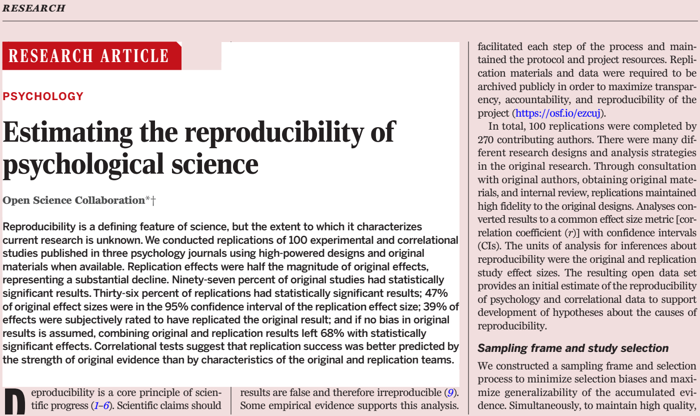
Chapter 6

Overfitting



RESEARCH

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Reproducibility is a core principle of scientific progress (1–6). Scientific claims should

results are false and therefore irreproducible (9). Some empirical evidence supports this analysis.

facilitated each step of the process and maintained the protocol and project resources. Replication materials and data were required to be archived publicly in order to maximize transparency, accountability, and reproducibility of the project (https://osf.io/ezcuj).

In total, 100 replications were completed by 270 contributing authors. There were many different research designs and analysis strategies in the original research. Through consultation with original authors, obtaining original materials, and internal review, replications maintained high fidelity to the original designs. Analyses converted results to a common effect size metric [correlation coefficient (r)] with confidence intervals (CIs). The units of analysis for inferences about reproducibility were the original and replication study effect sizes. The resulting open data set provides an initial estimate of the reproducibility of psychology and correlational data to support development of hypotheses about the causes of reproducibility.

Sampling frame and study selection

We constructed a sampling frame and selection process to minimize selection biases and maximize generalizability of the accumulated evidence. Simultaneously, to maintain high quality,

Figure: Science, 28 Aug 2015: 35 of 97 experiment could be reproduced!

Overfitting

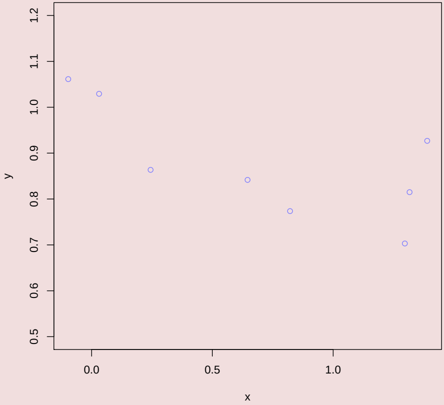


Figure: the data

Overfitting

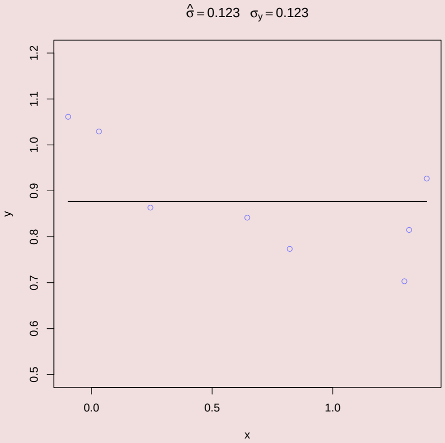


Figure: MAP estimate of $\mu = \beta_0$

Overfitting

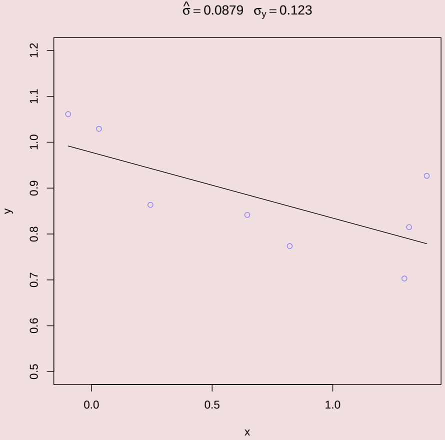


Figure: MAP estimate of $\mu = \beta_0 + x\beta_1$

Overfitting

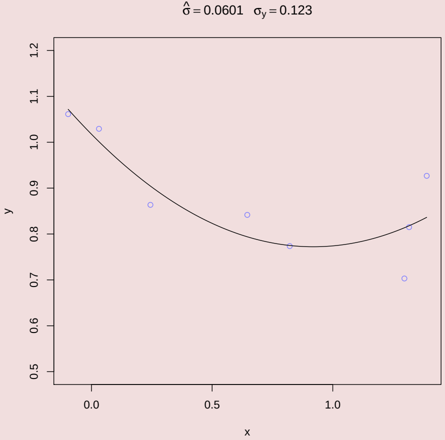


Figure: MAP estimate of $\mu = \beta_0 + x\beta_1 + x^2\beta_2$

Overfitting

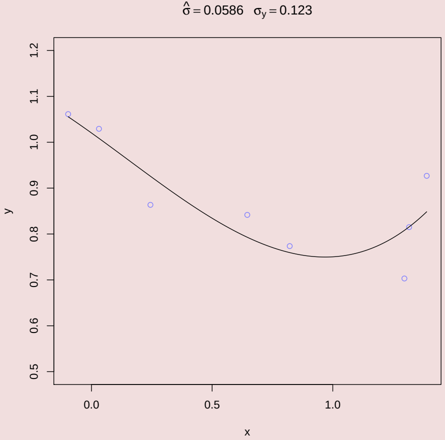


Figure: MAP estimate of $\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3$

Overfitting

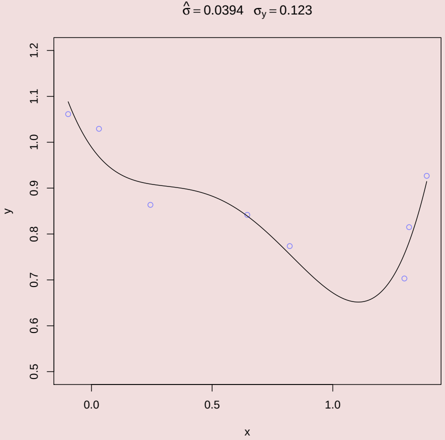


Figure: MAP estimate of $\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4$

Overfitting

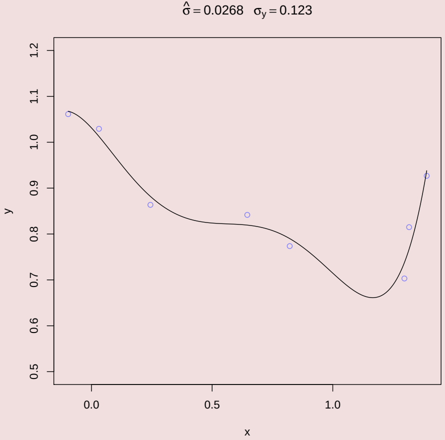


Figure: MAP estimate of $\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4 + x^5\beta_5$

Overfitting

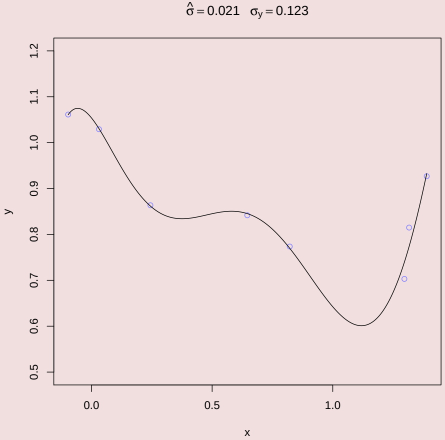


Figure: MAP estimate of
$\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4 + x^5\beta_5 + x^6\beta_6$

Removing one observations

- Loop over all data $j = 1, \ldots, n$:
- Remove one observation, y_j.
- Estimate β using $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n$.
- Estimate σ using $\sqrt{\frac{1}{n} \sum (y_j - \hat{y}_j)^2}$.

Leave one out

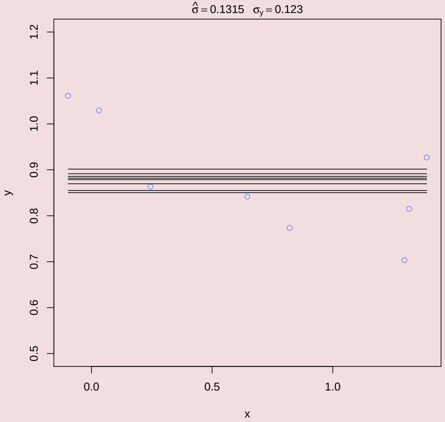


Figure: Leave one out estimate of $\mu = \beta_0$

Leave one out

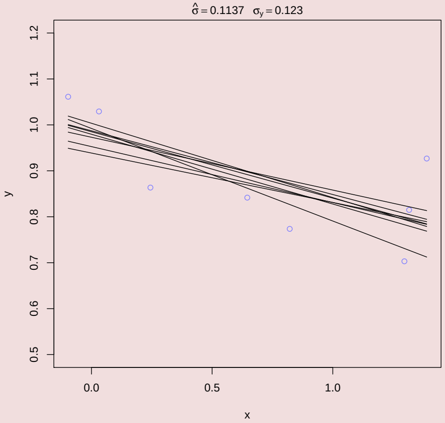


Figure: Leave one out of $\mu = \beta_0 + x\beta_1$

Leave one out

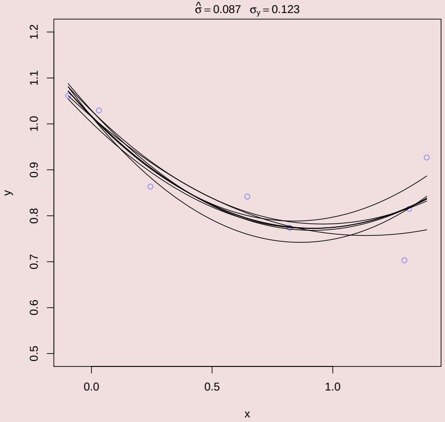


Figure: Leave one out of $\mu = \beta_0 + x\beta_1 + x^2\beta_2$

Leave one out

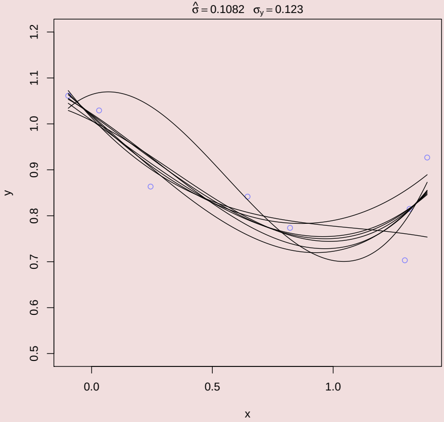


Figure: Leave one out of $\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3$

Leave one out

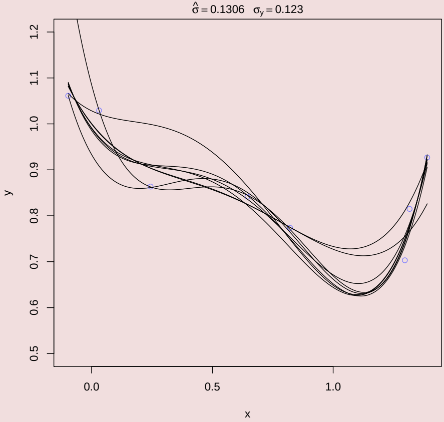


Figure: Leave one out of $\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4$

Leave one out

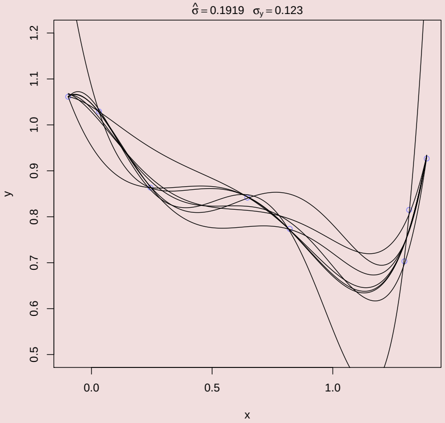


Figure: Leave one out of $\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4 + x^5\beta_5$

Leave one out

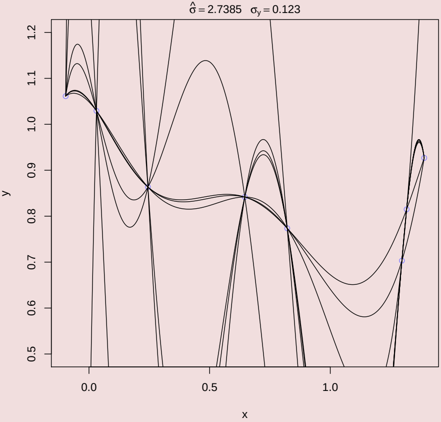


Figure: Leave one out of
$\mu = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4 + x^5\beta_5 + x^6\beta_6$

Beyond linear model

- For linear, there exists many different measure R^2_{adj}.

Beyond linear model

- For linear, there exists many different measure R^2_{adj}.
- For a general density/probability function what indicates a good fit?

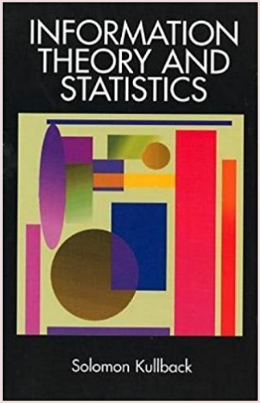
Beyond linear model

- For linear, there exists many different measure R^2_{adj}.
- For a general density/probability function what indicates a good fit?
- Answer: $p(y_i)$, or equivalently $\log(p(y_i))$

AIC

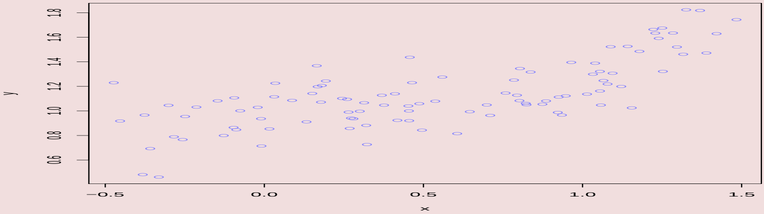
- $AIC = -2\sum_{i=1}^{n}\log(p(y_i)) + 2d$,
  d the number of parameters in the model.
- The smaller the better.

Information theory, Kullback-Leibler divergence



- Theoretical reasons why AIC good, based on information theory.
- Beyond the scope of the course, interested read Kullback book.

AIC, back to linear model



- True model:

$$y_i = 1 + x0.1 - x^2 0.3 + 0.4x^3 + \epsilon_i,$$

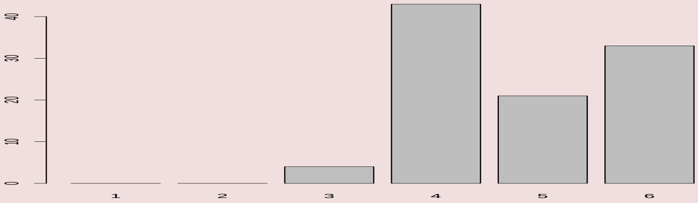
$\epsilon_i \sim N(0, 0.2)$

- Models with different number of parameters (d)

$$y_i = \alpha_0 + \sum_{i=1}^{d-1} x^i \beta_i + \epsilon_i.$$

- Choosing model by best (smallest) AIC.

repeated experiment, in sample

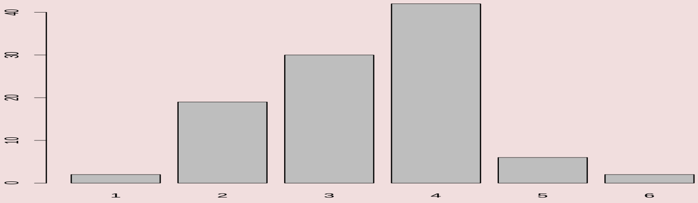


- Simulate 100 observations, repeat 101 times.
- Record the best model by AIC.

Cross-validation

- Split data into two or three sets.
- Training data - fit the parameters.
- Test data - evaluate performance.
- Evaluation data (not used here).

repeated experiment



- Simulate 100 observations, repeat 101 times.
- Split the data into two part, 60% training, 40% testing.
- Fit the parameters on the training data.
- Choose model on the testing data.

Bayesian

- The AIC is not well suited for the Bayesian modeling.
- Priors can affect over-fitting vs under-fitting.
- Possible to choose prior so they learn less from the data.
- Leaving, model selection for a slide to examine prior effect on the posterior distribution.

Regularization

Model:

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta x_i \\
\alpha &\sim N(0, \sigma_\beta) \\
\beta &\sim N(0, \sigma_\beta) \\
\sigma &\sim (0, 10)
\end{aligned}
$$

- What effect does σ_β on the posterior distribution.

Regularization, in Machine learning

Model:

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma) \\
\mu_i &= \alpha + \sum_{j=1}^{d} x_{ij}\beta_j \\
\alpha &\sim N(0, \sigma_\beta) \\
\beta_j &\sim N(0, \sigma_\beta),\ j = 1, \ldots, d \\
\sigma &\sim (0, 10)
\end{aligned}
$$

- σ_β is a hyperparameter.
- In Machine learning: choose the optimal σ_β gives best prediction.

Regularization

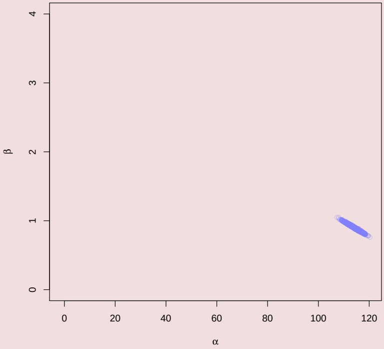


Figure: Posterior samples for $\sigma_\beta = 100$

Regularization

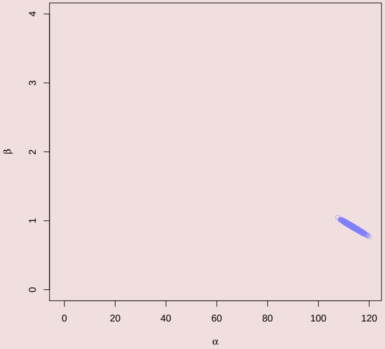


Figure: Posterior samples for $\sigma_\beta = 10$

Regularization

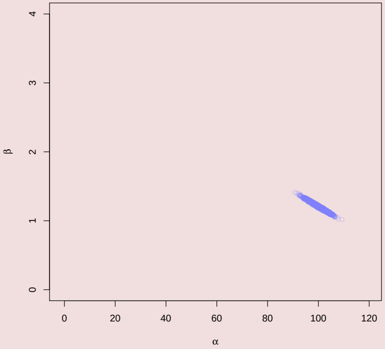


Figure: Posterior samples for $\sigma_\beta = 1$

Regularization

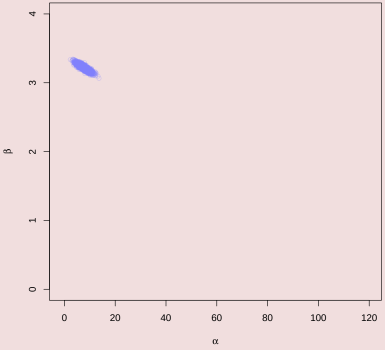


Figure: Posterior samples for $\sigma_\beta = 0.1$

Regularization

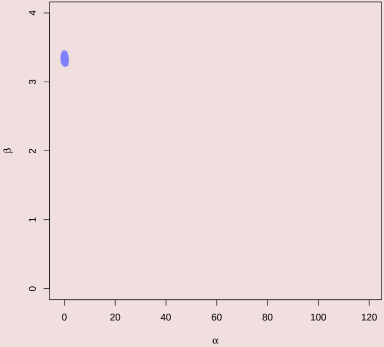


Figure: Posterior samples for $\sigma_\beta = 0.01$

Regularization

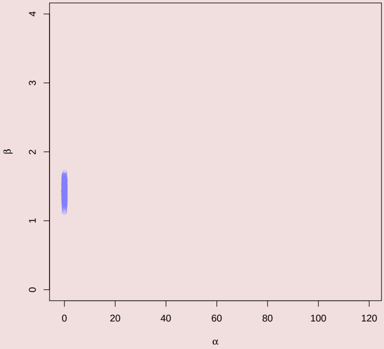


Figure: Posterior samples for $\sigma_\beta = 0.001$

Regularization

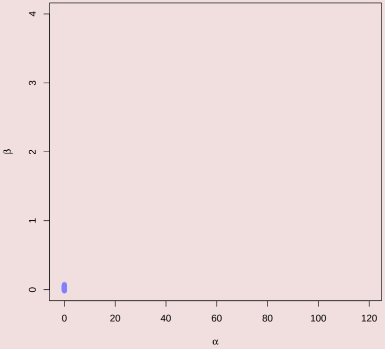


Figure: Posterior samples for $\sigma_\beta = 0.0001$

Regularization

- Prior makes the posterior conservative.
- How to compare models?
- For $\sigma_\beta = 0.0001$ and $\sigma_\beta = 100$ same number of parameters.

WAIC

WAIC balances how well a model predicts, with flexibility of the model

- How well does a model predict the data?

$$\begin{aligned}\sum_{j=1}^{n}\log(p(y_j|y_1,\ldots,y_n)) &= \sum_{j=1}^{n}\log(\mathbb{E}[p(y_j|\alpha,\beta,\sigma,y_1,\ldots,y_n)])\\ &= \sum_{j=1}^{n}\log\left(\int p(y_j|\tilde{\alpha},\tilde{\beta},\tilde{\sigma})\cdot\right.\\ &\qquad \left.\cdot p(\tilde{\alpha},\tilde{\beta},\tilde{\sigma}|y_1,\ldots,y_n)\,d\tilde{\alpha}\,d\tilde{\beta}\,d\tilde{\sigma}\right)\end{aligned}$$

- How flexible is the model?

$$p_{WAIC} = \sum_{j=1}^{n}\mathbb{V}[\log(p(y_j|\alpha,\beta,\sigma,y_1,\ldots,y_n))]$$

WAIC code

Building the model:

```
 data(milk)
 d           <- milk[ complete.cases(milk), ]
 d$neocortex <- d$neocortex.perc / 100
 model <- map(
alist(kcal.per.g ~ dnorm(mu, sigma),
mu <- alpha + bn * neocortex,
alpha ~ dnorm(0,10),
bn    ~ dnorm(0,10)),
data = d
             )
 n.sim <- 1000
 post <- extract.samples(model, n = n.sim)
```

WAIC code

Computing $\log(p(y_j|\alpha^{(i)}, \beta^{(i)}, \sigma^{(i)}))$ for data $j = 1, \ldots, n$ and samples $i = 1, \ldots, n_{sim}$:

```
ll <- sapply( 1:n.sim ,
              function(j){
                          mu <- post$alpha[j] + d$neocortex * post$bn[j]
                          dnorm(d$kcal.per.g, mu, post$sigma[j], log=T )} )
```

WAIC code, deviation

Approximating:

$$\sum_{j=1}^{n} \log(\mathbb{E}[p(y_j|\alpha,\beta,\sigma,y_1,\ldots,y_n)]) = \sum_{j=1}^{n} \log\left(\int p(y_j|\tilde{\alpha},\tilde{\beta},\tilde{\sigma})p(\tilde{\alpha},\tilde{\beta},\tilde{\sigma})|y_1,\ldots,y_n)\,d\tilde{\alpha}\,d\tilde{\beta}\,d\tilde{\sigma}\right)$$

with

$$lppd = \sum_{j=1}^{n} \log\left(\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} p(y_j|\alpha^{(i)},\beta^{(i)},\sigma^{(i)})\right)$$

```
n <- nrow(d)
lppd <- sum( sapply(1:n, function(j) log_sum_exp(ll[j,]) - log(n.sim)))
```

WAIC code, p_{WAIC}

Approximating:

$$\sum_{j=1}^{n} \mathbb{V}[\log(p(y_j|\alpha, \beta, \sigma, y_1, \ldots, y_n))]$$

with the variance of the function $\log(p(y_j|\alpha^{(i)}, \beta^{(i)}, \sigma^{(i)})$

```
pWAIC <- sum(sapply(1:n, function(j) var(ll[j,])))
```

milk and brains

Model:

$$k_i \sim N(\mu_i, \sigma),$$
$$\mu_i = \alpha + n_i\beta_n + \log(m_i)\beta_m$$

- k_i - calories per gram milk.
- m_i - weight of the ape.
- n_i - neocortex percentage.

Priors see book.

Cebus apella

0.89 kcal/g
68% neocortex

Eulemur fulvus

© anthony_m CC BY-NC-ND 2.0 http://creativecommons.org/licenses/by-nc-nd/2.0/

0.49 kcal/g
55% neocortex

WAIC code, p_{WAIC}

Model1

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma) \\
\mu_i &= \alpha + nero\beta_n \\
\alpha &\sim N(0, \sigma_\beta = 10) \\
\beta_n &\sim N(0, \sigma_\beta = 10)
\end{aligned}
$$

Model2

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma) \\
\mu_i &= \alpha + nero\beta_n \\
\alpha &\sim N(0, \sigma_\beta = 0.1) \\
\beta_n &\sim N(0, \sigma_\beta = 0.1)
\end{aligned}
$$

WAIC code, p_{WAIC}

Model1

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma) \\ \mu_i &= \alpha + nero\beta_n \\ \alpha &\sim N(0, \sigma_\beta = 10) \\ \beta_n &\sim N(0, \sigma_\beta = 10) \end{aligned}$$

Results:

$$-2lppd \approx -12$$

$$2p_{WAIC} \approx 6$$

$$WAIC = -2llpd + 2p_{WAIC} = -6$$

Model2

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma) \\ \mu_i &= \alpha + nero\beta_n \\ \alpha &\sim N(0, \sigma_\beta = 0.1) \\ \beta_n &\sim N(0, \sigma_\beta = 0.1) \end{aligned}$$

Results:

$$-2lppd \approx 36$$

$$2p_{WAIC} \approx 1$$

$$WAIC = -2llpd + 2p_{WAIC} = 37$$

WAIC model comparision

Comparing models

```
                        model1 <- map(
              alist(kcal.per.g ~ dnorm(mu, sigma),
              mu <- alpha,
              alpha ~ dnorm(0,10)),
              data = d
              )
model2 <- map(
              alist(kcal.per.g ~ dnorm(mu, sigma),
              mu   <- alpha + bn * neocortex,
              alpha ~ dnorm(0,10),
              bn     ~ dnorm(0,10)),
              data = d
              )
model3 <- map(
              alist(kcal.per.g ~ dnorm(mu, sigma),
              mu   <- alpha + bm * log(mass),
              alpha ~ dnorm(0,10),
              bm     ~ dnorm(0,10),
              sigma ~ dunif(0,10)),
              data = d
              )
model4 <- map(
              alist(kcal.per.g ~ dnorm(mu, sigma),
              mu   <- alpha + bn * neocortex + bm * log(mass),
              alpha ~ dnorm(0,10),
              bn     ~ dnorm(0,10),
              bm     ~ dnorm(0,10)),
              data = d
              )
```

WAIC model comparision

Comparing models

```
WAIC(model1)
```

```
[1] -7.410979
attr(,"lppd")
[1] 5.938131
attr(,"pWAIC")
[1] 2.232642
attr(,"se")
[1] 4.902928
```

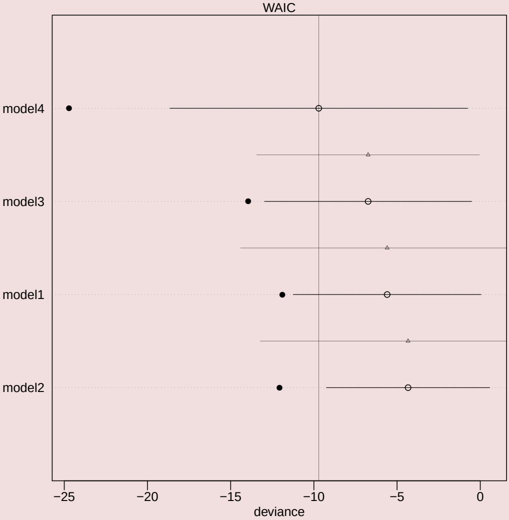
WAIC model comparision

Comparing models

```
models_milk<-compare(model1,model2,model3,  model4)
print(models_milk)
```

```
        WAIC  pWAIC  dWAIC  weight    SE   dSE
model4  -9.7    7.5    0.0    0.70  8.95    NA
model3  -6.7    3.6    3.0    0.16  6.22  6.70
model1  -5.6    3.1    4.1    0.09  5.64  8.82
model2  -4.3    3.9    5.4    0.05  4.91  8.88
```

WAIC model comparision



WAIC model comparision

- $WAIC_{min}$ is the smallest $WAIC$ of all compared models.
- $dWAIC_i = WAIC_i - WAIC_{min}$, gives the weight of a model

$$w_i = \frac{\exp(-\frac{1}{2} dWAIC_i)}{\sum_{j=1}^{m} \exp(-\frac{1}{2} dWAIC_j)}$$

Example data

Divorce rate in the US Model:

$$d_i \sim N(\mu_i, \sigma),$$
$$\mu_i = \alpha + r_i\beta_r + a_i\beta_a$$

- d_i - divorce rate by state
- r_i - marriage rate by state
- a_i - median age at marriage

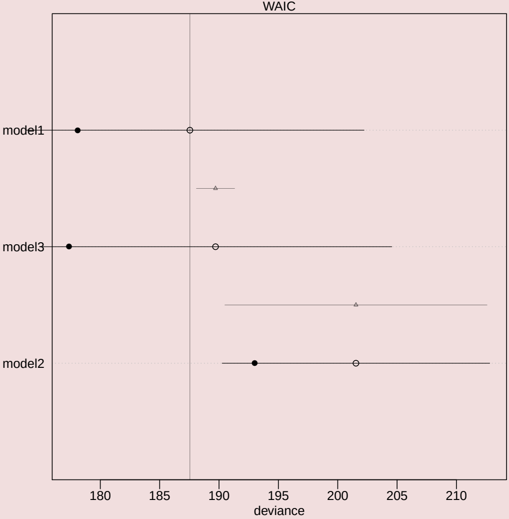
WAIC model comparision

Comparing models

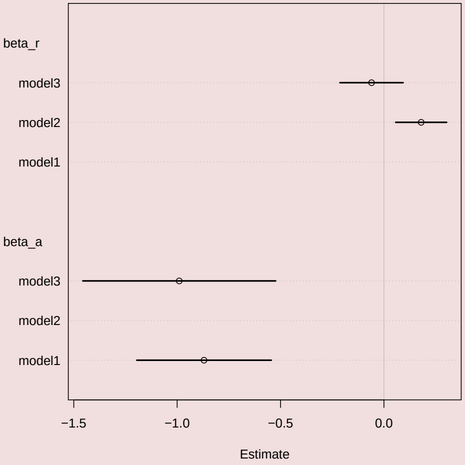
```
compare(model1, model2, model3)
```

```
        WAIC pWAIC dWAIC weight    SE   dSE
model1 187.7   4.8   0.0   0.78 14.73    NA
model3 190.2   6.4   2.5   0.22 15.29  1.60
model2 201.0   4.0  13.4   0.00 11.03 11.06
```

WAIC model comparision



WAIC model comparision


beta_r
model3
model2
model1
beta_a
model3
model2
model1
Estimate