Outline

Last week

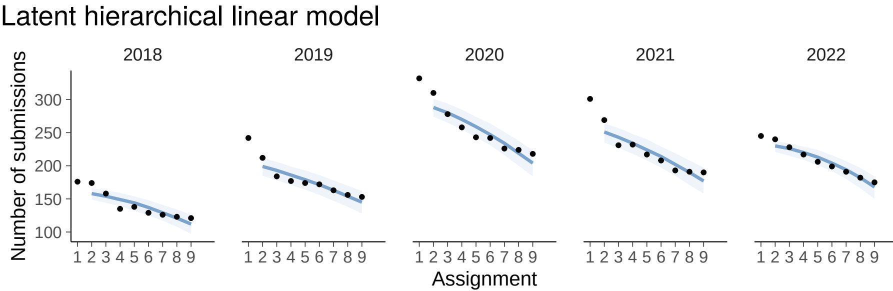
- What is cross-validation
- LOO-PIT checking
- Fast cross-validation with PSIS
- LOO model comparison and selection (elpd_diff, se)

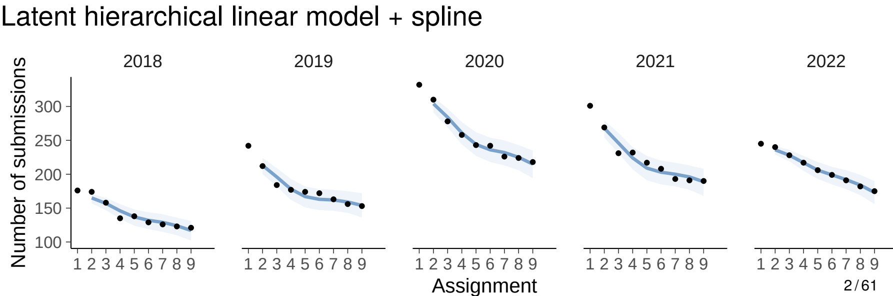
This week

- Model comparison with LOO-CV
- When is cross-validation applicable?
- K-fold cross-validation
- Related methods (WAIC, *IC, BF)
- Hypothesis testing
- Potential overfitting

Student retention – Posterior predictive distributions

with `tidybayes`

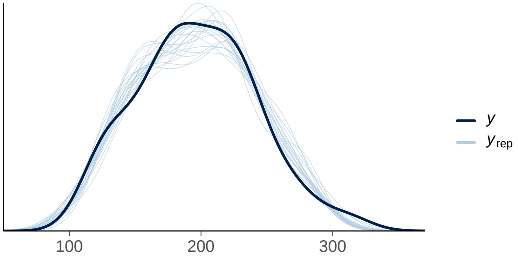




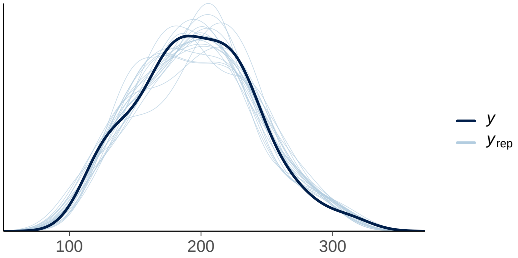
Student retention – Marginal PPC

```
pp_check(fit, ndraws=100)
```

Latent hierarchical linear model

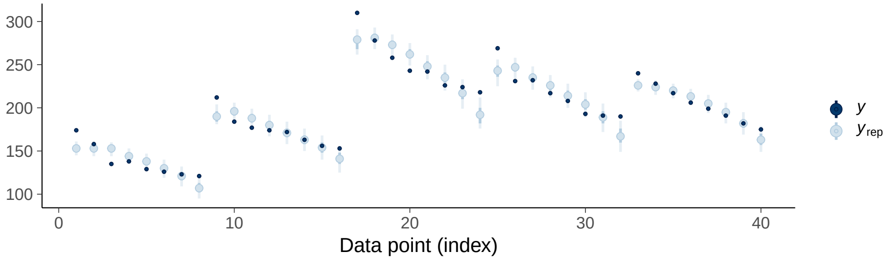


Latent hierarchical linear model + spline

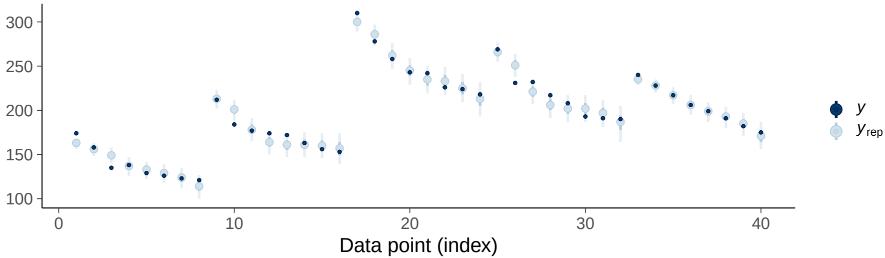


Student retention – LOO intervals

LOO predictive intervals – latent hierarchical linear

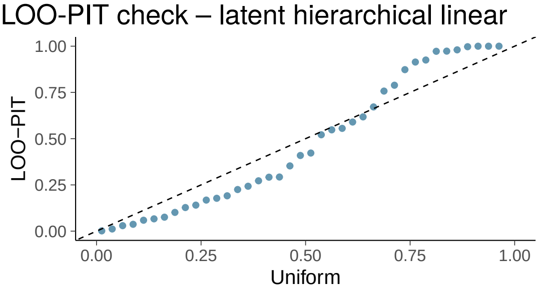


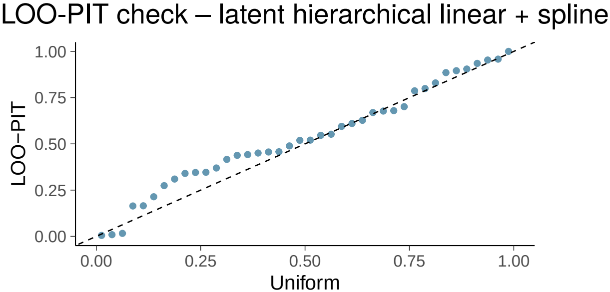
LOO predictive intervals – latent hierarchical linear + spline



Student retention – LOO-PIT checking

```
pp_check(fit, type = "loo_pit_qq", ndraws=4000)
```





Student retention – R^2

Latent hierarchical linear vs. latent hierarchical linear + spline

```
> loo_R2(fit4) |> round(digits=2)
   Estimate Est.Error Q2.5 Q97.5
R2     0.92      0.02 0.88  0.95

> loo_R2(fit6) |> round(digits=2)
   Estimate Est.Error Q2.5 Q97.5
R2     0.97      0.01 0.95  0.98
```

R^2 measures the goodness of the mean of the predictive distribution

Gelman, Goodrich, Gabry, and Vehtari (2019). R-squared for Bayesian regression models. *The American Statistician*, 73(3):307-309.

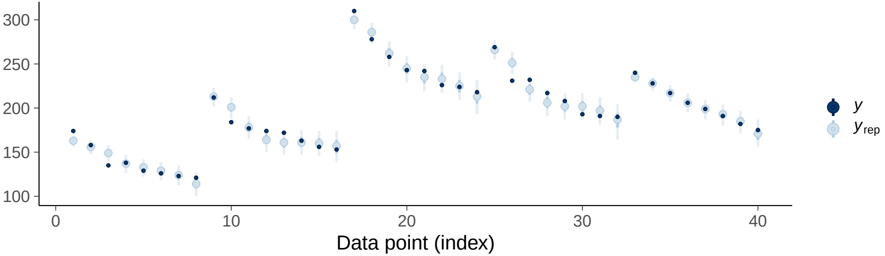
Student retention – log score – elpd

- information theoretical goodness of the whole distribution
- elpd = expected log predictive density (probability)
- elpd_loo = estimated with LOO predictive densities / probs
  $\sum_{n=1}^{N} \log p(y_i | x_i, x_{-i}, y_{-i})$

Student retention – log score – elpd

- information theoretical goodness of the whole distribution
- elpd = expected log predictive density (probability)
- elpd_loo = estimated with LOO predictive densities / probs
  $\sum_{n=1}^{N} \log p(y_i | x_i, x_{-i}, y_{-i})$

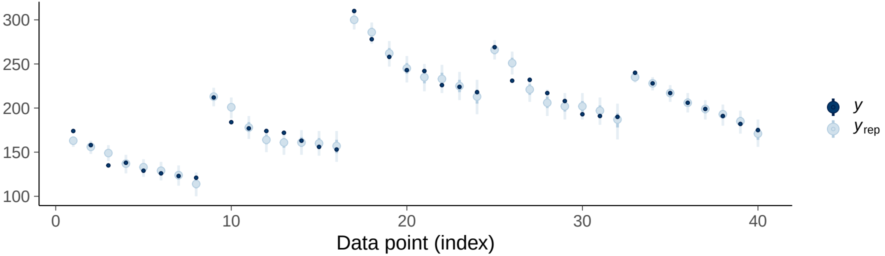
LOO predictive intervals – latent hierarchical linear + spline



Student retention – log score – elpd

- information theoretical goodness of the whole distribution
- elpd = expected log predictive density (probability)
- elpd_loo = estimated with LOO predictive densities / probs
  $\sum_{n=1}^{N} \log p(y_i|x_i, x_{-i}, y_{-i})$

LOO predictive intervals – latent hierarchical linear + spline

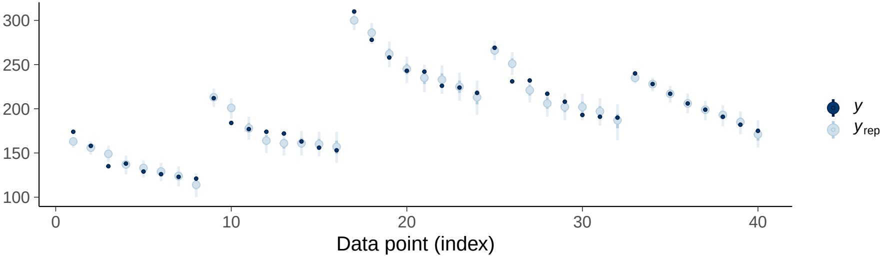


-8.4 -5.6 -2.9 -2.9 -2.8 -3.0 -4.0 -3.2 -3.9 -3.2 -3.4 -3.2 -2.9 -3.9 -3.4 -3.4 -3.2 -2.7 -2.8 -3.1
-2.5 -2.8 -2.9 -3.4 -5.4 -3.7 -3.1 -3.3 -3.5 -3.2 -3.5 -3.5 -6.6 -3.8 -3.7 -3.4 -2.5 -2.8 -2.9 -3.3

Student retention – log score – elpd

- information theoretical goodness of the whole distribution
- elpd = expected log predictive density (probability)
- elpd_loo = estimated with LOO predictive densities / probs
  $\sum_{n=1}^{N} \log p(y_i|x_i, x_{-i}, y_{-i})$

LOO predictive intervals – latent hierarchical linear + spline



-8.4 -5.6 -2.9 -2.9 -2.8 -3.0 -4.0 -3.2 -3.9 -3.2 -3.4 -3.2 -2.9 -3.9 -3.4 -3.4 -3.2 -2.7 -2.8 -3.1
-2.5 -2.8 -2.9 -3.4 -5.4 -3.7 -3.1 -3.3 -3.5 -3.2 -3.5 -3.5 -6.6 -3.8 -3.7 -3.4 -2.5 -2.8 -2.9 -3.3

$\sum$ = -141.7

Student retention – elpd_loo

Latent hierarchical linear + spline

```
> loo(fit6)

Computed from 4000 by 40 log-likelihood matrix

         Estimate   SE
elpd_loo   -141.7  7.2
p_loo        10.9  2.5
```

Student retention – elpd_loo

Latent hierarchical linear + spline

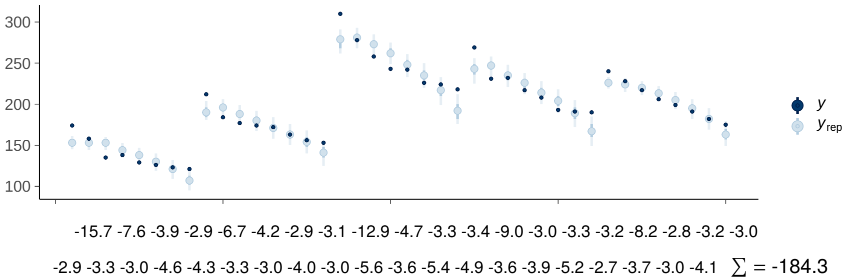
```
> loo(fit6)

Computed from 4000 by 40 log-likelihood matrix

         Estimate   SE
elpd_loo   -141.7  7.2
p_loo        10.9  2.5
```

Latent hierarchical linear

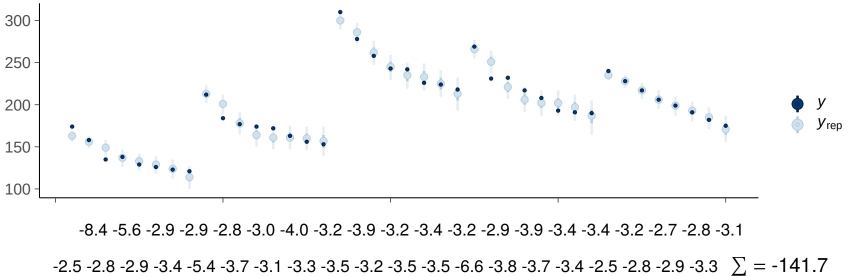
```
> loo(fit4)

Computed from 4000 by 40 log-likelihood matrix

         Estimate   SE
elpd_loo   -184.3 17.3
p_loo        24.3  5.8
```

Student retention – log score – elpd

LOO predictive intervals – latent hierarchical linear

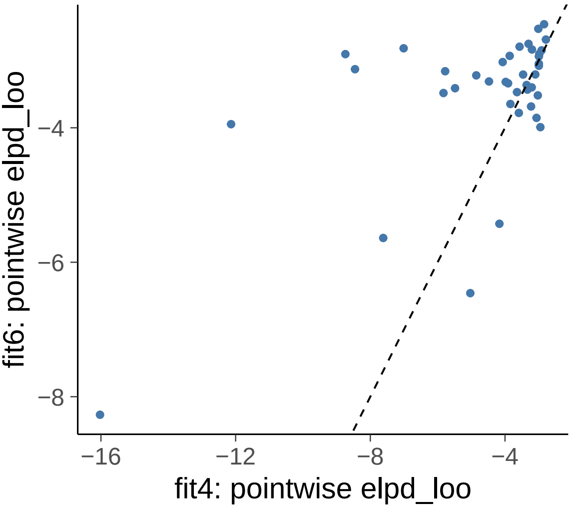


LOO predictive intervals – latent hierarchical linear + spline



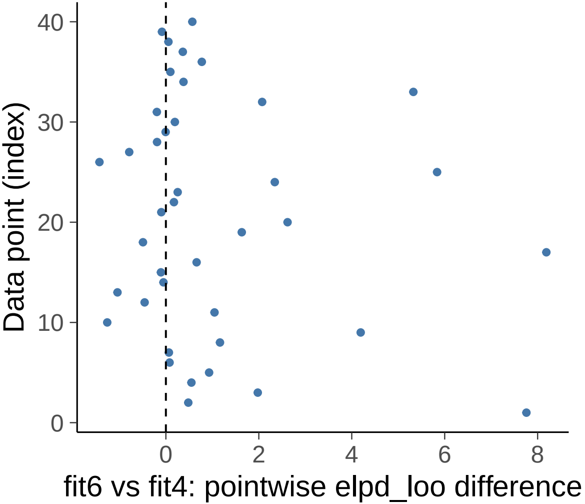
Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



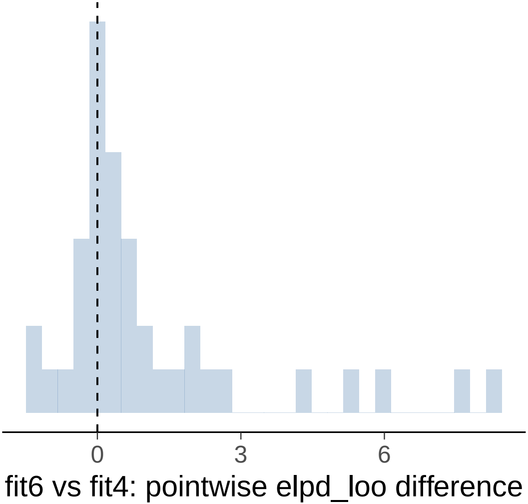
Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



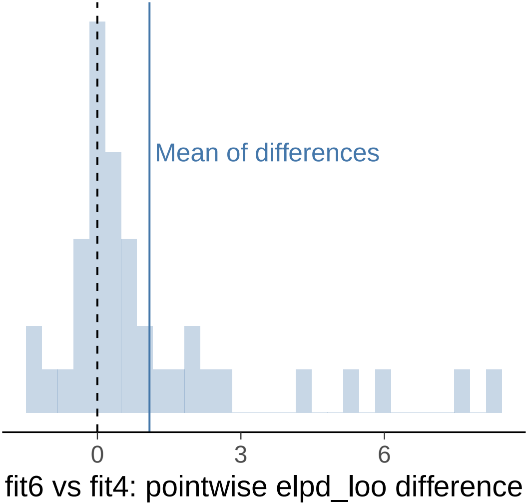
Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



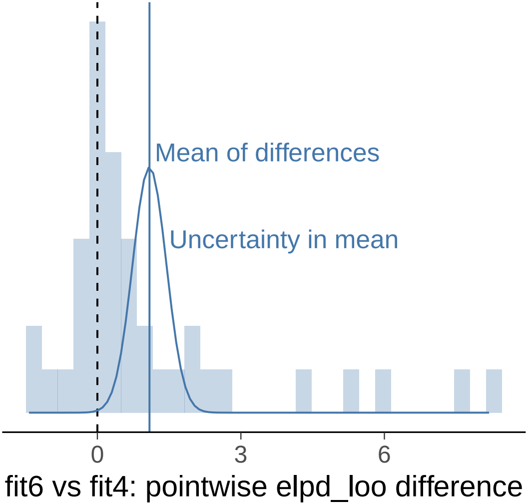
Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



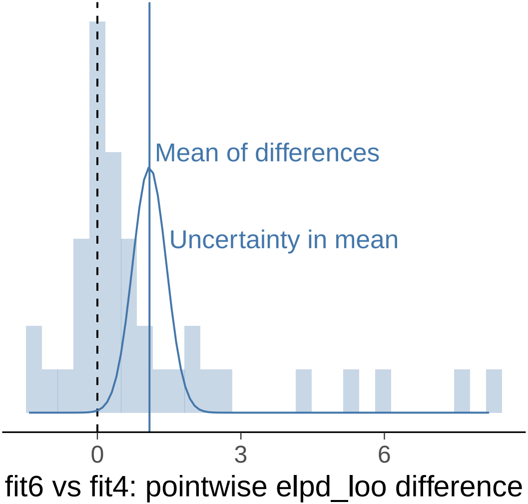
Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



Student retention – elpd_loo

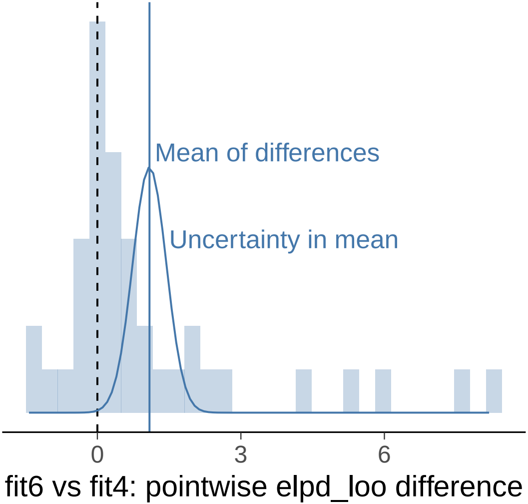
Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



mean ≈ 1.07

Student retention – elpd_loo

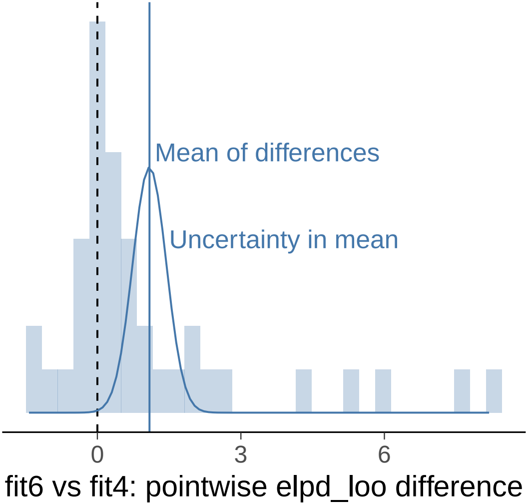
Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



mean ≈ 1.07
sd ≈ 2.26

Student retention – elpd_loo

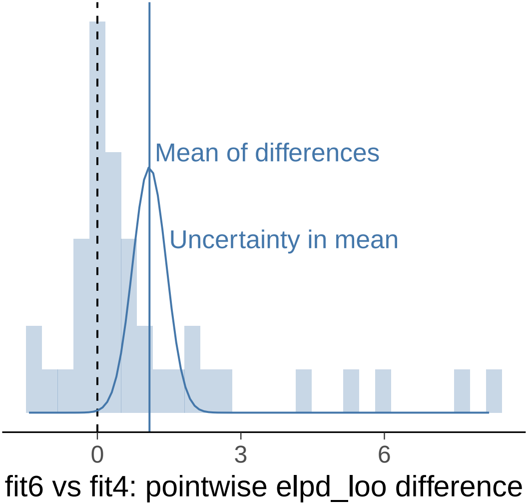
Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



mean ≈ 1.07
sd ≈ 2.26
SE = sd/$\sqrt{40} \approx 0.36$

Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)

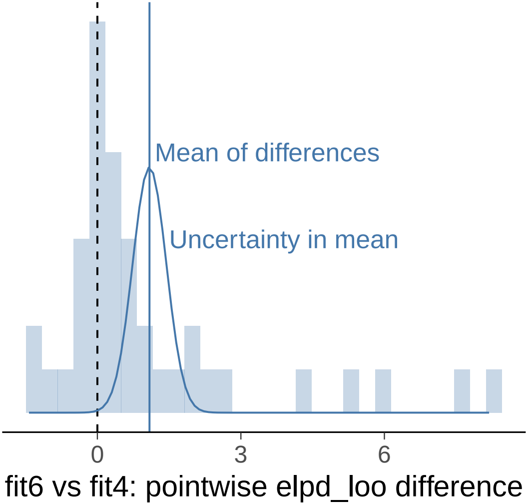


mean ≈ 1.07
sd ≈ 2.26
SE = sd/$\sqrt{40} \approx 0.36$

sum ≈ 42.6

Student retention – elpd_loo

Latent hierarchical linear (fit4) vs latent hierarchical linear + spline (fit6)



mean ≈ 1.07
sd ≈ 2.26
SE = sd/$\sqrt{40} \approx 0.36$

sum ≈ 42.6
SE = sd$*\sqrt{40} \approx 14.3$

Student retention – elpd_loo

Latent hierarchical linear + spline

```
> loo(fit6)
         Estimate   SE
elpd_loo   -141.7  7.2
p_loo        10.9  2.5
```

Latent hierarchical linear

```
> loo(fit4)
         Estimate   SE
elpd_loo   -184.3 17.3
p_loo        23.8  5.7
```

```
> loo_compare(loo(fit4), loo(fit6))
     elpd_diff se_diff
fit6   0.0       0.0
fit4 -42.6      14.3
```

LOO difference uncertainty estimate (SE) reliability

1. The models make very similar predictions

2. The models are misspecified with outliers in the data

3. The number of observations is small

Sivula, Magnusson, Matamoros, and Vehtari (2022). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296v3*.

LOO difference uncertainty estimate (SE) reliability

1. The models make very similar predictions
   - if |elpd_loo| < 4, SE is not reliable, but the difference is small anyway
   - selecting a "wrong" model has small cost
   - in nested case, the skewness favors the simpler model
2. The models are misspecified with outliers in the data
3. The number of observations is small

Sivula, Magnusson, Matamoros, and Vehtari (2022). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296v3*.

LOO difference uncertainty estimate (SE) reliability

1. The models make very similar predictions
   - if |elpd_loo| < 4, SE is not reliable, but the difference is small anyway
   - selecting a "wrong" model has small cost
   - in nested case, the skewness favors the simpler model
2. The models are misspecified with outliers in the data
   - in nested case, the bias favors the simpler model
   - model checking and model extension to avoid misspecified models (Bayesian workflow)
3. The number of observations is small

Sivula, Magnusson, Matamoros, and Vehtari (2022). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296v3*.

LOO difference uncertainty estimate (SE) reliability

1. The models make very similar predictions
   - if |elpd_loo| < 4, SE is not reliable, but the difference is small anyway
   - selecting a "wrong" model has small cost
   - in nested case, the skewness favors the simpler model
2. The models are misspecified with outliers in the data
   - in nested case, the bias favors the simpler model
   - model checking and model extension to avoid misspecified models (Bayesian workflow)
3. The number of observations is small
   - in nested case the skewness favors the simpler model
   - any inference with small n is difficult
   - if |elpd_loo| > 4, model is well specified, and $n > 100$ then the normal approximation is good

Sivula, Magnusson, Matamoros, and Vehtari (2022). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296v3*.

Log score and elpd_loo

- Log score is not easily interpretable
- but is information theoretically good utility for the goodness of the whole distribution
- and thus is useful in model comparison

Log score and elpd_loo

- Interpretation in discrete case
  - log probability

Log score and elpd_loo

- Interpretation in discrete case
  - log probability
- For example
  - $\frac{1}{N}\sum_{n=1}^{N}\exp(\text{elpd}_{\text{loo},n}) \approx 4\%$ probability that we predict the observed value

Log score and elpd_loo

- Interpretation in discrete case
  - log probability
- For example
  - $\frac{1}{N}\sum_{n=1}^{N}\exp(\mathrm{elpd}_{\mathrm{loo},n}) \approx 4\%$ probability that we predict the observed value
  - compare to guessing uniformly from the data range [121,310] having $1/(310-121+1) \approx 0.5\%$ probability

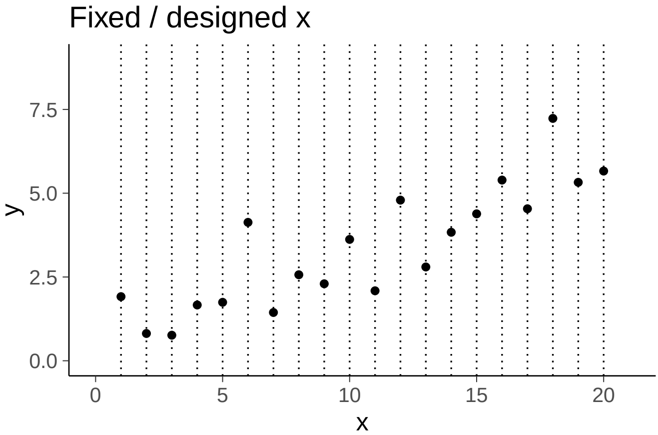
Log score and elpd_loo

- Interpretation in discrete case
  - log probability
- For example
  - $\frac{1}{N}\sum_{n=1}^{N}\exp(\mathrm{elpd}_{\mathrm{loo},n}) \approx 4\%$ probability that we predict the observed value
  - compare to guessing uniformly from the data range [121,310] having $1/(310-121+1) \approx 0.5\%$ probability (log score -210)

Log score and elpd_loo

- Interpretation in discrete case
  - log probability
- For example
  - $\frac{1}{N}\sum_{n=1}^{N}\exp(\mathrm{elpd}_{\mathrm{loo},n}) \approx 4\%$ probability that we predict the observed value
  - compare to guessing uniformly from the data range [121,310] having $1/(310-121+1) \approx 0.5\%$ probability (log score -210)
- Interpretation in continuous case
  - can be compared to a simple reference distribution

Assumptions about the future observations



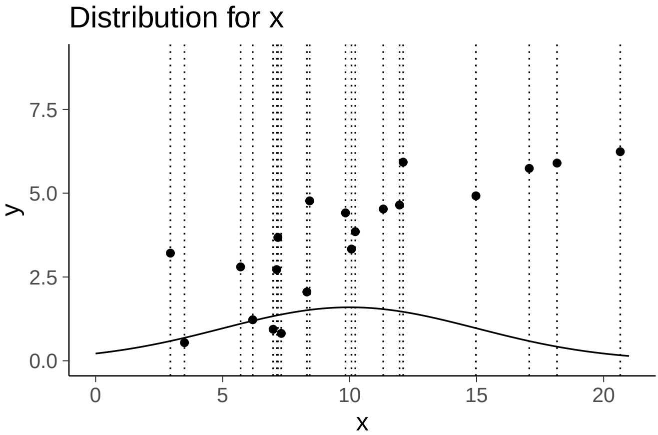
elpd_loo $= \sum_{i=1}^{20} \log p(y_i \mid x_i, x_{-i}, y_{-i}) \approx -29.5$

SE $= \mathrm{sd}(\log p(y_i \mid x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$

LOO is ok for fixed / designed x. SE is uncertainty about $y \mid x$.

see Vehtari & Ojanen (2012) and CV-FAQ

Assumptions about the future observations



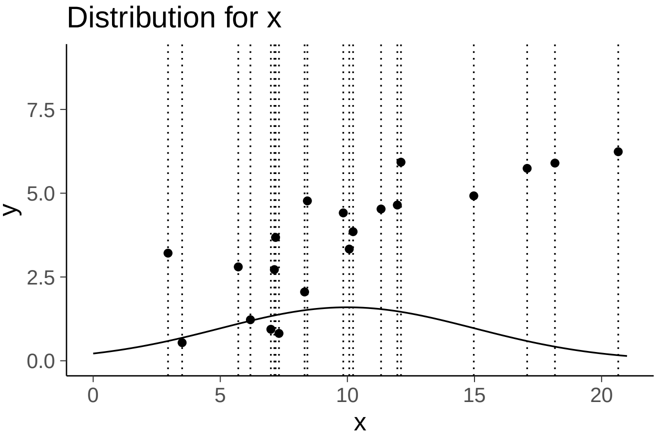
elpd_loo $= \sum_{i=1}^{20} \log p(y_i \mid x_i, x_{-i}, y_{-i}) \approx -29.5$

SE $= \mathrm{sd}(\log p(y_i \mid x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$

LOO is ok for random x. SE is uncertainty about $y \mid x$ and x.

see Vehtari & Ojanen (2012) and CV-FAQ

Assumptions about the future observations



elpd_loo $= \sum_{i=1}^{20} \log p(y_i \mid x_i, x_{-i}, y_{-i}) \approx -29.5$

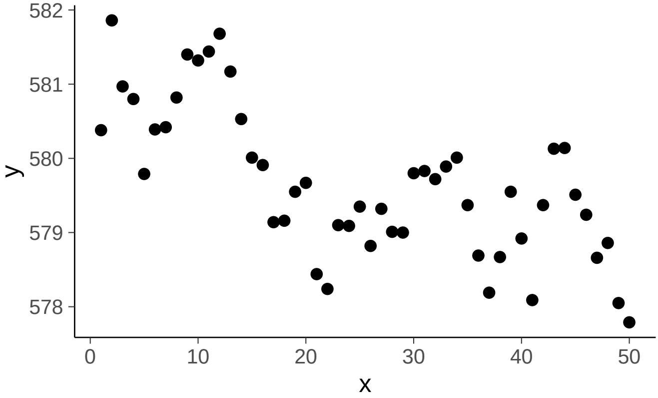
SE $= \mathrm{sd}(\log p(y_i \mid x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$

LOO is ok for random x. SE is uncertainty about $y \mid x$ and x.

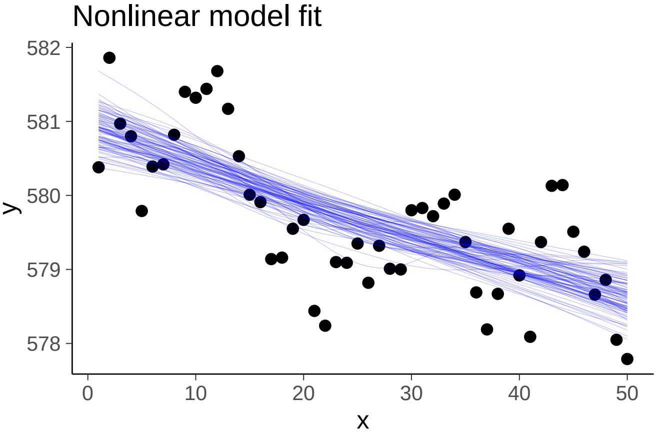
Covariate shift handled with importance weighting or modelling

see Vehtari & Ojanen (2012) and CV-FAQ

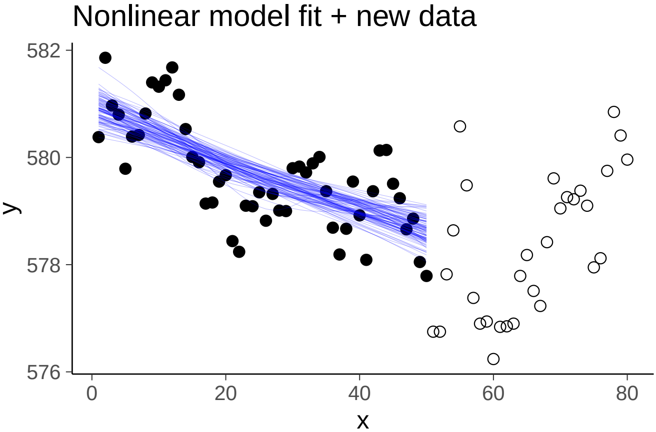
Interpolation vs extrapolation



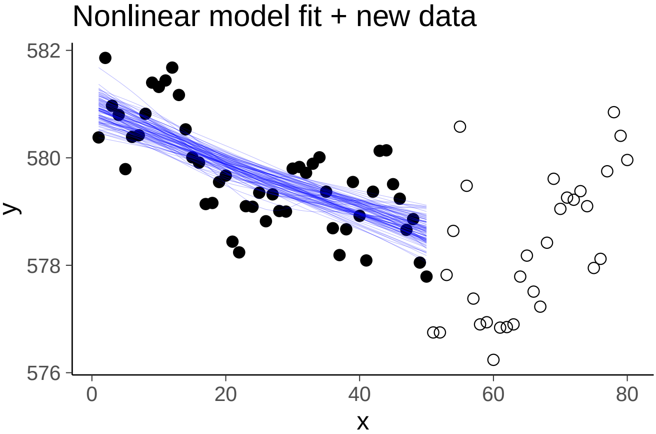
Interpolation vs extrapolation



Interpolation vs extrapolation

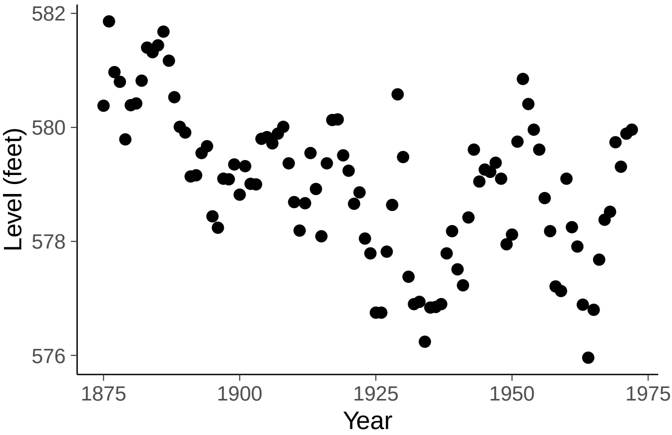


Interpolation vs extrapolation



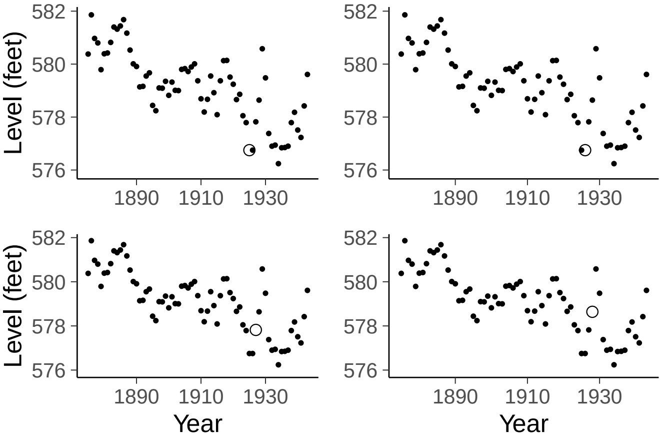
- Extrapolation is more difficult item<5> In high dimensional case mostly extrapolation

Cross-validation for time series?



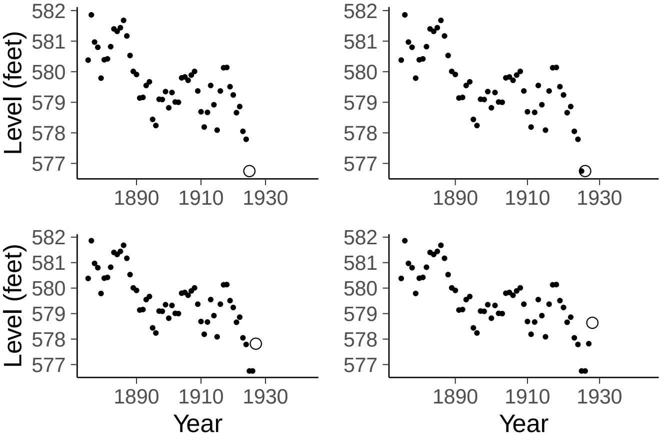
Can LOO or other cross-validation be used with time series?

Cross-validation for time series



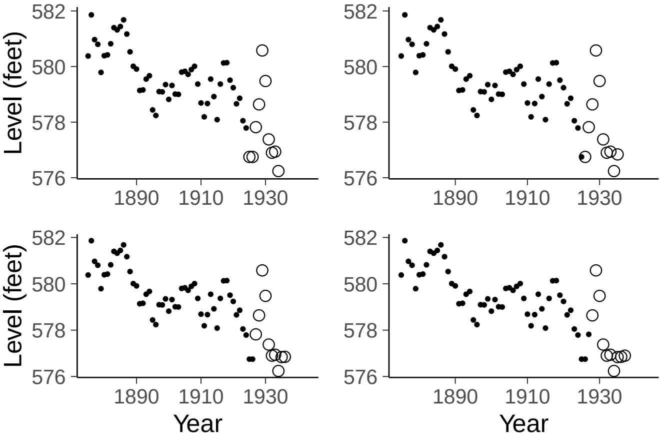
Leave-one-out cross-validation is ok for assessing conditional model

Cross-validation for time series



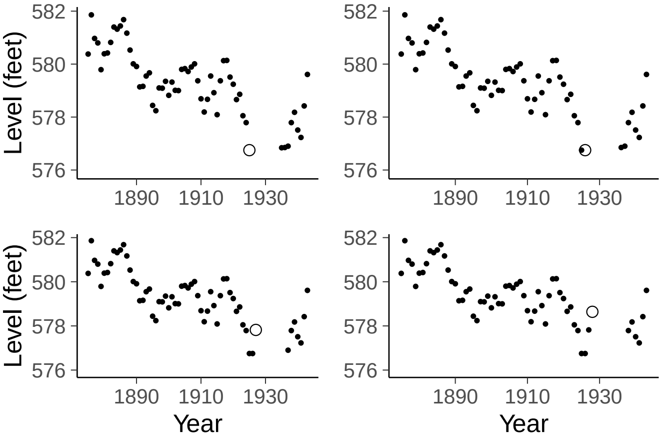
Leave-future-out (LFO) cross-validation is better for predicting future

Cross-validation for time series



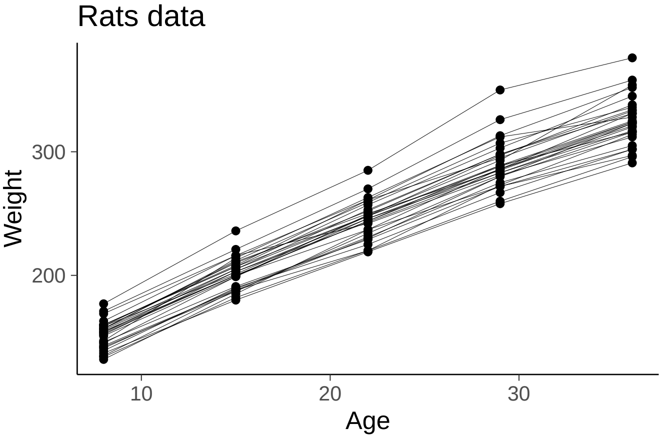
m-step-ahead cross-validation is better for predicting further future

Cross-validation for time series



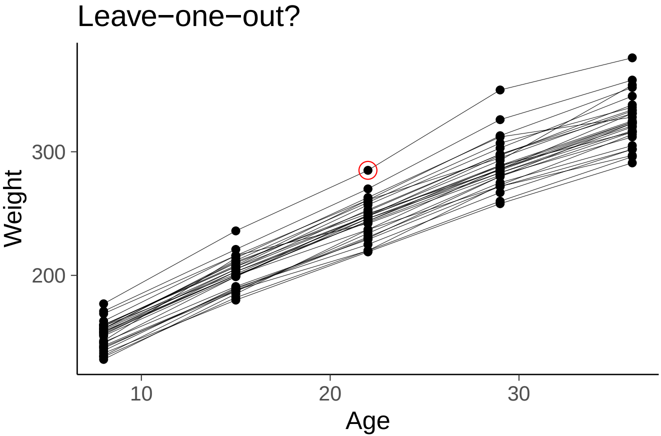
m-step-ahead leave-a-block-out cross-validation

Cross-validation for hierarchical data



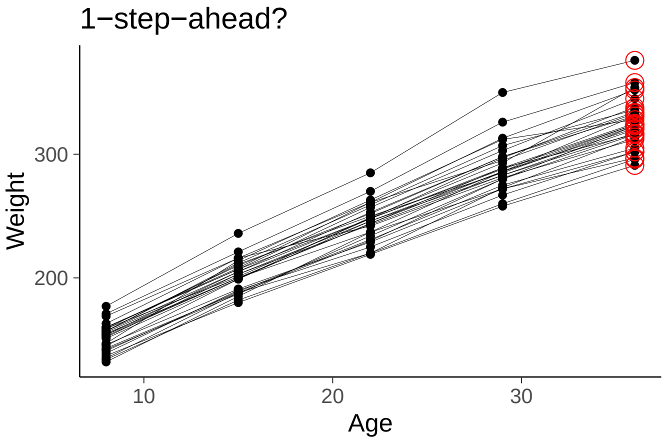
Can LOO or other cross-validation be used with hierarchical data?

Cross-validation for hierarchical data



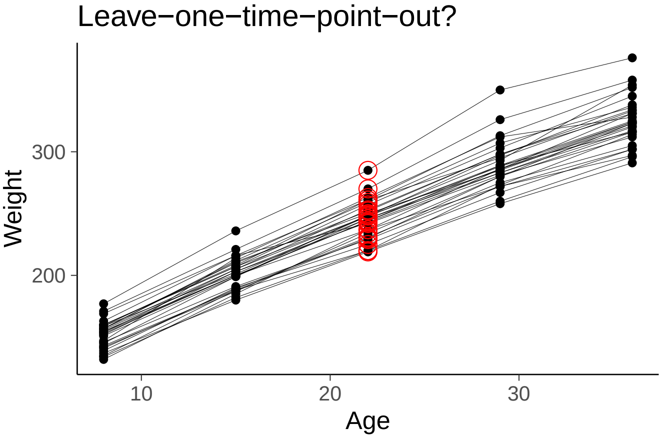
Yes!

Cross-validation for hierarchical data



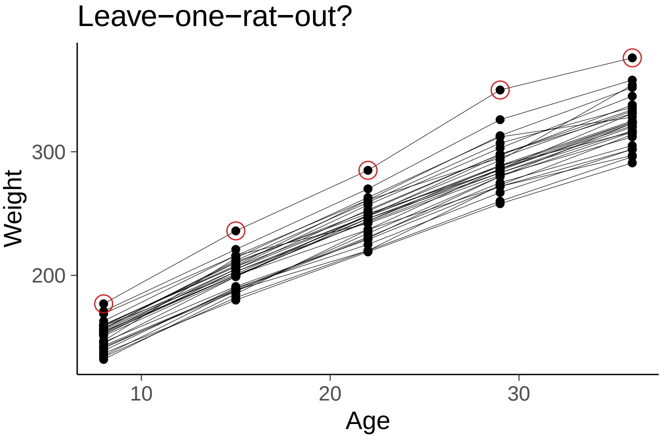
Yes!

Cross-validation for hierarchical data



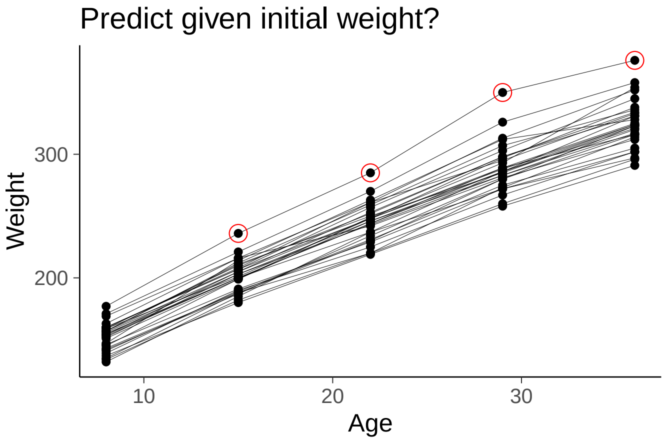
Yes!

Cross-validation for hierarchical data



Yes!

Cross-validation for hierarchical data



Yes!

Summary of data generating mechanisms and prediction tasks

- You have to make some assumptions on data generating mechanism
- Use the knowledge of the prediction task if available
- Cross-validation can be used to analyse different parts, even if there is no clear prediction task

see Vehtari & Ojanen (2012) and CV-FAQ

Pareto smoothed importance sampling CV variants

- PSIS-LOO for hierarchical models
  - leave-one-group out is challenging for PSIS-LOO
  - Stan demo of the challenges and integrated LOO at https://users.aalto.fi/~ave/modelselection/roaches.html
  - see also Merkel, Furr and Rabe-Hesketh (2018)

Pareto smoothed importance sampling CV variants

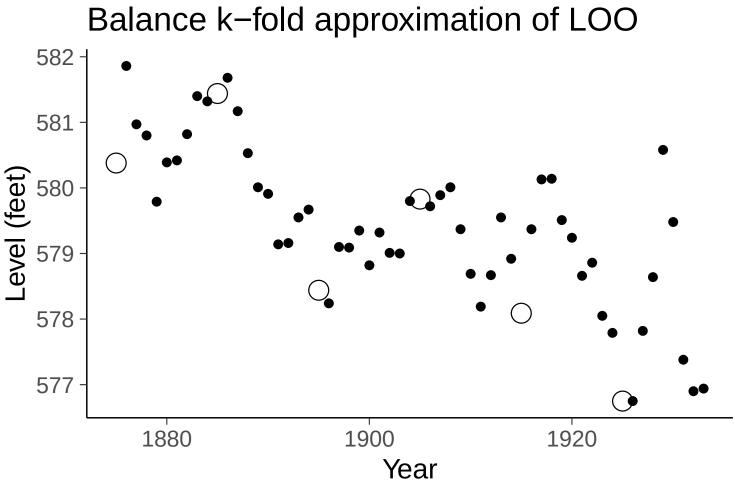
- PSIS-LOO for hierarchical models
  - leave-one-group out is challenging for PSIS-LOO
  - Stan demo of the challenges and integrated LOO at https://users.aalto.fi/~ave/modelselection/roaches.html
  - see also Merkel, Furr and Rabe-Hesketh (2018)
- PSIS-LOO for non-factorized models
  - mc-stan.org/loo/articles/loo2-non-factorizable.html

Pareto smoothed importance sampling CV variants

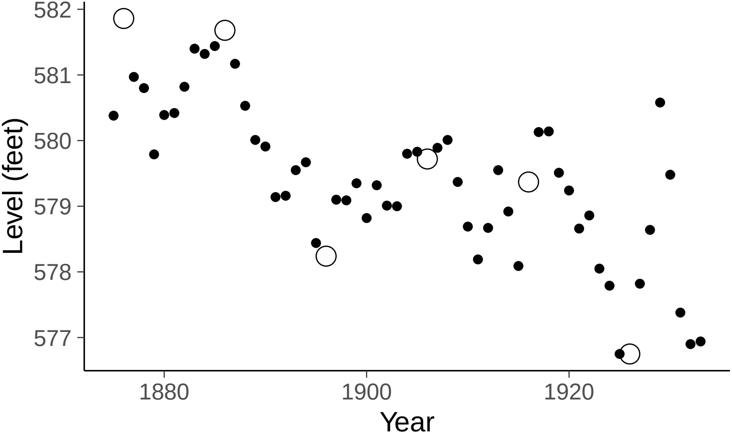
- PSIS-LOO for hierarchical models
  - leave-one-group out is challenging for PSIS-LOO
  - Stan demo of the challenges and integrated LOO at https://users.aalto.fi/~ave/modelselection/roaches.html
  - see also Merkel, Furr and Rabe-Hesketh (2018)
- PSIS-LOO for non-factorized models
  - mc-stan.org/loo/articles/loo2-non-factorizable.html
- PSIS-LOO for time series
  - Approximate leave-future-out cross-validation (LFO-CV) mc-stan.org/loo/articles/loo2-lfo.html

K-fold cross-validation

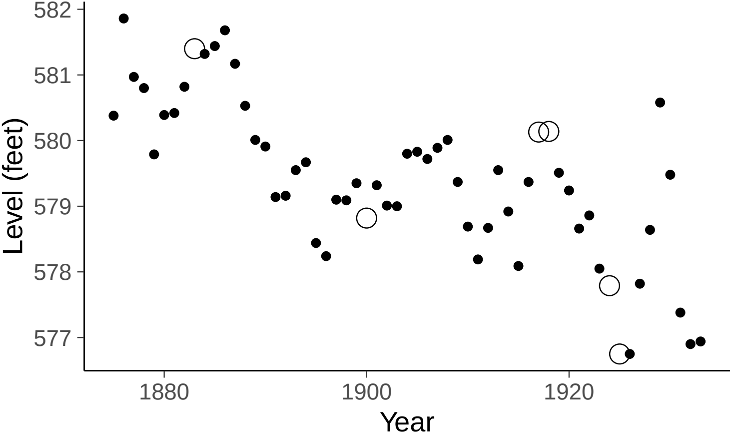
- K-fold cross-validation can approximate LOO
  - the same use cases as with LOO
- K-fold cross-validation can be used for hierarchical models
  - good for leave-one-group-out
- K-fold cross-validation can be used for time series
  - with leave-block-out


Balance k−fold approximation of LOO
Level (feet)
582
581
580
579
578
577
1880
1900
1920
Year

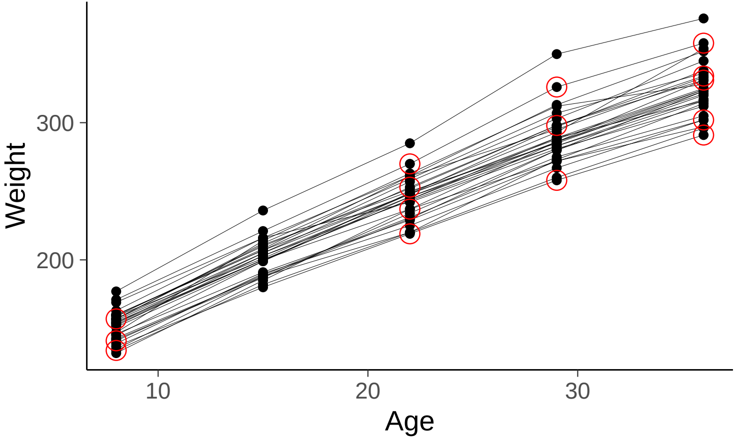
Balance k−fold approximation of LOO



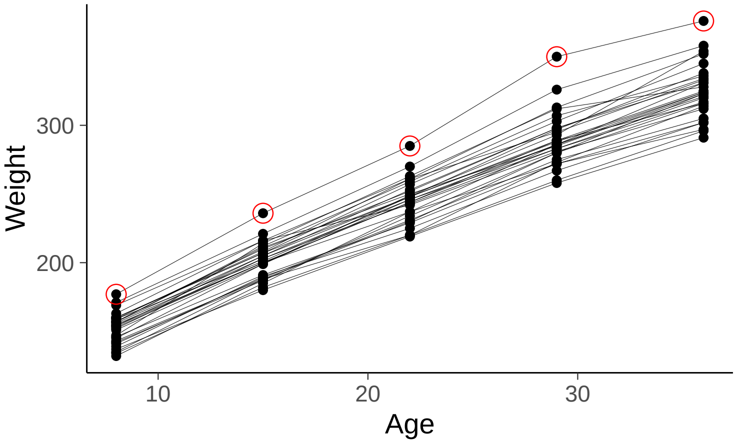
Random k−fold approximation of LOO



Random kfold approximation of LOO



Leave−one−rat−out



K-fold-CV code

- RStan, CmdStanR
  See vignette http://mc-stan.org/loo/articles/loo2-elpd.html
- RStanARM, brms
  `kfold(fit)`
- Alternative data divisions
  `kfold_split_random()`
  `kfold_split_balanced()`
  `kfold_split_stratified()`

looic?

```
> loo(fit6)

Computed from 4000 by 40 log-likelihood matrix

         Estimate   SE
elpd_loo   -141.7  7.2
p_loo        10.9  2.5
looic       283.4 14.4
------
Monte Carlo SE of elpd_loo is 0.1.
```

- `loo` output shows also looic
- for historical non-Bayesian reasons it's -2 * elpd_loo
  - connection to deviance and information criteria
  - you can just ignore it (I'd prefer it would not be shown)

Information criteria

Information criteria estimate predictive performance, too

- AIC uses maximum likelihood estimate for prediction

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Information criteria

Information criteria estimate predictive performance, too

- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Information criteria

Information criteria estimate predictive performance, too

- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction
- BIC is a simple approximation for marginal likelihood

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Information criteria

Information criteria estimate predictive performance, too

- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction
- BIC is a simple approximation for marginal likelihood
- TIC, NIC, RIC, PIC, BPIC, QIC, AICc, ...

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Information criteria

Information criteria estimate predictive performance, too

- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction
- BIC is a simple approximation for marginal likelihood
- TIC, NIC, RIC, PIC, BPIC, QIC, AICc, ...
- WAIC is the only Bayesian information criterion

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

WAIC vs PSIS-LOO

- WAIC has the same target and assumptions as LOO

Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

WAIC vs PSIS-LOO

- WAIC has the same target and assumptions as LOO
- PSIS-LOO is more accurate

Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

WAIC vs PSIS-LOO

- WAIC has the same target and assumptions as LOO
- PSIS-LOO is more accurate
- PSIS-LOO has much better diagnostics

Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

WAIC vs PSIS-LOO

- WAIC has the same target and assumptions as LOO
- PSIS-LOO is more accurate
- PSIS-LOO has much better diagnostics
- moment matching and reloo are natural improvements for PSIS-LOO

Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

WAIC vs PSIS-LOO

- WAIC has the same target and assumptions as LOO
- PSIS-LOO is more accurate
- PSIS-LOO has much better diagnostics
- moment matching and reloo are natural improvements for PSIS-LOO
- LOO makes the prediction assumption more clear, which helps if K-fold-CV is needed instead

Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

WAIC vs PSIS-LOO

- WAIC has the same target and assumptions as LOO
- PSIS-LOO is more accurate
- PSIS-LOO has much better diagnostics
- moment matching and reloo are natural improvements for PSIS-LOO
- LOO makes the prediction assumption more clear, which helps if K-fold-CV is needed instead
- Multiplying by -2 doesn't give any benefit (Watanabe didn't multiply by -2)

Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Marginal likelihood and Bayes factor

Bayes Factor $\frac{p(y|M_1)}{p(y|M_2)}$

Marginal likelihood $p(y|M_1) = \int p(y|\theta, M_1)p(\theta|M_1)d\theta$

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Marginal likelihood and Bayes factor

Bayes Factor $\frac{p(y|M_1)}{p(y|M_2)}$

Marginal likelihood $p(y|M_1) = \int p(y|\theta, M_1)p(\theta|M_1)d\theta$

Marginal likelihood with chain rule:
$p(y|M_1) = p(y_1|M_1)p(y_2|y_1, M_1), \ldots, p(y_n|y_1, \ldots, y_{n-1}, M_1)$

Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Marginal likelihood and Bayes factor

Bayes Factor $\frac{p(y|M_1)}{p(y|M_2)}$

Marginal likelihood $p(y|M_1) = \int p(y|\theta, M_1)p(\theta|M_1)d\theta$

Marginal likelihood with chain rule:
$p(y|M_1) = p(y_1|M_1)p(y_2|y_1, M_1), \ldots, p(y_n|y_1, \ldots, y_{n-1}, M_1)$
where
$p(y_1|M_1) = \int p(y_1|\theta, M_1)p(\theta|M_1)d\theta$
$p(y_2|y_1, M_1) = \int p(y_2|\theta, M_1)p(\theta|y_1, M_1)d\theta$
$\ldots$
$p(y_n|y_1, \ldots, y_{n-1}, M_1) = \int p(y_n|\theta, M_1)p(\theta|y_1, \ldots, y_{n-1}, M_1)d\theta$

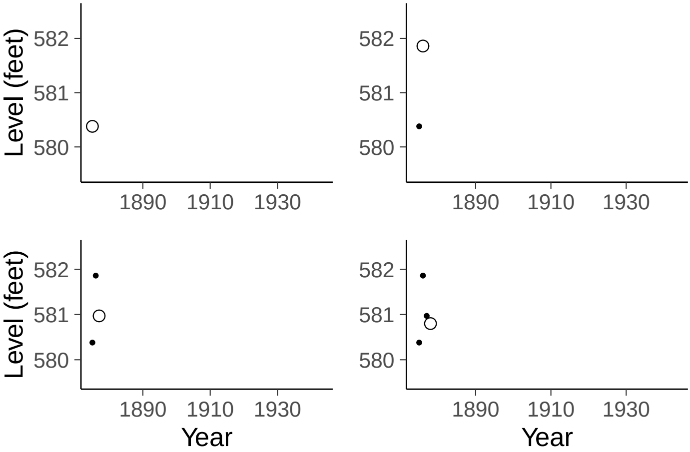
Vehtari & Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142-228.

Marginal likelihood / Bayes factor

- Like leave-future-out 1-step-ahead cross-validation but starting with 0 observations

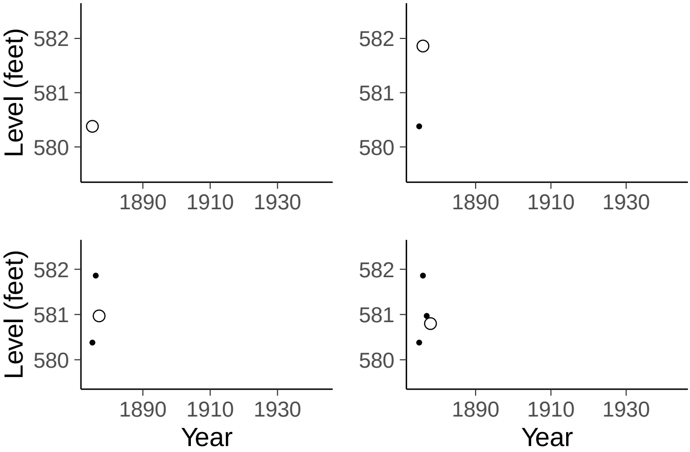
Marginal likelihood / Bayes factor

- Like leave-future-out 1-step-ahead cross-validation but starting with 0 observations



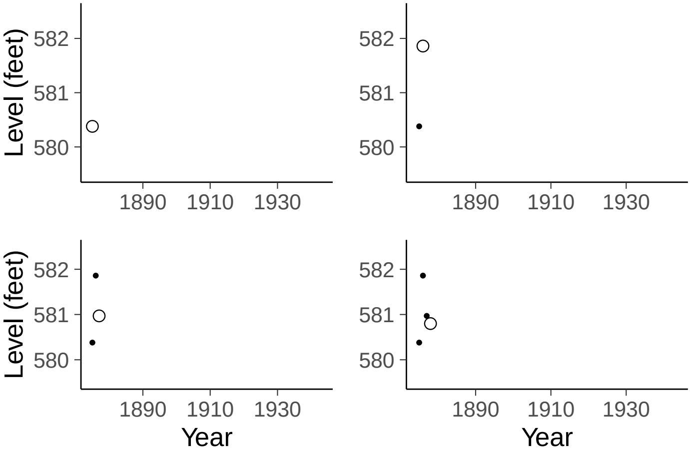
Marginal likelihood / Bayes factor

- Like leave-future-out 1-step-ahead cross-validation but starting with 0 observations
  - which makes it very sensitive to prior



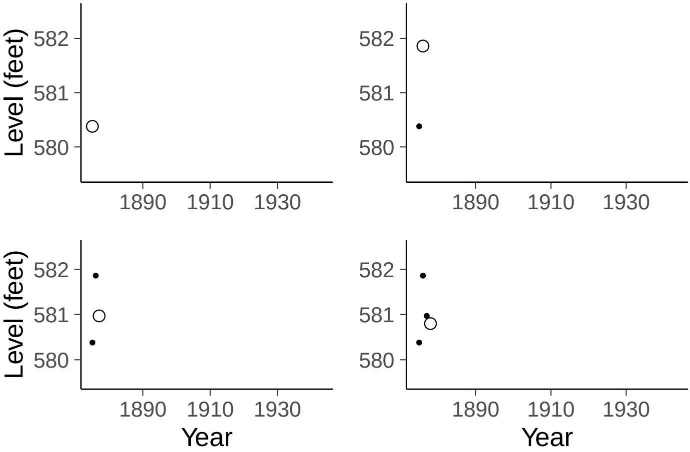
Marginal likelihood / Bayes factor

- Like leave-future-out 1-step-ahead cross-validation but starting with 0 observations
  - which makes it very sensitive to prior and
  - unstable in case of misspecified models



Marginal likelihood / Bayes factor

- Like leave-future-out 1-step-ahead cross-validation but starting with 0 observations
  - which makes it very sensitive to prior and
  - unstable in case of misspecified models also asymptotically



Marginal likelihood / Bayes factor

- Like leave-future-out 1-step-ahead cross-validation but starting with 0 observations
  - which makes it very sensitive to prior and
  - unstable in case of misspecified models also asymptotically
- Oelrich, Ding, Magnusson, Vehtari, and Villani (2020). When are Bayesian model probabilities overconfident? *arXiv:2003.04026*.

Predictive model selection

- Student retention
  - latent hierarchical linear vs.
  - latent hierarchical linear + spline

  is a good example where predictive model selection is useful

Sometimes cross-validation is not needed

- In a simple nested case, often easier and more accurate to analyze posterior distribution of more complex model directly
  - instead of comparing

    Model 1: $y \sim \mathrm{normal}(\alpha, \sigma)$

    vs

    Model 2: $y \sim \mathrm{normal}(\alpha + \beta x, \sigma)$

    look at the posterior of β directly

Common statistical tests as Bayesian models

- Most common statistical tests are linear models

test	model	formula
t-test	mean of data	`y ~ 1`
paired t-test	mean of diffs	`(y1 - y2) ~ 1`
Pearson correl.	linear model	`y ~ 1 + x`
two-sample t-test	group means	`y ~ 1 + gid`
ANOVA	hier. model	`y ~ 1 + (1 | gid)`

...

Common statistical tests as Bayesian models

- Most common statistical tests are linear models

test	model	formula
t-test	mean of data	`y ~ 1`
paired t-test	mean of diffs	`(y1 - y2) ~ 1`
Pearson correl.	linear model	`y ~ 1 + x`
two-sample t-test	group means	`y ~ 1 + gid`
ANOVA	hier. model	`y ~ 1 + (1 | gid)`

...

- Possible to extend, e.g., with group specific variances and and different distributions such t- or Poisson distribution
  - and go beyond named tests

Common statistical tests as Bayesian models

- Most common statistical tests are linear models

test	model	formula
t-test	mean of data	`y ~ 1`
paired t-test	mean of diffs	`(y1 - y2) ~ 1`
Pearson correl.	linear model	`y ~ 1 + x`
two-sample t-test	group means	`y ~ 1 + gid`
ANOVA	hier. model	`y ~ 1 + (1 | gid)`

...

- Possible to extend, e.g., with group specific variances and and different distributions such t- or Poisson distribution
  - and go beyond named tests
- See longer list and illustrations (with `lm`) at https://lindeloev.github.io/tests-as-linear/ and with rstanarm in Regression and other stories

Beta blockers

- An experiment was performed to estimate the effect of beta-blockers on mortality of cardiac patients
- A group of patients were randomly assigned to treatment and control groups:
  - out of 674 patients receiving the control, 39 died
  - out of 680 receiving the treatment, 22 died

Beta blockers

- An experiment was performed to estimate the effect of beta-blockers on mortality of cardiac patients
- A group of patients were randomly assigned to treatment and control groups:
  - out of 674 patients receiving the control, 39 died
  - out of 680 receiving the treatment, 22 died

```
d_bin2 <- data.frame(N = c(674, 680),
                     y = c(39,22),
                     grp2 = c(0,1))

fitb1 <- brm(y | trials(N) ~ 1, family = binomial(), data = d_bin2)

fitb2 <- brm(y | trials(N) ~ 1 + grp2, family = binomial(), data = d_bin2)
```

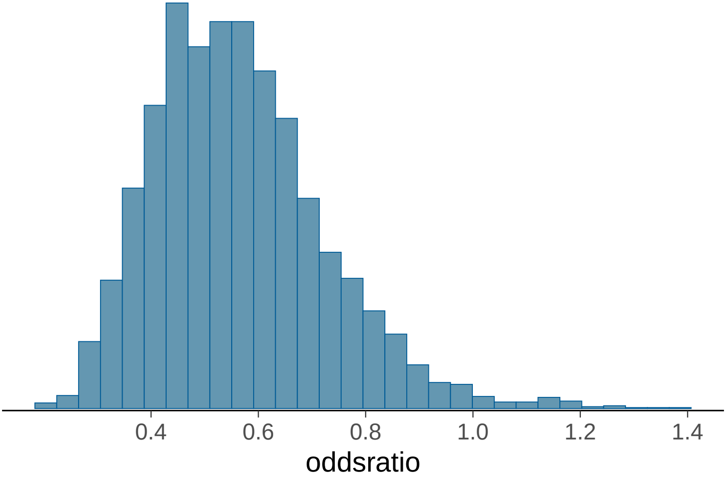
Beta blockers

- An experiment was performed to estimate the effect of beta-blockers on mortality of cardiac patients
- A group of patients were randomly assigned to treatment and control groups:
  - out of 674 patients receiving the control, 39 died
  - out of 680 receiving the treatment, 22 died

```
d_bin2 <- data.frame(N = c(674, 680),
                     y = c(39,22),
                     grp2 = c(0,1))

fitb1 <- brm(y | trials(N) ~ 1, family = binomial(), data = d_bin2)

fitb2 <- brm(y | trials(N) ~ 1 + grp2, family = binomial(), data = d_bin2)

> loo_compare(loo(fitb1),loo(fitb2))
      elpd_diff se_diff
fitb2  0.0       0.0
fitb1 -1.6       2.3
```

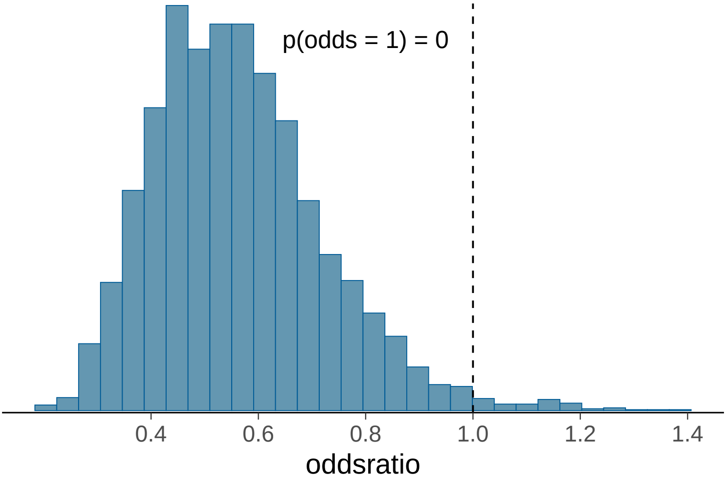
Posterior inference

- Instead of model selection, report full posterior and
  - compare to expert information
  - combine with utility/cost function



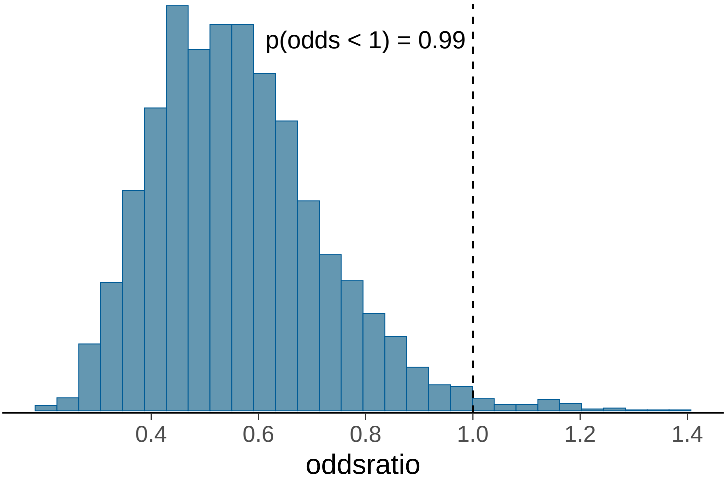
Posterior inference

- Instead of model selection, report full posterior
  - for continuous posterior there is zero probability that e.g. treatment effect is exactly zero



Posterior inference

- Instead of model selection, report full posterior
  - for continuous posterior we could report the probability that we know the sign of the effect

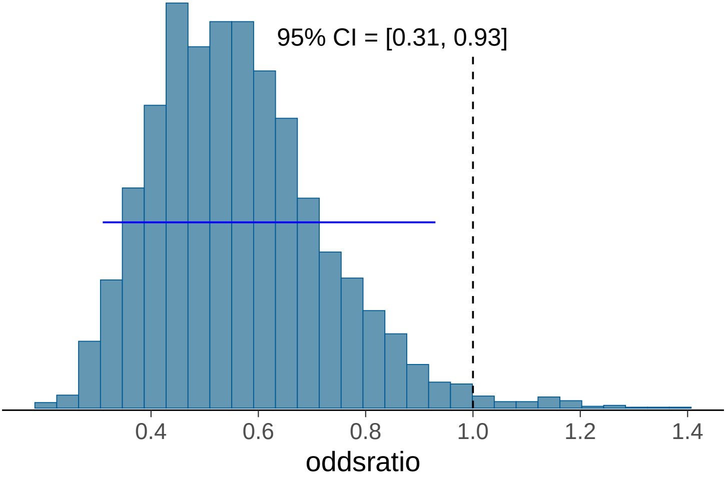


Bayesian hypothesis testing

- Sometimes people want to make a dichotomous choice
    - model selection
    - hypothesis testing

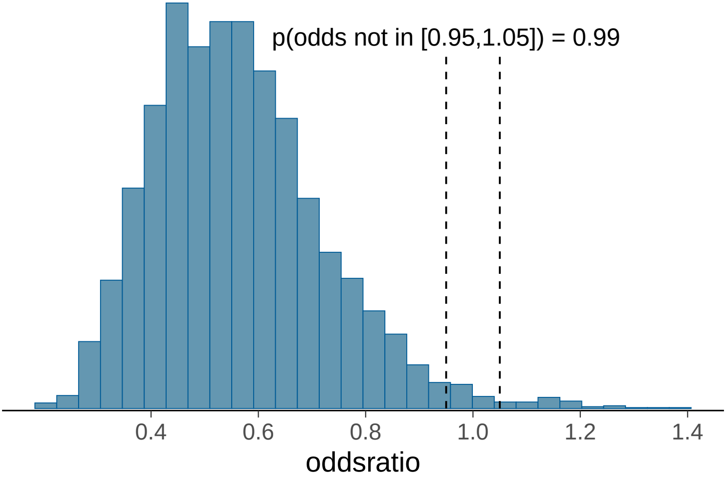
Bayesian hypothesis testing

- Instead of model selection, report full posterior and
  - for continuous posterior some people compare whether posterior interval includes null case



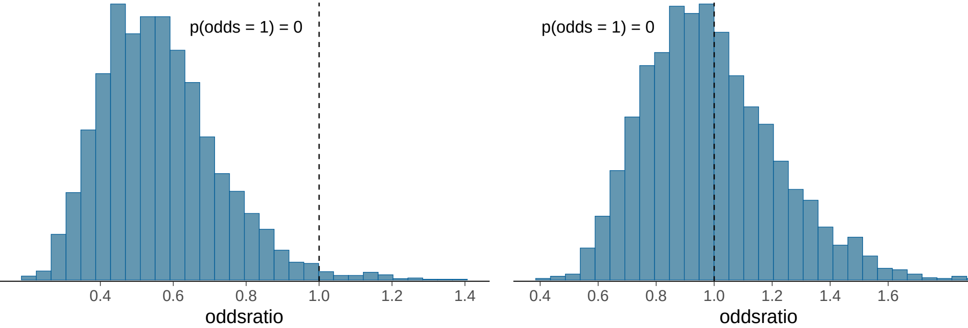
Bayesian hypothesis testing

- Equivalence testing (region of practical equivalence)
  - what is the probability that the effect is closer than ϵ to null, where ϵ is based on what is practically useful effect size



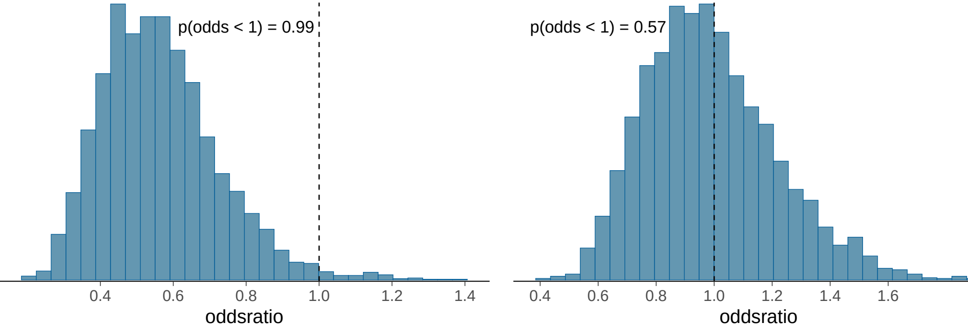
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior there is zero probability that e.g. treatment effect is exactly zero



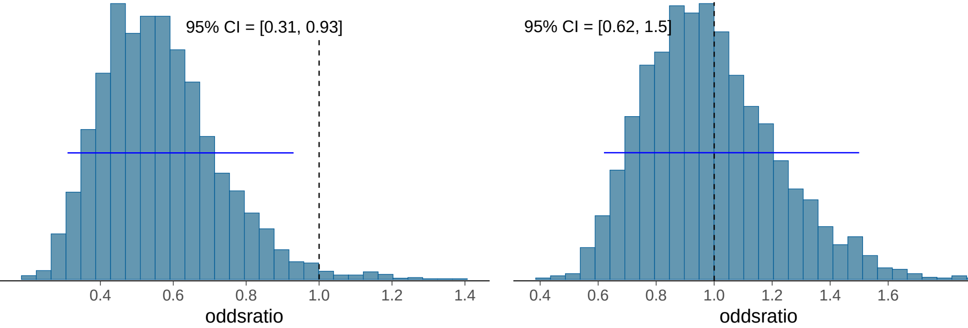
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior we could compute the probability that we know the sign of the effect



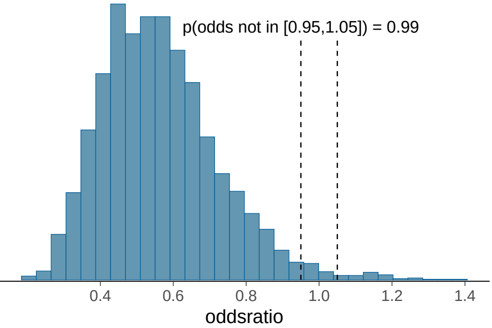
Bayesian hypothesis testing

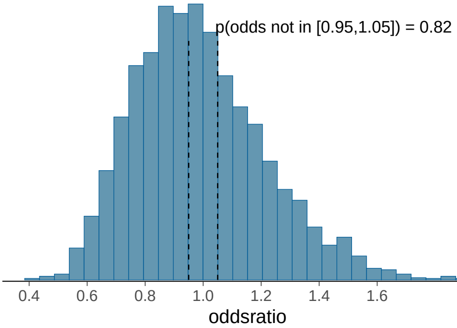
- Instead of hypothesis testing, report full posterior
  - for continuous posterior some people compare whether posterior interval includes null case



Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)



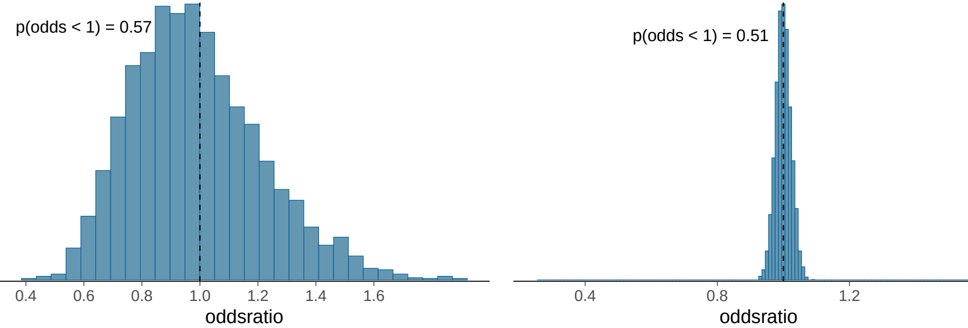


Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior there is zero probability that e.g. treatment effect is exactly zero

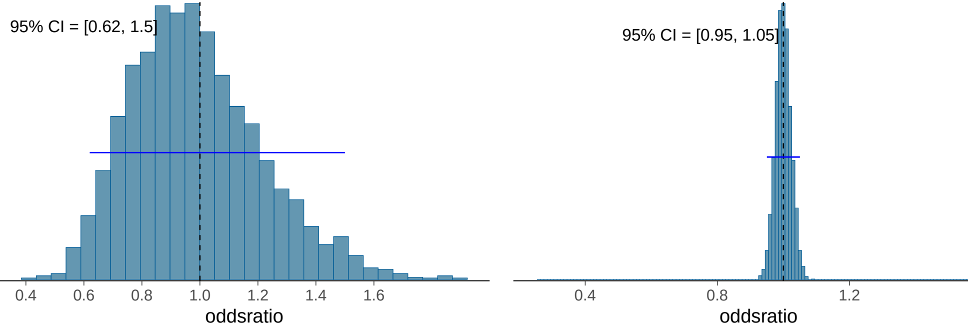
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior we could compute the probability that we know the sign of the effect



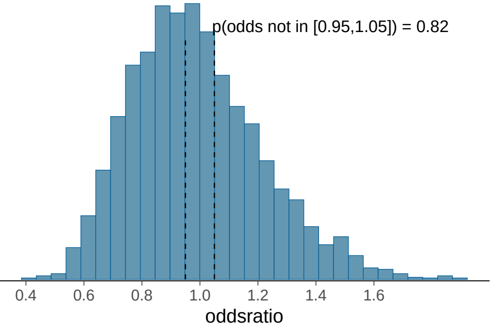
Bayesian hypothesis testing

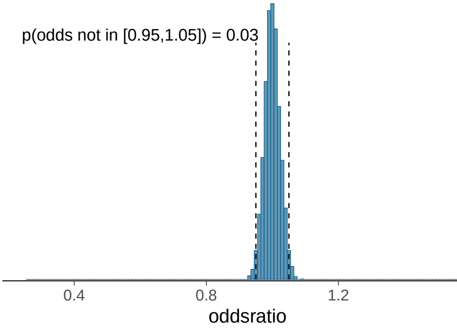
- Instead of hypothesis testing, report full posterior
  - for continuous posterior some people compare whether posterior interval includes null case



Bayesian hypothesis testing

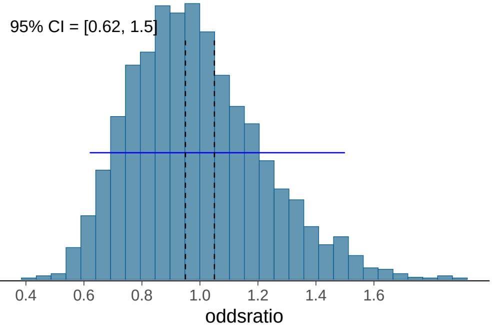
- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)

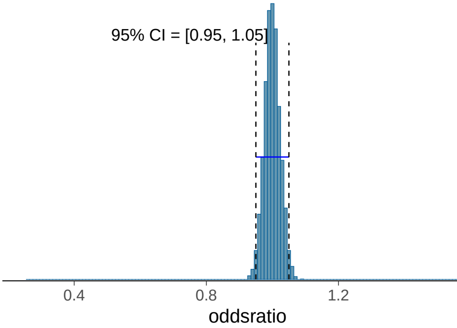




Bayesian hypothesis testing

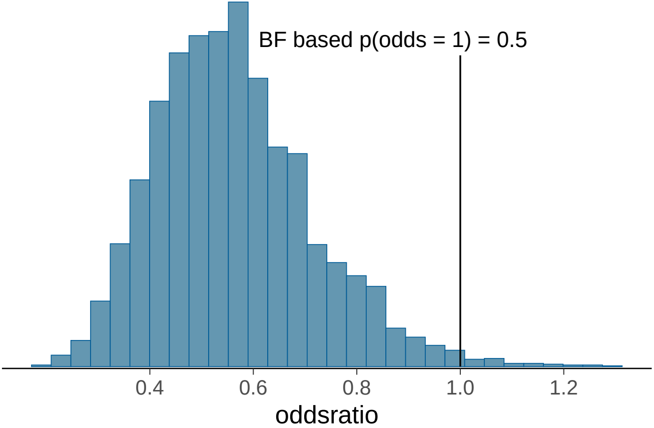
- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)





Bayesian hypothesis testing

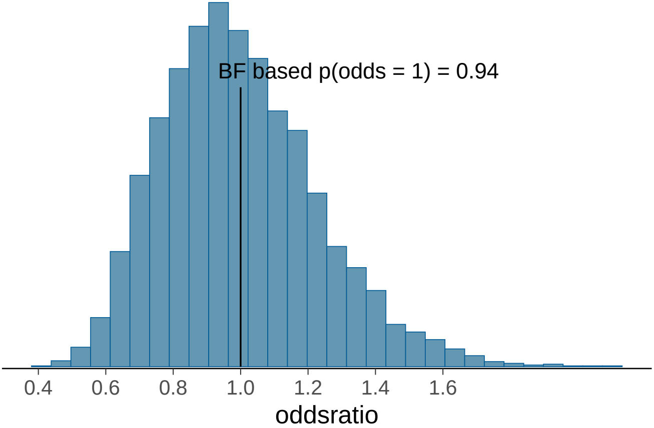
- Bayes factor
  - null model has, e.g., the treatment effect fixed to 0
  - assumes that there is non-zero probability that the treatment effect can be exactly zero (point mass)
  - requires posterior inference for the null model, too



with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

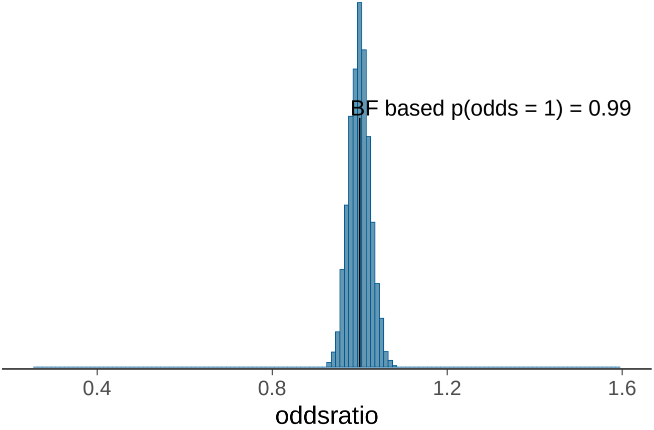
- Bayes factor
  - null model has, e.g., the treatment effect fixed to 0
  - assumes that there is non-zero probability that the treatment effect can be exactly zero (point mass)
  - requires posterior inference for the null model, too



with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

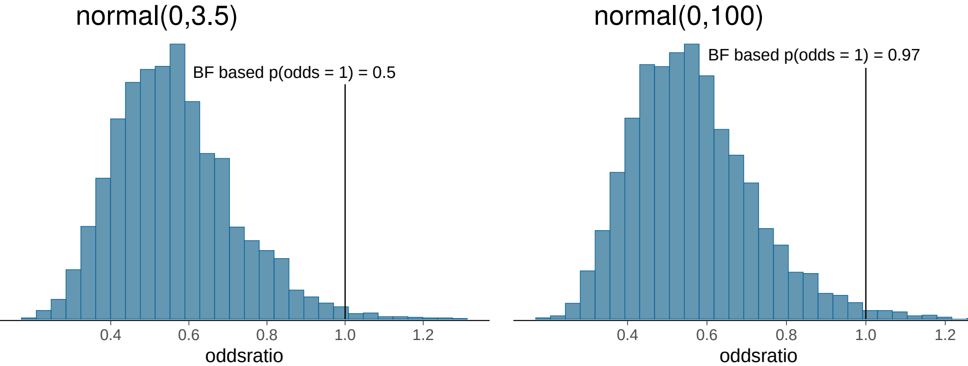
- Bayes factor
  - null model has, e.g., the treatment effect fixed to 0
  - assumes that there is non-zero probability that the treatment effect can be exactly zero (point mass)
  - requires posterior inference for the null model, too



with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

- Bayes factor
  - sensitive to the prior choice even when the posterior is not



with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

- Predictive performance
  - is there difference in predictive performance with, e.g., treatment effect fixed to zero or unknown treatment effect
  - requires posterior inference for the null model or projection from the full to null
  - looking at the posterior is better if parameters are independent

Bayesian hypothesis testing

- Predictive performance
  - is there difference in predictive performance with, e.g., treatment effect fixed to zero or unknown treatment effect
  - requires posterior inference for the null model or projection from the full to null
  - looking at the posterior is better if parameters are independent

In the beta blockers example

- Leave-one-person-out works, but is less efficient than looking at the posterior (see https://users.aalto.fi/~ave/modelselection/betablockers.html)

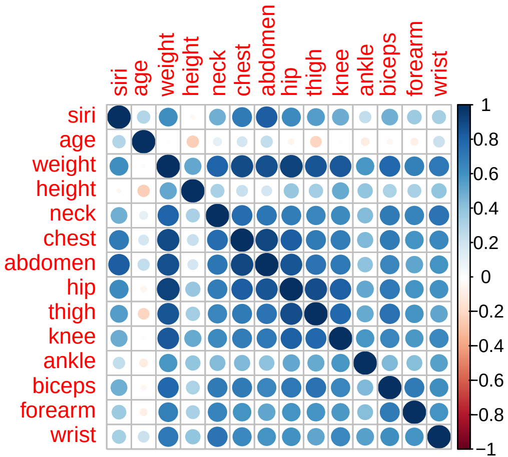
```
> loo_compare(loo(fitb1),loo(fitb2))
      elpd_diff se_diff
fitb2  0.0       0.0
fitb1 -1.6       2.3
```

Bodyfat: many predictors

- Predict bodyfat percentage
- The reference value (siri) is obtained by immersing person in water. $n = 251$.
- Which measurements to use in the future?

Bodyfat: many predictors

- Predict bodyfat percentage
- The reference value (siri) is obtained by immersing person in water. $n = 251$.
- Which measurements to use in the future?



Prediction

- Goal: prediction

Prediction

- Goal: prediction
- Use all the predictors and sensible prior

Prediction

- Goal: prediction
- Use all the predictors and sensible prior
  - no model selection needed

Predictive performance based variable selection

- Goal:
  - minimize future measurement cost
  - easier explainability of the model

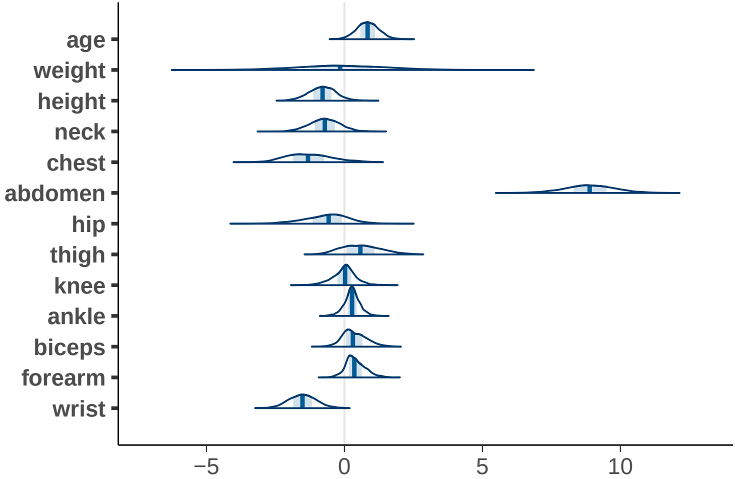
Predictive performance based variable selection

- Goal:
  - minimize future measurement cost
  - easier explainability of the model
- Select the minimal number of covariates with similar predictive performance as the full model

Hypothesis testing and posterior dependencies

Looking at the marginal posterior $p(\beta < 0)$ can be misleading when there are many parameters

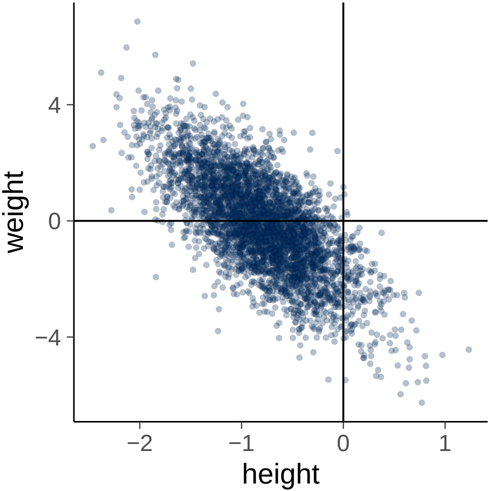
Marginal posteriors of coefficients in bodyfat example



Hypothesis testing and posterior dependencies

Looking at the marginal posterior(s) can be misleading when there are many parameters

Bivariate marginal of weight and height



Hypothesis testing and posterior dependencies

In bodyfat example, starting from full model

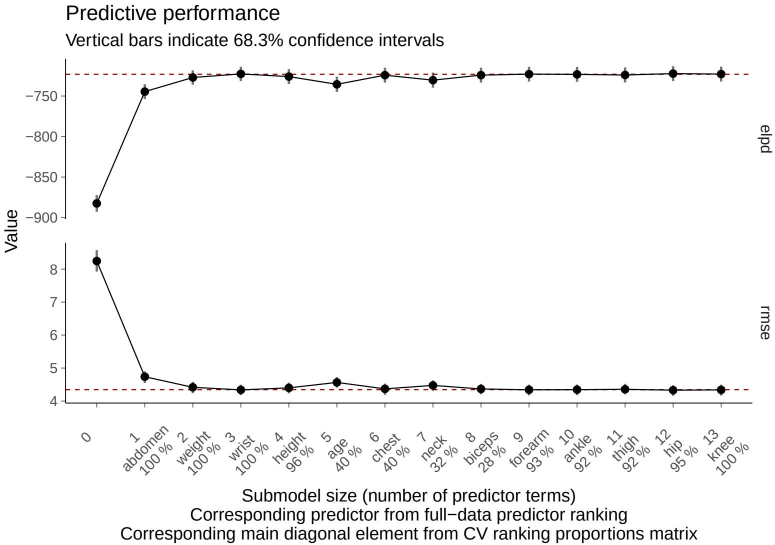
- BF in favor of removing weight (p=0.92)
- LOO in favor of removing weight (p=0.99)

In bodyfat example, starting from model y $\sim$ abdomen

- BF in favor of adding weight (p=1.0)
- LOO in favor of adding weight (p=1.0)

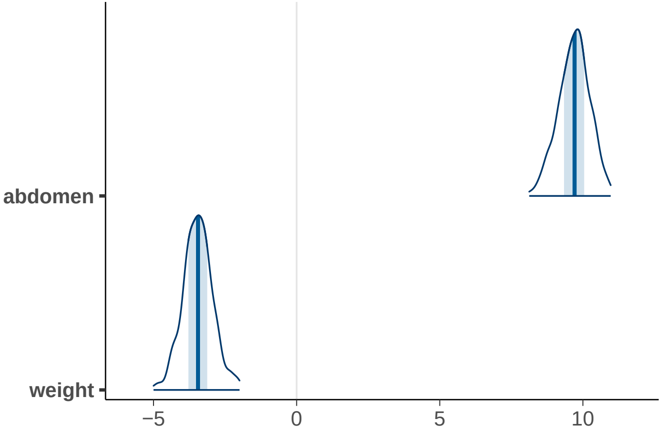
Predictive performance based variable selection

Projection predictive variable selection selects the minimal set of variables with similar predictive performance as the full model



Projected posterior

Projection predictive variable selection selects the minimal set of variables with similar predictive performance as the full model



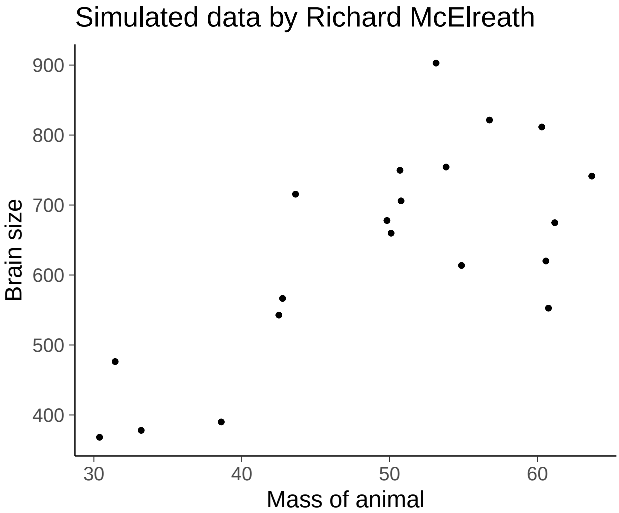
More about projpred in the end of the course

Model selection needed to avoid overfitting?

- Classic example is polynomial model with increasing number of components
  - overfits also with Bayesian inference and weak priors

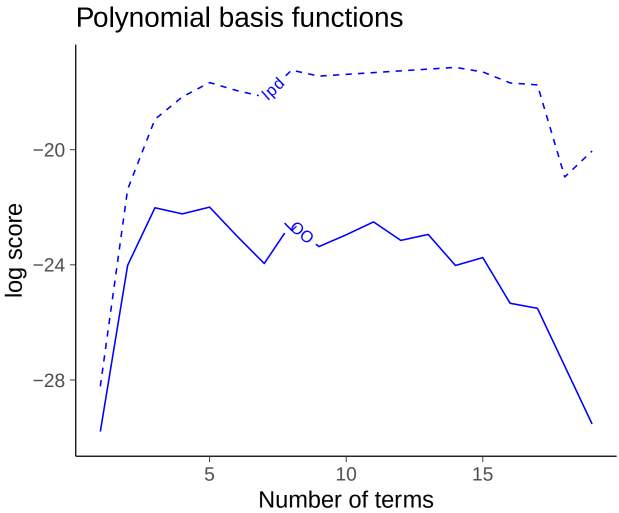
Model selection needed to avoid overfitting?

- Classic example is polynomial model with increasing number of components
  - overfits also with Bayesian inference and weak priors



Model selection needed to avoid overfitting?

- Classic example is polynomial model with increasing number of components
  - overfits also with Bayesian inference and weak priors



Model selection needed to avoid overfitting?

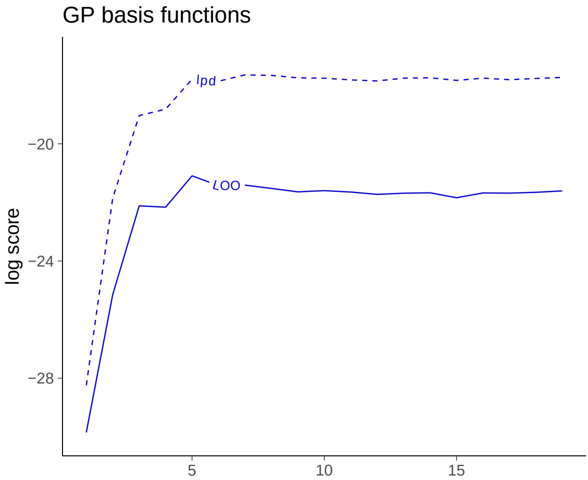
- Gaussian process can be used as a prior on function space
  - GP can be approximated with basis functions

Model selection needed to avoid overfitting?

- Gaussian process can be used as a prior on function space
  - GP can be approximated with basis functions
  - more basis functions makes the approximation more accurate, but doesn't inflate the prior on function space

Model is not needed to avoid overfitting

- Gaussian process can be used as a prior on function space
  - GP can be approximated with basis functions
  - more basis functions makes the approximation more accurate, but doesn't inflate the prior on function space

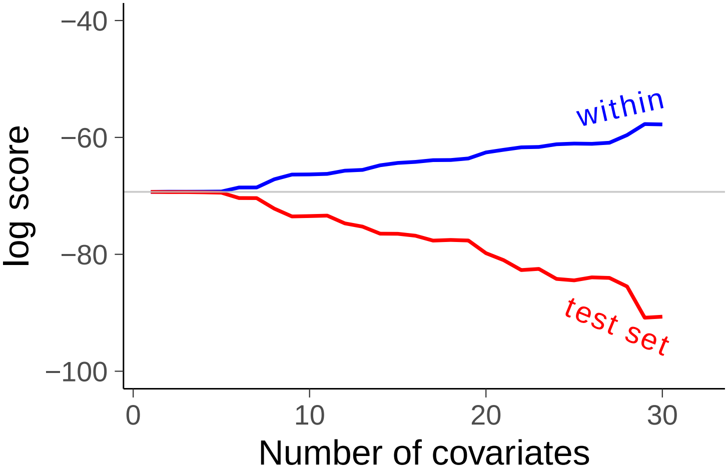


Model selection needed to avoid overfitting?

logistic regression: 30 **completely irrelevant** variables,
100 observations

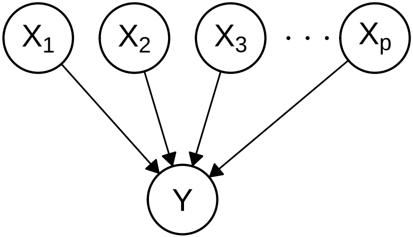
Model selection needed to avoid overfitting?

logistic regression: 30 **completely irrelevant** variables,
100 observations



Prior on parameters vs predictions

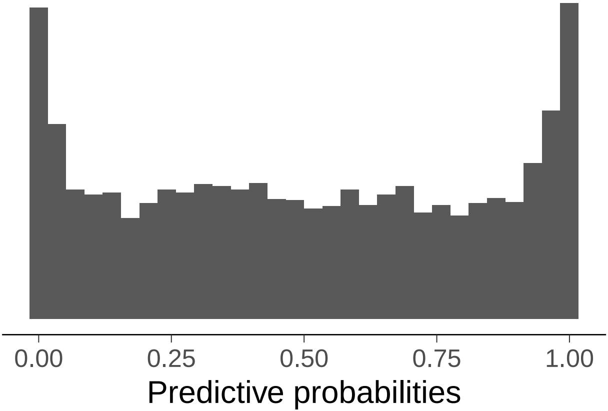
N(0,3) prior on each coefficient



Prior on parameters vs predictions

N(0,3) prior on each coefficient

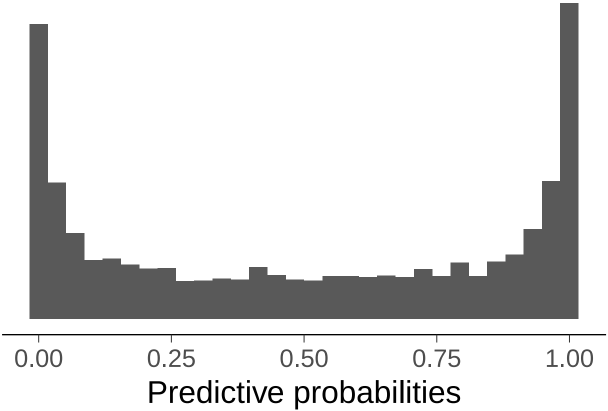
1 variable



Prior on parameters vs predictions

N(0,3) prior on each coefficient

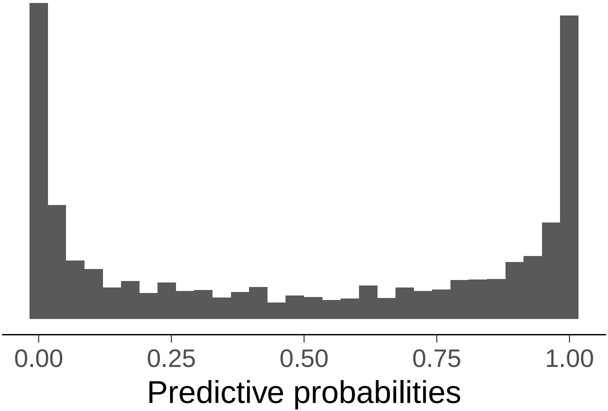
2 variables



Prior on parameters vs predictions

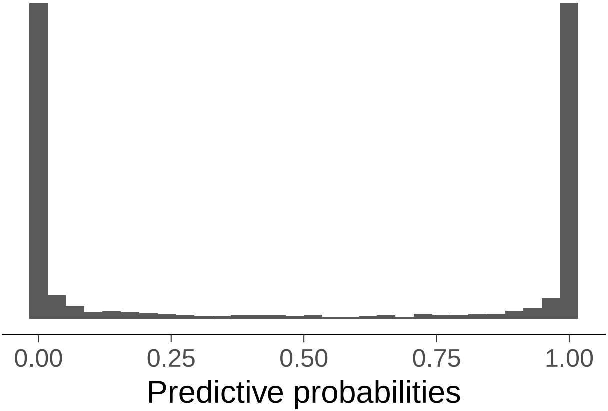
N(0,3) prior on each coefficient

3 variables



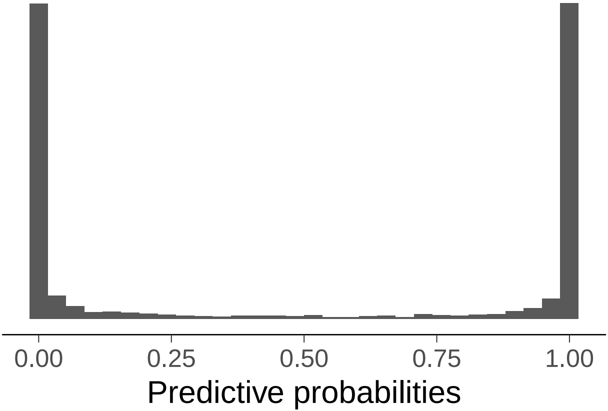
Prior on parameters vs predictions

N(0,3) prior on each coefficient
30 variables



Prior on parameters vs predictions

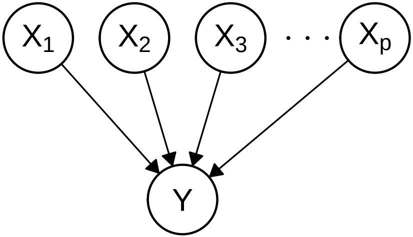
N(0,3) prior on each coefficient
30 variables



A weak prior on parameters can be a strong prior on predictions that favors overfitting

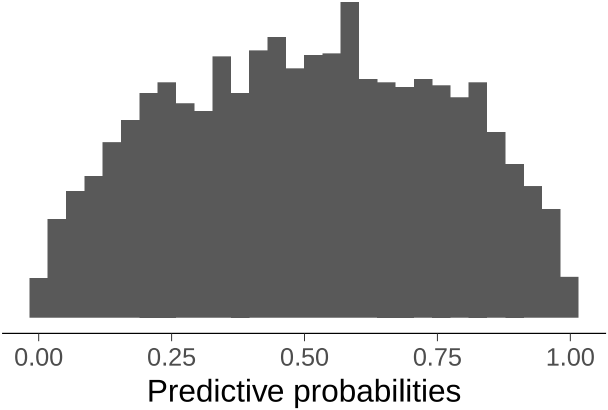
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient



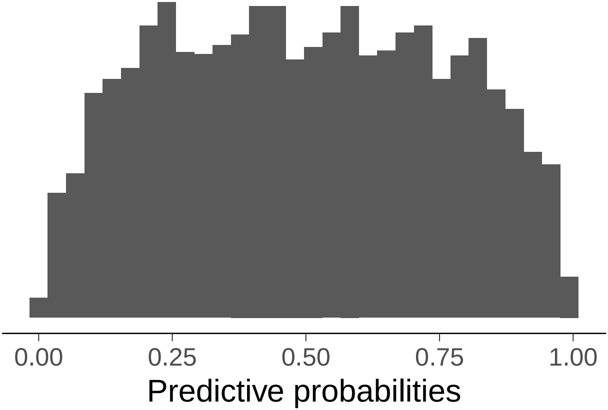
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
1 variable



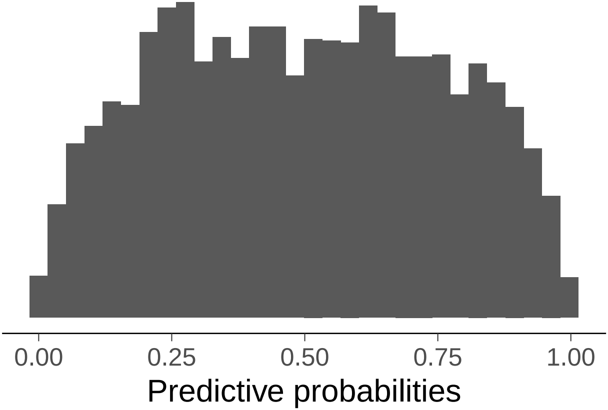
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
2 variables



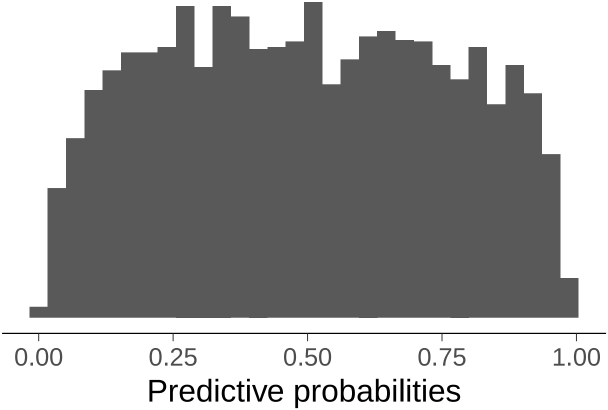
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
3 variables



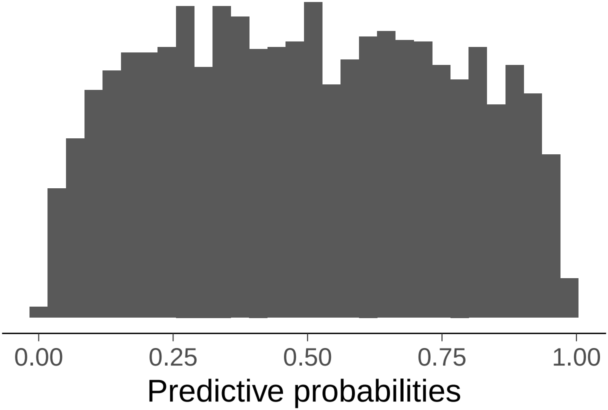
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
30 variables



Better priors

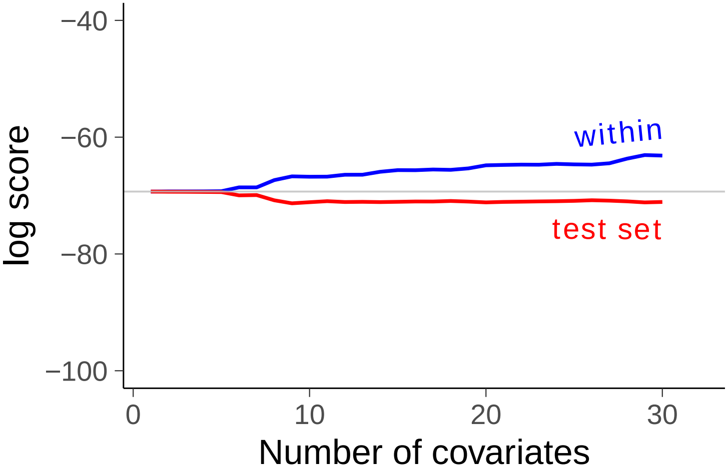
$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
30 variables



Prior on predictions (almost) fixed when the model gets bigger

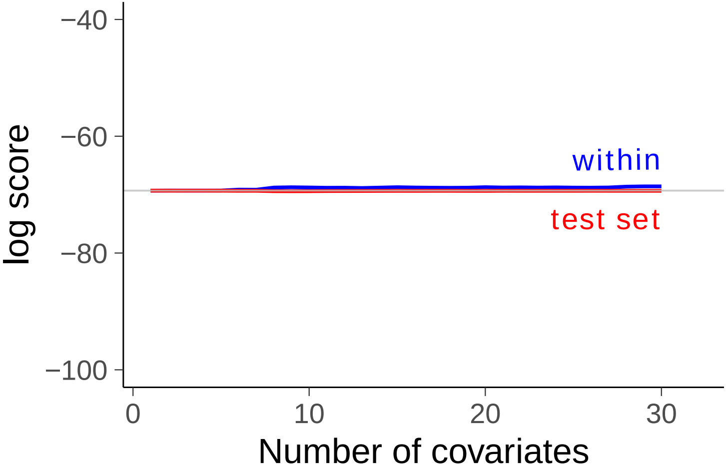
Better priors, no overfitting

logistic regression: 30 **completely irrelevant** variables, 100 observations, $N(0, \frac{1}{\sqrt{p}})$ prior



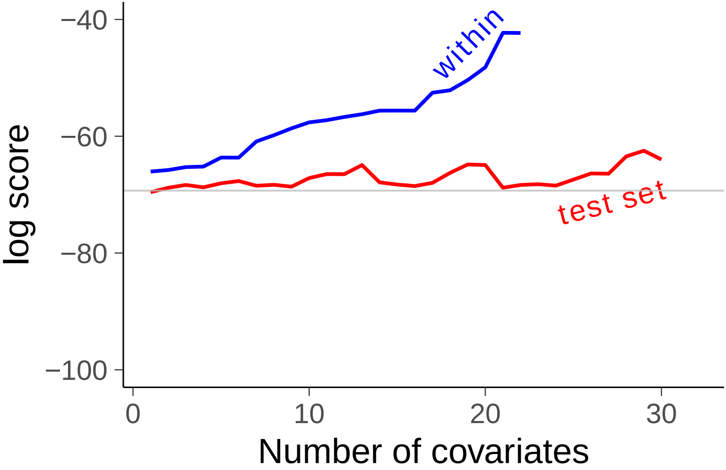
Better priors, no overfitting

logistic regression: 30 **completely irrelevant** variables,
100 observations, regularized horseshoe prior



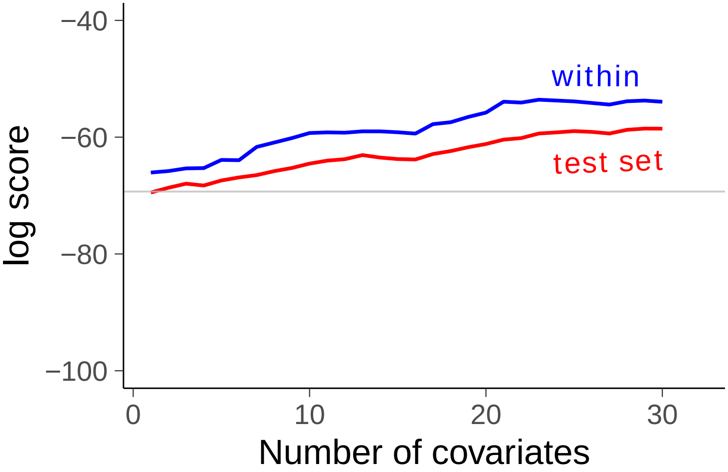
Many weak effects, wide prior on parameters

logistic regression: 30 **weakly relevant** variables,
100 observations, N(0,3) prior



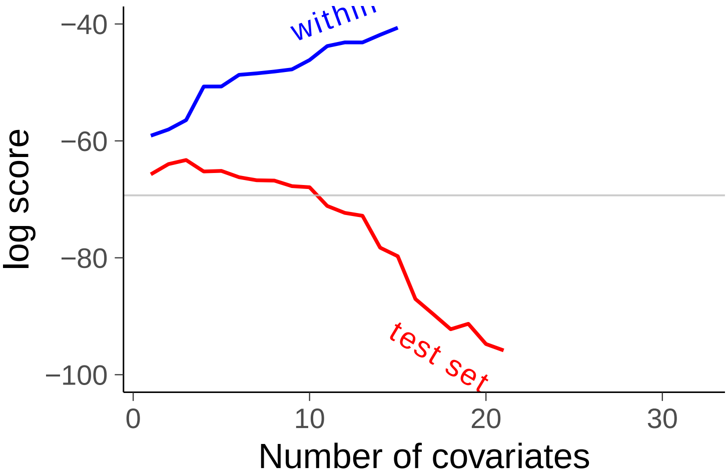
Many weak effects, better prior

logistic regression: 30 **weakly relevant** variables,
100 observations, N(0, $\frac{1}{\sqrt{p}}$) prior



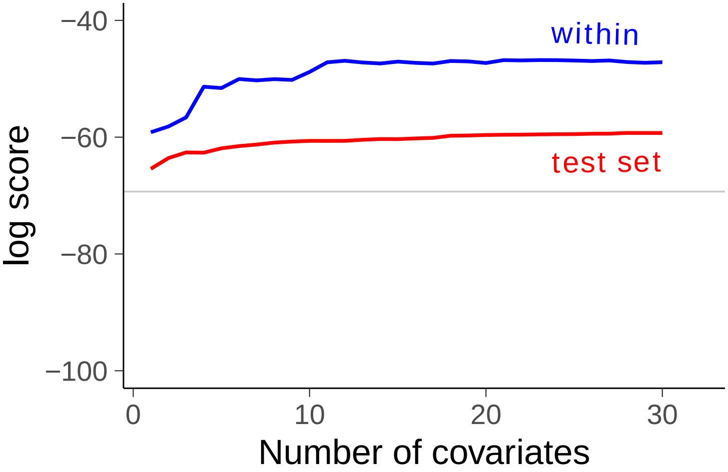
Correlating variables, wide prior on parameters

logistic regression: 30 **correlating relevant** variables,
100 observations, N(0,3) prior



Correlating variables, better prior

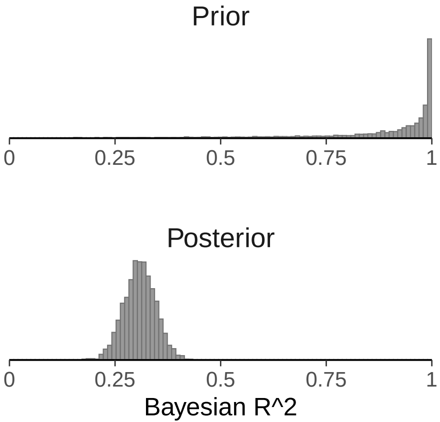
logistic regression: 30 **correlating relevant** variables,
100 observations N(0, $\frac{1}{\sqrt{p}}$) prior



Implied prior on R^2

Regression and Other Stories, Section 12.7 Models for regression coefficients:

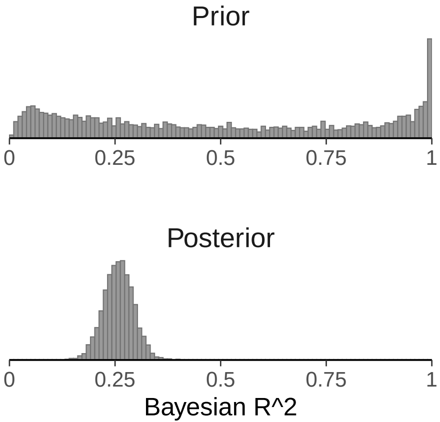
Wide prior on coefficients favors overfitting



Implied prior on R^2

Regression and Other Stories, Section 12.7 Models for regression coefficients:

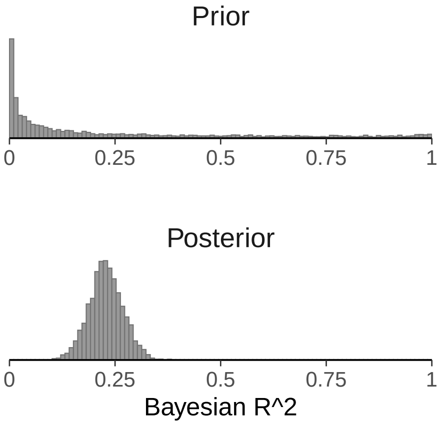
Scaled prior on coefficients



Implied prior on R^2

Regression and Other Stories, Section 12.7 Models for regression coefficients:

Regularized horseshoe prior on coefficients



Better priors

For example:

- scaled: many weak effects
- regularized horseshoe, R2-D2: only some relevant
- R2-D2: defined directly for R^2
- PCA-type: highly correlating variables

$p \gg n$

- With good priors, possible to have more variables than observations
- e.g. $p = 22283, n = 85$ demonstrated by Piironen, Paasiniemi, Vehtari (2020)

Variable selection

Variable selection

1. is not needed to avoid overfitting
2. can be used to reduce costs and improve explainability

Model selection can overfit

- Selection induced bias in cross-validation
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the CV estimate for the selected model is biased
  - recognized already, e.g., by Stone (1974)

Model selection can overfit

- Selection induced bias in cross-validation
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the CV estimate for the selected model is biased
  - recognized already, e.g., by Stone (1974)
- Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models

Model selection can overfit

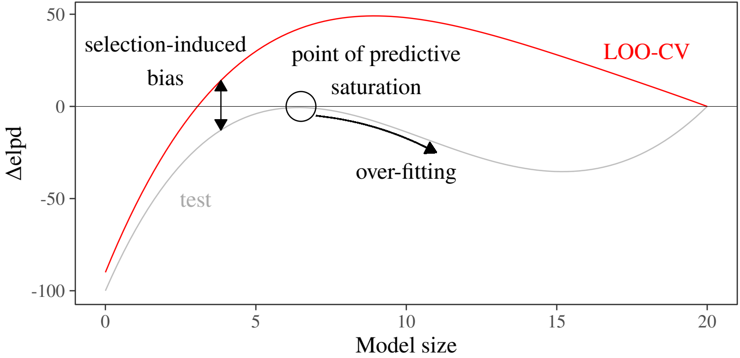
- Selection induced bias in cross-validation
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the CV estimate for the selected model is biased
  - recognized already, e.g., by Stone (1974)
- Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models
- Bigger problem if there is a large number of models as in covariate selection

Model selection can overfit

- Variable selection with forward selection
  - start with null model
  - add the variable improving the predictive performance most
  - add the next variable improving... and so on

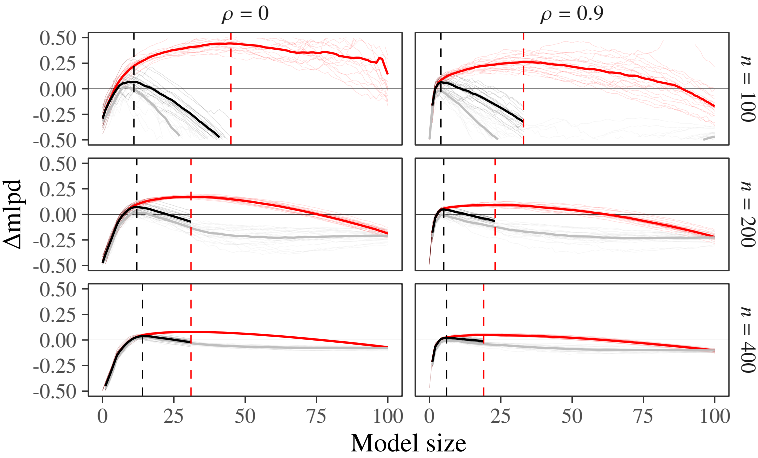
Model selection can overfit

- Variable selection with forward selection
  - start with null model
  - add the variable improving the predictive performance most
  - add the next variable improving... and so on



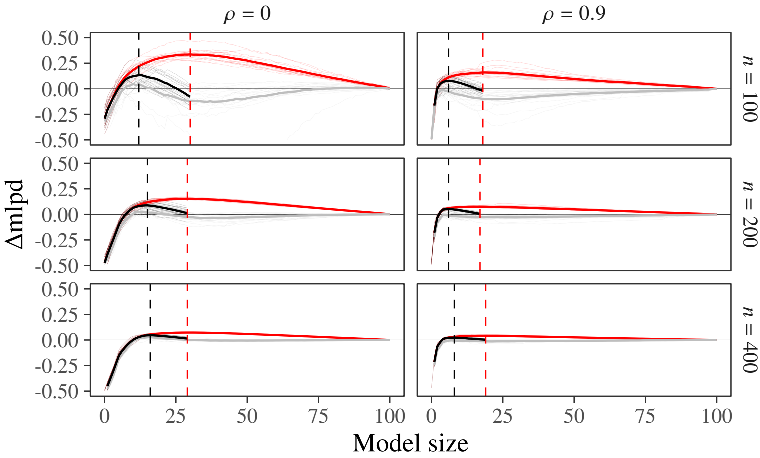
Model selection can overfit

Wide normal prior



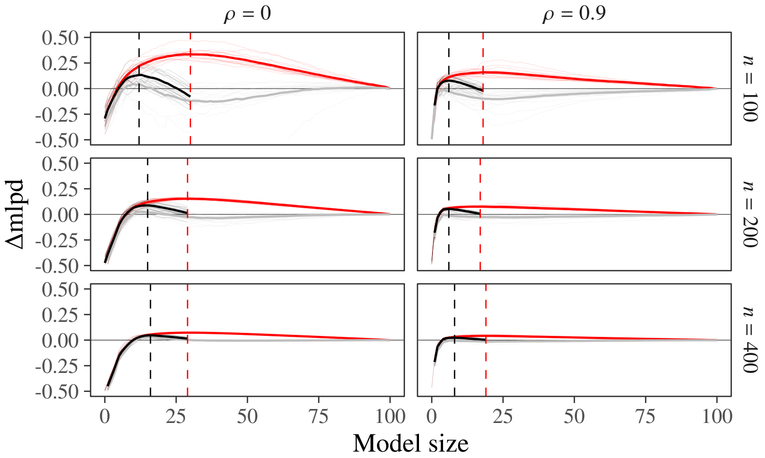
Model selection can overfit

R2D2 prior reduces overfit in model selection



Model selection can overfit

R2D2 prior reduces overfit in model selection



Reminder: variable selection is not needed with good priors to get good predictive performance, but may be useful for other purposes

Model averaging

- Prefer continuous model expansion

Model averaging

- Prefer continuous model expansion
- If needed integrate over the model space = model averaging

Model averaging

- Prefer continuous model expansion
- If needed integrate over the model space = model averaging
- Bayesian model averaging is just the usual integration over unknowns

Model averaging

- Prefer continuous model expansion
- If needed integrate over the model space = model averaging
- Bayesian model averaging is just the usual integration over unknowns
- Bayesian stacking may work better than BMA in case of misspecified models or small data
  - Yao, Vehtari, Simpson, and Gelman (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917-1003

Cross-validation and model selection

- Cross-validation can be used for model selection if
  - small number of models
  - the difference between models is clear

Cross-validation and model selection

- Cross-validation can be used for model selection if
  - small number of models
  - the difference between models is clear
- Be careful if using cross-validation to choose from a large set of models
  - selection process can lead to severe overfitting

Cross-validation and model selection

- Cross-validation can be used for model selection if
  - small number of models
  - the difference between models is clear
- Be careful if using cross-validation to choose from a large set of models
  - selection process can lead to severe overfitting
- Overfitting in selection process is not unique for cross-validation

Take-home messages

- It's good to think predictions of observables, because observables are the only ones we can observe
- Cross-validation can simulate predicting and observing new data
- Cross-validation is good if you don't trust your model
- Different variants of cross-validation are useful in different scenarios
- Cross-validation has high variance, and **if** you trust your model you can beat cross-validation in accuracy

Take-home messages

- It's good to think predictions of observables, because observables are the only ones we can observe
- Cross-validation can simulate predicting and observing new data
- Cross-validation is good if you don't trust your model
- Different variants of cross-validation are useful in different scenarios
- Cross-validation has high variance, and **if** you trust your model you can beat cross-validation in accuracy

Take-home messages

- It's good to think predictions of observables, because observables are the only ones we can observe
- Cross-validation can simulate predicting and observing new data
- Cross-validation is good if you don't trust your model
- Different variants of cross-validation are useful in different scenarios
- Cross-validation has high variance, and **if** you trust your model you can beat cross-validation in accuracy

Take-home messages

- It's good to think predictions of observables, because observables are the only ones we can observe
- Cross-validation can simulate predicting and observing new data
- Cross-validation is good if you don't trust your model
- Different variants of cross-validation are useful in different scenarios
- Cross-validation has high variance, and **if** you trust your model you can beat cross-validation in accuracy

Take-home messages

- It’s good to think predictions of observables, because observables are the only ones we can observe
- Cross-validation can simulate predicting and observing new data
- Cross-validation is good if you don’t trust your model
- Different variants of cross-validation are useful in different scenarios
- Cross-validation has high variance, and **if** you trust your model you can beat cross-validation in accuracy