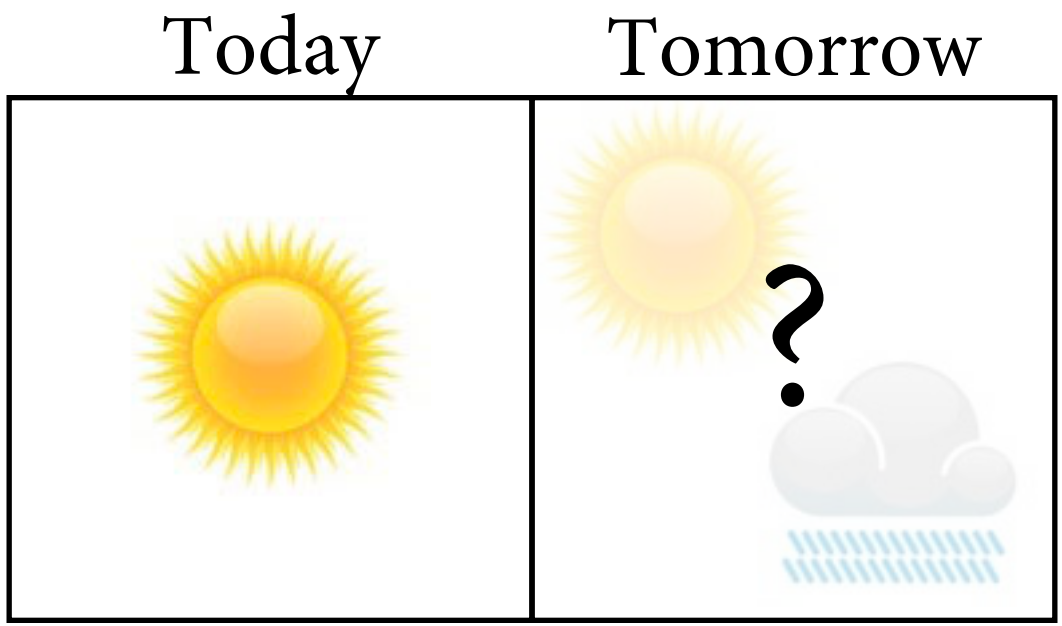
Statistical Rethinking
Winter 2019

Lecture 08 / Week 4

Ulysses' Compass

Information theory

- Machine prediction obeys **information theory**
- *Information*: Reduction in uncertainty caused by learning an outcome.



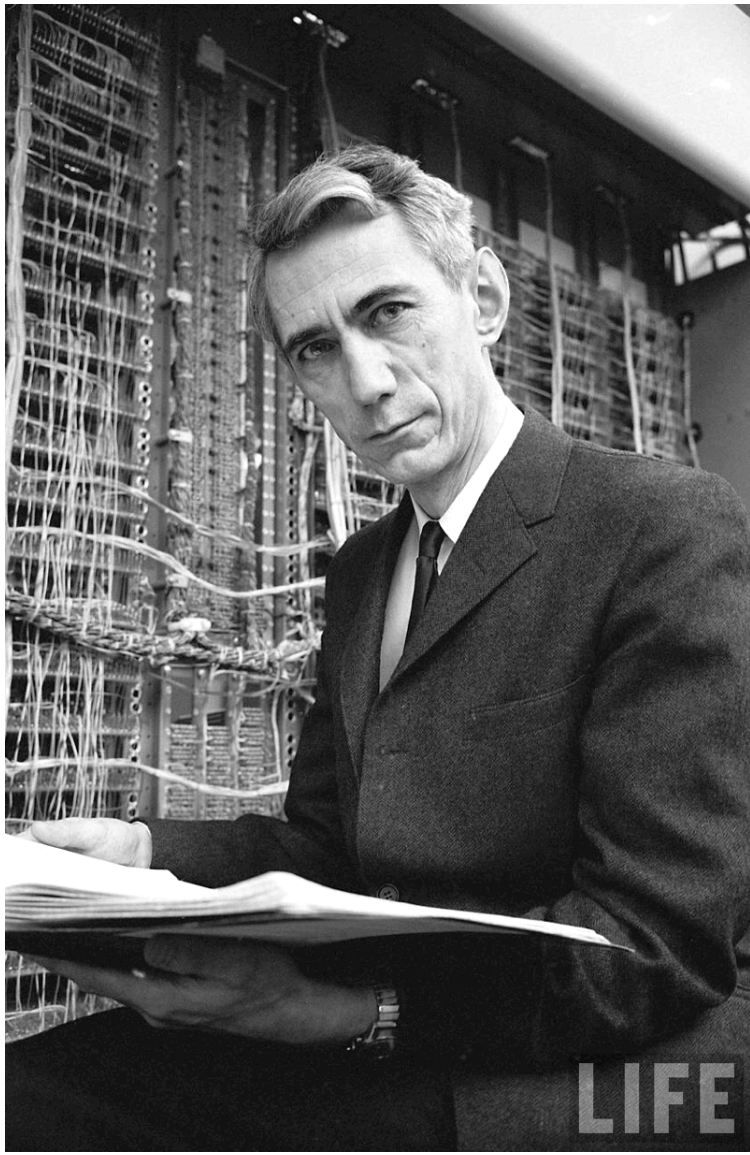
	Today	Tomorrow	
Los Angeles		?	
Glasgow		?	
New York		?	

Information entropy

- 1948, Claude Shannon derived *information entropy*:

$$H(p) = -\operatorname{E}\log(p_i) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Uncertainty in a probability distribution is average (minus) log-probability of an event.



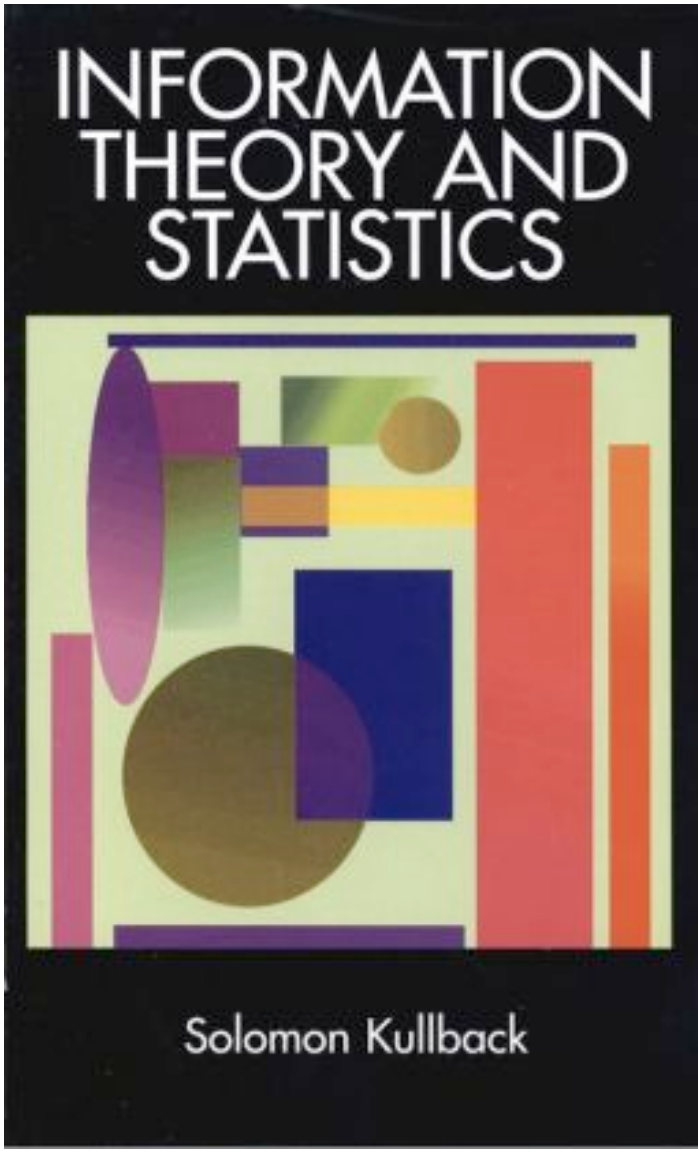
Shannon (1916–2001)

Entropy to accuracy

- Two probability distributions: p, q
- p is true, q is model
- How accurate is q, for describing p?
- Distance from q to p: *Divergence*

$$D_{\mathrm{KL}}(p, q) = \sum_i p_i \big(\log(p_i) - \log(q_i) \big)$$

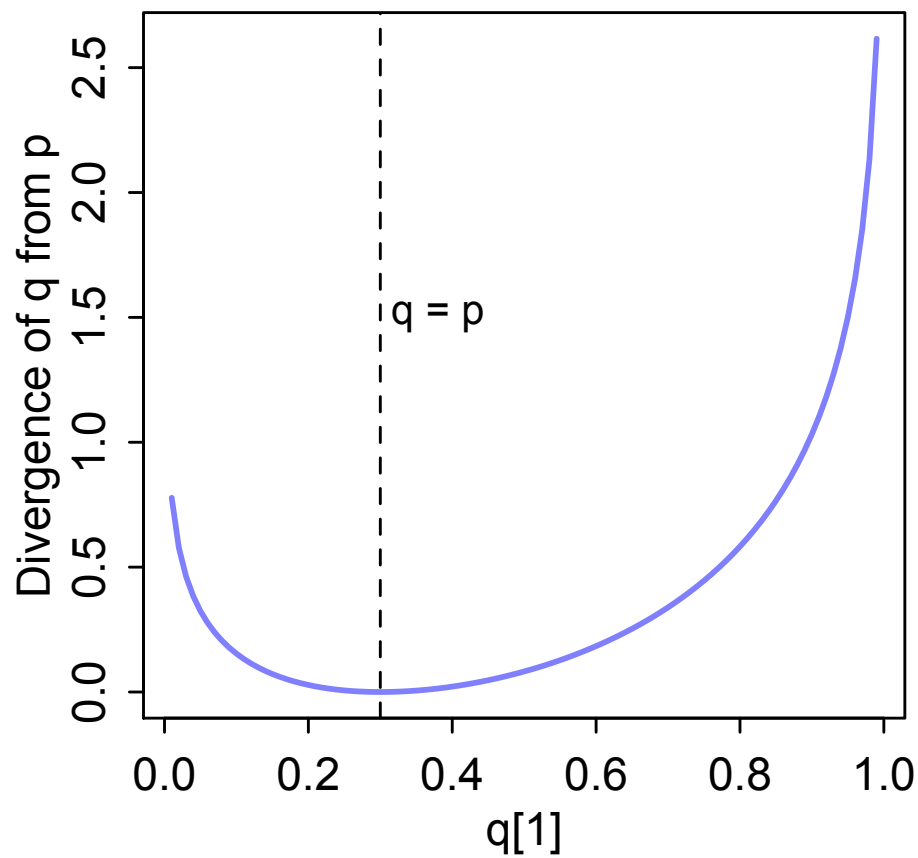
Distance from q to p is the average difference in log-probability.



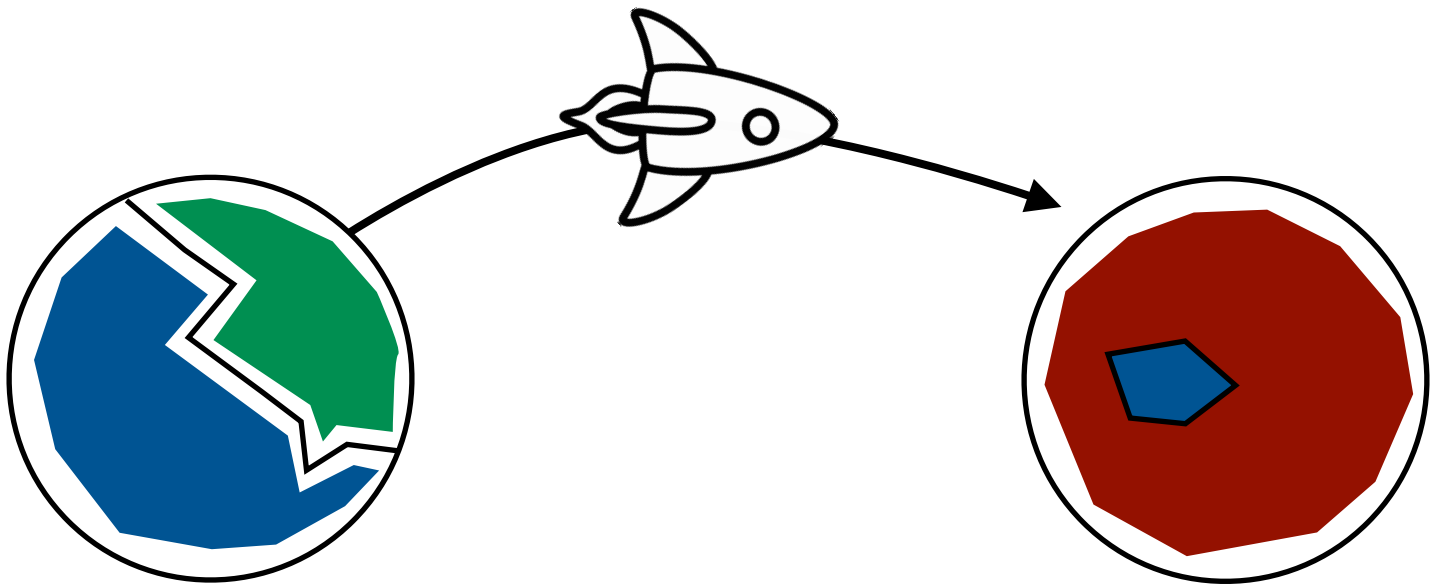
Computing divergence

$$D_{\mathrm{KL}}(p,q) = \sum_i p_i \big(\log(p_i) - \log(q_i) \big)$$

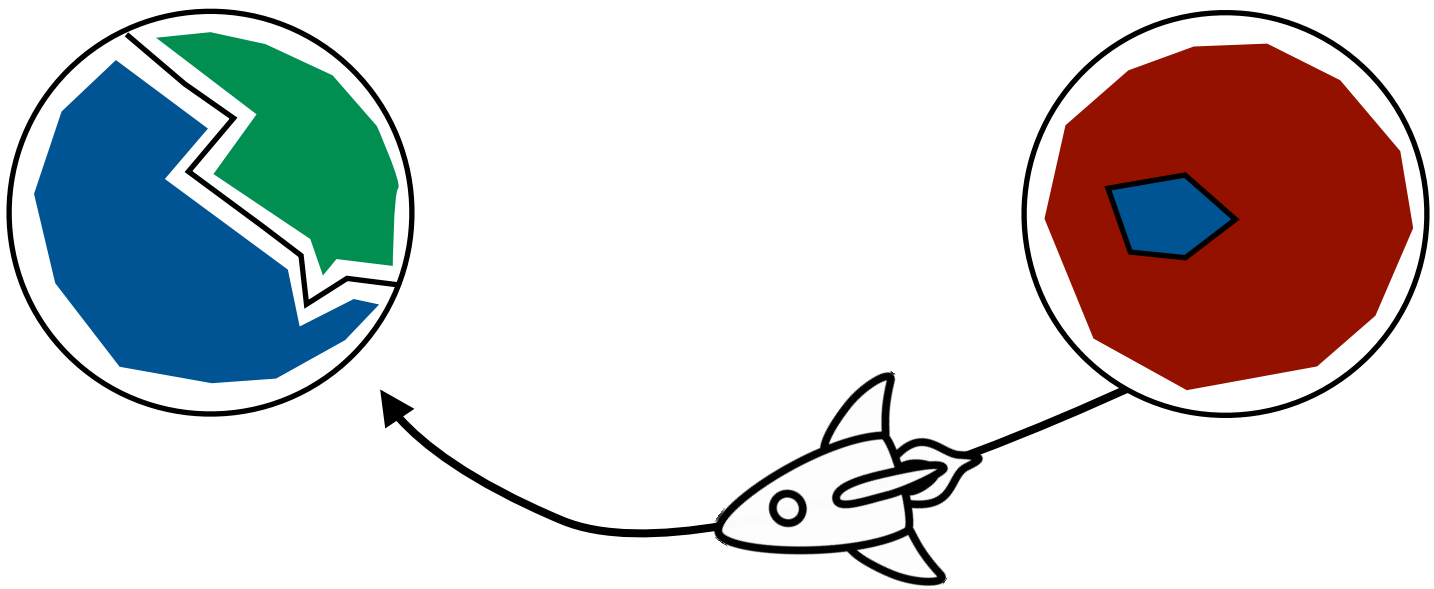
```
p <- c(0.3,0.7)

DKL <- function(p,q)
    sum(p*(log(p)-log(q)))

q1seq <- seq(from=0.01,to=0.99,by=0.01)
DKLseq <- sapply(q1seq,
    function(q1) DKL(p,c(q1,1-q1)) )
plot( q1seq , DKLseq )
```



Divergence is not symmetric!



Divergence is not symmetric!



Estimating Divergence

- Use **log-score**: Sum of log probabilities of each observation

$$S(q) = \sum_i \log(q_i)$$

- In practice, need to average over posterior:

$$\text{lppd}(y, \Theta) = \sum_i \log \frac{1}{S} \sum_s p(y_i|\Theta_s)$$

log-pointwise-predictive-density

Everybody overfits (sometimes)

- Common to scale log-score by –2, "Deviance"
  - Smaller values are better
- A meta-model of forecasting:
  - Two samples: *training* and *testing*, size N
  - Fit model to *training* sample, get D_{train}
  - Use posterior from *training* to compute D_{test}
  - Difference $D_{\text{test}} - D_{\text{train}}$ is overfitting

Everybody overfits

Data generating model:

$$y_i \sim \text{Normal}(\mu_i, 1)$$
$$\mu_i = (0.15)x_{1,i} - (0.4)x_{2,i}$$

Models fit to data:
(flat priors)

$$\mu_i = \alpha$$
$$\mu_i = \alpha + \beta_1 x_{1,i}$$
$$\mu_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$
$$\mu_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$$
$$\mu_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i}$$

Everybody overfits

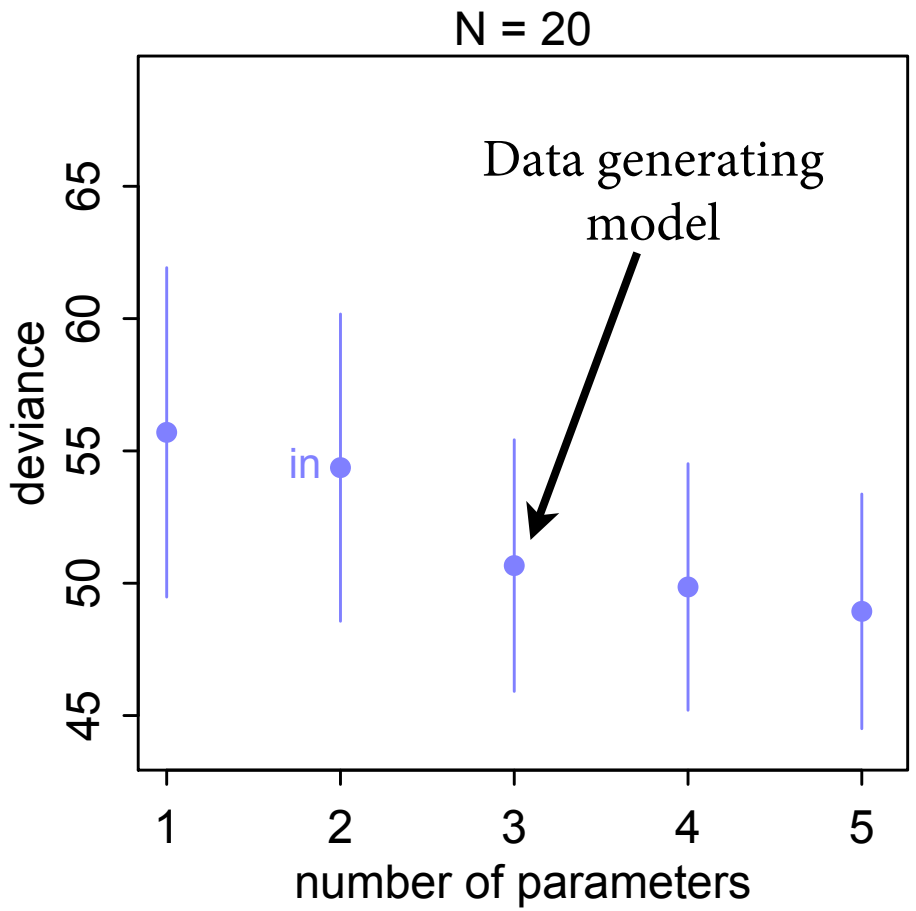


Figure 7.7

Everybody overfits

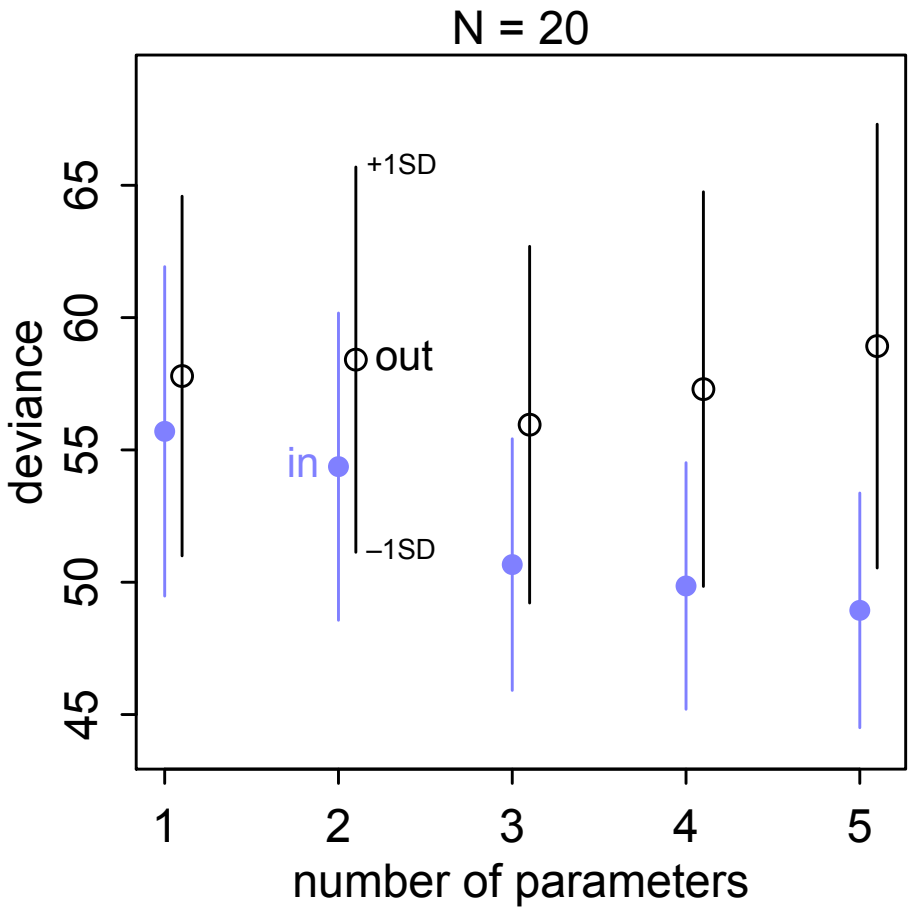


Figure 7.7

Everybody overfits

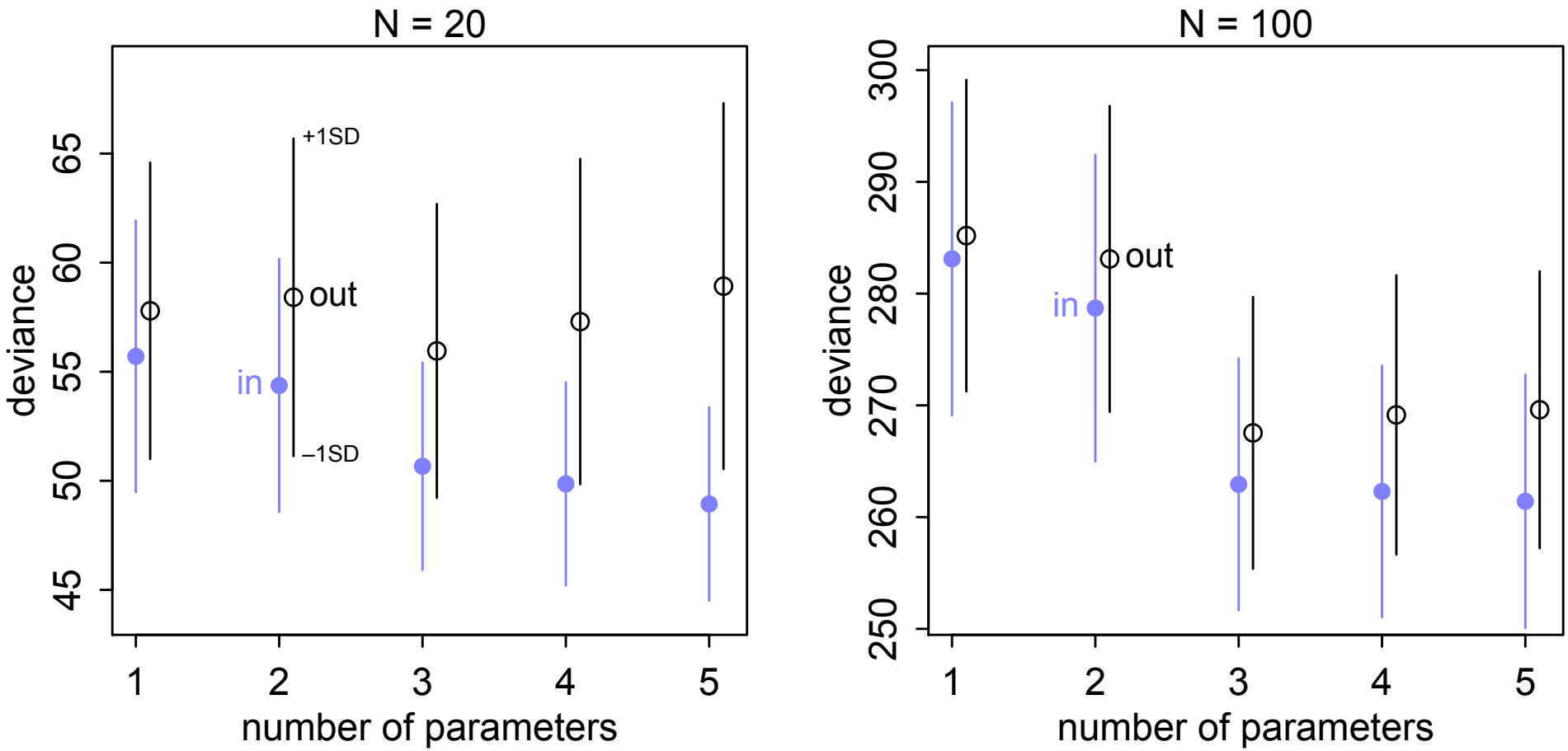


Figure 7.7

Regularization

- **Must be skeptical of the sample!**
- Use informative, conservative priors to reduce overfitting => model learns less from sample
- But if too skeptical, model learns too little
- Such priors are *regularizing*

Regularization

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim \text{Normal}(0, 100)$$
prior ⇢ $$\beta \sim \text{Normal}(0, 1)$$
$$\sigma \sim \text{Uniform}(0, 10)$$

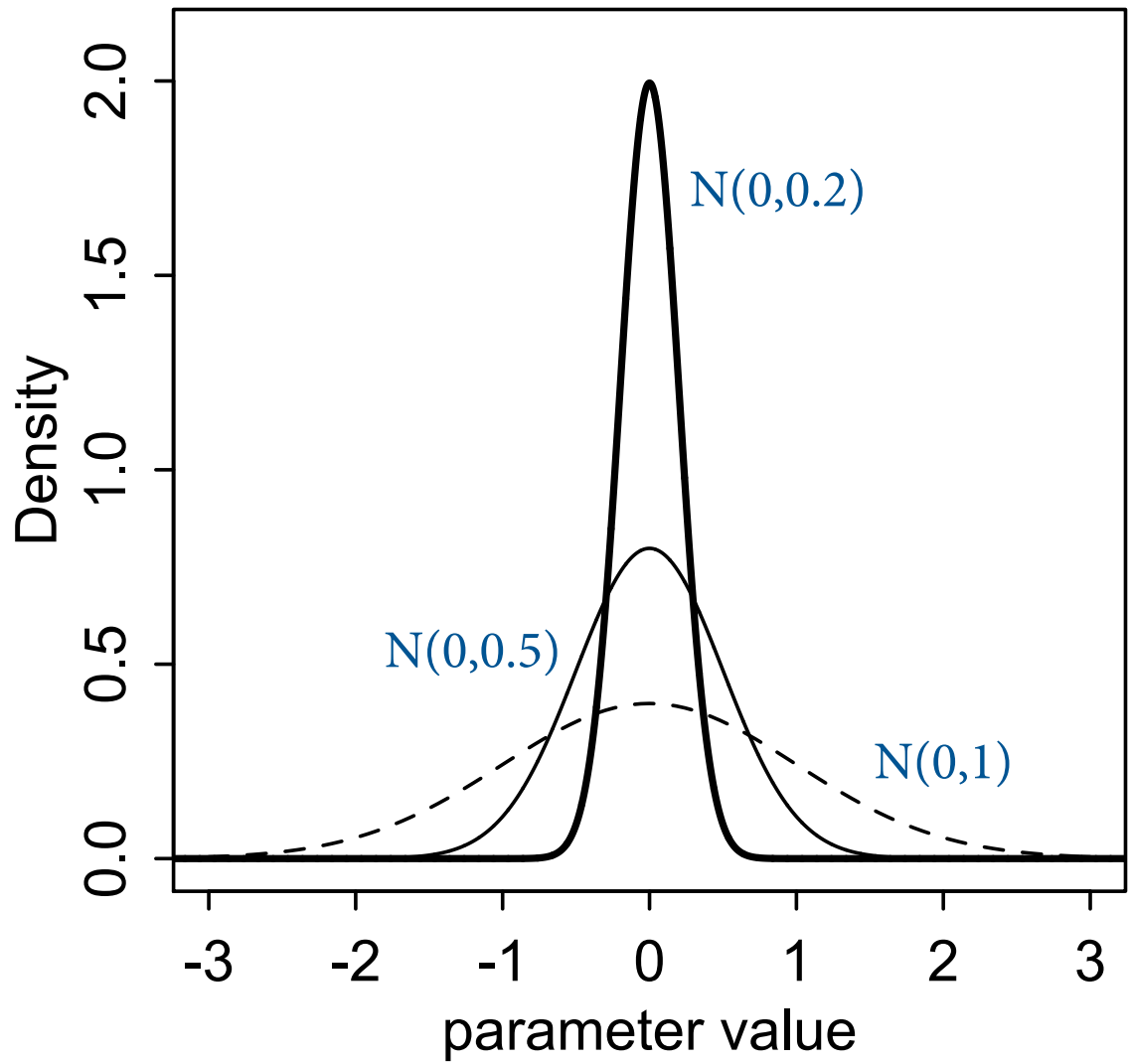
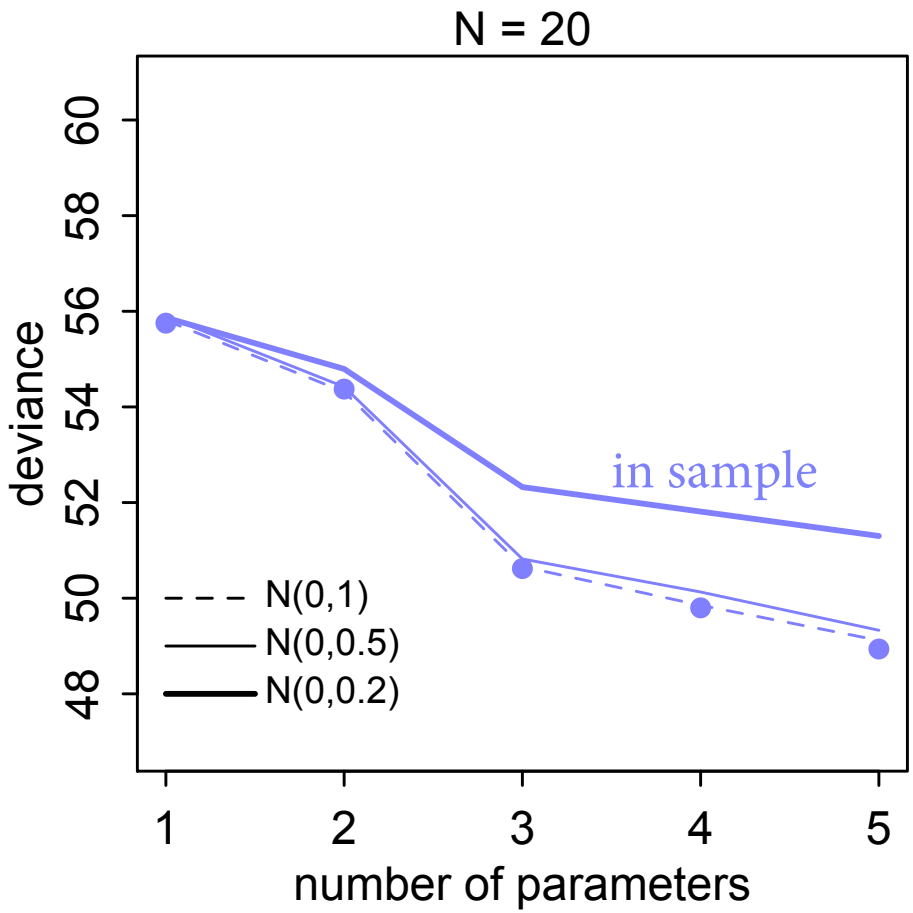


Figure 7.8

Regularization



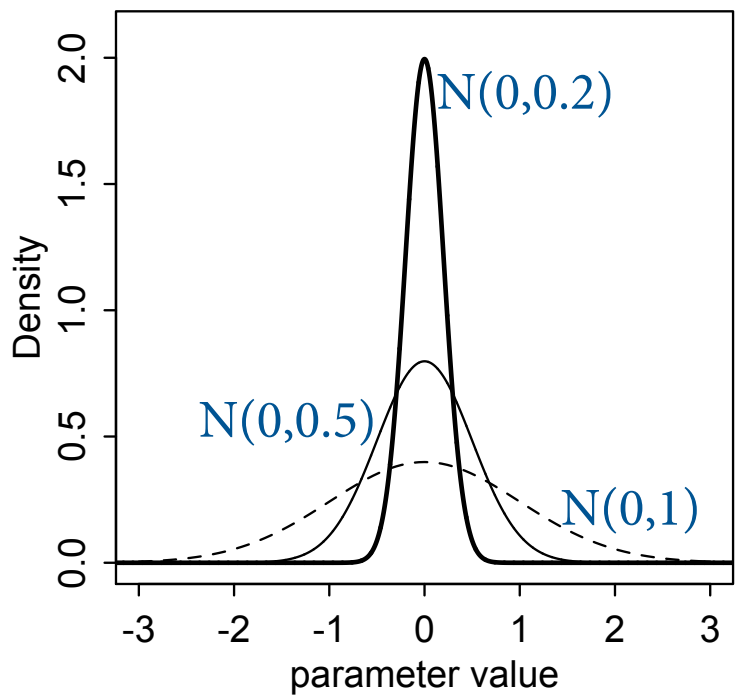
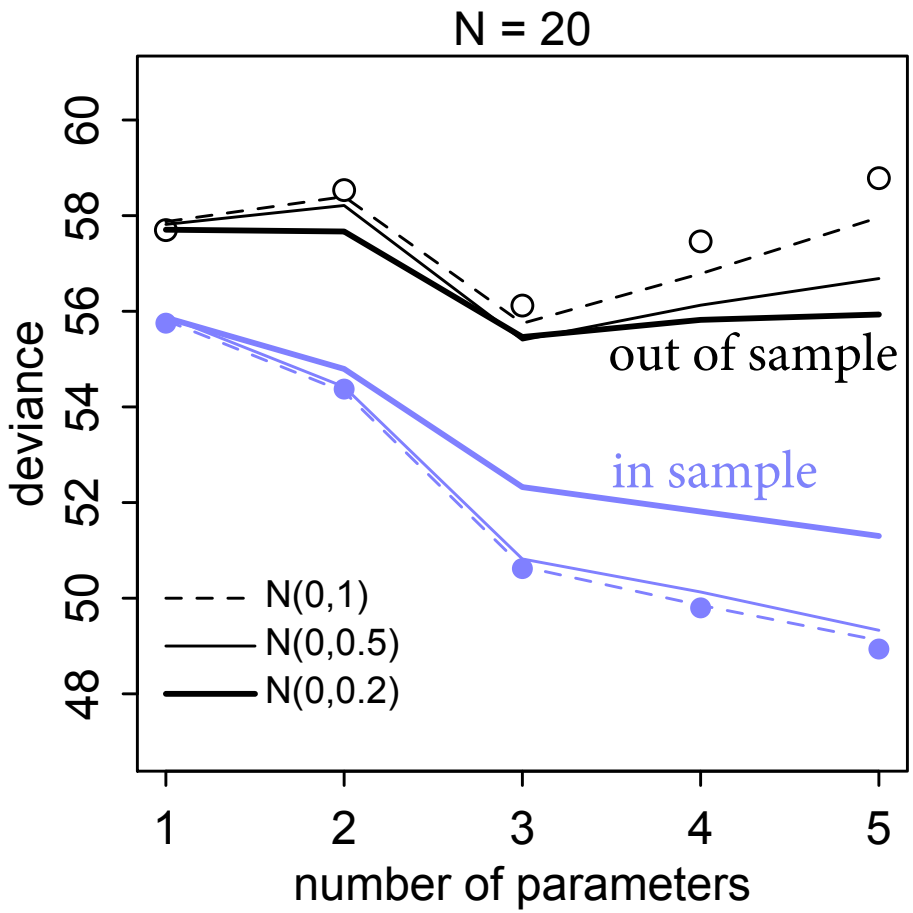


Figure 7.9

Regularization



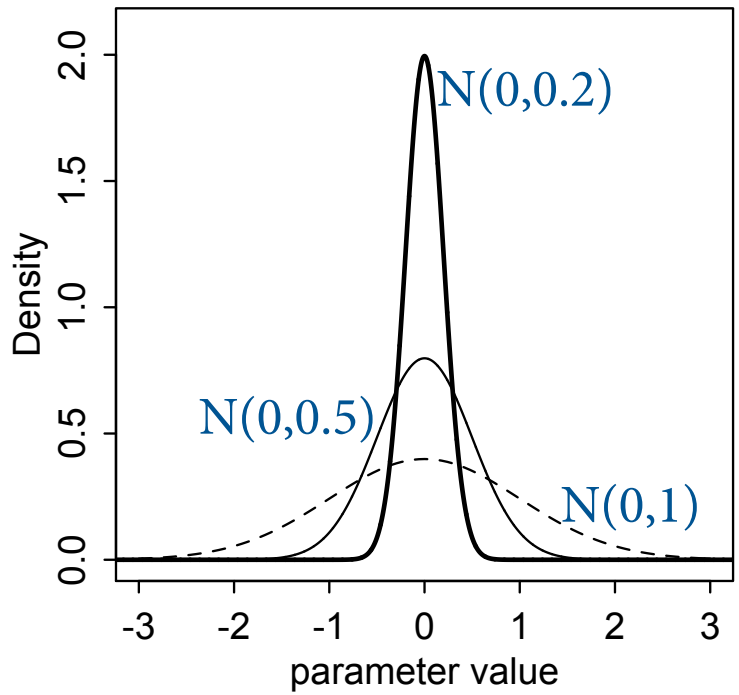


Figure 7.9

Regularization

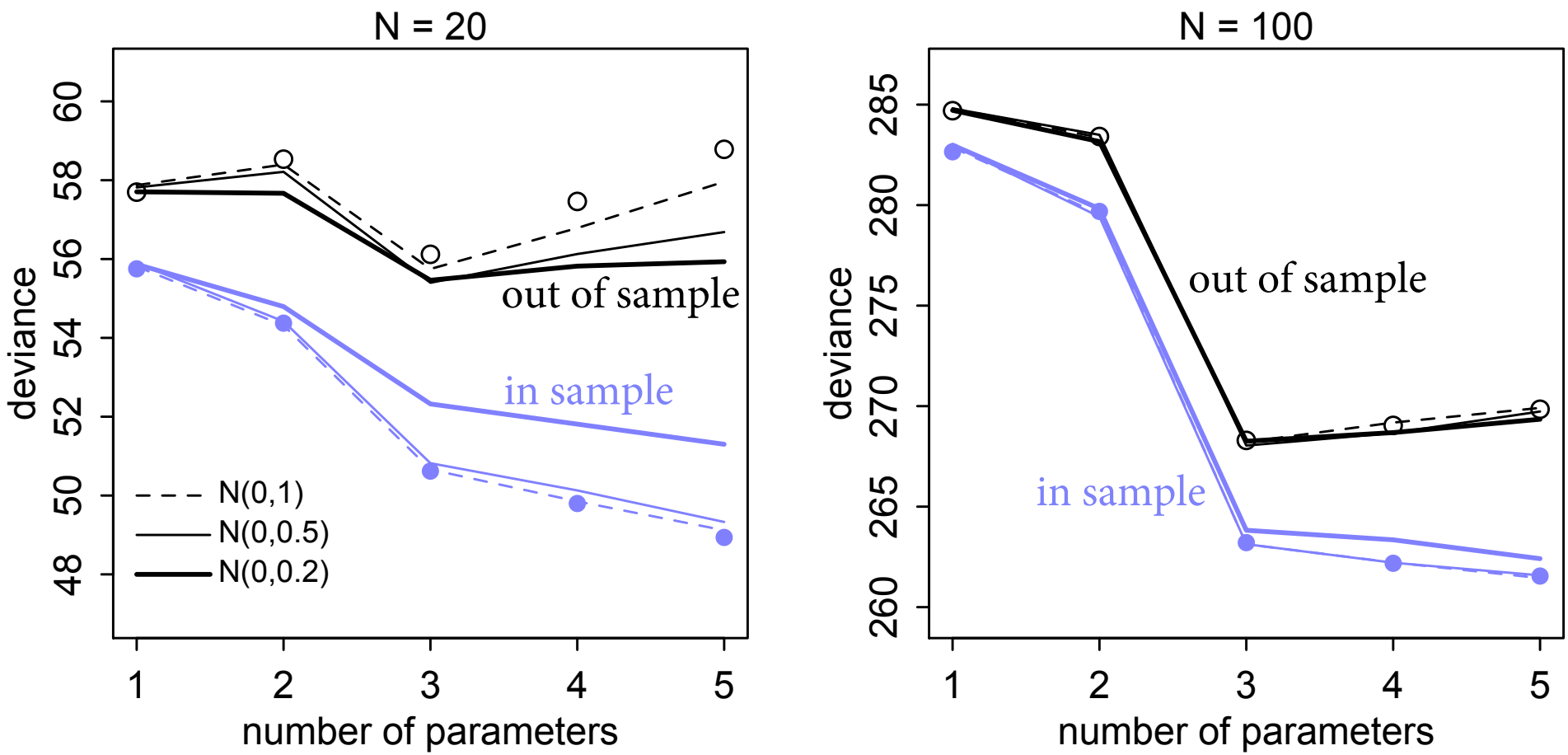


Figure 7.9

Regularization Rare in Science

- Why don't more scientists regularize?
- Never taught it
- Makes significant results rarer
- Most scientists judged not on predictive accuracy

Cross-validation & Information criteria

- Can we estimate out-of-sample deviance?
- In theory: Cross-validation
- Also in theory: Information criteria
- Both tend to perform similarly

Cross-validation

- Leave out some observations
- Train on remaining; score on those left out
- Average over many leave-out sets is estimate of out-of-sample accuracy

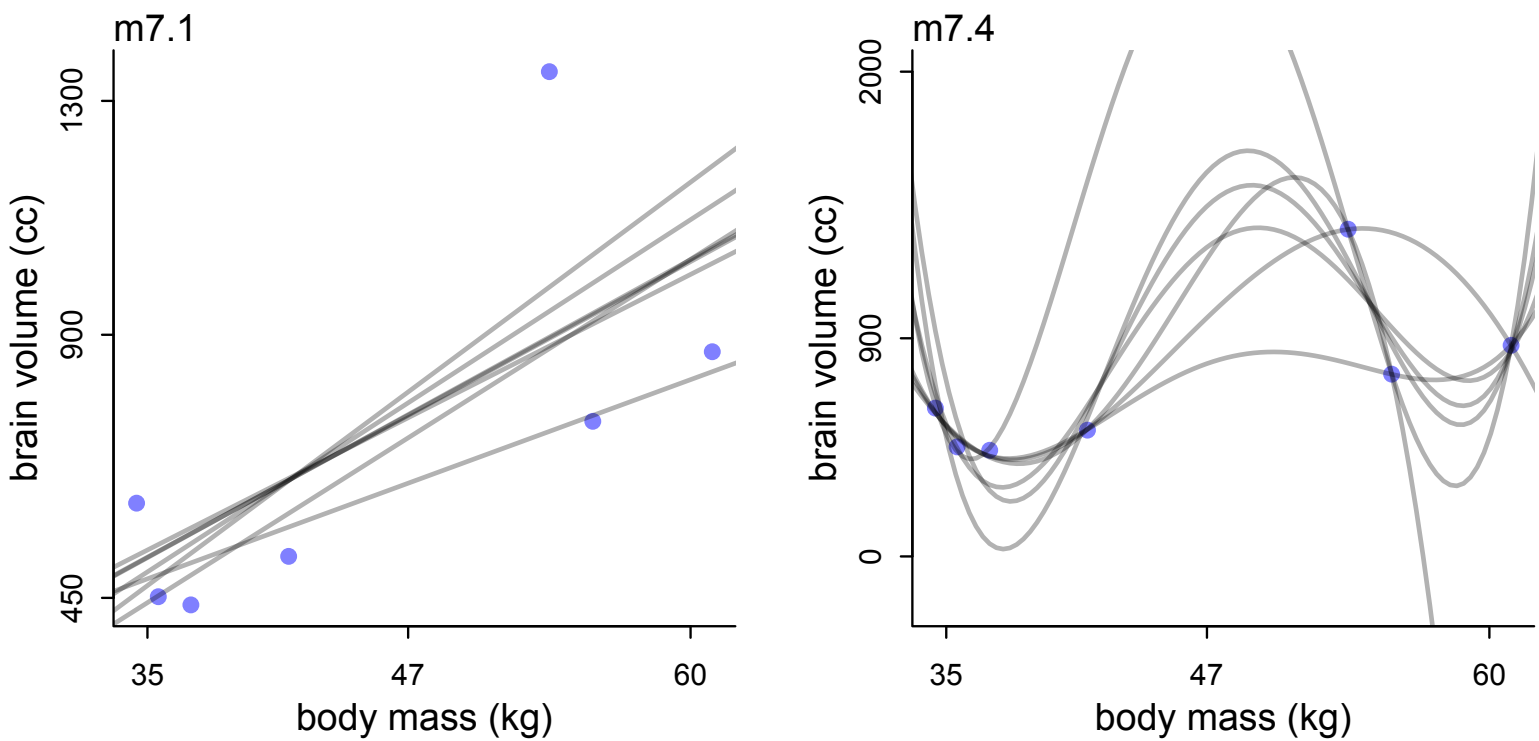


Figure 7.5

Smooth Cross-validation

- Most common: Leave-one-out
- Very expensive!
- Useful approximation: Importance sampling (IS)
- More useful: Pareto-smoothed importance sampling (PSIS)
- PSIS-LOO accurate, lots of useful diagnostics
- `LOO` function in rethinking
- See also `loo` package

Prof Aki Vehtari (Helsinki), smooth estimator

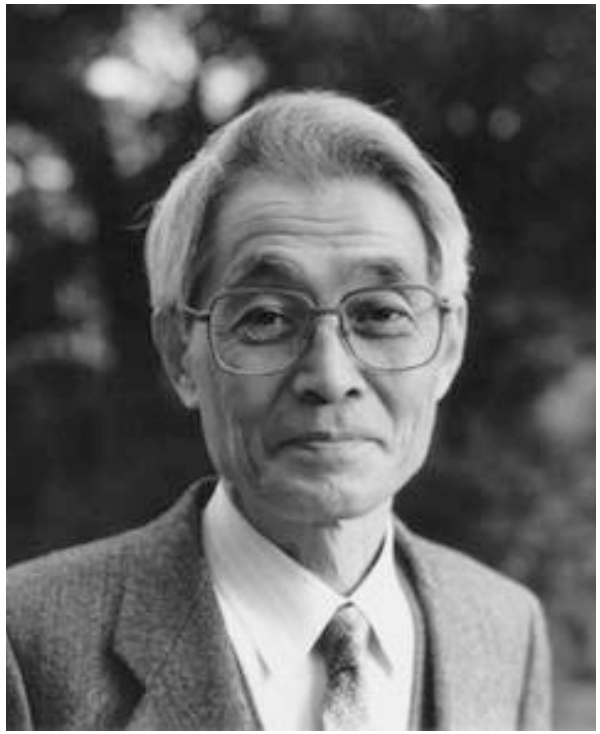
Akaike information criterion

[ah–ka–ee–kay]

- Estimate K-L Distance in theory
- Most famous is the first, AIC
- Under some strict conditions:

$$\mathrm{AIC} = D_{\text{train}} + 2k \approx \mathrm{E}\, D_{\text{test}}$$

k is parameter count



Hirotugu Akaike

赤池弘次

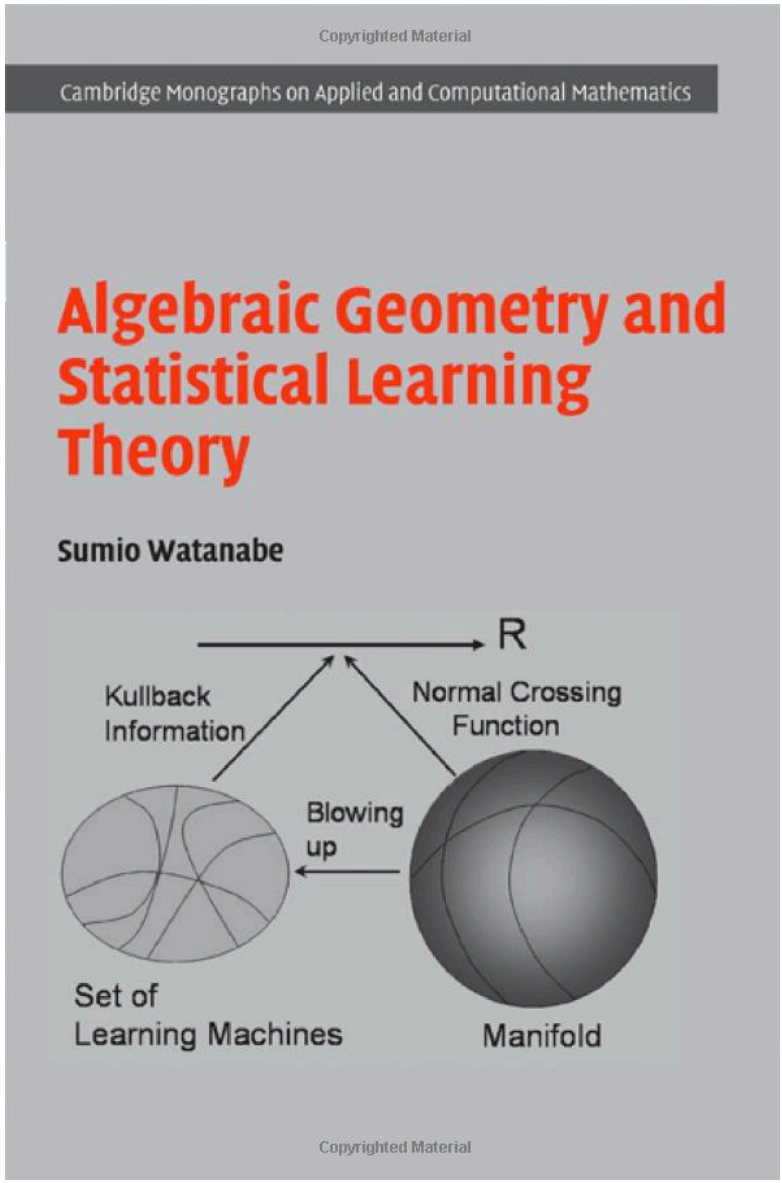
(1927–2009)

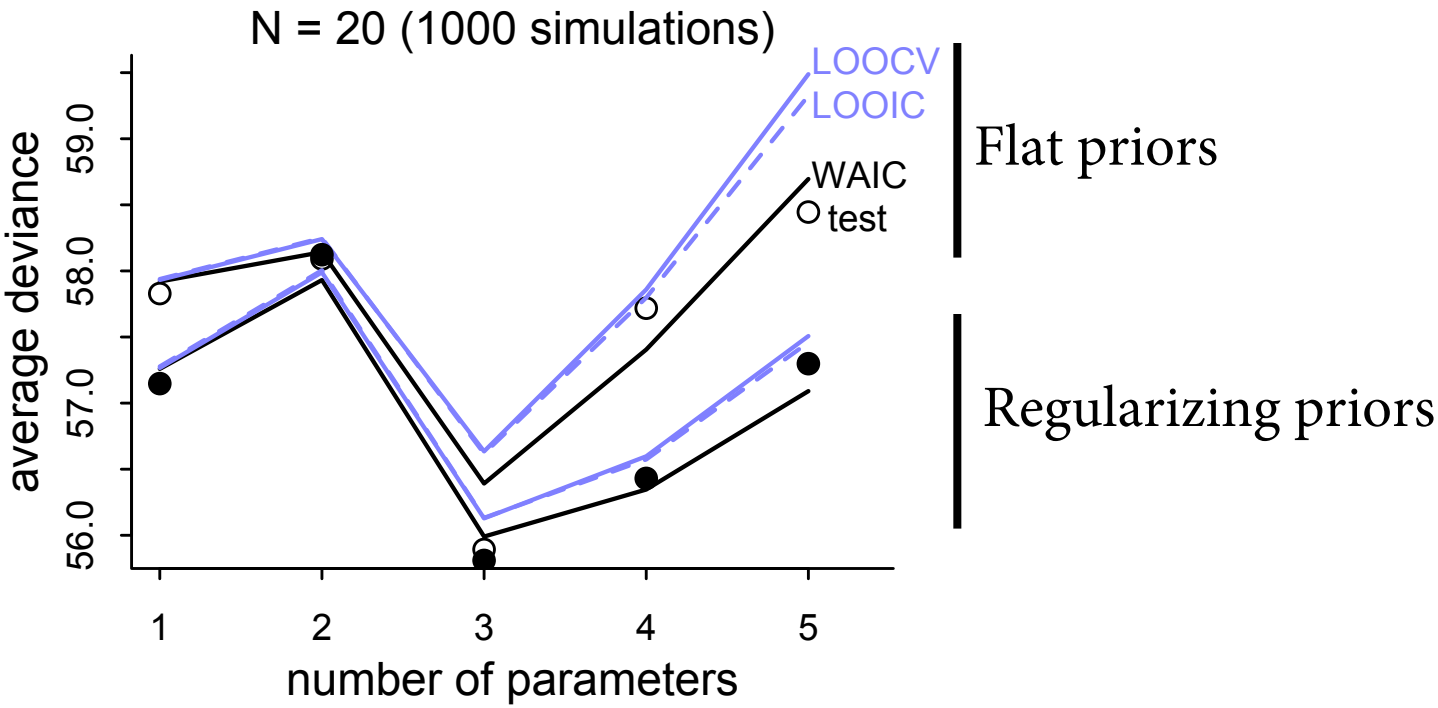
Widely Applicable IC

- AIC of historical interest now
- Widely Applicable Information Criterion (WAIC)
  - Sumio Watanabe 2010

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

- Does not assume Gaussian posterior
- `WAIC` function in rethinking





Compare out-of-sample only:

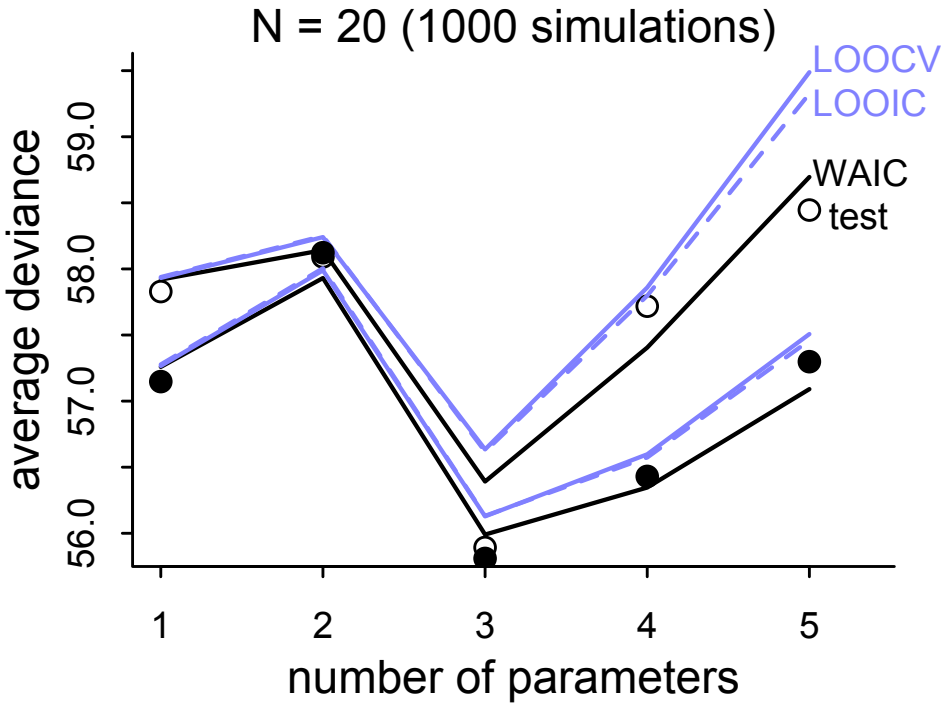
LOOCV: Actual leave-one-out CV

LOOIC: PSIS-LOO

WAIC: Widely Applicable IC

Points are actual scores

Figure 7.10


N = 20 (1000 simulations)
average deviance
56.0
57.0
58.0
59.0
1
2
3
4
5
number of parameters
LOOCV
LOOIC
WAIC
test

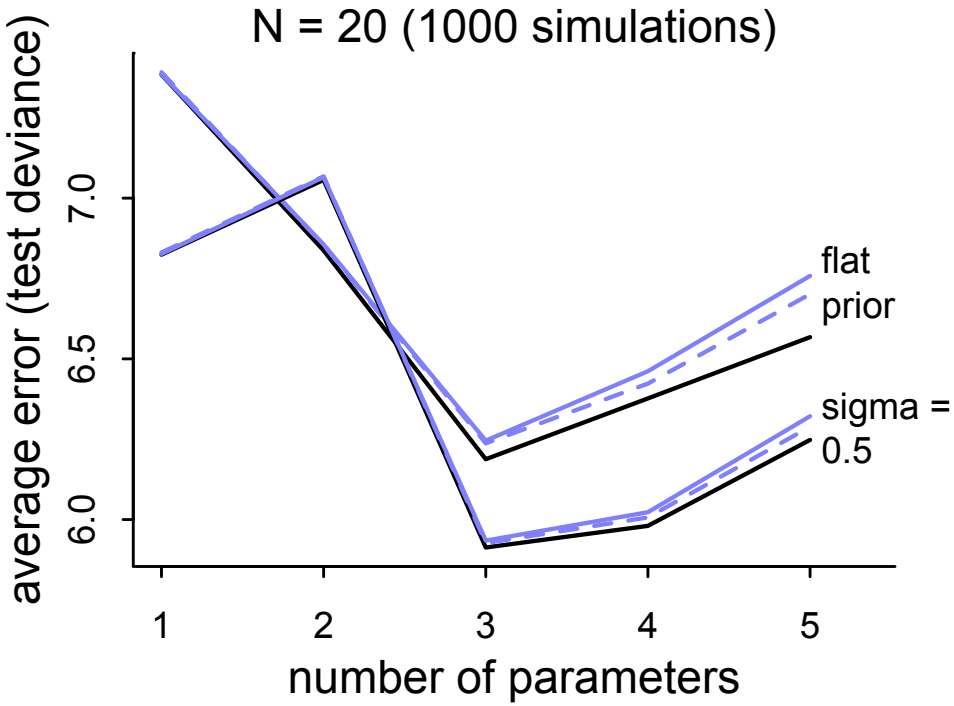

N = 20 (1000 simulations)
average error (test deviance)
6.0
6.5
7.0
1
2
3
4
5
number of parameters
flat
prior
sigma =
0.5

Figure 7.10

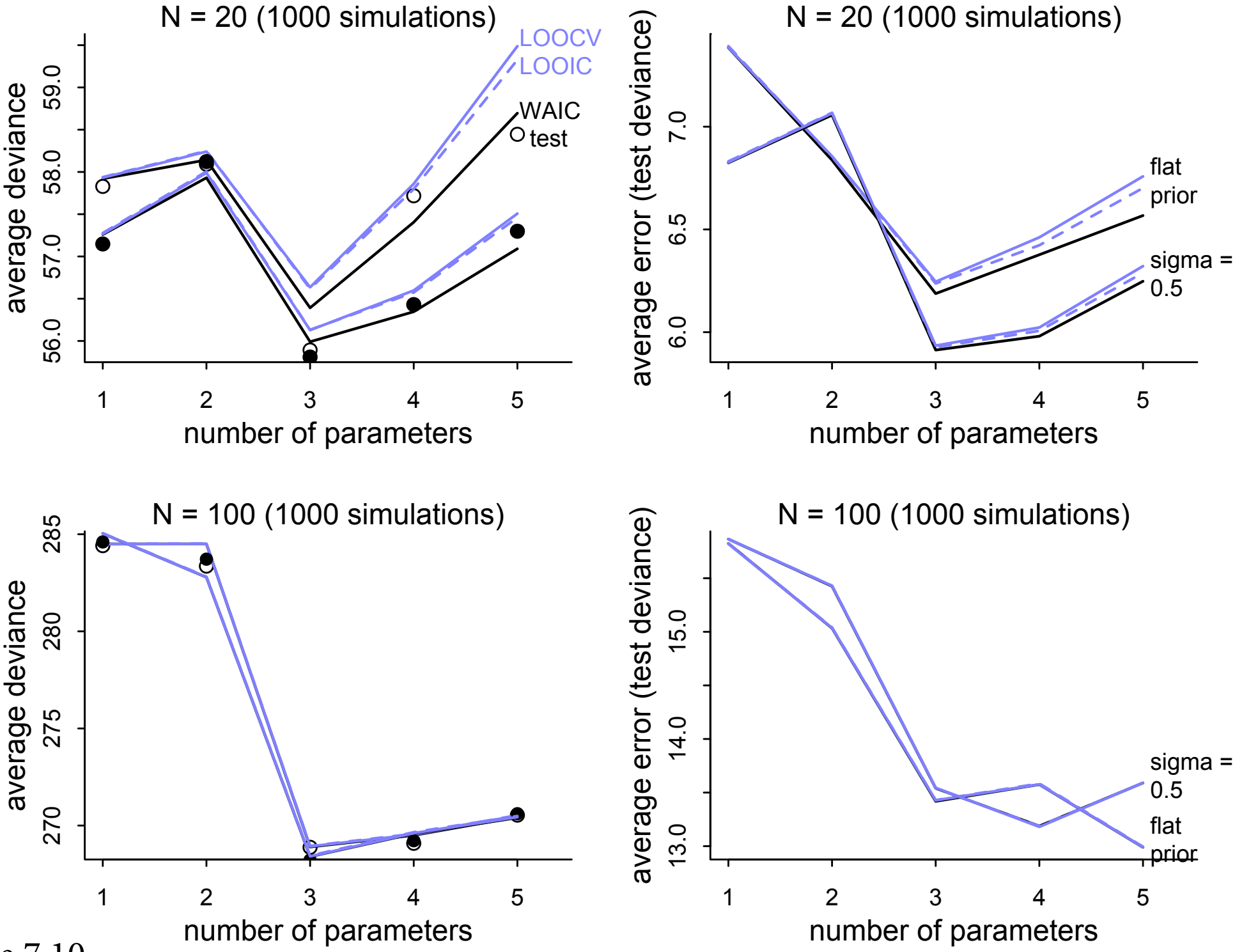
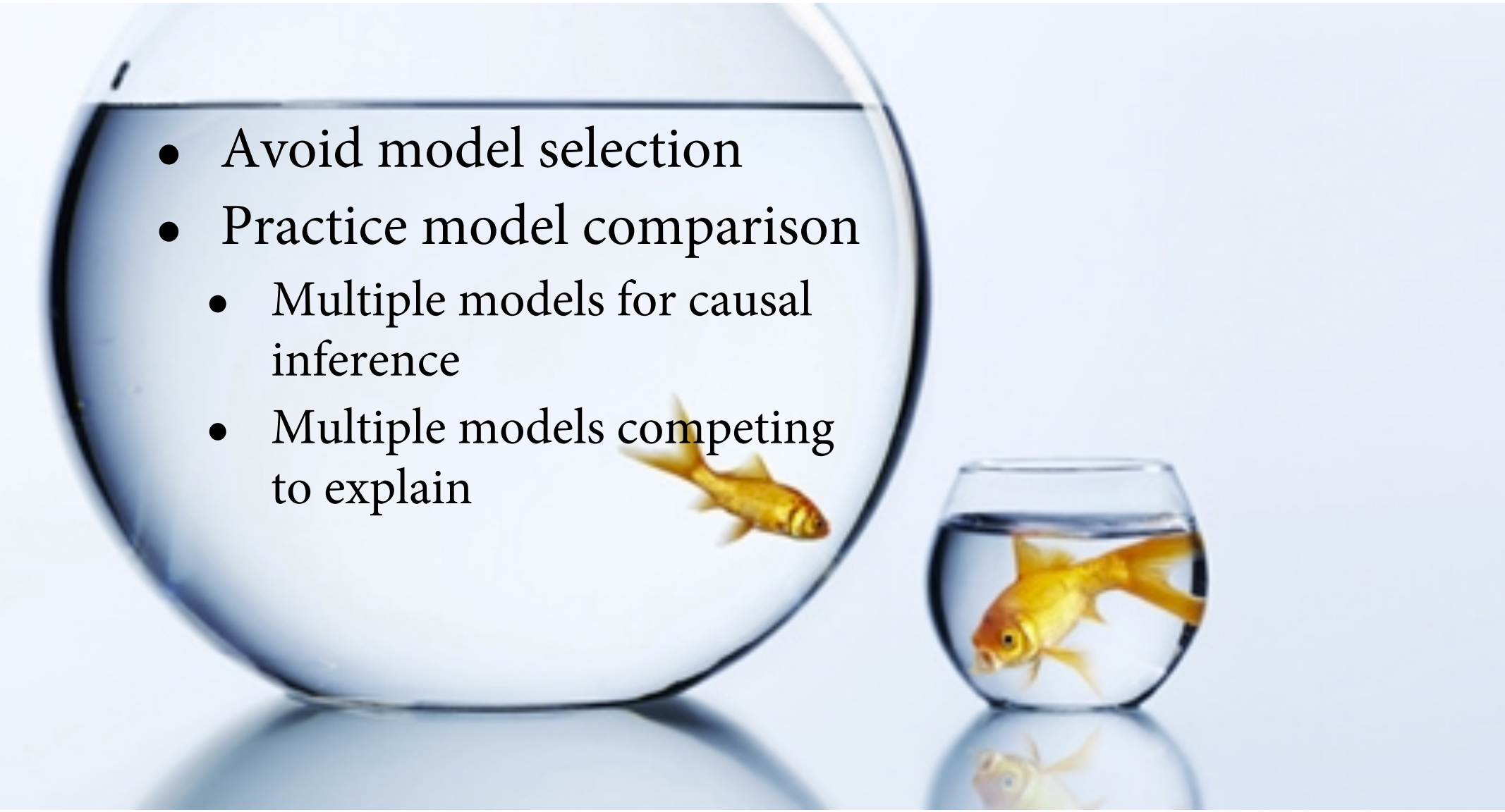

N = 20 (1000 simulations)
average deviance
56.0
57.0
58.0
59.0
LOOCV
LOOIC
WAIC
test
1
2
3
4
5
number of parameters
N = 20 (1000 simulations)
average error (test deviance)
6.0
6.5
7.0
flat
prior
sigma =
0.5
1
2
3
4
5
number of parameters
N = 100 (1000 simulations)
average deviance
270
275
280
285
1
2
3
4
5
number of parameters
N = 100 (1000 simulations)
average error (test deviance)
13.0
14.0
15.0
sigma =
0.5
flat
prior
1
2
3
4
5
number of parameters

Figure 7.10

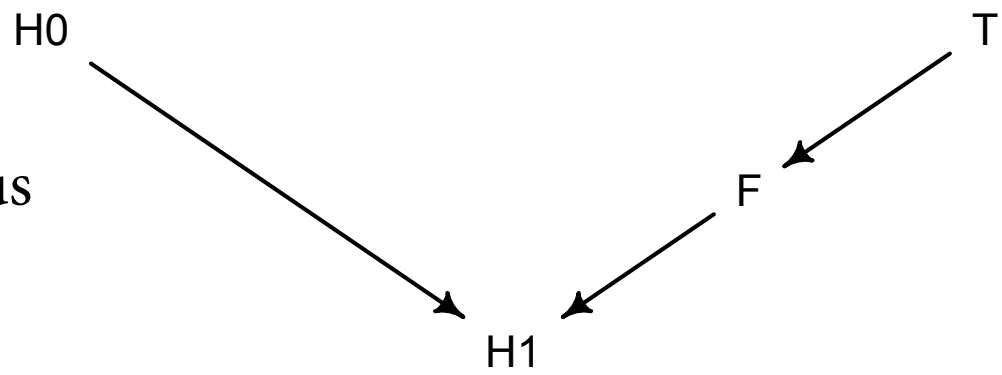
Using CV & WAIC



- Avoid model selection
- Practice model comparison
  - Multiple models for causal inference
  - Multiple models competing to explain

Example: Model Mis-selection

- Model comparison is not causal inference
- Recall plant/fungus example from last week
- Three models
  - m6.6: intercept only
  - m6.7: treatment + fungus
  - m6.8: treatment only



R code 7.27

```
set.seed(77)
compare( m6.6 , m6.7 , m6.8 )
```

```
      WAIC pWAIC dWAIC weight    SE   dSE
m6.7 361.9   3.8   0.0      1 14.26    NA
m6.8 402.8   2.6  40.9      0 11.28 10.48
m6.6 405.9   1.6  44.0      0 11.66 12.23
```

R code 7.27

```
set.seed(77)
compare( m6.6 , m6.7 , m6.8 )
```

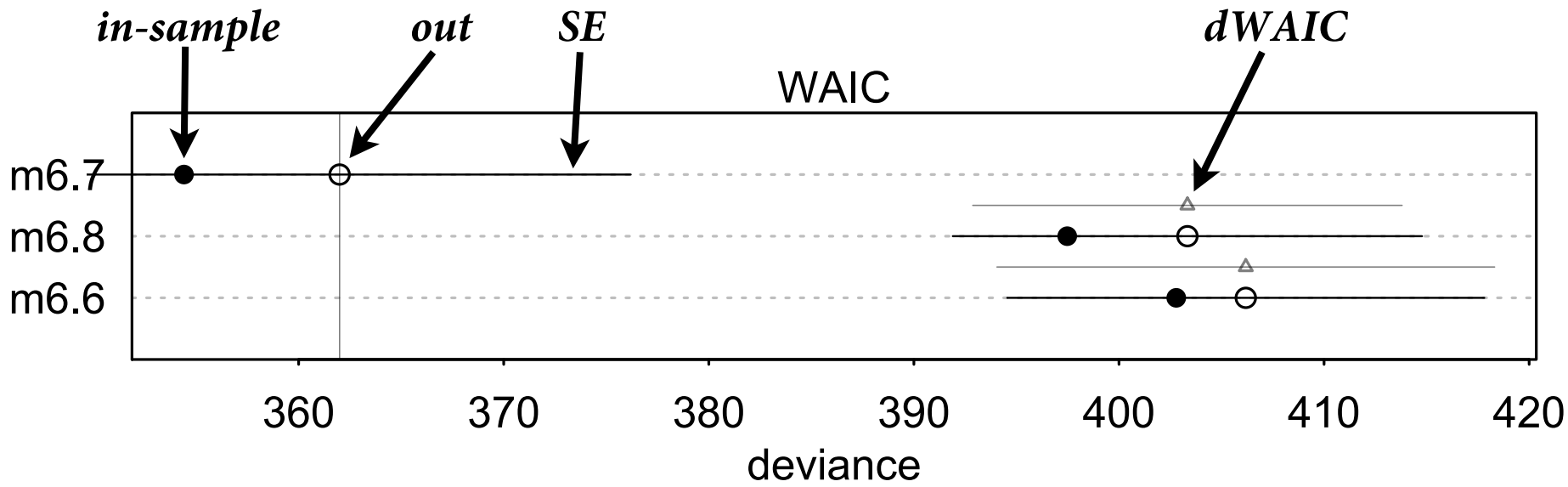
```
                 WAIC pWAIC dWAIC weight    SE   dSE
treat + fungus m6.7 361.9   3.8   0.0       1 14.26    NA
       fungus  m6.8 402.8   2.6  40.9       0 11.28 10.48
     intercept m6.6 405.9   1.6  44.0       0 11.66 12.23
```

- WAIC: estimated out-of-sample log-score
- pWAIC: penalty, "effective number of parameters"
- dWAIC: difference from top model
- weight: Akaike weight — see text for details
- SE: Standard error of WAIC
- dSE: Standard error of dWAIC

R code 7.27

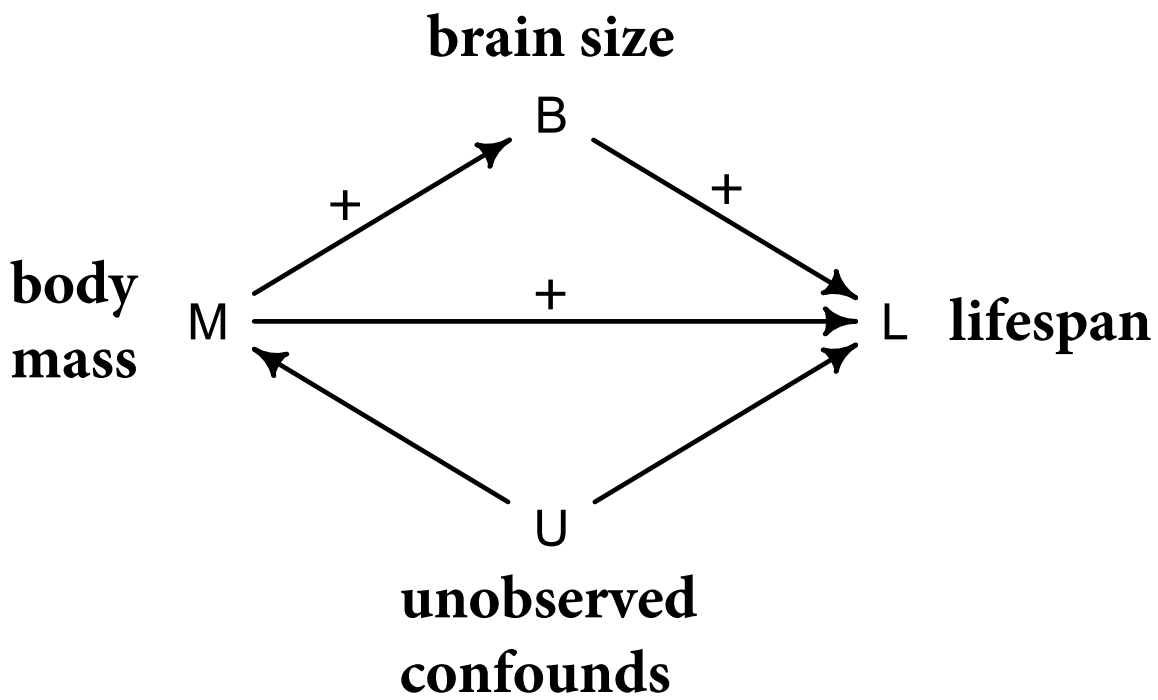
```
set.seed(77)
compare( m6.6 , m6.7 , m6.8 )
```

		WAIC	pWAIC	dWAIC	weight	SE	dSE
treat + fungus	m6.7	361.9	3.8	0.0	1	14.26	NA
fungus	m6.8	402.8	2.6	40.9	0	11.28	10.48
intercept	m6.6	405.9	1.6	44.0	0	11.66	12.23



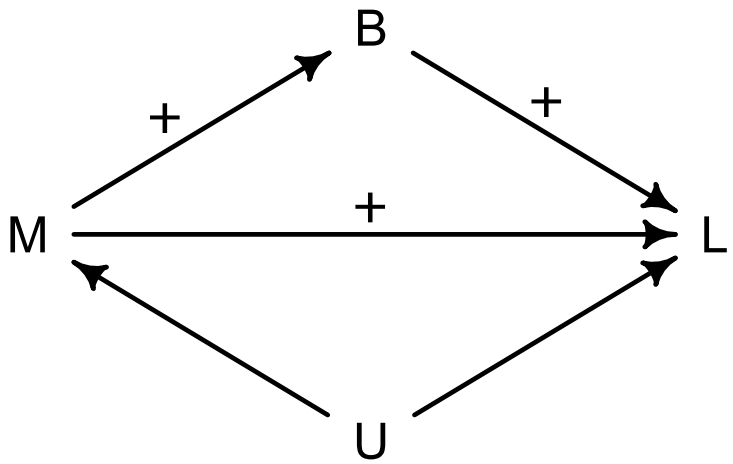
Something About *Cebus*

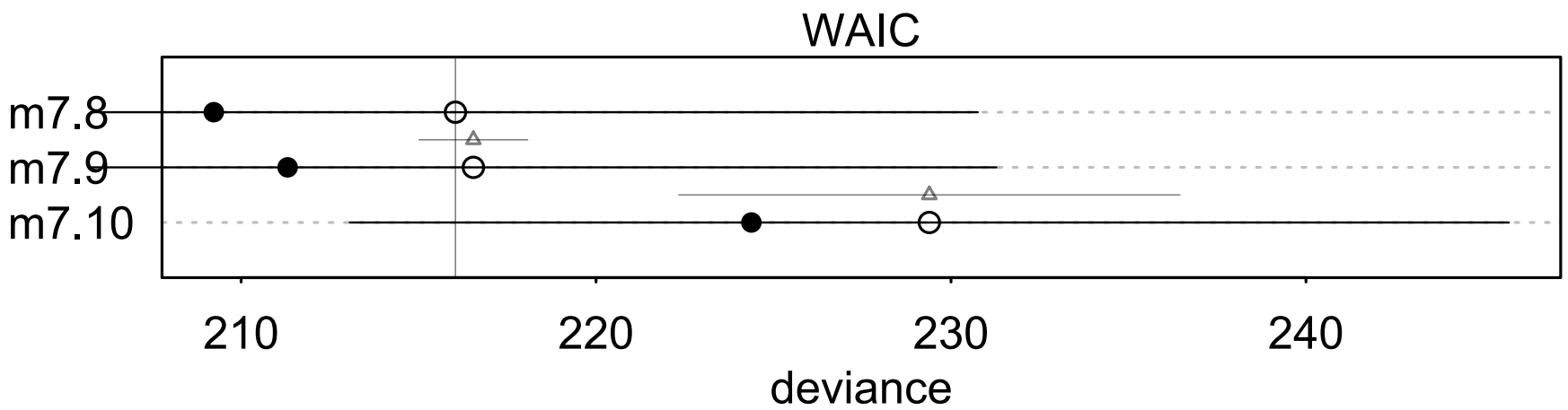
- Why do primates live a long time?
- Consider:



Something About *Cebus*

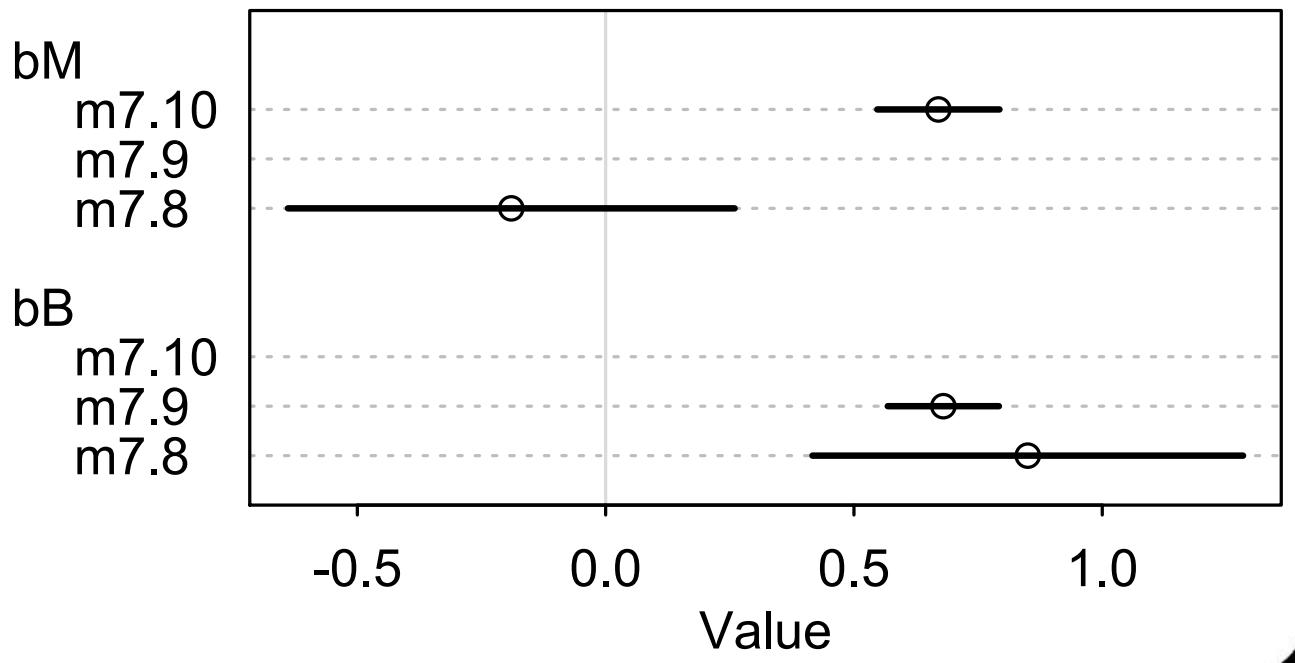
- 112 primate species
- Three models:
  - m7.8: log L ~ log M + log B
  - m7.9: log L ~ log B
  - m7.10: log L ~ log M
- Funny stuff happens

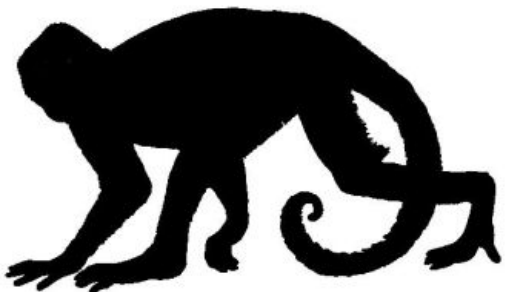




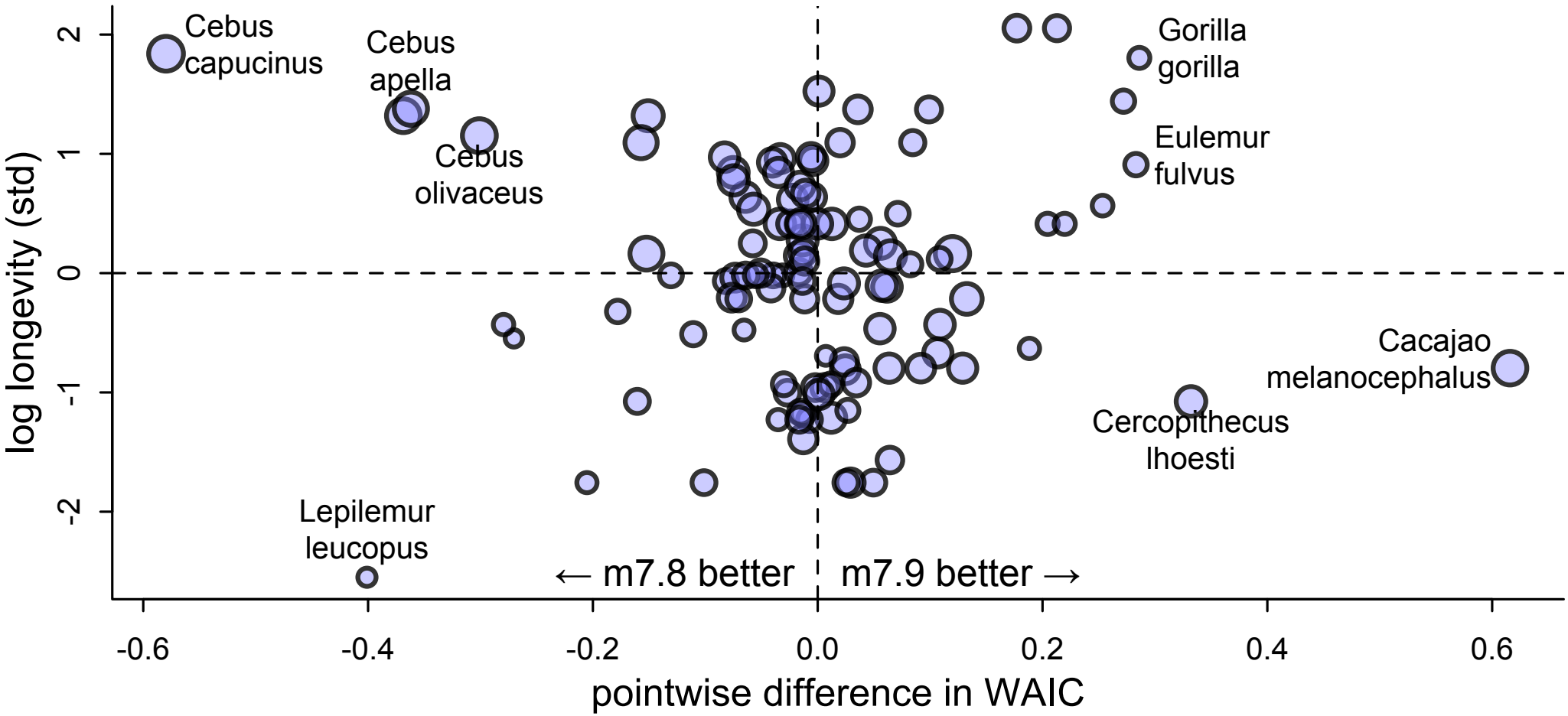
Something About *Cebus*

- Inspect the posterior distributions for answers
- Why does body mass go *negative* in joint model?





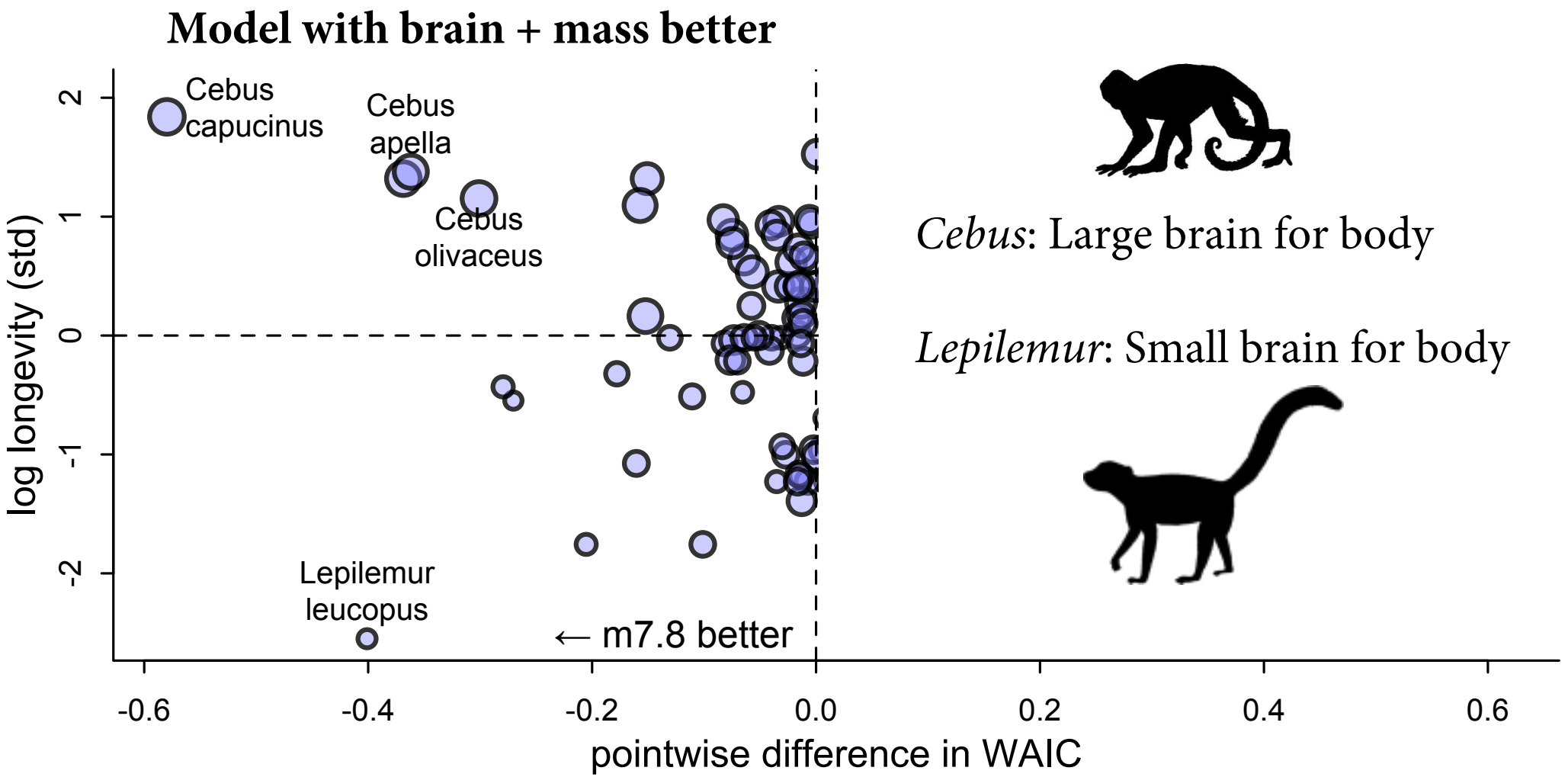
Pointwise perspective



Point size proportional to abs diff brain z-score – body z-score

Figure 7.12

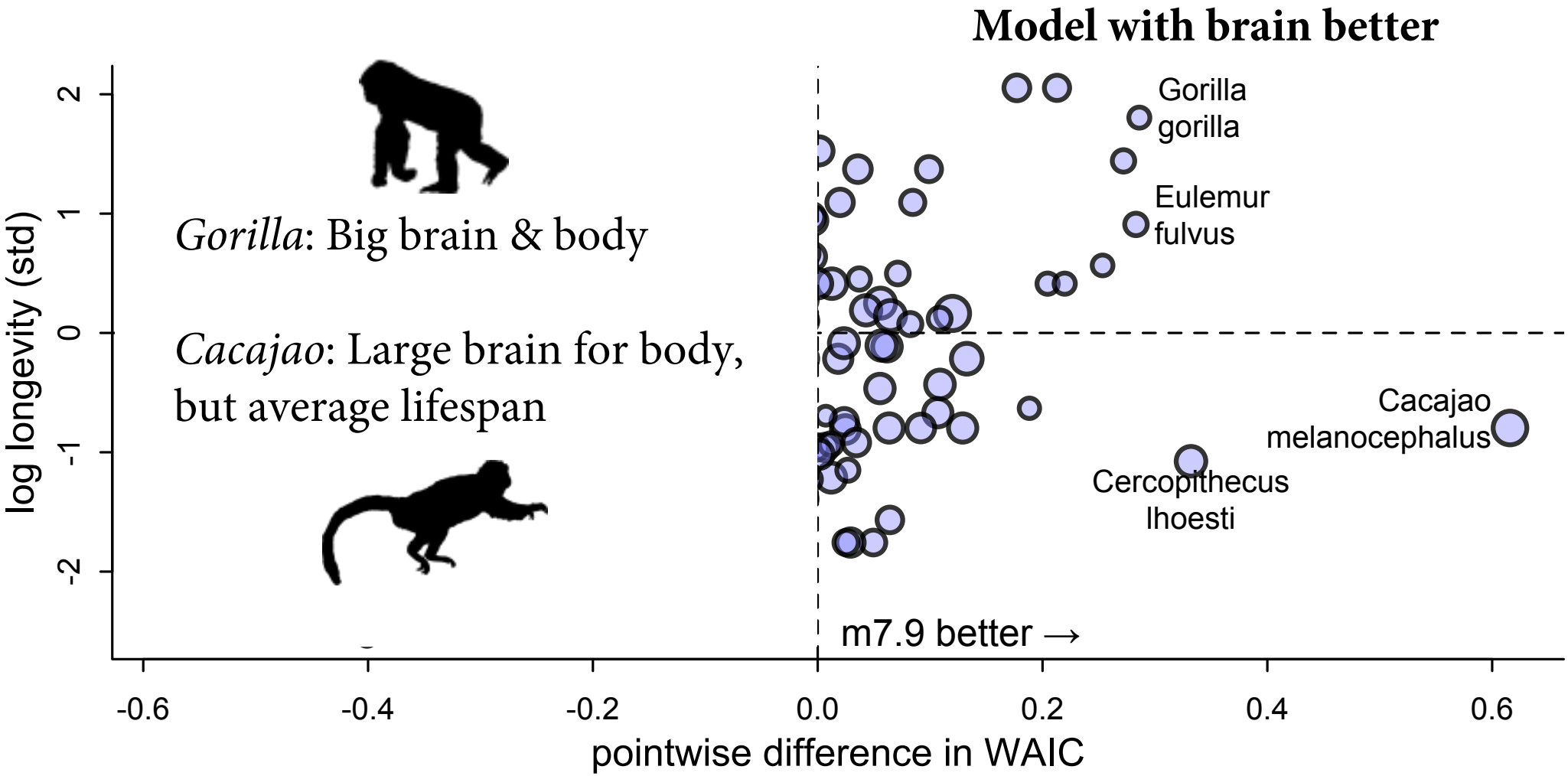
Pointwise perspective



Point size proportional to abs diff brain z-score – body z-score

Figure 7.12

Pointwise perspective

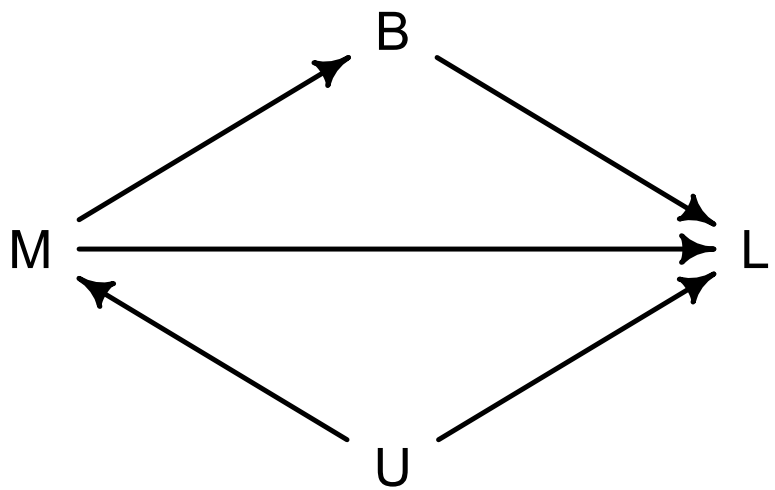


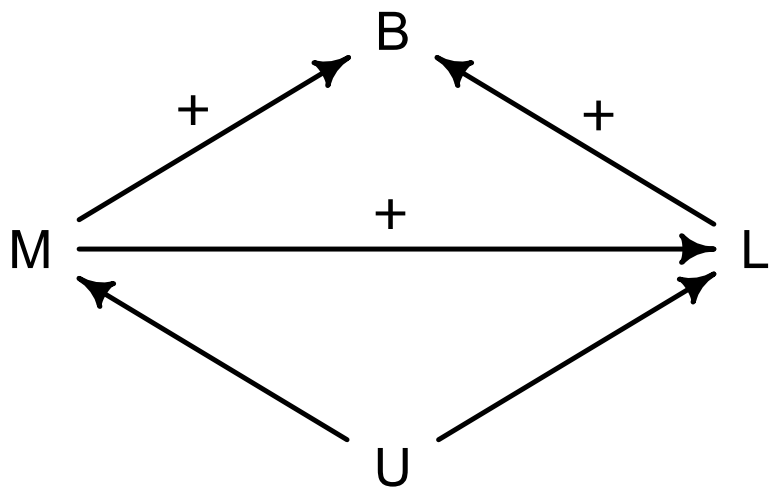
Point size proportional to abs diff brain z-score – body z-score

Figure 7.12

Cebus Collider

- Another idea: Reversal of body size coefficient consistent with **collider bias** (**ominous music**)
- Conditioning on brain opens backdoor path M –> B <– L



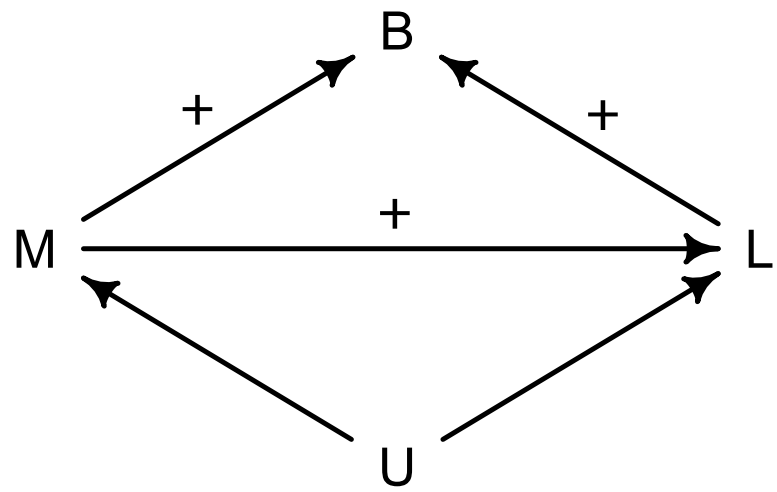


Cebus Collider

R code 7.45

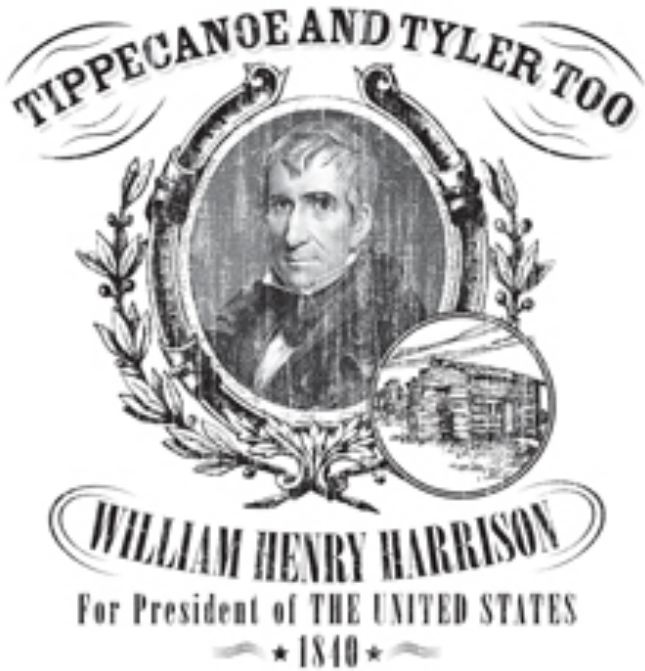
```
m7.11 <- quap(
    alist(
        log_B ~ dnorm( mu , sigma ),
        mu <- a + bM*log_M + bL*log_L,
        a ~ dnorm(0,0.1),
        bM ~ dnorm(0,0.5),
        bL ~ dnorm(0,0.5),
        sigma ~ dexp(1)
    ) , data=d2 )
precis( m7.11 )
```

```
        mean   sd  5.5% 94.5%
a      -0.05 0.02 -0.07 -0.02
bM      0.94 0.03  0.90  0.98
bL      0.12 0.03  0.07  0.16
sigma   0.19 0.01  0.17  0.21
```



Curse of Tippecanoe

- 1840–1960: Every US president elected in year ending in digit "0" died in office
  - W. H. Harrison first, "Old Tippecanoe"
  - Lincoln, Garfield, McKinley, Harding, FD Roosevelt
  - J. F. Kennedy last, assassinated in 1963
  - Reagan broke the curse!
- Trying all possible models: A formula for overfitting
  - Be thoughtful
  - Be honest: Admit data exploration



Onwards!

- Be patient: We'll keep using these tools in future
- Homework 4: Entropy, happiness, and foxes
- Next week: Interactions, MCMC
- Coming up: Maximum entropy, generalized linear models, multilevel models