know the value of future observations yet

Uncertainty

The main contribution of information theory to statistics is a measurable notion of uncertainty.

What is uncertainty?

- We don't know the value of future observations yet
- However, we know something about them (predictive distribution)
- The more "flat" the probability distribution, the more uncertainty

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?

- On any given day, what's the weather like in Seattle?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?
- On any given day, what's the weather like in Seattle?
- On any given day, what's the weather like in Chicago?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?
- On any given day, what's the weather like in Seattle?
- On any given day, what's the weather like in Chicago?

Not much uncertainty for Tucson or Seattle; a lot for Chicago

Information entropy

Measurement for uncertainty: *information entropy.* Introduced by Claude Shannon (1947) at Bell Labs; named for similarity to thermodynamic entropy.

If p is any probability distribution:

$$H(p) = -\sum_i p_i \log_2(p_i)$$

- The base 2 is a convention, and sets the "units" of uncertainty to "bits"; one bit is the amount of uncertainty in a fair coin flip or yes/no question. Natural log also used sometimes (nats); base 10 historically (bans).
- Difficult to interpret in isolation; but comparison across distributions with the same sample space is useful
- Key property: maximized by flat distributions.

Information entropy

Where did this definition come from? Three assumptions/targets:

- The measure of uncertainty should be continuous, so that small changes to the probabilities make small changes to uncertainty
- More possible outcomes should mean more uncertainty
- Uncertainty about independent observations should be additive

This determines the function up to a constant (i.e. a unit of measurement)

Entropy and encoding

History: symbol codes

- Goal: encode information (e.g. text messages) into sequences of bits (0/1)
- Assign a bit string (called a code word) to each symbol in the alphabet
- How many bits does each symbol need?

Entropy and encoding

History: symbol codes

- Goal: encode information (e.g. text messages) into sequences of bits (0/1)
- Assign a bit string (called a code word) to each symbol in the alphabet
- How many bits does each symbol need?
- Exploit symbol frequencies: assign shorter code words to more common symbols
- Theoretical minimum *average* length: entropy of the frequency distribution

Encoding and statistical models

Why should a theory of encoding and compression inform us about statistics?

- Think of a statistical model as a way of "compressing" the data
  - Have N data points, fit a model

$$\hat{y} = \alpha + \beta x$$

  - This model equation is a simpler way of summarizing the data (which show the "full" relationship between x and y)
  - Evaluating this equation at specific x decompresses the data
- Most extreme case of overfitting: if we have N parameters, the model could compress the data losslessly (what we saw with the polynomials before)

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

$$D_{KL}(p, q) = \sum p_i (\log_2 p_i - \log_2 q_i)$$

Interpretation:

- p is the true outcome distribution
- q is the model predictive distribution
- How much additional uncertainty have we introduced by using q to approximate p?

Potential for surprise

A nice interpretation of KL divergence is as a "potential for surprise." (The idea of surprise as a measurable quantity is all over information theory.)

Imagine two scenarios:

- You raise a dog in Chicago, and then you move here to Tucson
- You raise a dog in Tucson, and then you move to Chicago

Potential for surprise

The weather in Chicago is highly variable:

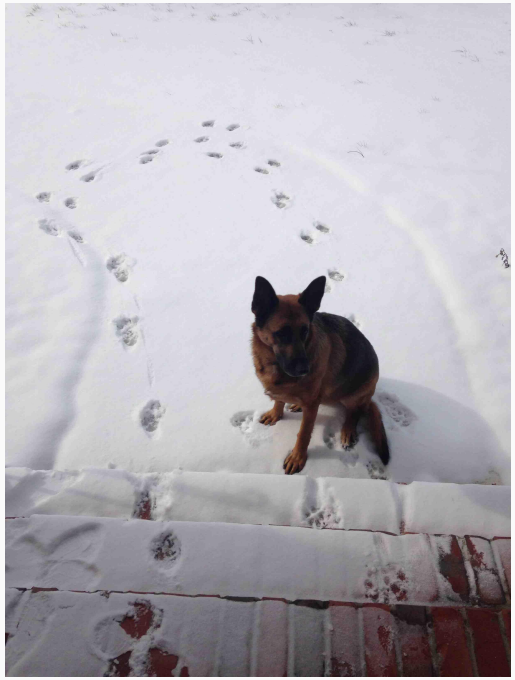
- It's hot and humid in the summer
- It's bitterly cold in the winter
- Sometimes it just oscillates between the two on a daily basis

Your Chicagoan dog has experienced all kinds of weather, and will be comfortable in the heat and the cold

Potential for surprise

As previously noted, the weather in Tucson is pretty consistent.

Your Tucsonan dog, upon moving to Chicago:



Chicago weather is much more surprising to a Tucsonan than the other way around.

Asymmetry in KL divergence

This is reflected by the asymmetry in KL divergence.

City	Tucson	Chicago
p_{hot}	0.95	0.5
p_{cold}	0.05	0.5

$$D_{KL}(\text{Tuc}, \text{Chi}) \approx 0.714$$

$$D_{KL}(\text{Chi}, \text{Tuc}) \approx 1.198$$

Asymmetry in KL divergence

Points of statistical interpretation:

- A flat model is closer to a nonflat model than vice versa
- Advantage to simpler models: they have higher entropy
    - Informs a preference for simpler models
- Maximum-entropy distributions:
    - include the weakest assumptions (example: for a given mean/variance, a normal distribution maximizes entropy)
    - are closest to other distributions

Information criteria for scoring models

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

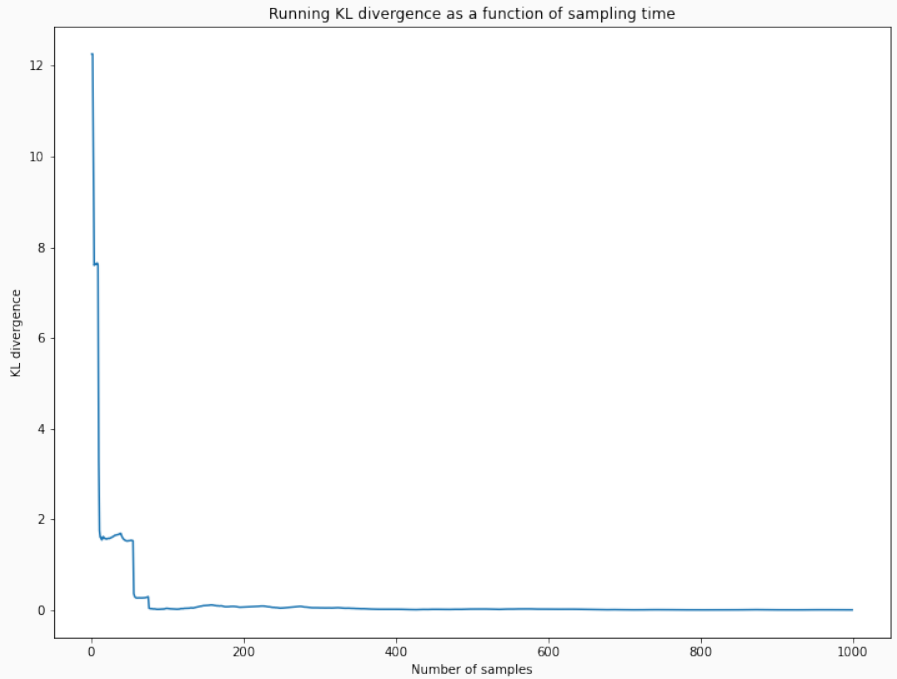
$$D_{KL}(p, q) = \sum p_i(\log_2 p_i - \log_2 q_i)$$

Interpretation:

- p is the true outcome distribution
- q is the model predictive distribution
- KL divergence measures a "distance" from the model q to the true p

KL divergence as a measure of error

Recall we used KL divergence as a way to measure convergence of our toy Metropolis algorithm:



So can we just use KL divergence to measure how close our model q is to the true outcome distribution p?

KL divergence for model comparison

We think of $D_{KL}(p, q)$ as measuring the distance from our model, q, to the truth, p.

Problem: we don't know p and never will!

But this isn't an obstacle for comparing models, because if we have two models q and r, then

$$D_{KL}(p, q) - D_{KL}(p, r) = \sum p_i(\log_2(r_i) - \log_2(q_i))$$

i.e. the $H(p)$ term drops out. We still don't know p_i, but we can estimate this from a sample of observations (because the observations are drawn from p_i)

Log score and deviance

Scoring models using log probabilities:

$$\log \text{score} \qquad S(q) = \sum_{\text{observations } i} \log(q_i)$$

Deviance:

$$D = -2S(q) = -2 \sum \log(q_i)$$

(What's the factor of -2 about? Mostly historical.)

In Bayesian world, the posterior isn't one model, it's a distribution of models – so we should average:

$$\text{lppd}(y, \theta) = \sum_i \log \left(\frac{1}{S} \sum_s p(y_i | \theta_s) \right)$$

(log pointwise predictive density). What's S? A sample from the posterior (e.g., MCMC)

Out-of-sample prediction error

lppd isn't enough on its own, though, because it only looks inside the sample

- Adding parameters nearly always improves fit within the sample
- Eventually, adding parameters reduces accuracy out of the sample (overfitting)
- How can we predict out-of-sample prediction accuracy?
  - Cross-validation
  - Information criteria

Goals

Since we use lppd to estimate the fit of our model, our goal with all of these tools is to estimate what our lppd will be on new data.

In other words, ultimately we are trying to estimate some form of:

$$\text{elpd} = \mathbb{E}(\log p(\tilde{y}|y))$$

the expected log predictive density of a new data point.

In some cases (e.g. with AIC) we'll calculate the expected deviance of a new data set of the same size (can work either way).

Overfitting in action

To demonstrate overfitting, we'll consider a few models fit to fake data. True data-generating process:

$$y_i \sim \text{Normal}(\mu_i, 1)$$
$$\mu_i = 0.15x_1 - 0.40x_2$$

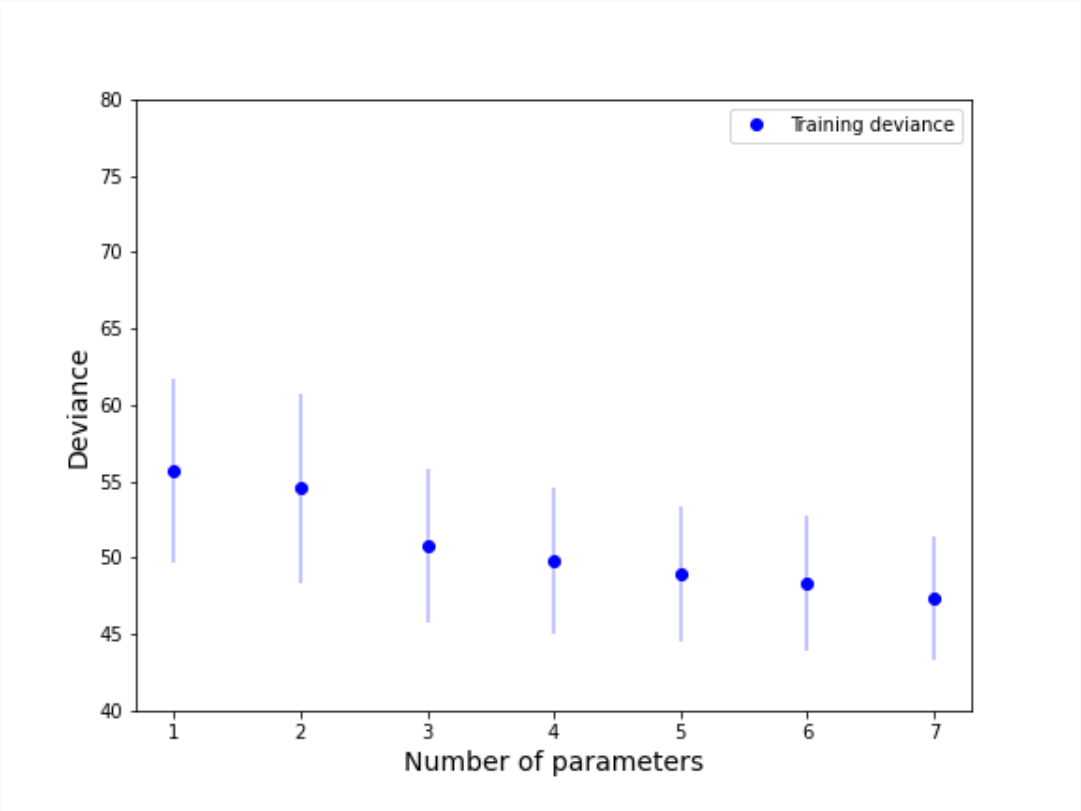
We'll fit models with the same likelihood and

$$\mu_i = \alpha$$
$$\mu_i = \alpha + \beta_1 x_1$$
$$\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2$$
$$\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$\dots$$

$x_i, i > 2$ are just noise

Overfitting in action

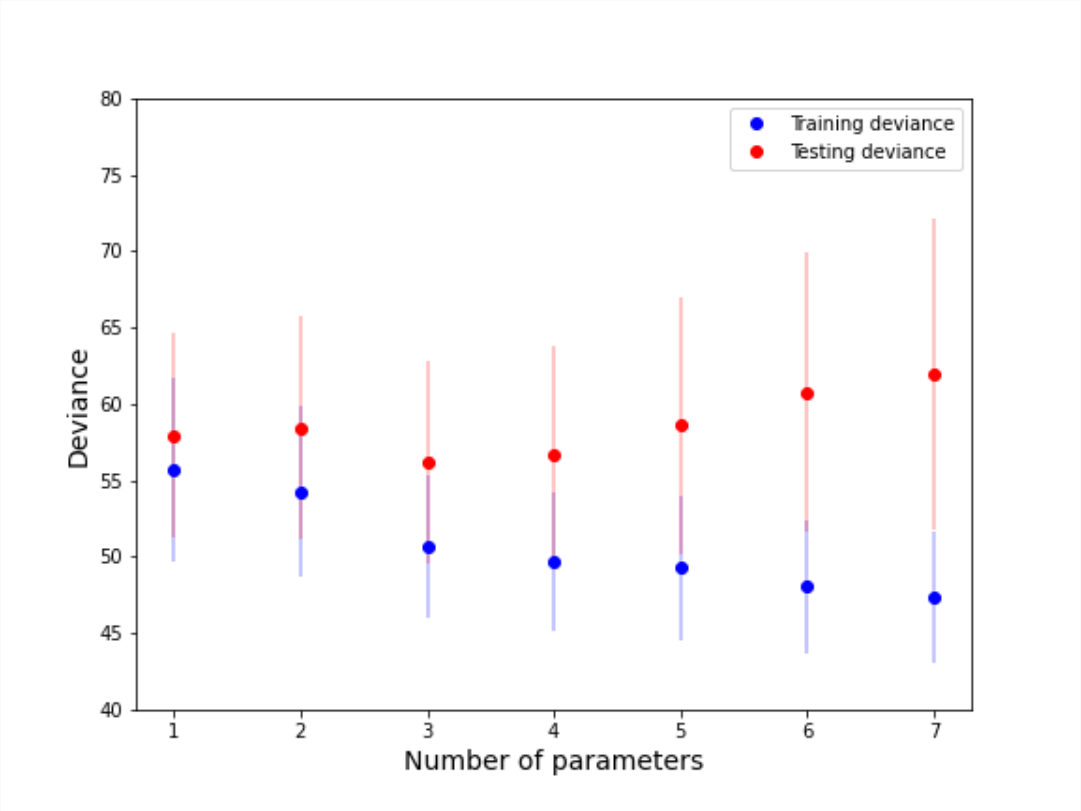
On the training set:



Remember: past 3 parameters, the predictors have no relationship to y in the true data generating process

Overfitting in action

Add in the testing set:



As expected, the additional parameters just make matters worse.

Akaike information criterion (AIC)

AIC: named for Akaike (but he called it "an information criterion"). Attempts to estimate the out-of-sample deviance.

Assuming a point estimate $\hat{\theta}$ for model parameters, calculate the log score and apply a penalty to correct for overfitting:

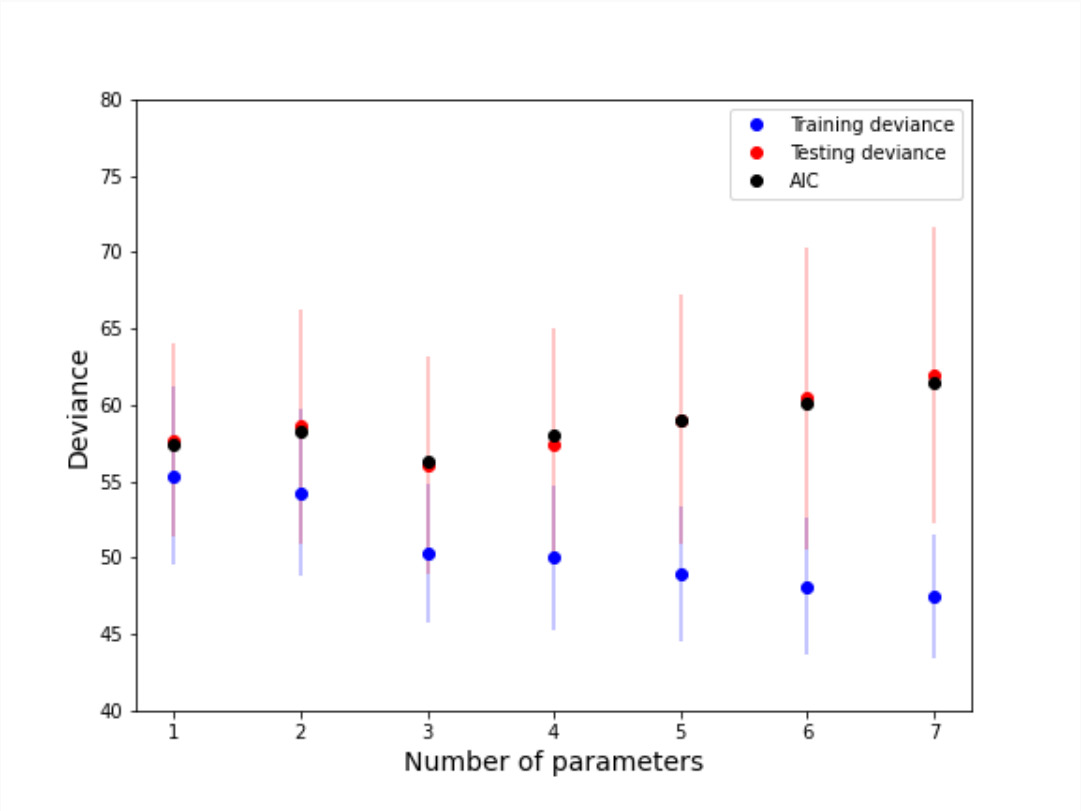
$$AIC = D_{\text{train}} + 2k$$

Here D_train is the training deviance: log predictive density evaluated at the maximum likelihood estimate, multiplied by -2.

k is the number of parameters. Assumes Gaussian posterior. Where it comes from: Taylor expansion around the posterior mode.

Overfitting in action

Adding the AIC:



We see that for this model, the AIC is a good estimate of the out-of-sample deviance. So, it is reasonable to choose the model with the smallest AIC.

Summary

Today:

- information theory / entropy
- A first information criterion: AIC

Next time:

- Better criteria for estimating out-of-sample error:
  - WAIC – a refinement of AIC
  - PSIS (or LOO-CV) – another estimate of out-of-sample error