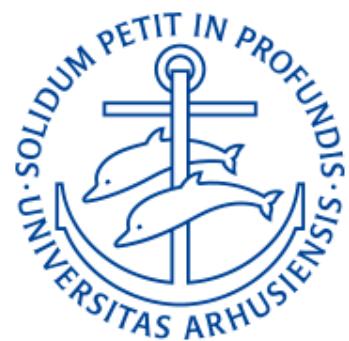


Methods 4 - 7

Chris Mathys



BSc Programme in Cognitive Science

Spring 2022

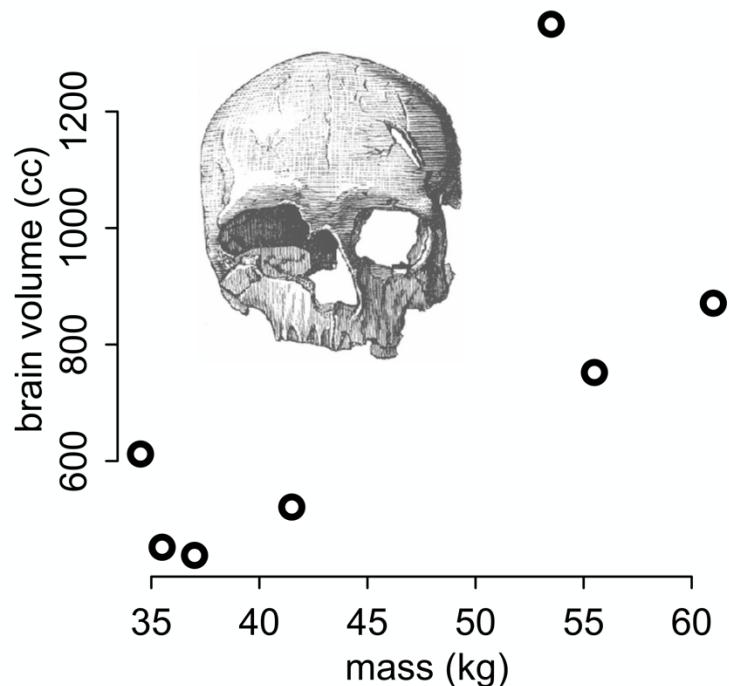
Problems of Prediction

What function describes these points?
(fitting, compression)

What function explains these points?
(causal inference)

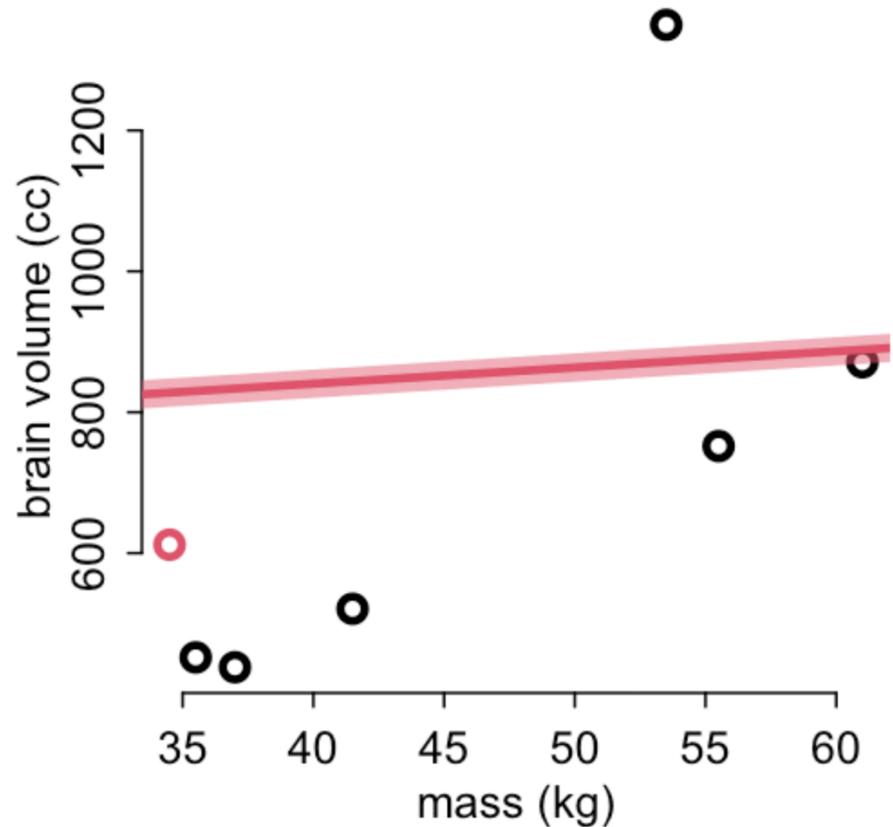
What would happen if we changed a point's mass? (intervention)

What is the next observation from the same process? (prediction)



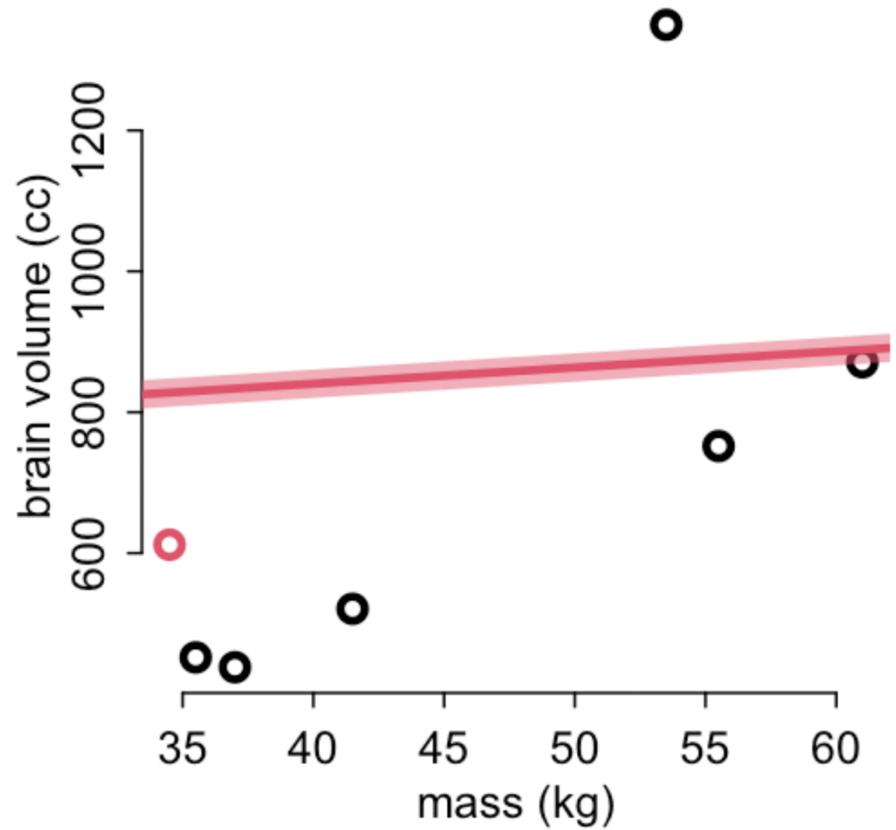
Leave-one-out cross-validation

- (1) Drop one point
- (2) Fit line to remaining
- (3) Predict dropped point
- (4) Repeat (1) with next point
- (5) Score is error on dropped



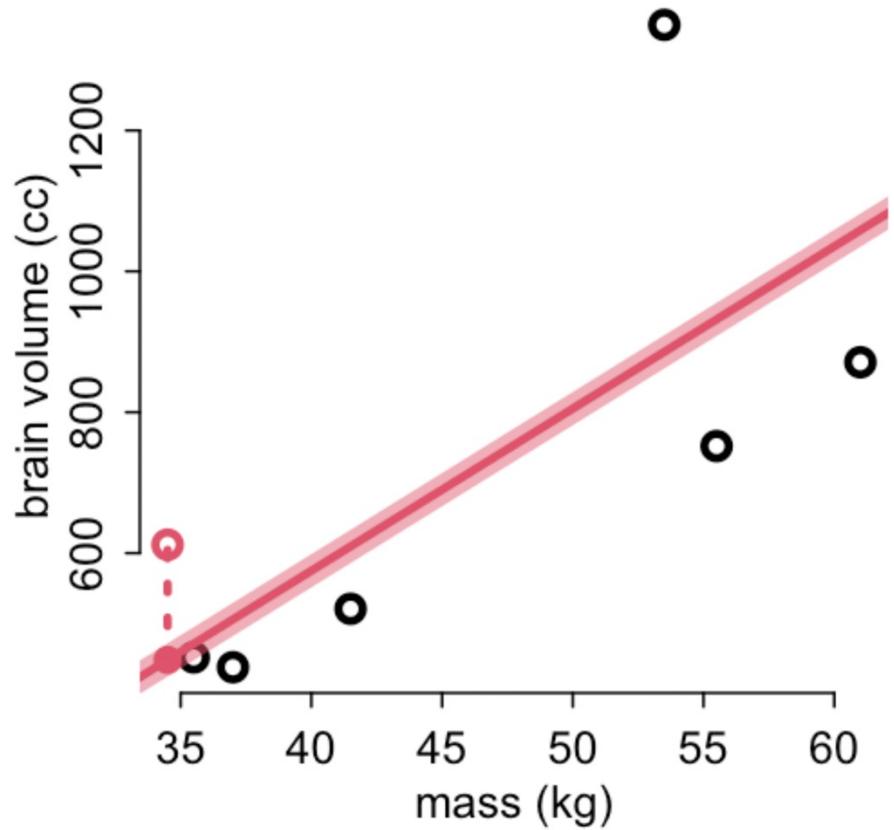
Leave-one-out cross-validation

- (1) Drop one point
- (2) Fit line to remaining
- (3) Predict dropped point
- (4) Repeat (1) with next point
- (5) Score is error on dropped



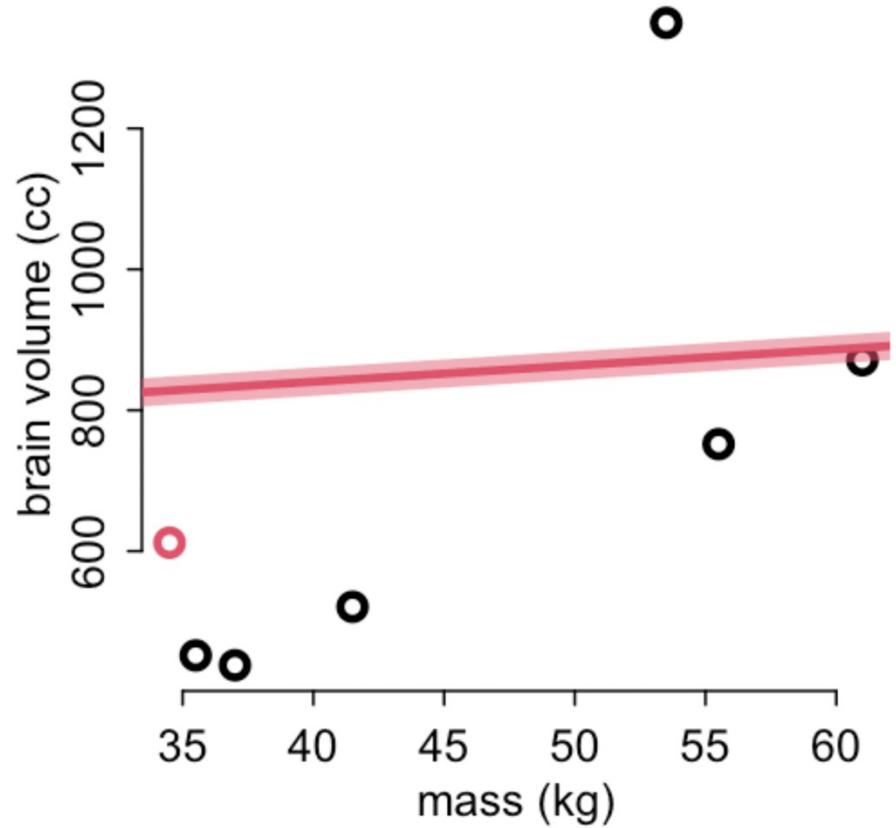
Leave-one-out cross-validation

- (1) Drop one point
- (2) Fit line to remaining
- (3) Predict dropped point
- (4) Repeat (1) with next point
- (5) Score is error on dropped



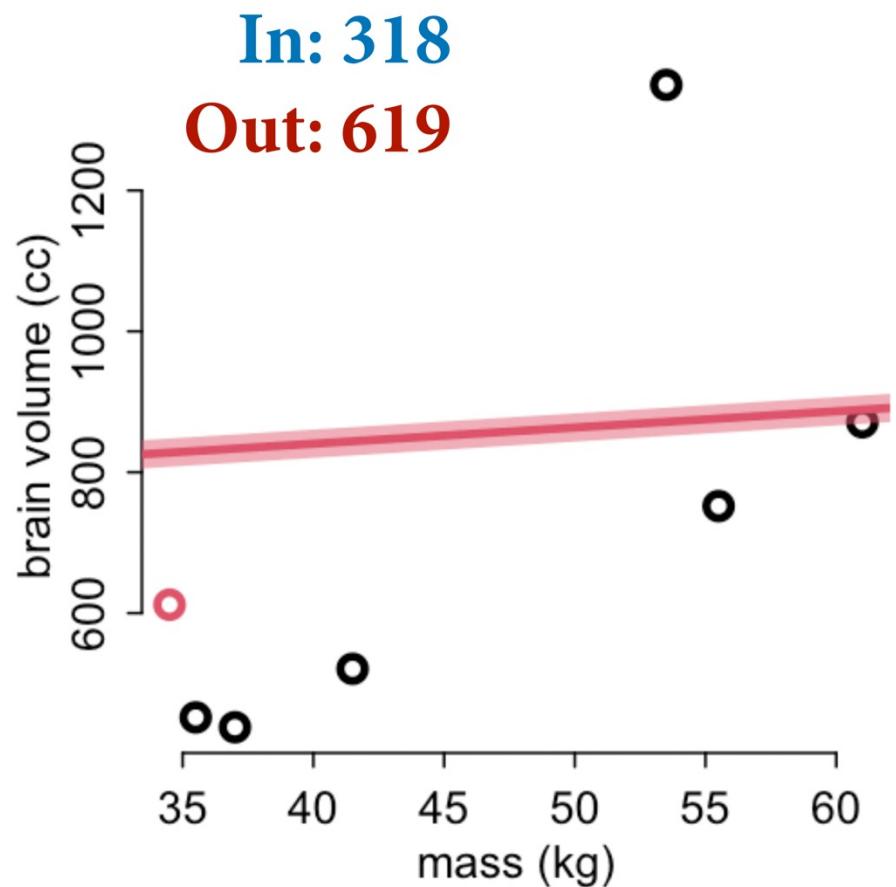
Leave-one-out cross-validation

- (1) Drop one point
- (2) Fit line to remaining
- (3) Predict dropped point
- (4) Repeat (1) with next point
- (5) Score is error on dropped



Leave-one-out cross-validation

- (1) Drop one point
- (2) Fit line to remaining
- (3) Predict dropped point
- (4) Repeat (1) with next point
- (5) Score is error on dropped

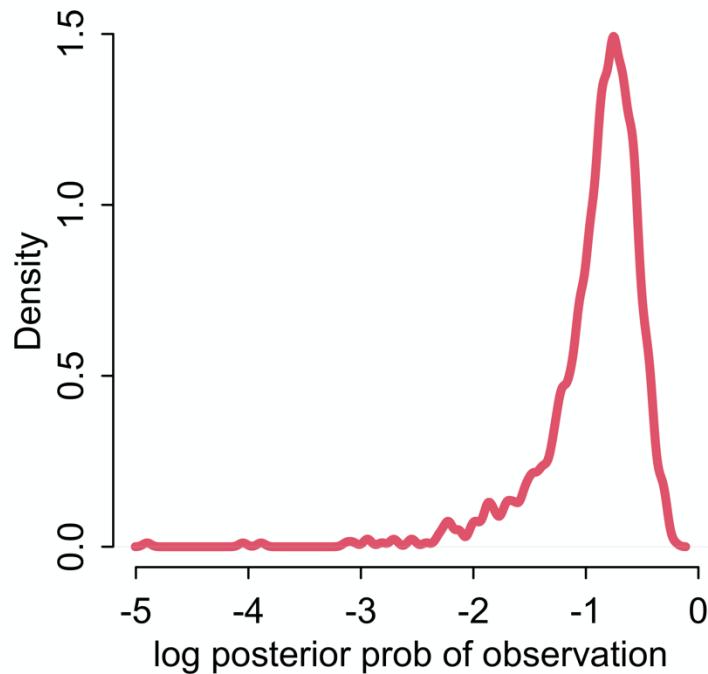


Bayesian Cross-Validation

We use the entire posterior, not just a point prediction

Cross-validation score is:

$$\text{lppd}_{\text{CV}} = \sum_{i=1}^N \frac{1}{S} \sum_{s=1}^S \log \Pr(y_i | \theta_{-i,s})$$



Bayesian Cross-Validation

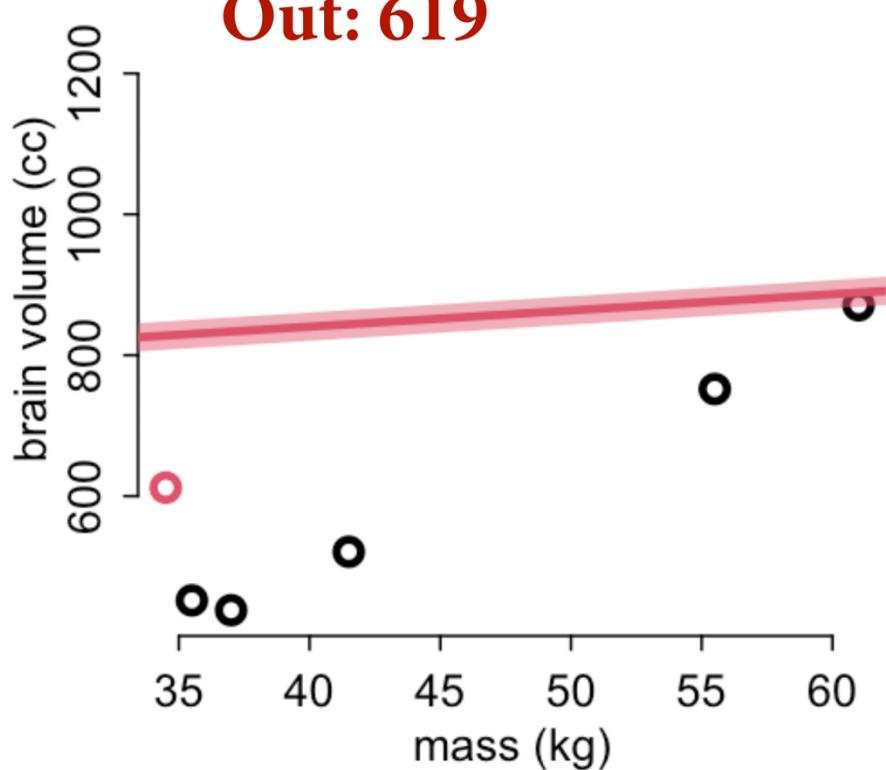
$$\text{lppd}_{\text{CV}} = \sum_{i=1}^N \frac{1}{S} \sum_{s=1}^S \underline{\log \Pr(y_i | \theta_{-i,s})}$$

Diagram illustrating the components of the Bayesian Cross-Validation formula:

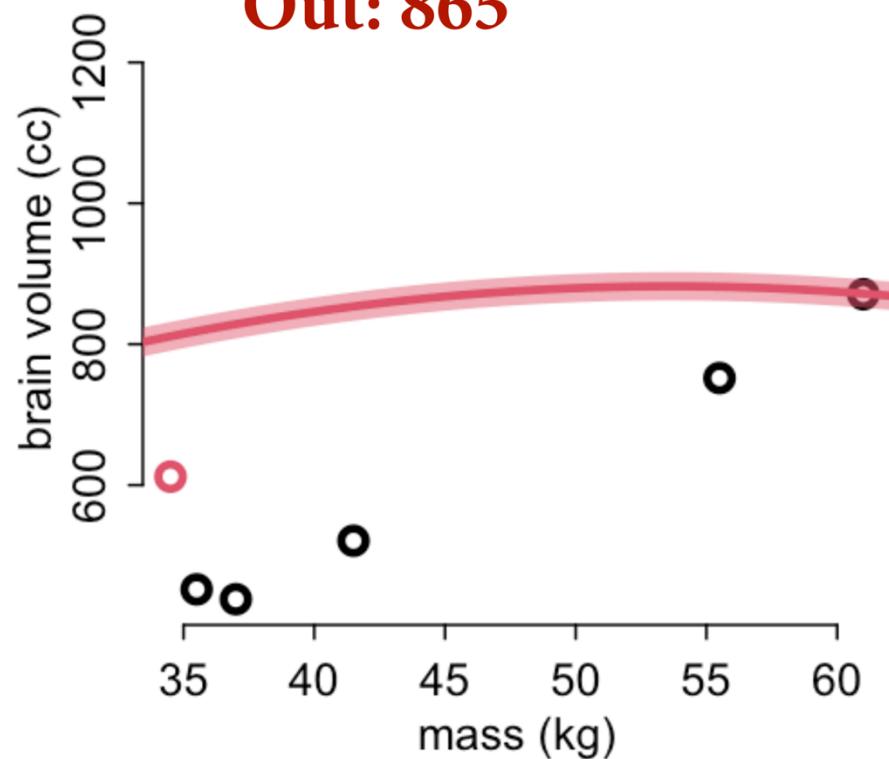
- N data points* (red text) points to the term $\sum_{i=1}^N$.
- S samples from posterior* (red text) points to the term $\sum_{s=1}^S$.
- log pointwise predictive density* (red text) points to the term $\log \Pr(y_i | \theta_{-i,s})$.
- average log probability for point i* (red text) points to the term $\frac{1}{S} \sum_{s=1}^S$.
- log probability of each point i, computed with posterior that omits point i* (red text) points to the term $\log \Pr(y_i | \theta_{-i,s})$.

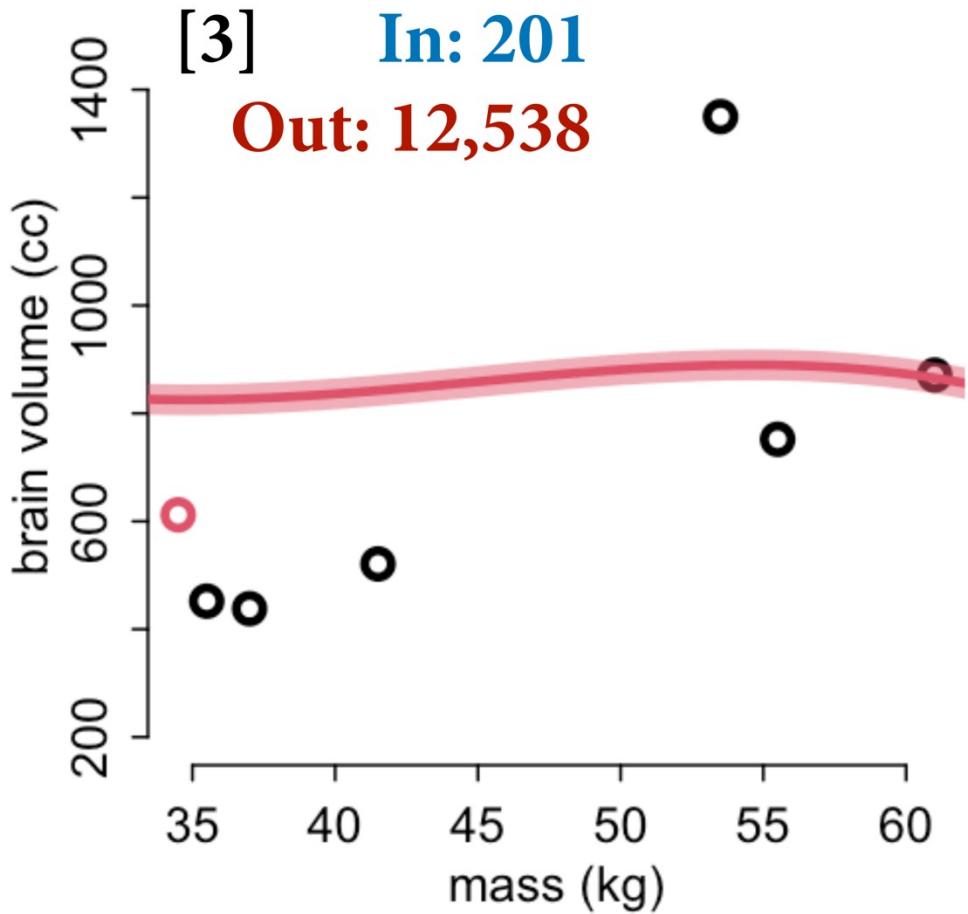
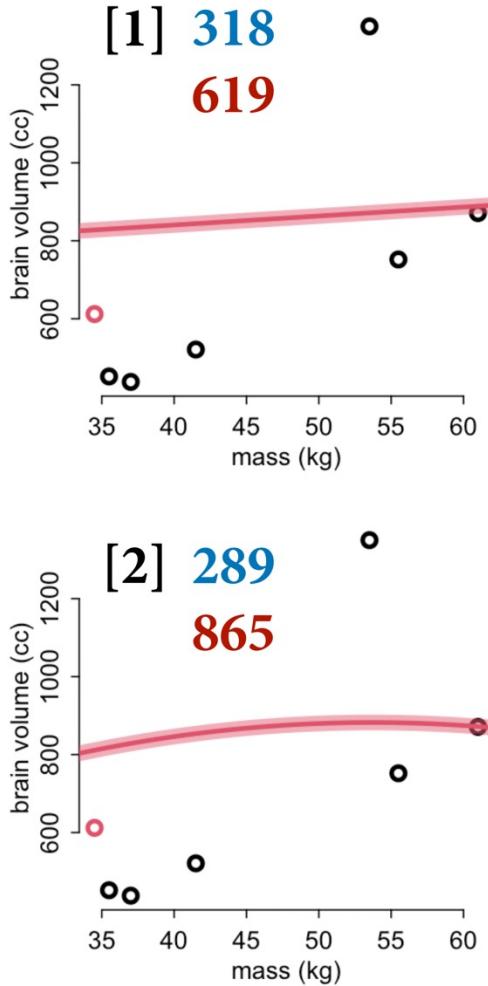
Pages 210 and 218

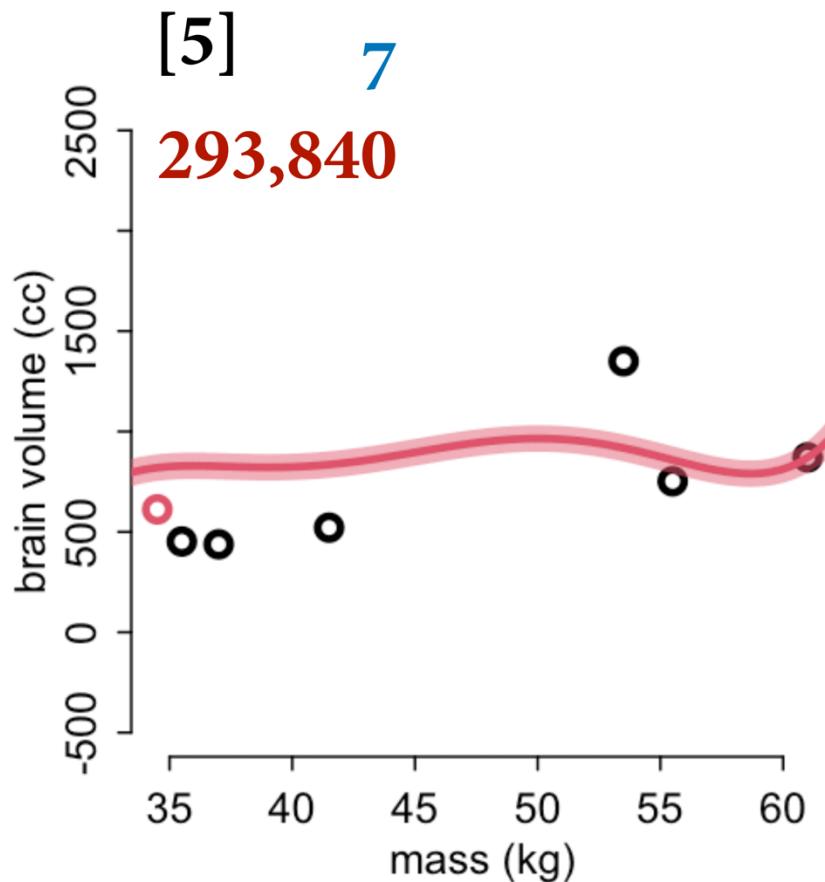
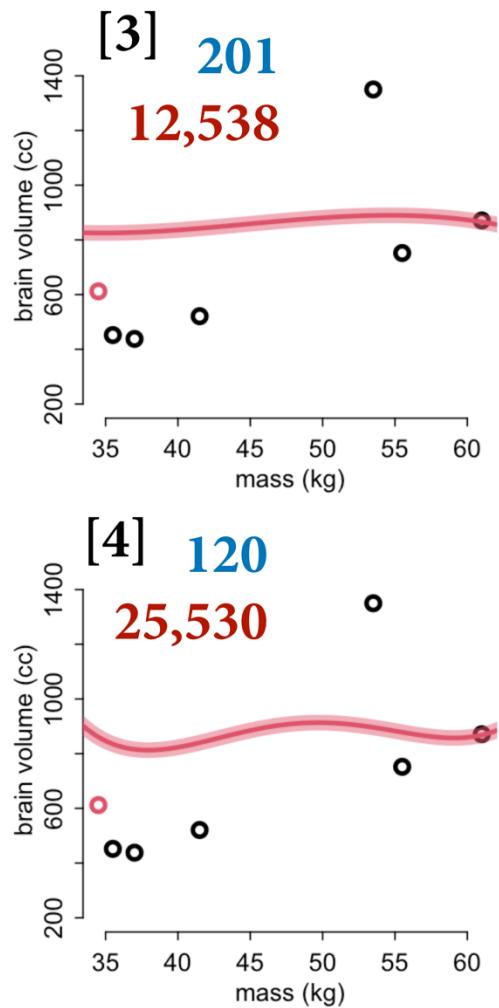
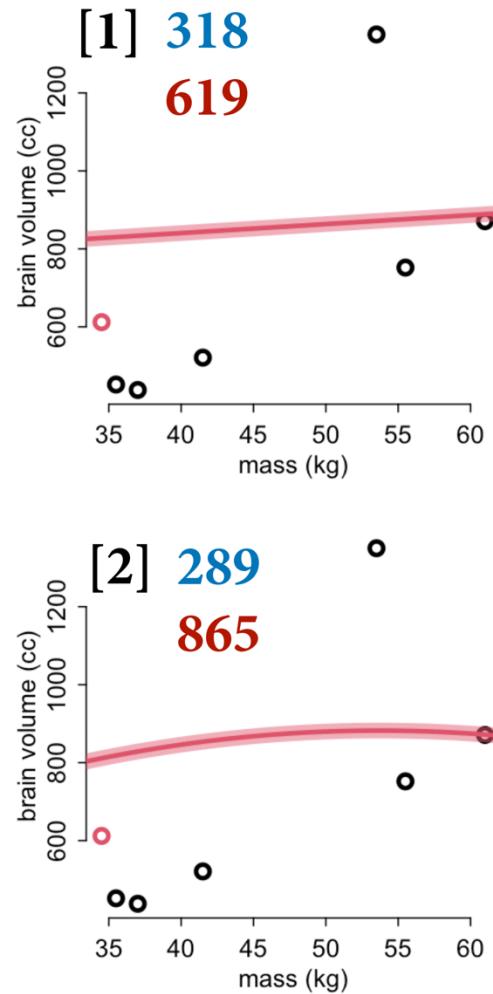
[1] In: 318
Out: 619



[2] In: 289
Out: 865





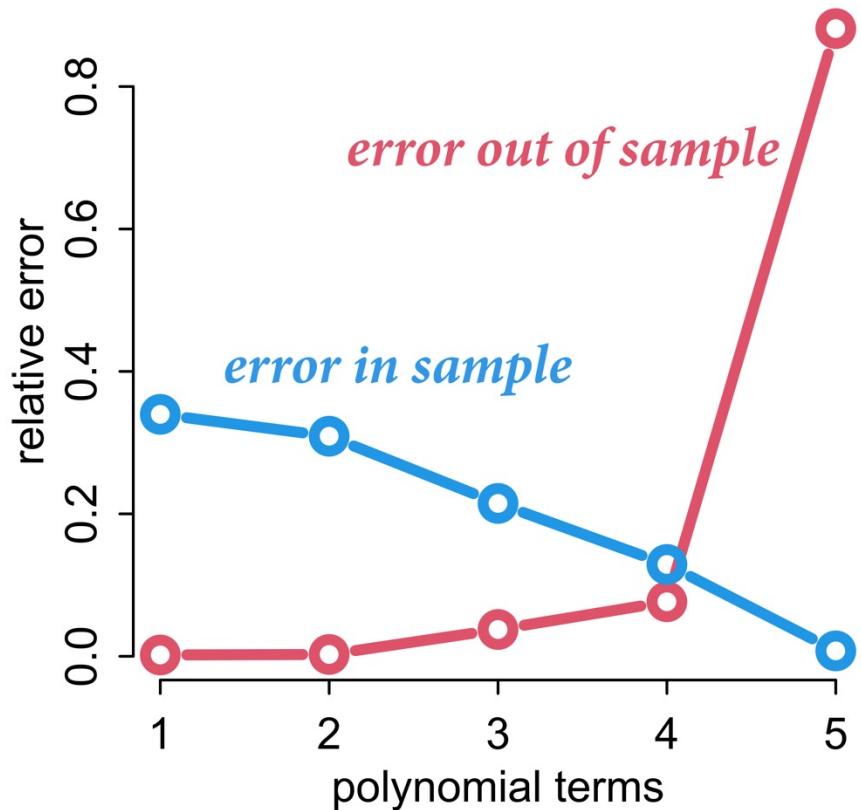


Cross-validation

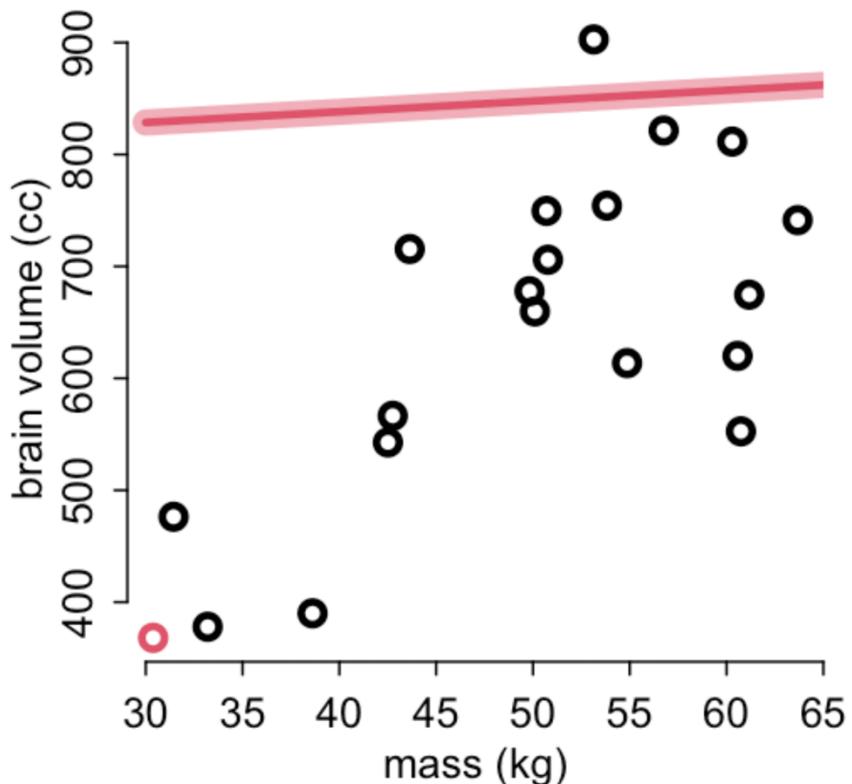
For simple models, increasing parameters improves fit to sample

But may reduce accuracy of predictions out of sample

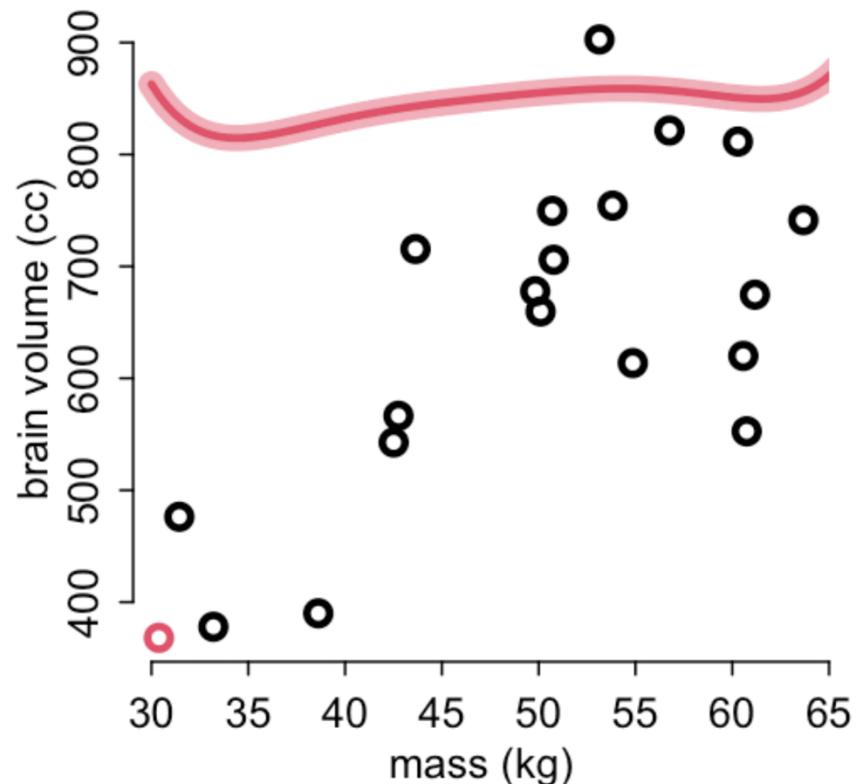
Most accurate model trades off flexibility with overfitting



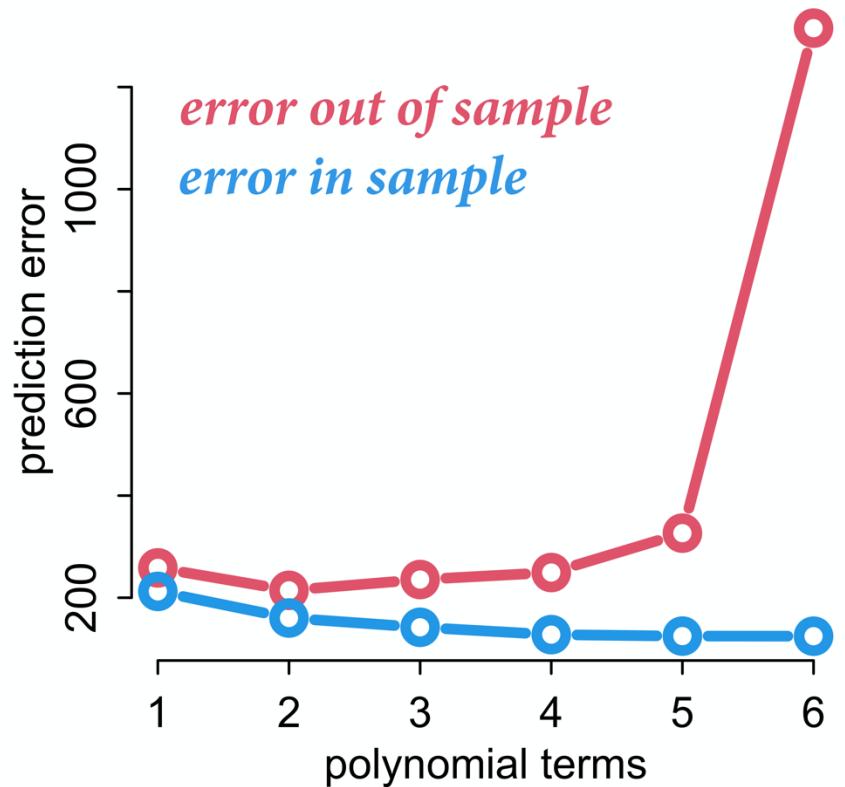
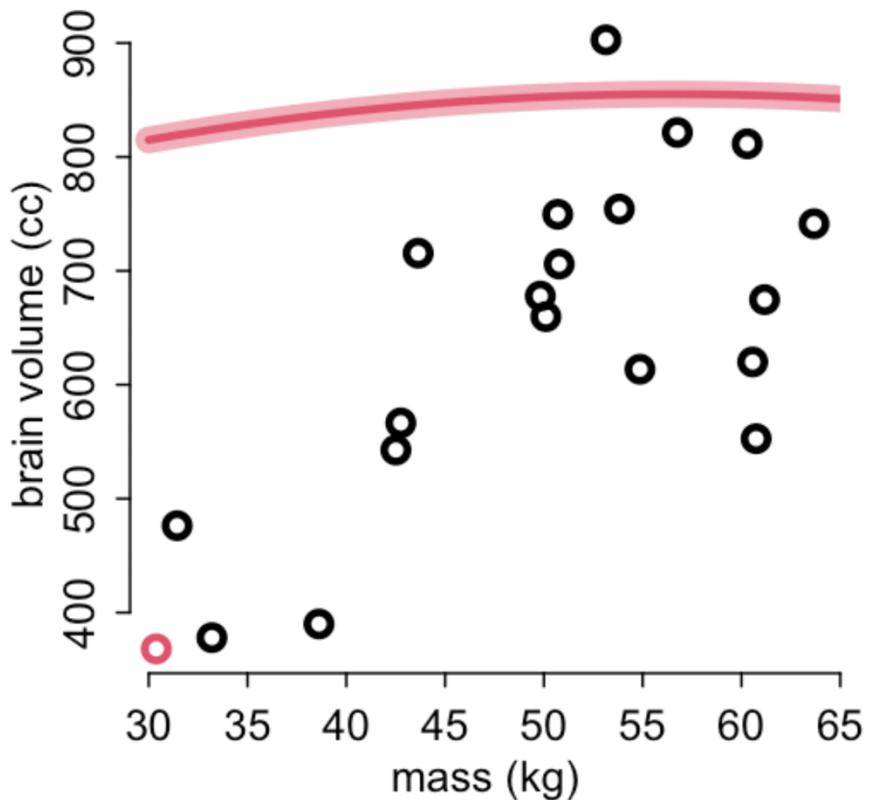
1st degree polynomial



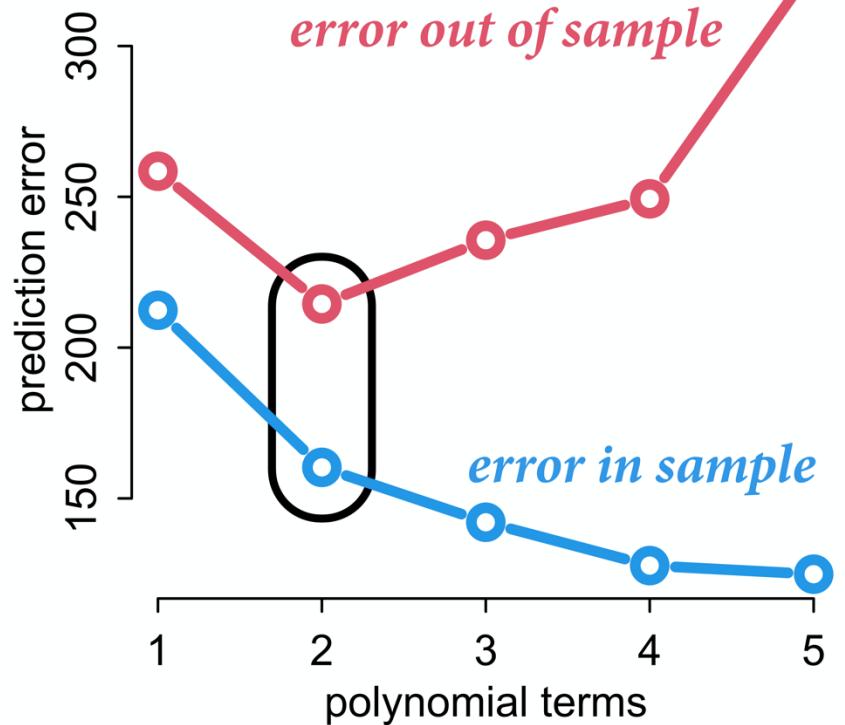
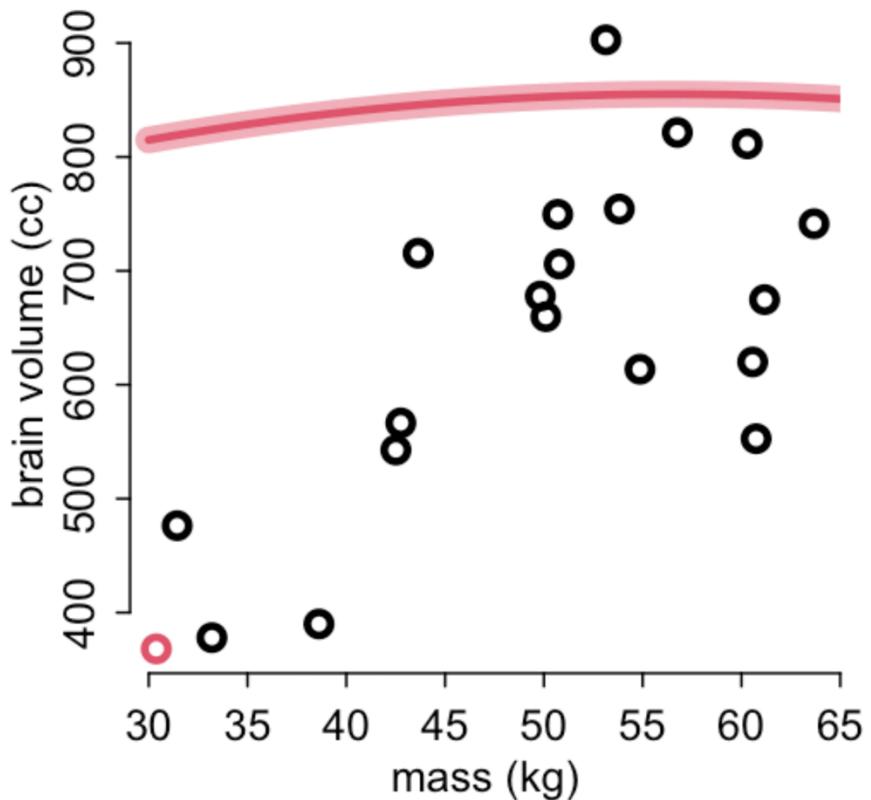
6th degree polynomial



2nd degree polynomial



2nd degree polynomial



Regularization

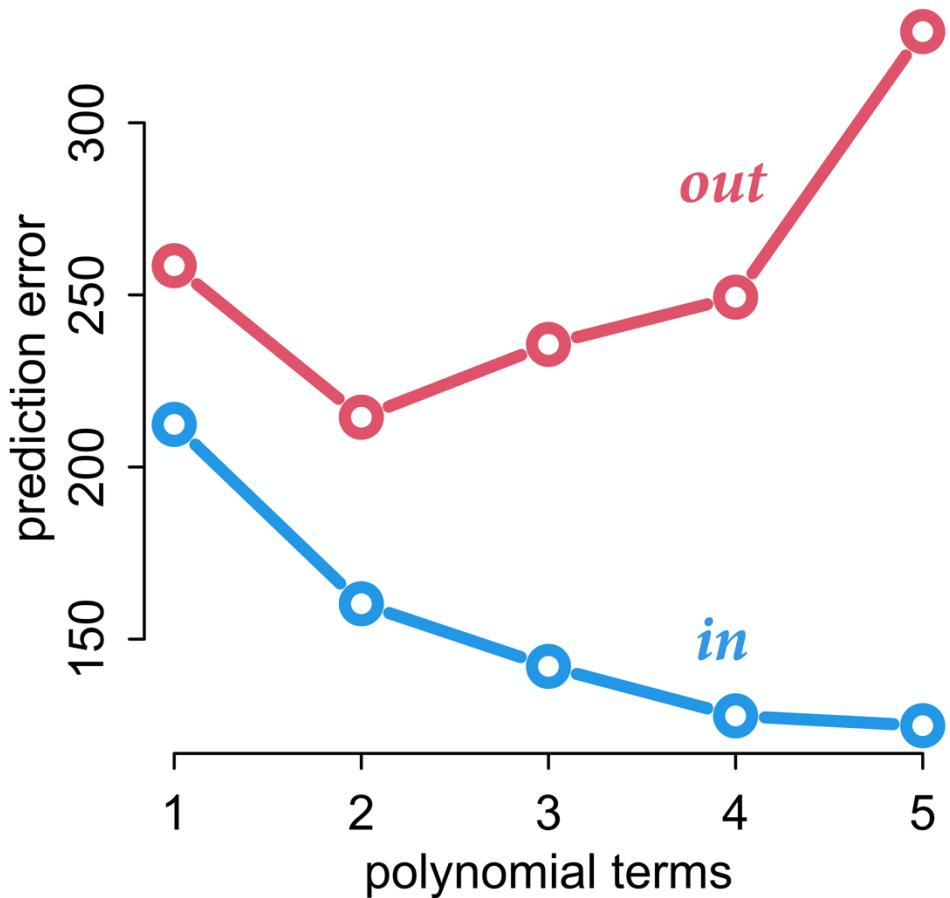
Overfitting depends upon the priors

Skeptical priors have tighter variance, reduce flexibility

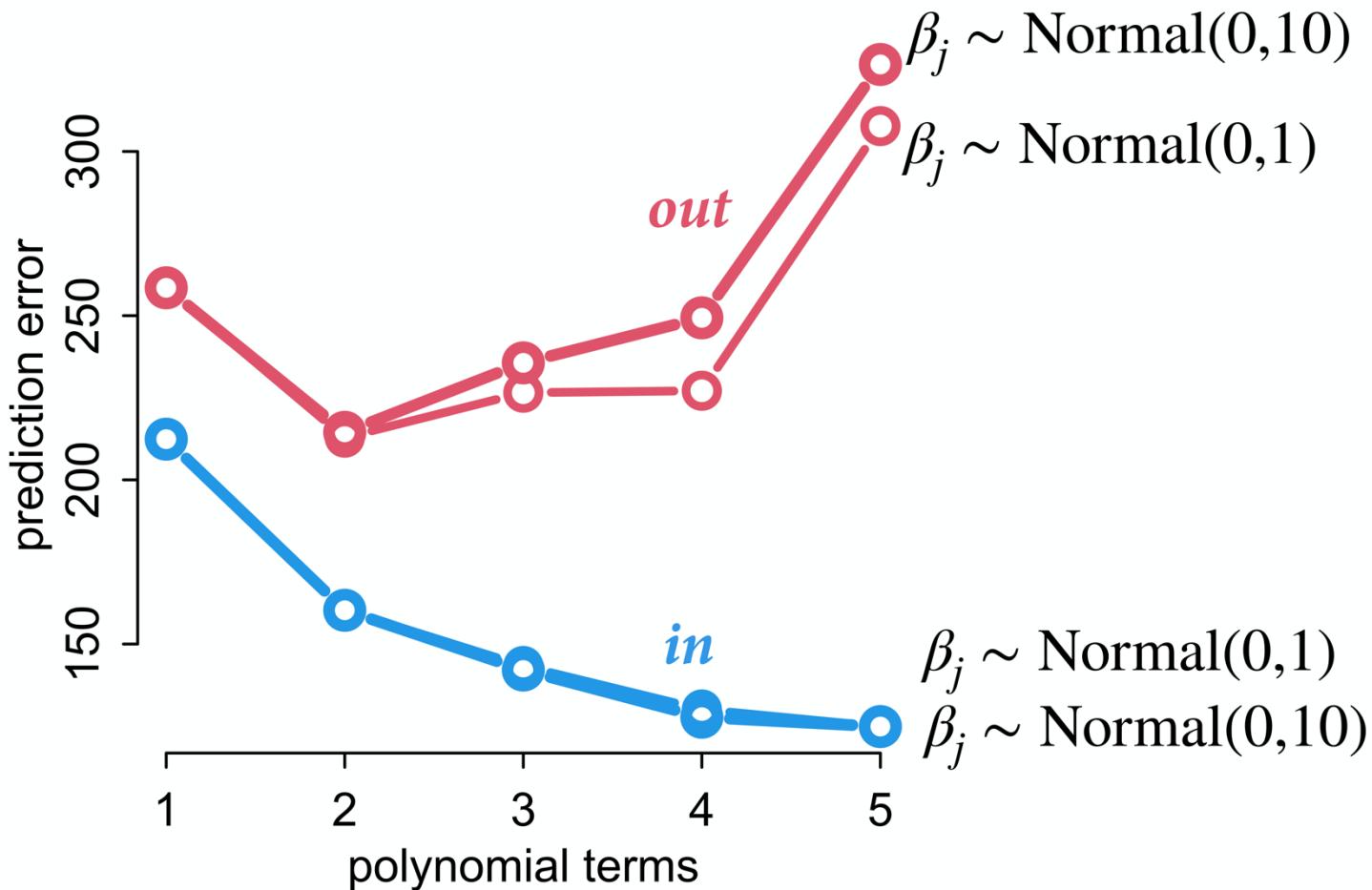
Regularization: Function finds regular features of process

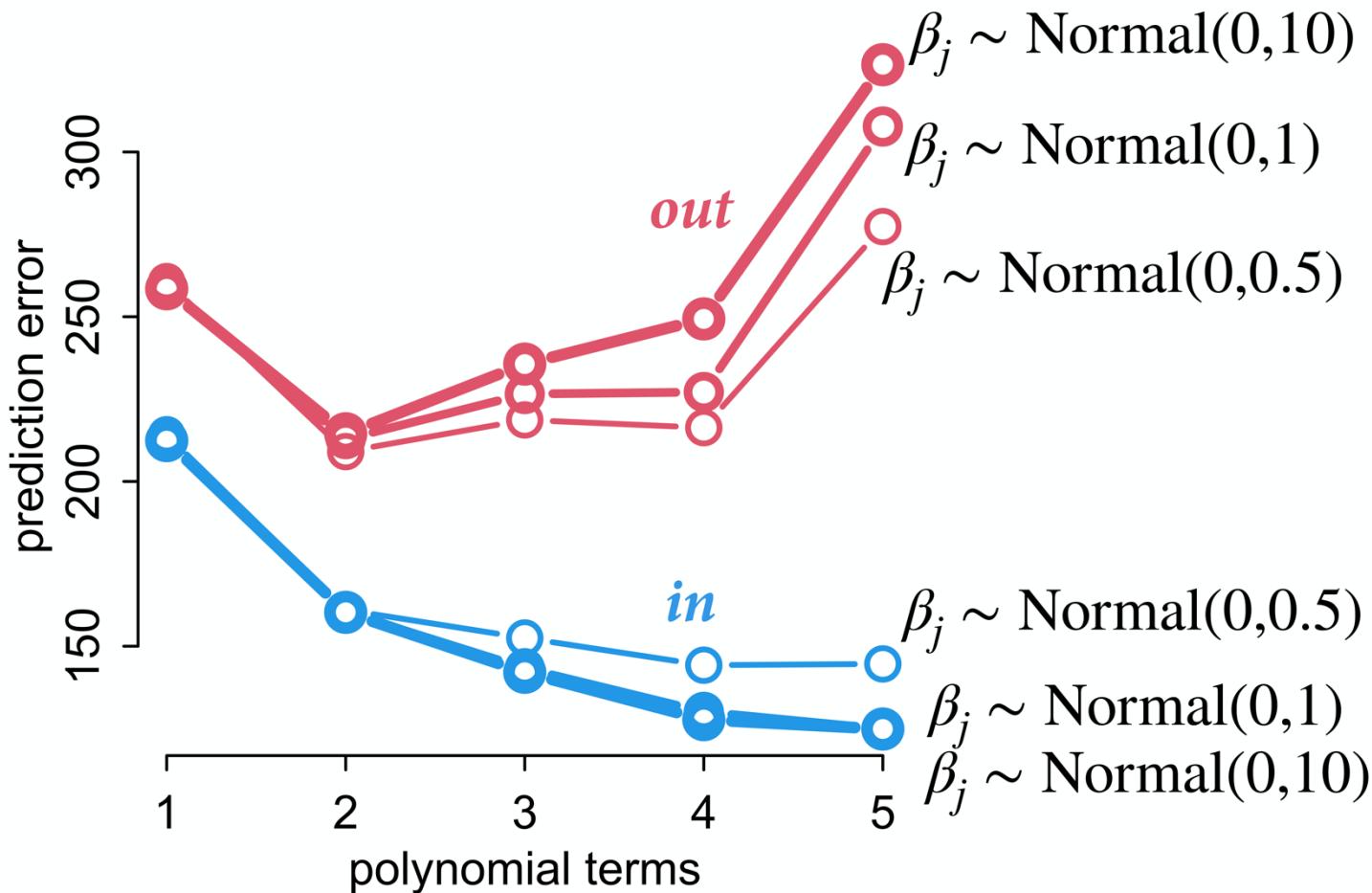
Good priors are often tighter than you think!

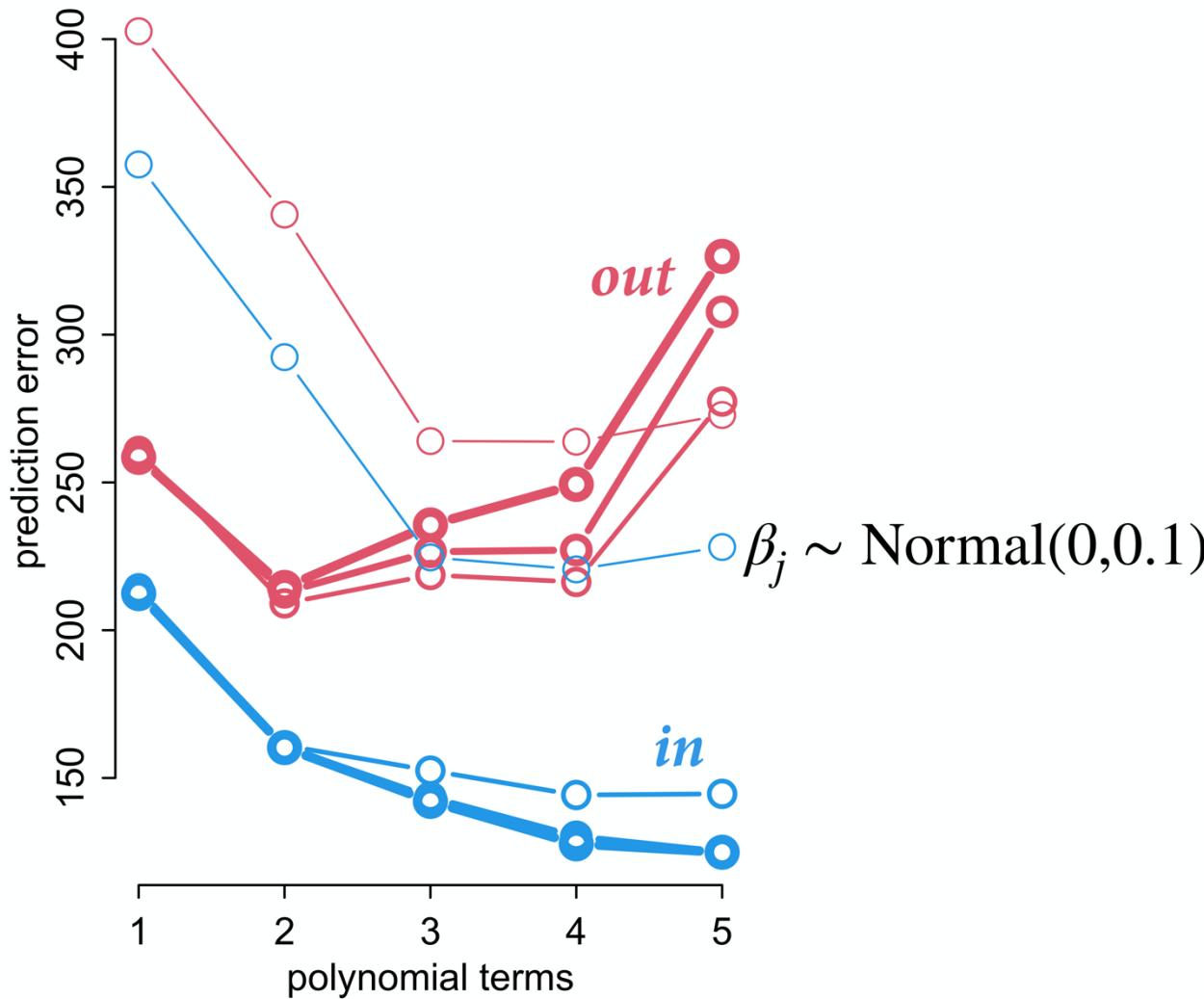




$$\mu_i = \alpha + \sum_{j=1}^m \beta_j x_i^j$$
$$\beta_j \sim \text{Normal}(0, 10)$$







Regularizing priors

How to choose width of prior?

For **causal inference**, use science

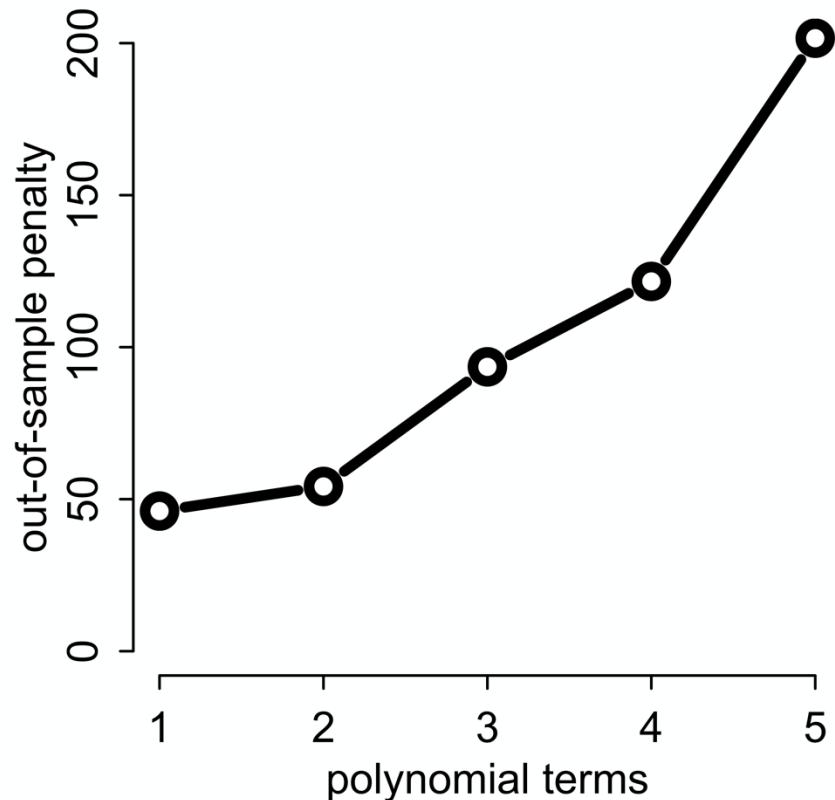
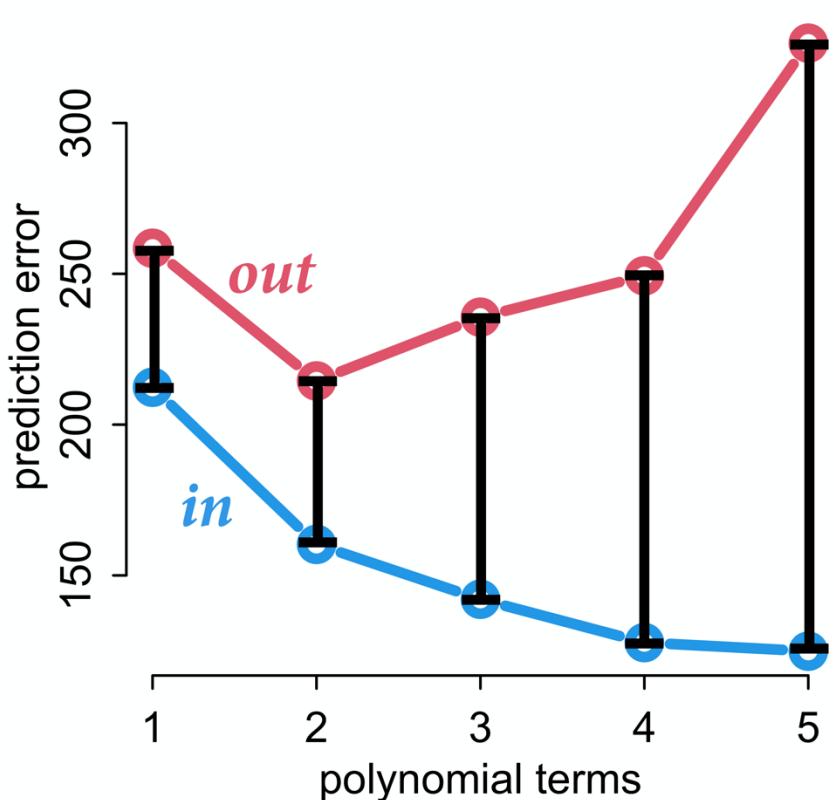
For **pure prediction**, can tune the prior using cross-validation

Many tasks are a mix of inference and prediction

No need to be perfect, just better



Prediction penalty



Penalty prediction

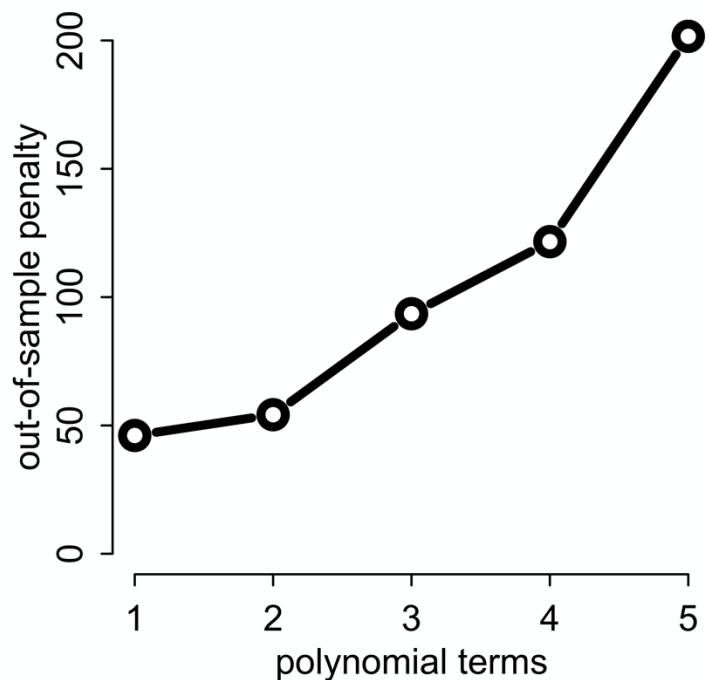
For N points, cross-validation
requires fitting N models

What if you could compute the
penalty from a single model fit?

Good news! You can:

Importance sampling (PSIS)

Information criteria (WAIC)

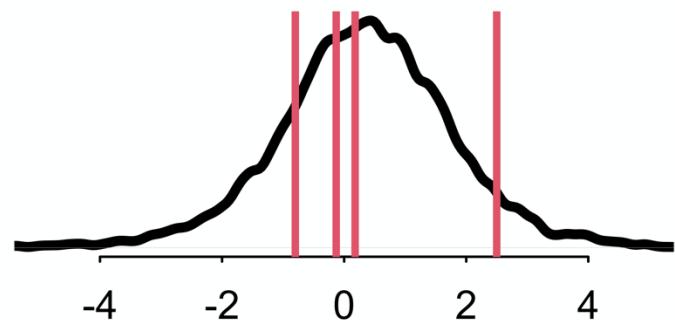


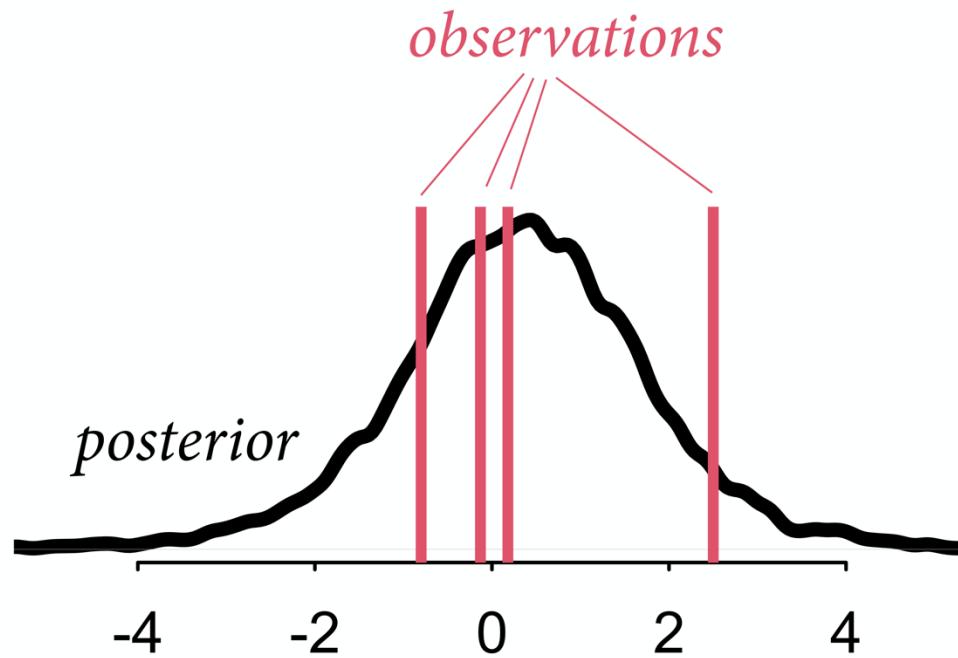
Importance Sampling

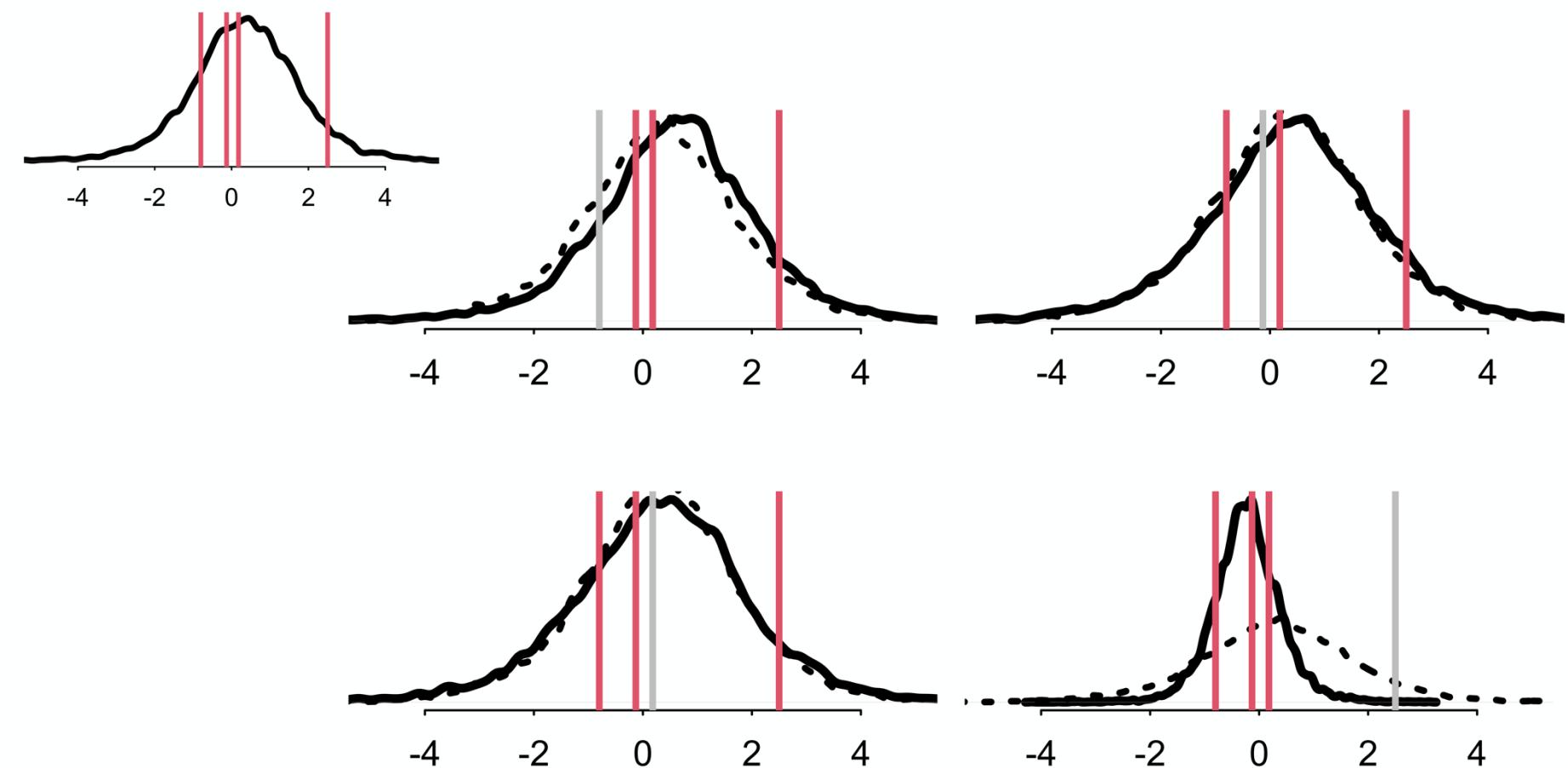
Importance sampling: Use a single posterior distribution for N points to sample from each posterior for $N-1$ points

Key idea: Point with **low** probability has a strong influence on posterior distribution

Can use pointwise probabilities to reweight samples from posterior







Smooth Importance Sampling

Importance sampling tends to be unreliable, has high variance

Pareto-smoothed importance sampling (PSIS) more stable (lower variance)

Useful diagnostics

Identifies important (high leverage) points (“outliers”)



Prof Aki Vehtari (Helsinki),
smooth estimator

Akaike information criterion

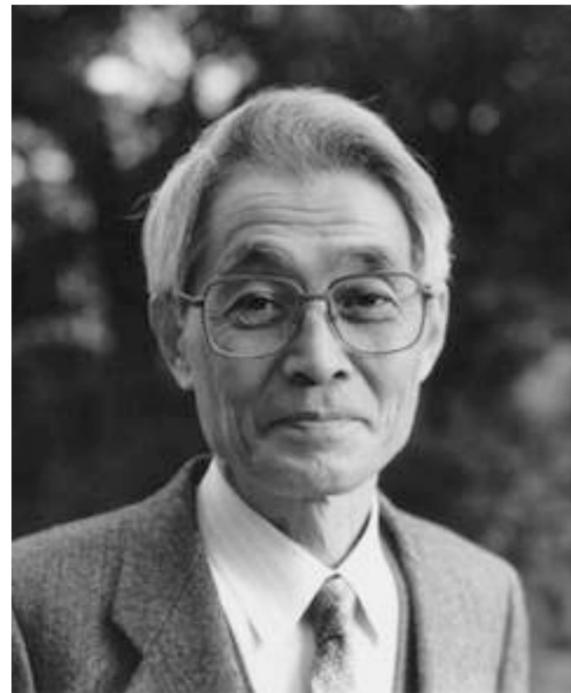
[ah-ka-ee-kay]

Estimate information-theoretic
measure of predictive accuracy
(K-L Distance)

For flat priors and large samples:

$$\text{AIC} = (-2) \times \text{lppd} + 2k$$

*log pointwise
predictive density* *number of
parameters*



Hirotugu Akaike (1927–2009)

赤池弘次

Widely Applicable IC

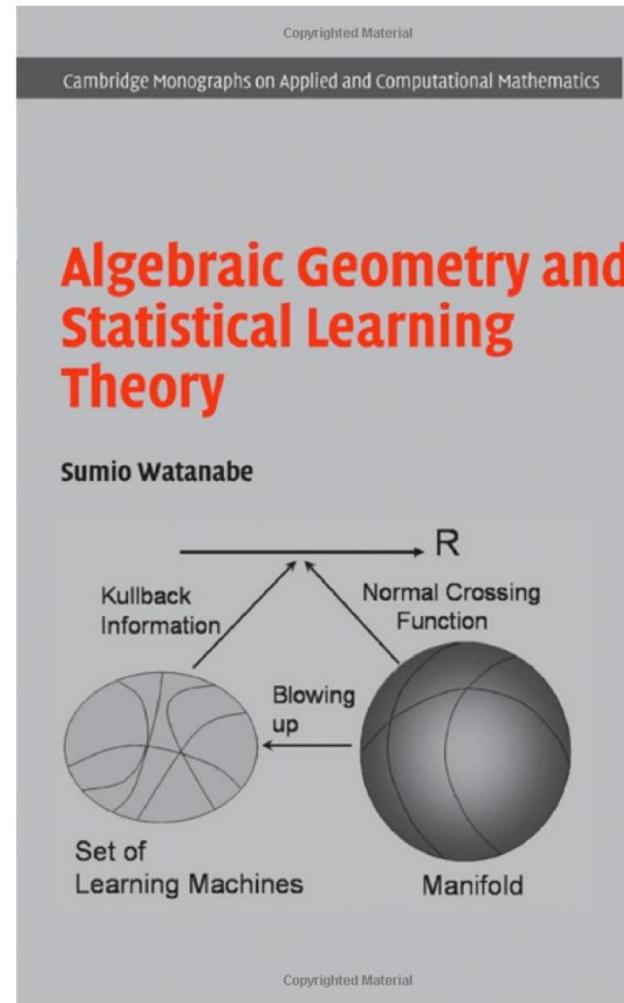
AIC of historical interest now

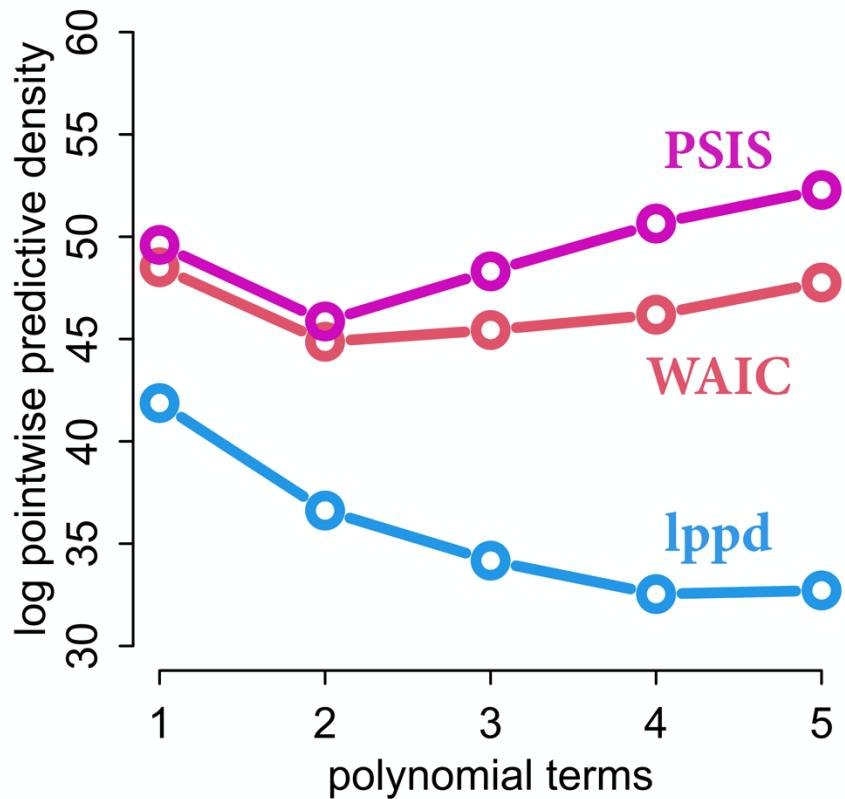
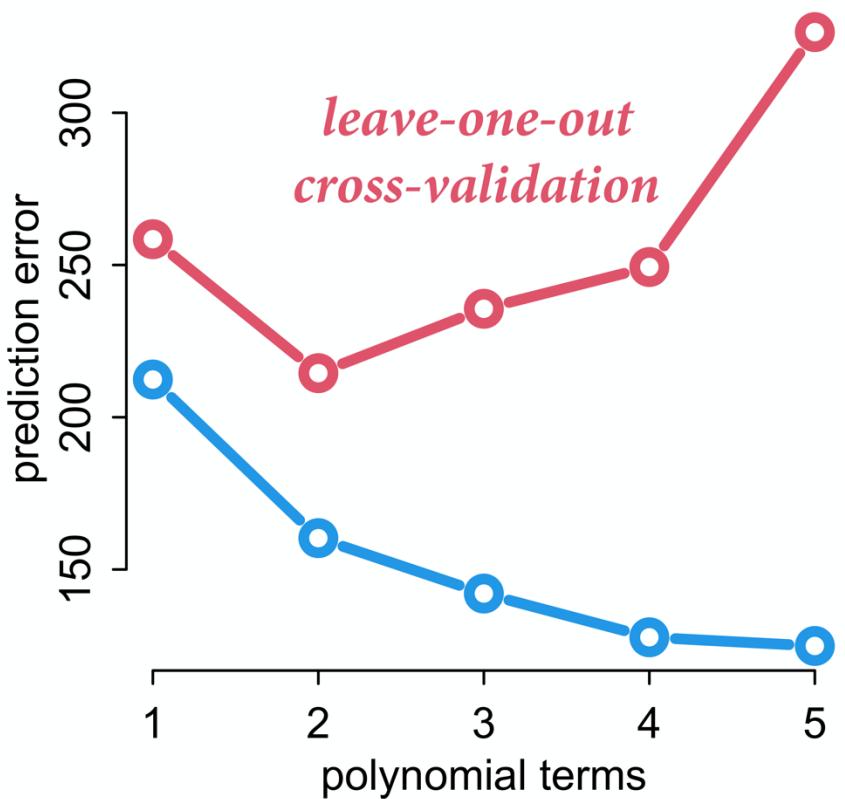
Widely Applicable Information Criterion (WAIC)

Sumio Watanabe (渡辺澄夫) 2010

$$\text{WAIC}(y, \Theta) = -2 \left(\text{lppd} - \underbrace{\sum_i \text{var}_\Theta \log p(y_i | \Theta)}_{\text{penalty term}} \right)$$

Very similar to PSIS score, but no automatic diagnostics





WAIC,PSIS,CV measure
overfitting

Regularization manages
overfitting

None directly address
causal inference

All important to
understanding how
statistical inference works

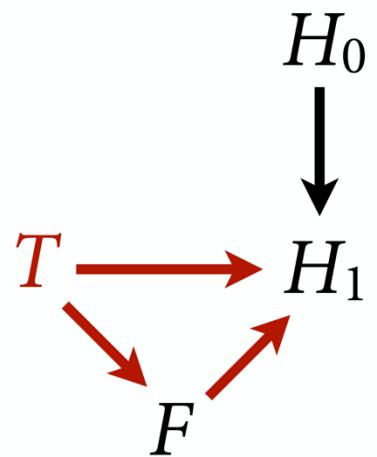


Model Mis-selection

Do not use predictive criteria (WAIC, PSIS, CV) to choose a causal estimate

Predictive criteria actually prefer confounds & colliders

Example: Plant growth experiment

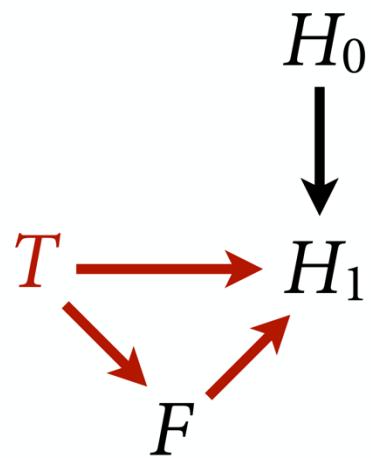


$$H_1 \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = H_0 \times p_i$$

$$p_i = \alpha + \beta_T T_i + \beta_F F_i$$

**Wrong adjustment set
for total causal effect of
treatment (blocks
mediating path)**



$$H_1 \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = H_0 \times p_i$$

$$p_i = \alpha + \beta_T T_i$$

**Correct adjustment set for
total causal effect of
treatment**

$$H_1 \sim \text{Normal}(\mu_i, \sigma)$$

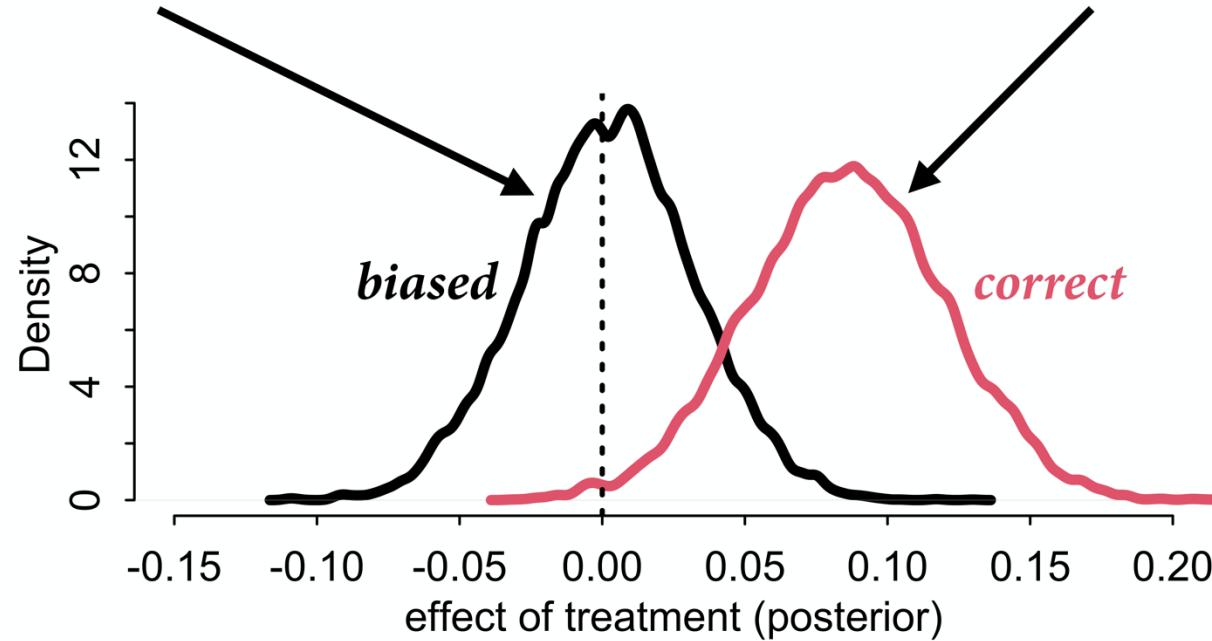
$$\mu_i = H_0 \times p_i$$

$$p_i = \alpha + \beta_T T_i + \beta_F F_i$$

$$H_1 \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = H_0 \times p_i$$

$$p_i = \alpha + \beta_T T_i$$



$$H_1 \sim \text{Normal}(\mu_i, \sigma)$$

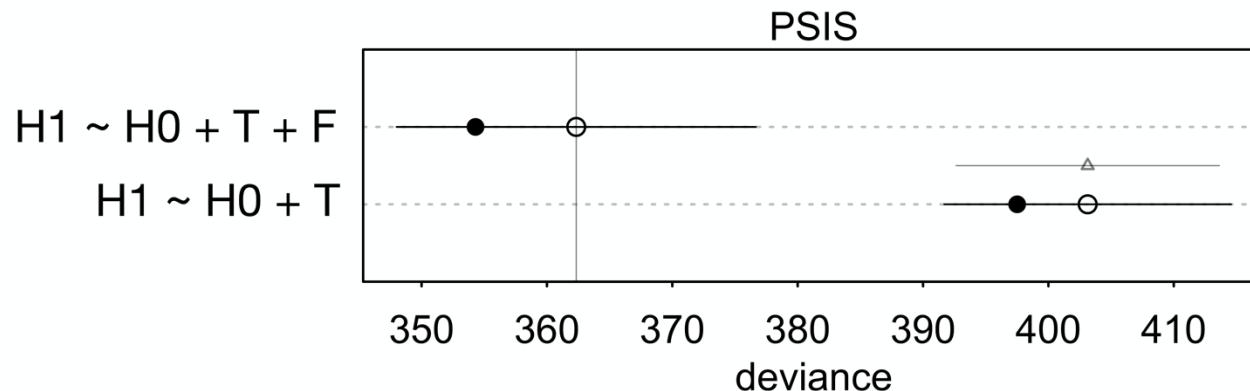
$$\mu_i = H_0 \times p_i$$

$$p_i = \alpha + \beta_T T_i + \beta_F F_i$$

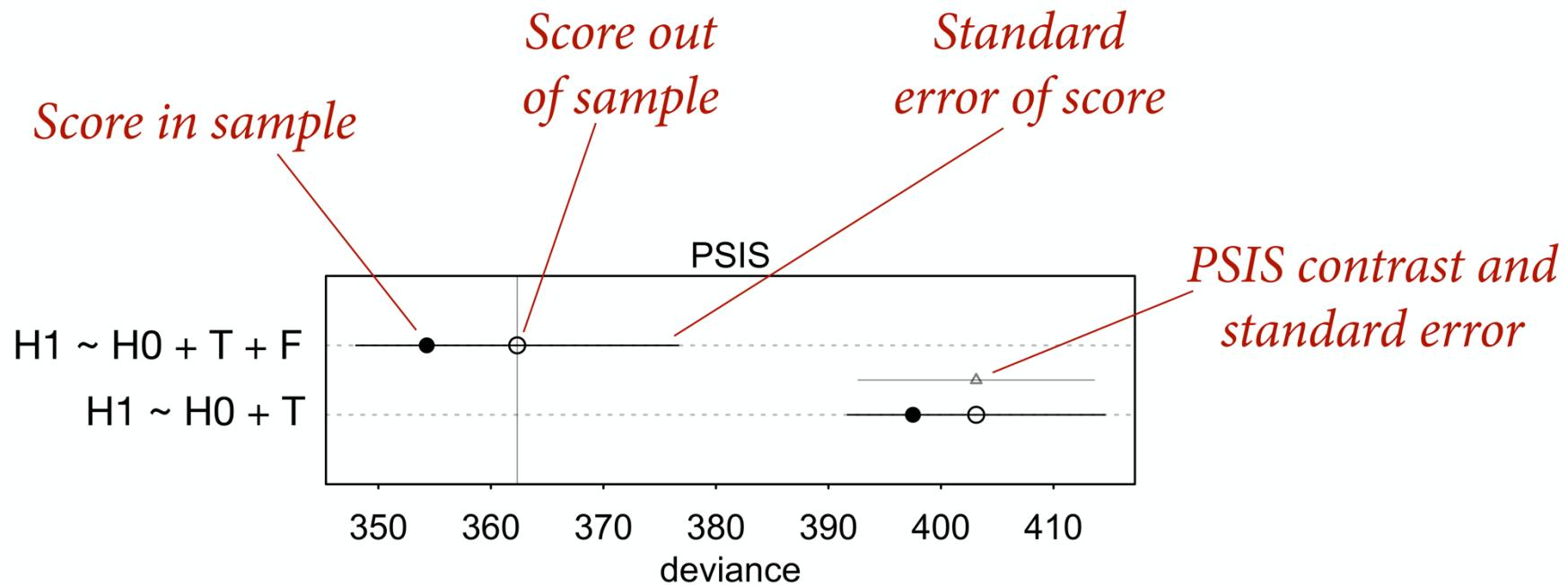
$$H_1 \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = H_0 \times p_i$$

$$p_i = \alpha + \beta_T T_i$$

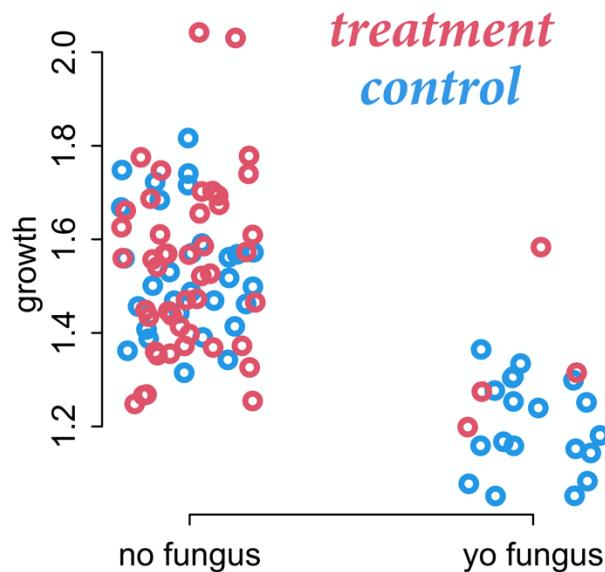
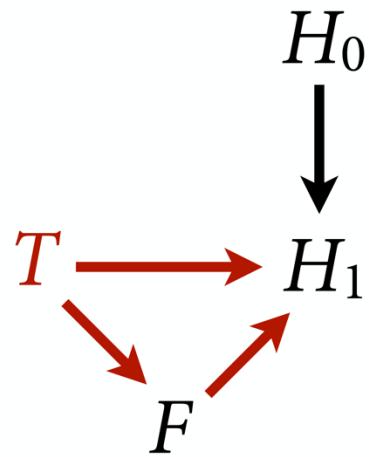


Wrong model wins at prediction

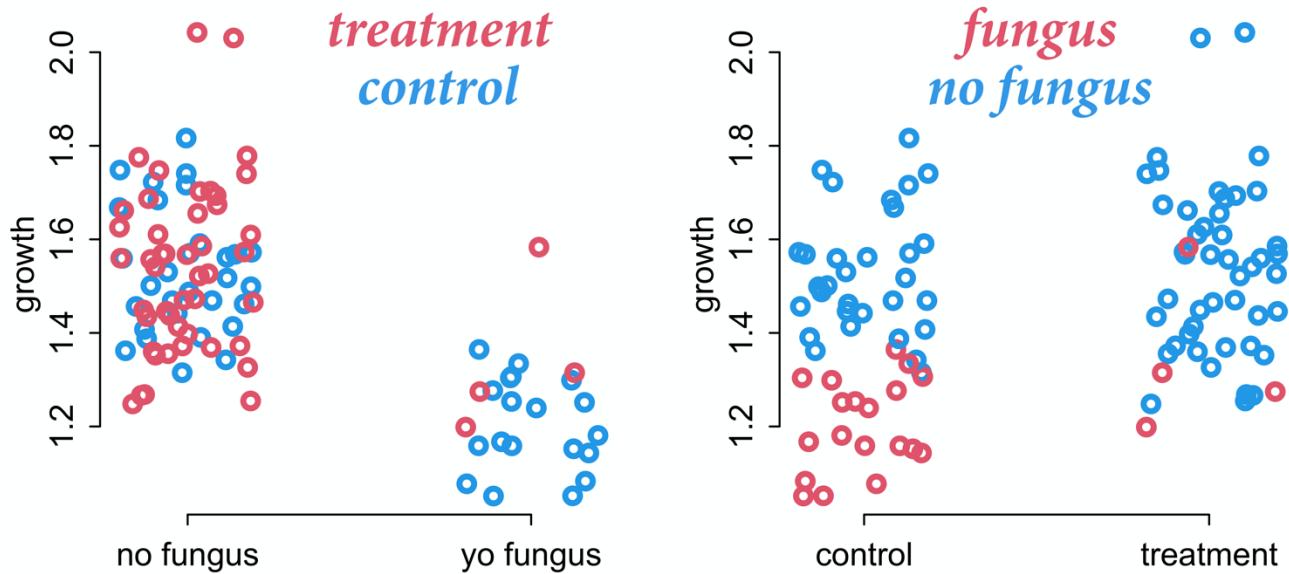
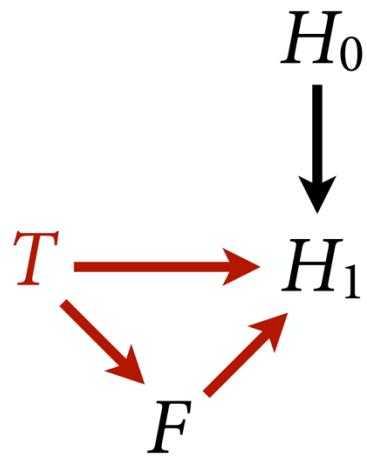


Wrong model wins at prediction

Why does the wrong model win at prediction?



Why does the wrong model win at prediction?



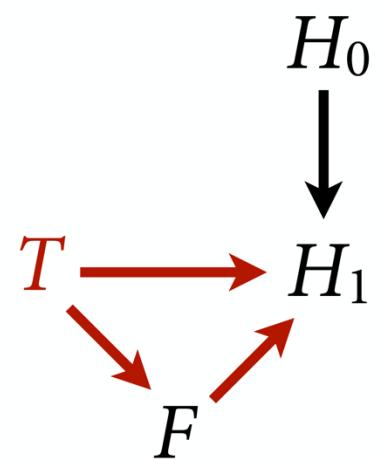
Fungus is in fact a better predictor than treatment

Model Mis-selection

Do not use predictive criteria (WAIC, PSIS, CV) to choose a causal estimate

However, many analyses are mixes of inferential and predictive chores

Still need help finding good functional descriptions while avoiding overfitting



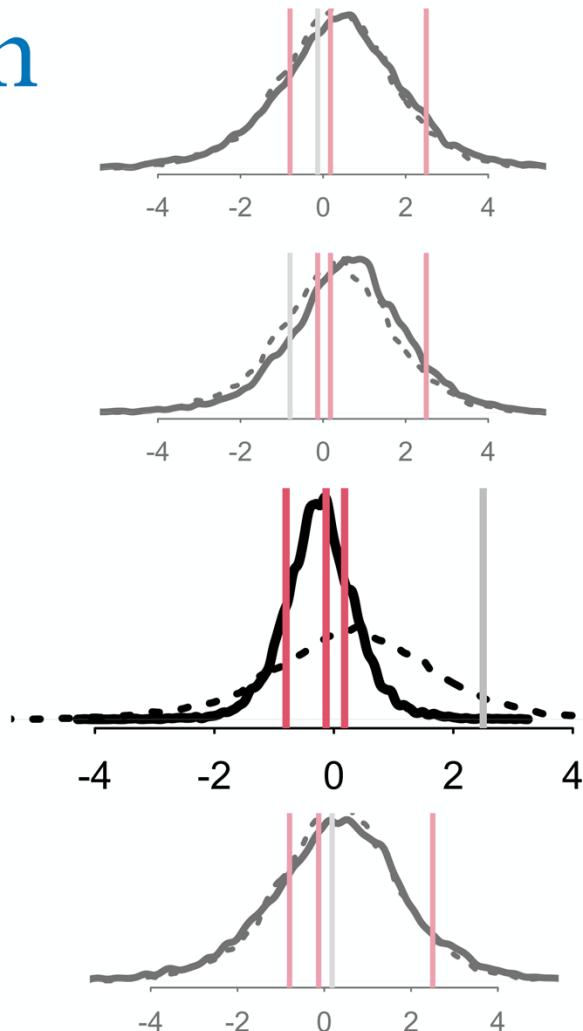
Outliers & Robust Regression

Some points are more influential than others

“Outliers”: Observations in the tails of predictive distribution

Outliers indicate predictions are possibly overconfident, unreliable

The model doesn't expect enough variation



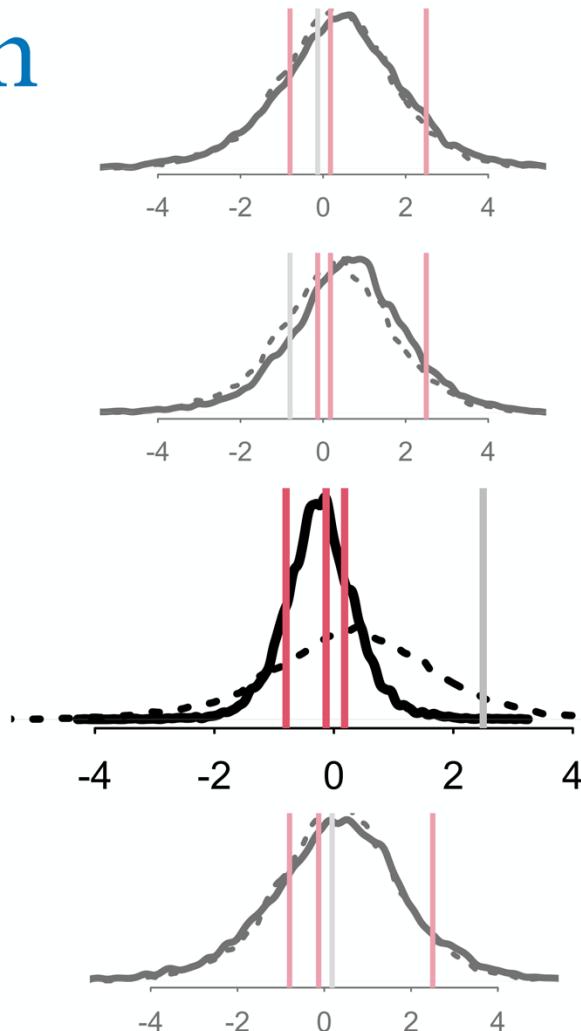
Outliers & Robust Regression

Dropping outliers is bad: Just ignores the problem; predictions are still bad!

It's the model that's wrong, not the data

First, quantify influence of each point

Second, use a mixture model (robust regression)



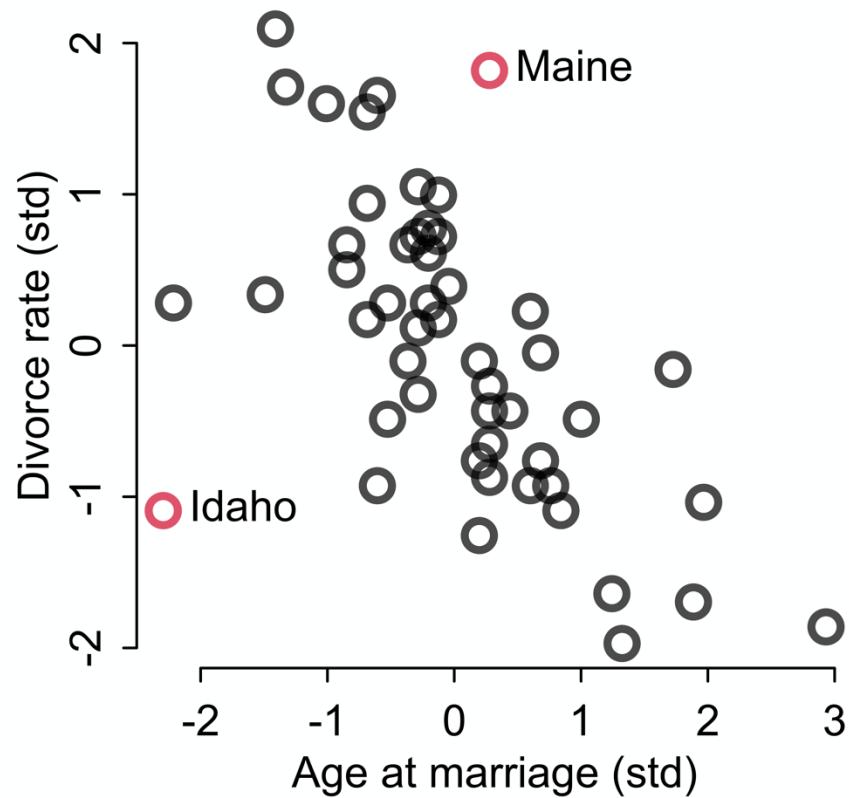
Outliers & Robust Regression

Divorce rate example

Maine and Idaho both highly unusual

Maine: high divorce for trend

Idaho: low divorce for trend

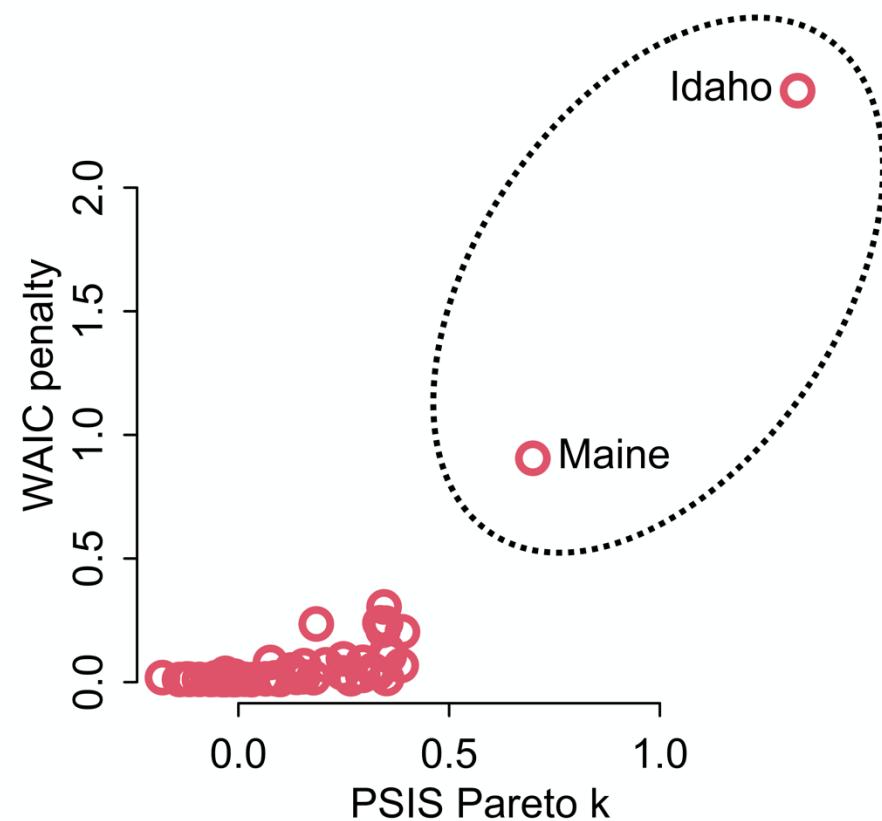


Outliers & Robust Regression

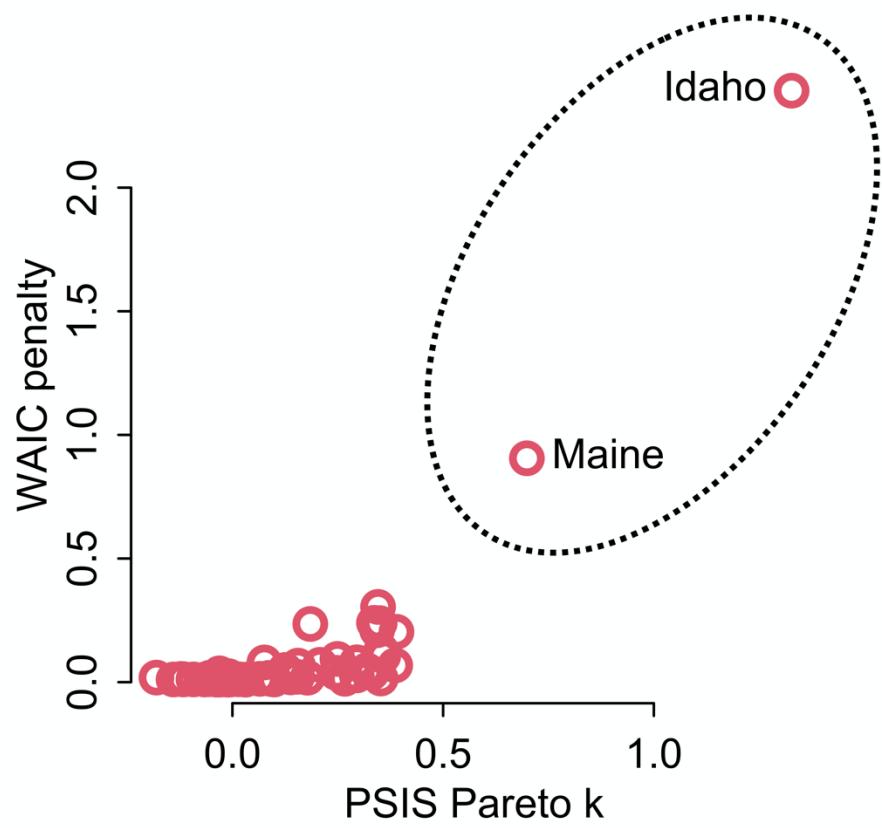
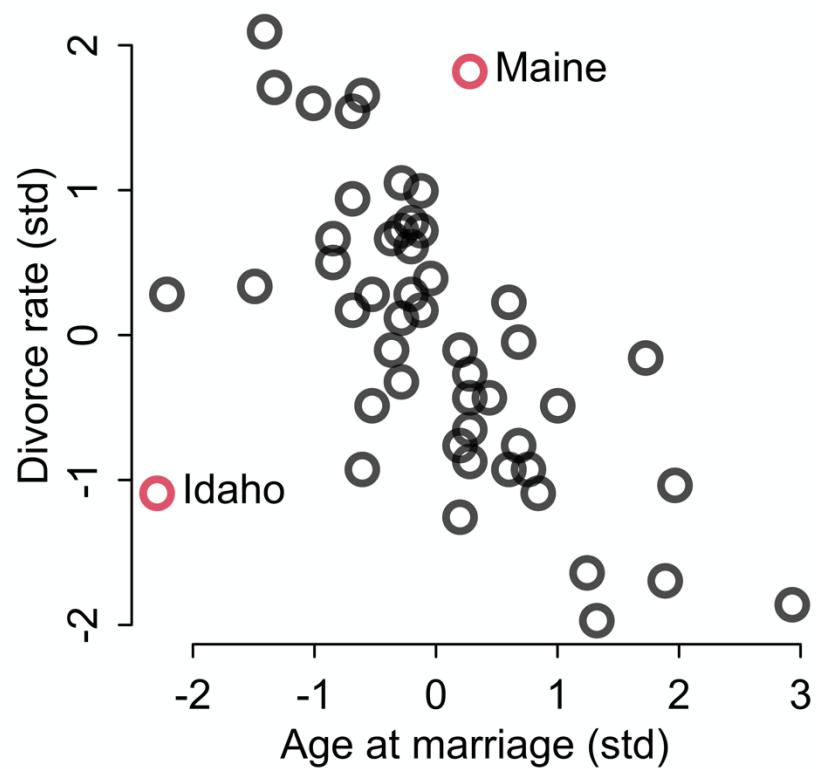
Quantify influence:

PSIS k statistic

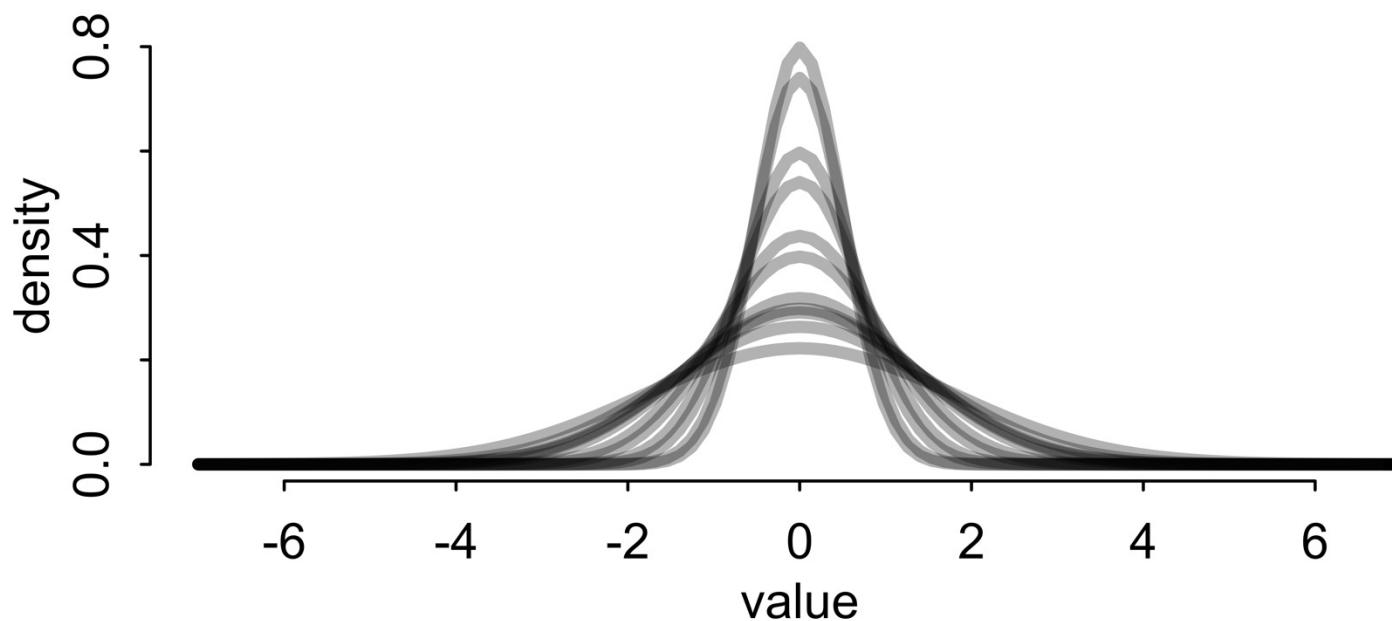
WAIC penalty term (“effective number of parameters”)



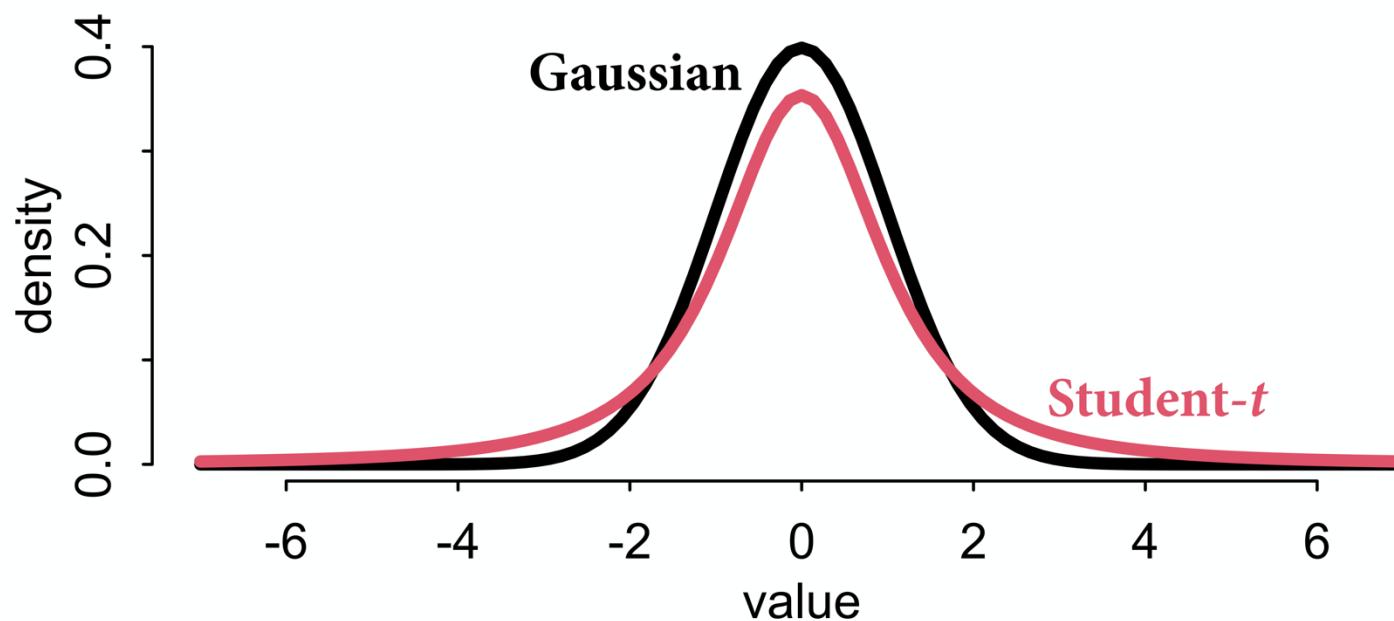
Outliers & Robust Regression



Mixing Gaussians



Mixing Gaussians



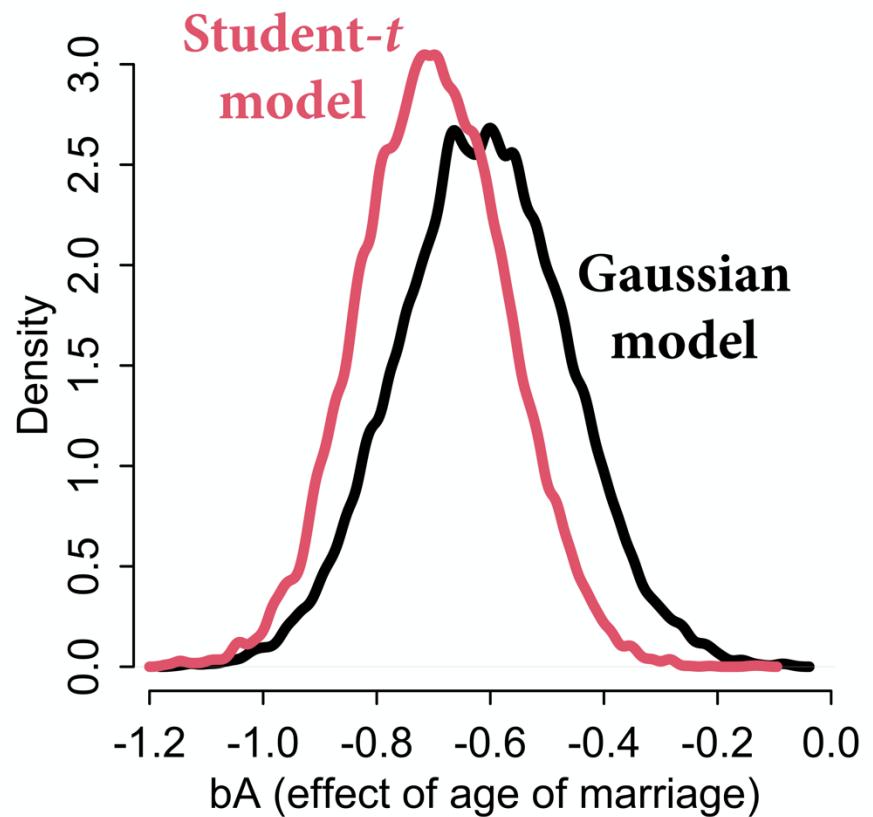
```
m5.3 <- quap(  
  alist(  
    D ~ dnorm( mu , sigma ) ,  
    mu <- a + bM*M + bA*A ,  
    a ~ dnorm( 0 , 0.2 ) ,  
    bM ~ dnorm( 0 , 0.5 ) ,  
    bA ~ dnorm( 0 , 0.5 ) ,  
    sigma ~ dexp( 1 )  
  ) , data = dat )  
  
m5.3t <- quap(  
  alist(  
    D ~ dstudent( 2 , mu , sigma ) ,  
    mu <- a + bM*M + bA*A ,  
    a ~ dnorm( 0 , 0.2 ) ,  
    bM ~ dnorm( 0 , 0.5 ) ,  
    bA ~ dnorm( 0 , 0.5 ) ,  
    sigma ~ dexp( 1 )  
  ) , data = dat )
```

```

m5.3 <- quap(
  alist(
    D ~ dnorm( mu , sigma ) ,
    mu <- a + bM*M + bA*A ,
    a ~ dnorm( 0 , 0.2 ) ,
    bM ~ dnorm( 0 , 0.5 ) ,
    bA ~ dnorm( 0 , 0.5 ) ,
    sigma ~ dexp( 1 )
  ) , data = dat )

m5.3t <- quap(
  alist(
    D ~ dstudent( 2 , mu , sigma ) ,
    mu <- a + bM*M + bA*A ,
    a ~ dnorm( 0 , 0.2 ) ,
    bM ~ dnorm( 0 , 0.5 ) ,
    bA ~ dnorm( 0 , 0.5 ) ,
    sigma ~ dexp( 1 )
  ) , data = dat )

```

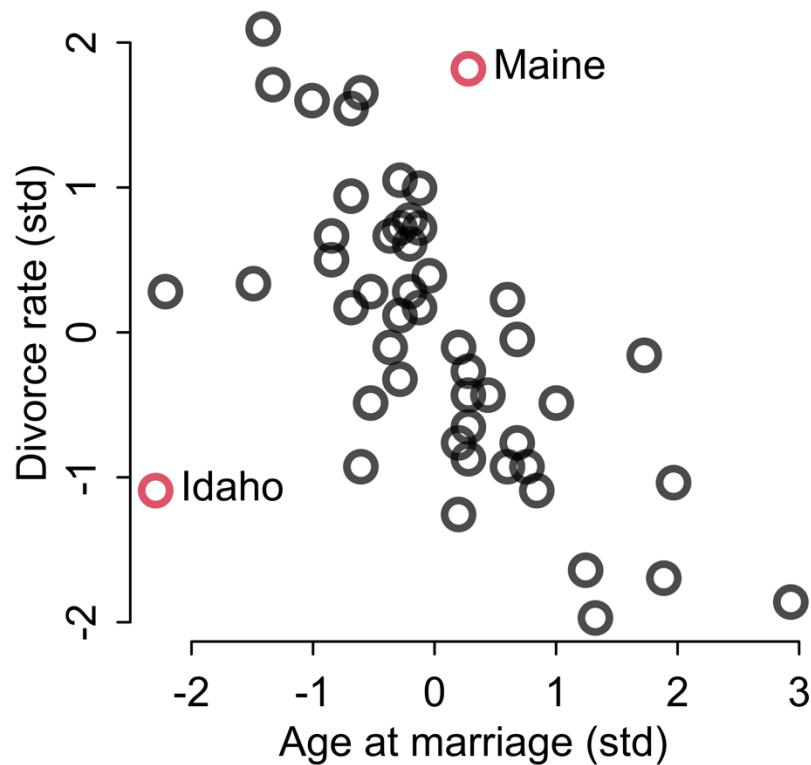


Robust Regressions

Unobserved heterogeneity =>
mixture of Gaussians

Thick tails means model is less
surprised by extreme values

Less surprise, possibly better
predictions if extreme values are
rare



Problems of Prediction

What is the next observation from the same process? (prediction)

Possible to make very good predictions without knowing causes

Optimizing prediction does not reliably reveal causes

Powerful tools (PSIS, regularization) for measuring and managing accuracy

