UPPSALA UNIVERSITET

Bayesian Statistics and Data Analysis
Lecture 8b

Måns Magnusson
Department of Statistics, Uppsala University
Thanks to Aki Vehtari, Aalto University

Section 1

Model assessment and selection

UPPSALA UNIVERSITET

Predictive performance

- Modeling complex phenomena with models that are simplified
  All models are wrong... but some are useful.

Predictive performance

- Modeling complex phenomena with models that are simplified
  All models are wrong... but some are useful.
- True predictive performance is found out by using it to make predictions and comparing predictions to true observations
  - external validation

Predictive performance

- Modeling complex phenomena with models that are simplified
  All models are wrong... but some are useful.
- True predictive performance is found out by using it to make predictions and comparing predictions to true observations
  - external validation
- Expected predictive performance
  - approximates the external validation

Predictive performance

- Model choice is a (model-)decision-theoretic problem (see next week)

Predictive performance

- Model choice is a (model-)decision-theoretic problem (see next week)
- Choose the model function to maximize our utility

Predictive performance

- Model choice is a (model-)decision-theoretic problem (see next week)
- Choose the model function to maximize our utility
- Application specific utility/cost functions are important
  - eg. money, life years, quality adjusted life years, etc.

Predictive performance

- Model choice is a (model-)decision-theoretic problem (see next week)
- Choose the model function to maximize our utility
- Application specific utility/cost functions are important
  - eg. money, life years, quality adjusted life years, etc.
- General utility: overall in the goodness of the predictive distribution
  - we don't know (yet) the application specific utility

then good information theoretically justified choice is log-score

$$\log p(y^{\mathrm{rep}}|y, M)$$

Background and Motivation

- Evaluate how model M *generalizes* to unseen data (the *expected log predictive density*):

$$\text{elpd}_M = \int \log p_M(\tilde{y}_i|y) p_{\text{true}}(\tilde{y}_i) d\tilde{y}_i \,,$$

where $\tilde{y}_i$ is an unseen observation generated from the true data generating process $p_t(\tilde{y}_i)$, and y are observed data.

Section 2

Cross-validation

UPPSALA UNIVERSITET

Background and Motivation

- Evaluate how model M *generalizes* to unseen data (the *expected log predictive density*):

$$\mathrm{elpd}_M = \int \log p_M(\tilde{y}_i | y) p_{\mathrm{true}}(\tilde{y}_i) d\tilde{y}_i \,,$$

where $\tilde{y}_i$ is an unseen observation generated from the true data generating process $p_t(\tilde{y}_i)$, and y are observed data.

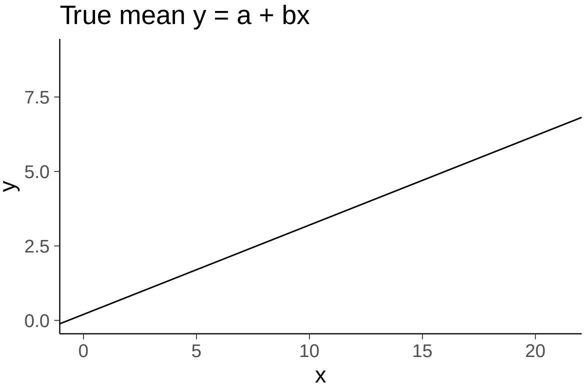
UPPSALA UNIVERSITET

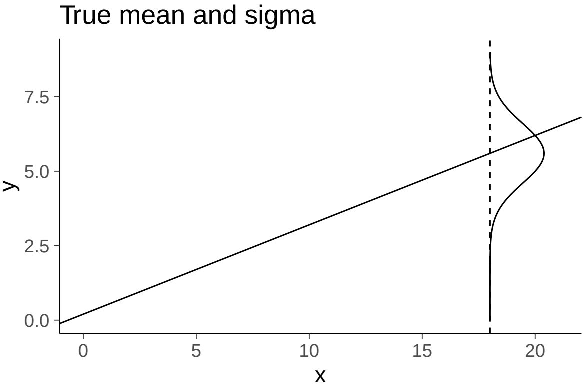
Leave-one-out cross-validation (LOO-CV)

- Hold out observation i and try to predict y_i based on $\mathbf{y}_{-i}$
- Estimation of elpd_M using leave-one-out cross-validation

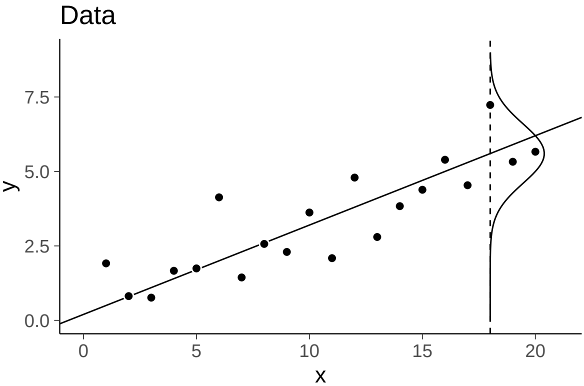
$$
\begin{aligned}
\text{elpd}_{\text{loo}} &= \sum_{i=1}^{n} \log p_M(y_i|\mathbf{y}_{-i}) \\
&= \sum_{i=1}^{n} \log \int p_M(y_i|\theta)p(\theta|\mathbf{y}_{-i})d\theta \\
&= \sum_{i=1}^{n} \pi_i
\end{aligned}
$$

- Pareto-Smoothed Importance Sampling (PSIS) can be used to efficiently compute each π_i (see Gelfand, 1996 and Vehtari et. al., 2019)

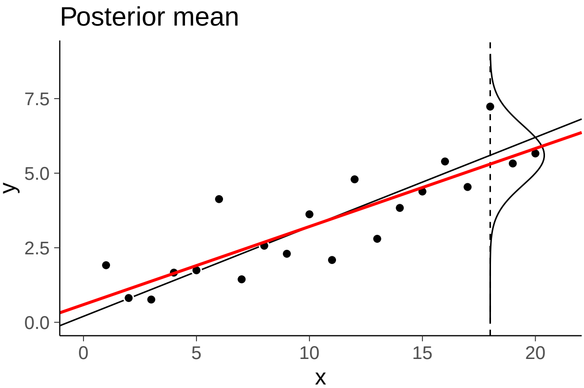

True mean y = a + bx
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

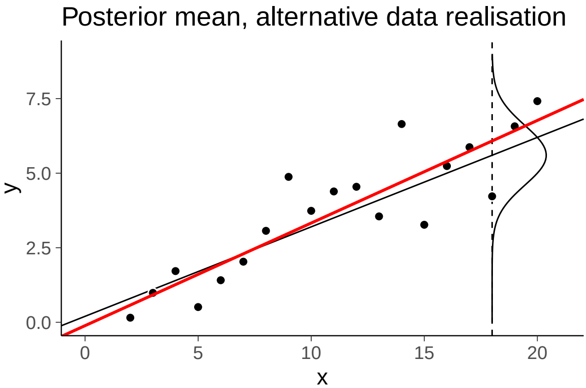

True mean and sigma
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

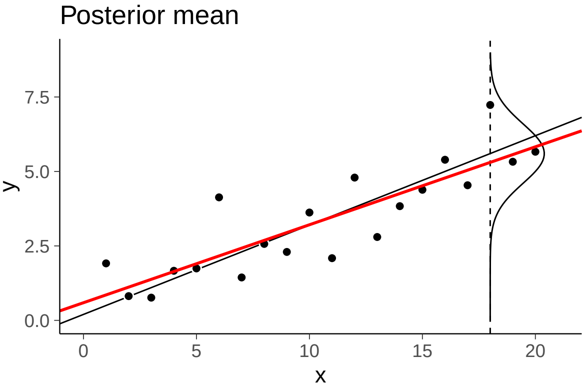
UPPSALA UNIVERSITET



UPPSALA UNIVERSITET

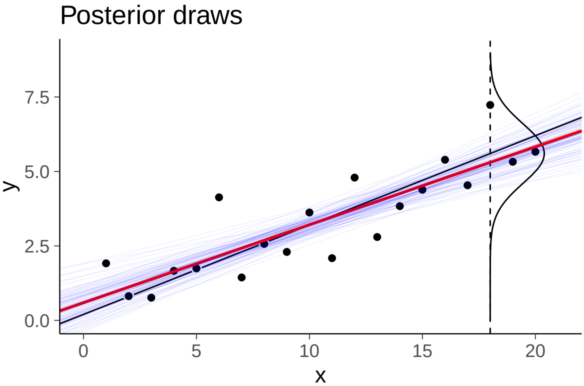


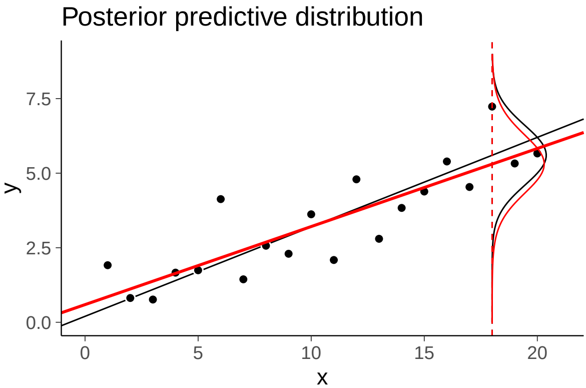

Posterior mean, alternative data realisation
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

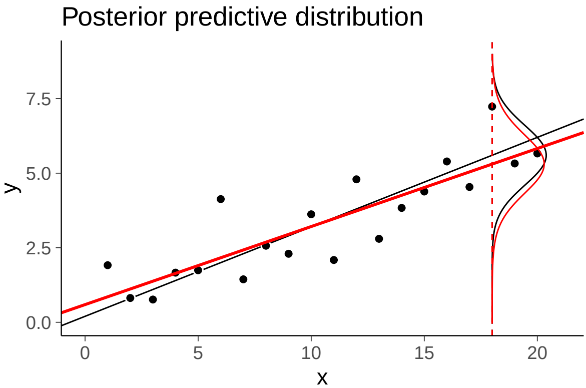

Posterior mean
y
7.5
5.0
2.5
0.0
x
0
5
10
15
20

UPPSALA UNIVERSITET

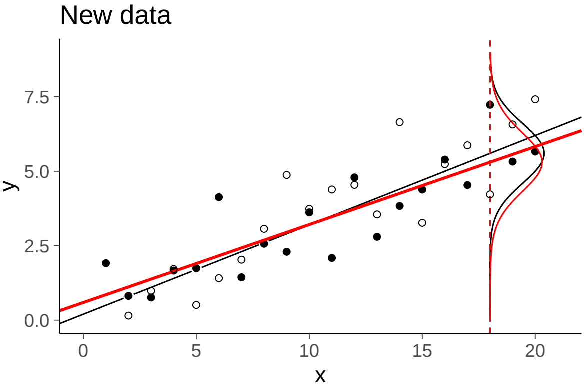
- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up

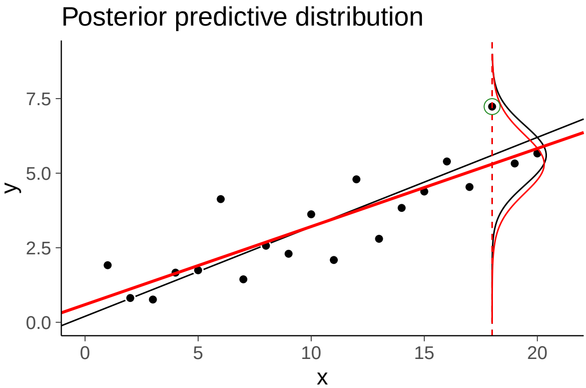



Posterior predictive distribution
y
x
0.0
2.5
5.0
7.5
0
5
10
15
20

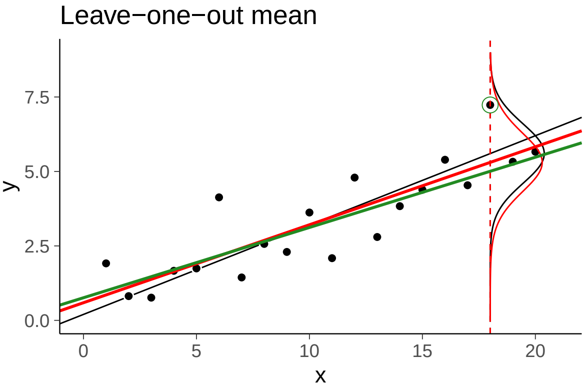


$$p(\tilde{y}|\tilde{x} = 18, x, y) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x, y)d\theta$$


New data
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

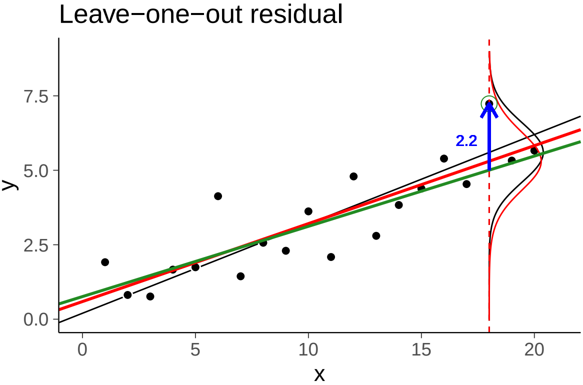

Posterior predictive distribution
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

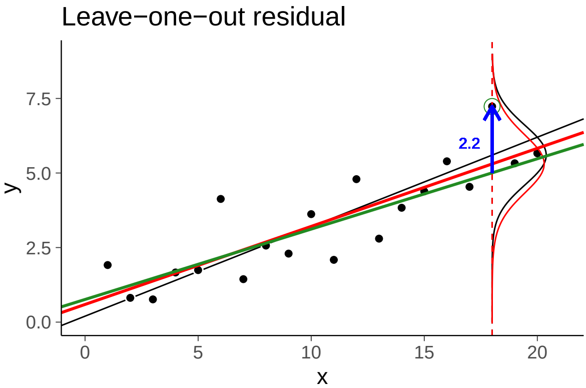
UPPSALA UNIVERSITET



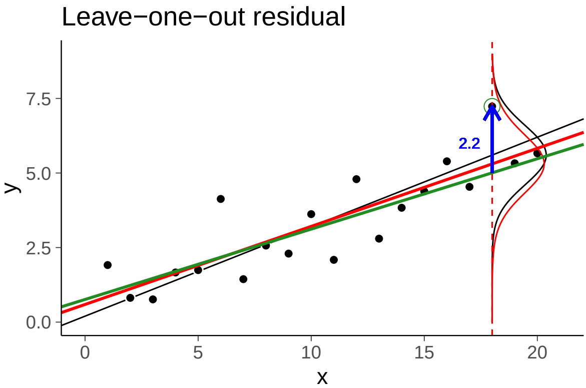
UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up





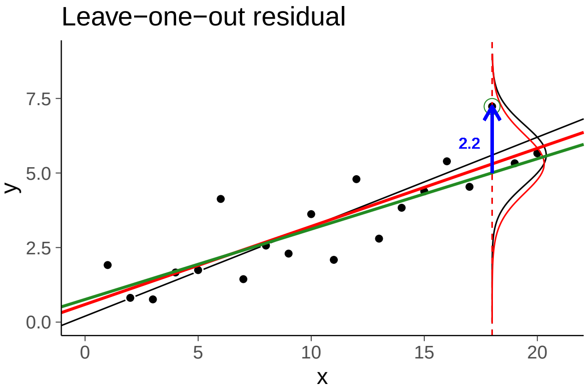
$$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$$



$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

Can be use to compute, e.g., RMSE, R^2, 90% error

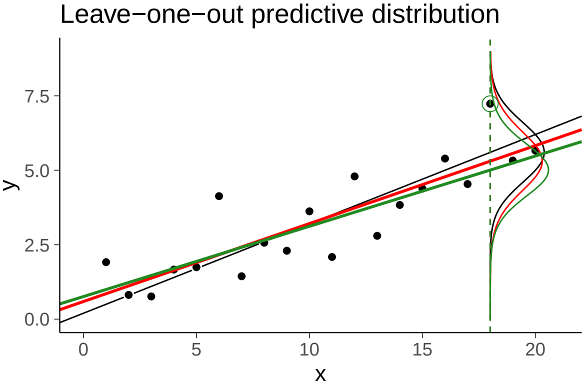
UPPSALA UNIVERSITET

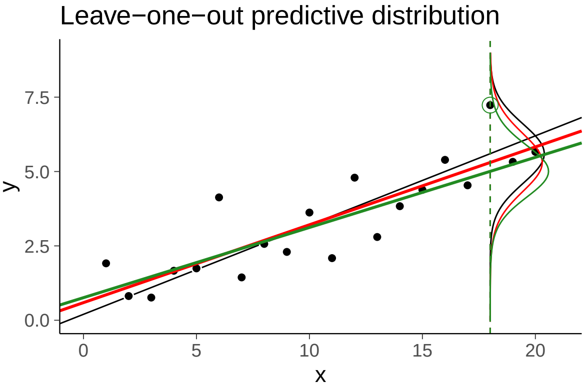


$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

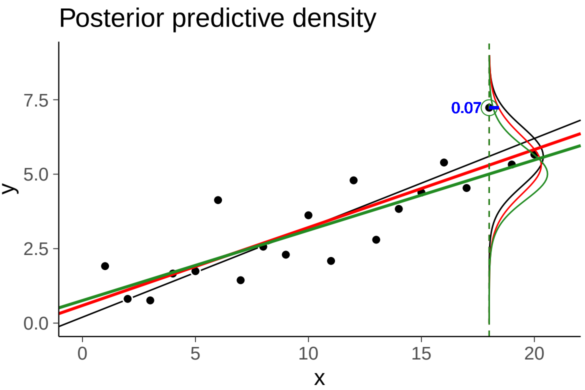
Can be use to compute, e.g., RMSE, R^2, 90% error

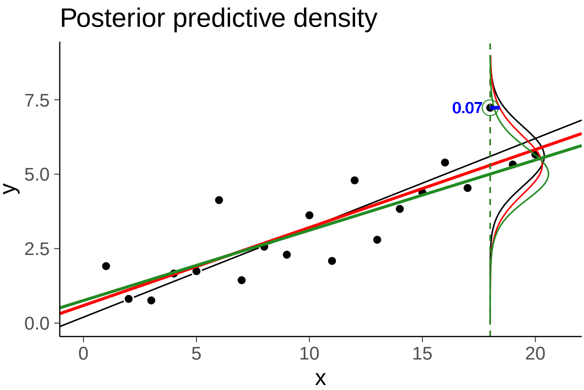
See LOO-R^2 at avehtari.github.io/bayes_R2/bayes_R2.html


Leave−one−out predictive distribution
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

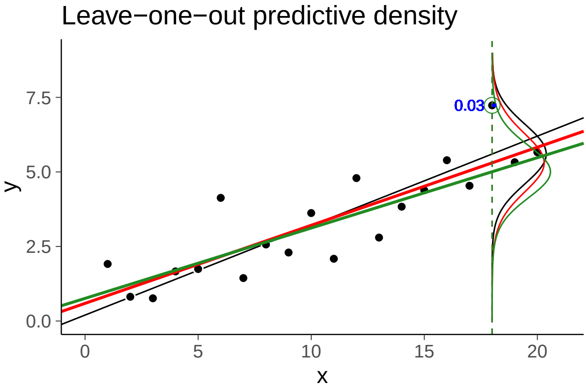


$$p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18}) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x_{-18}, y_{-18})d\theta$$


Posterior predictive density
0.07
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

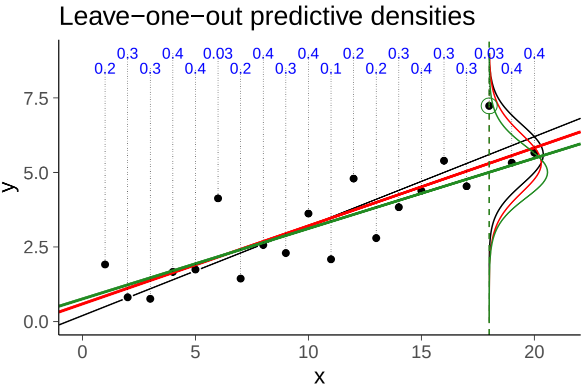


$$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$$



$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$

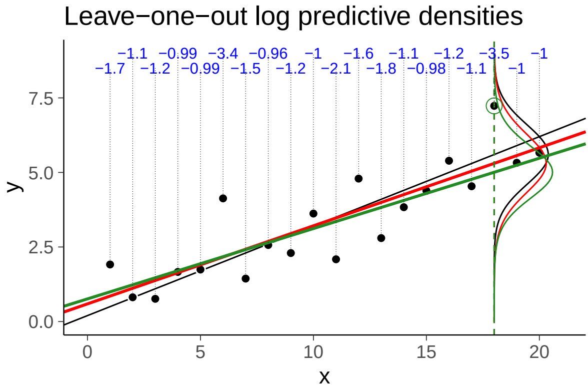
$p(\tilde{y} = y_{18} | \tilde{x} = 18, x_{-18}, y_{-18}) \approx 0.03$



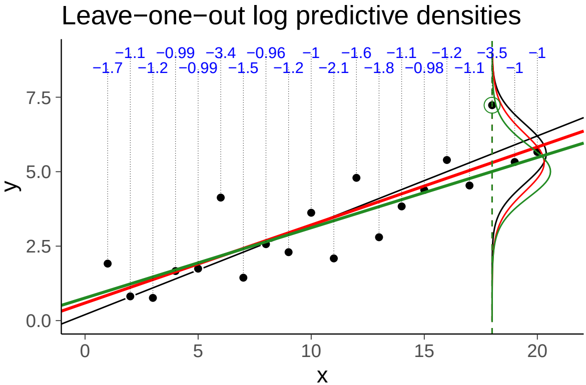
$p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

UPPSALA UNIVERSITET

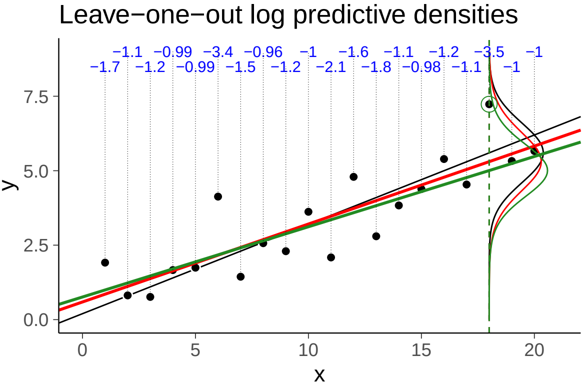
- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up



$$\log p(y_i | x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$$



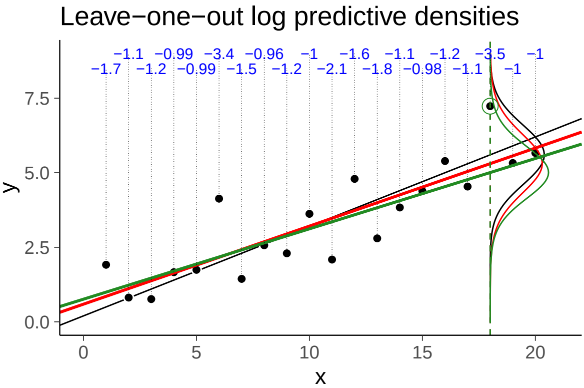
$\sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$



$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$$

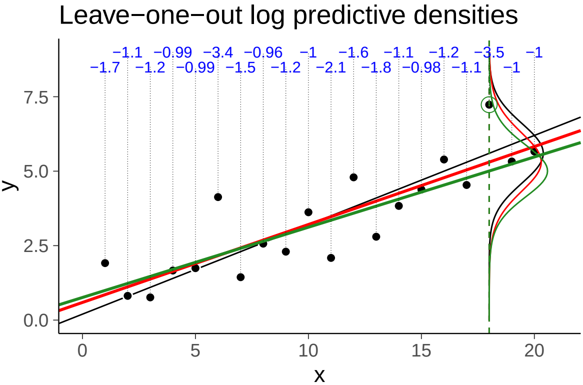
UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up



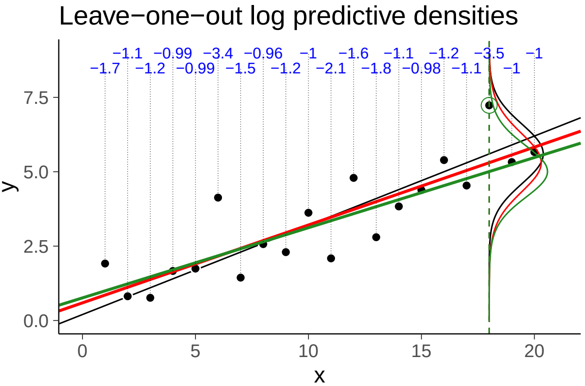
$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$

unbiased estimate of log posterior pred. density for new data



$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$$

$$\text{lpd} = \sum_{i=1}^{20} \log p(y_i|x_i, x, y) \approx -26.8$$



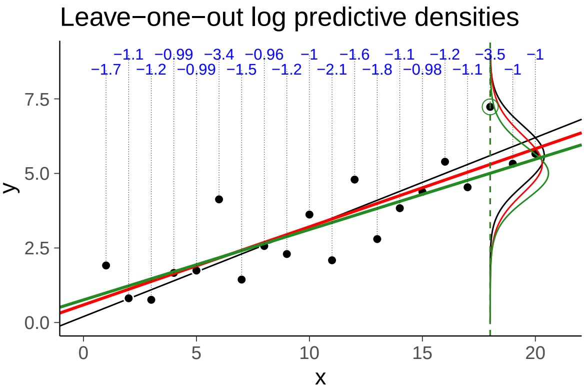
$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$$

$$\text{lpd} = \sum_{i=1}^{20} \log p(y_i | x_i, x, y) \approx -26.8$$

$$\text{p\_loo} = \text{lpd} - \text{elpd\_loo} \approx 2.7$$

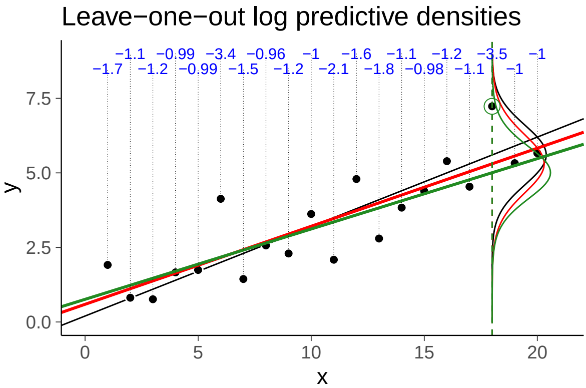
UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up



$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

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



$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$$

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

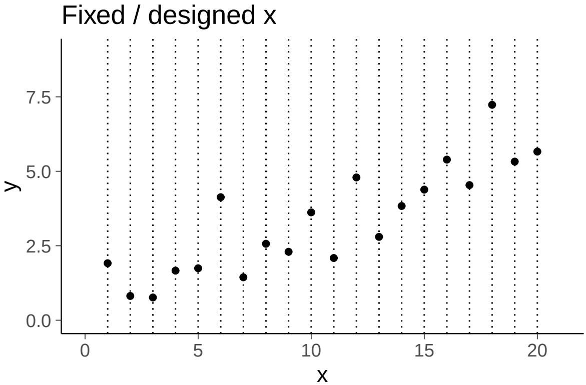
see Vehtari, Gelman & Gabry (2017a) and Vehtari & Ojanen (2012) for more

UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and loo
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up

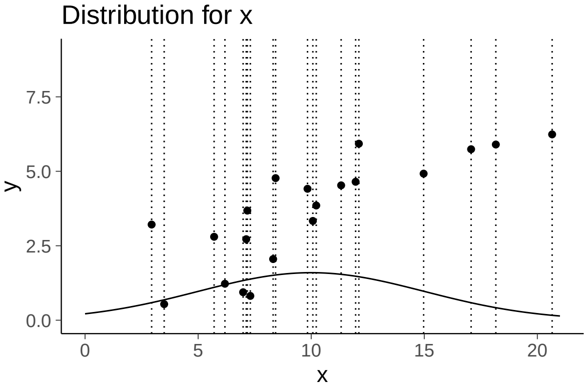
Stan and loo package

```
 Computed from 4000 by 20 log-likelihood matrix

         Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0
------
Monte Carlo SE of elpd_loo is 0.1.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     18    90.0%   899
 (0.5, 0.7]   (ok)        2    10.0%   459
   (0.7, 1]   (bad)       0     0.0%   <NA>
   (1, Inf)   (very bad)  0     0.0%   <NA>

All Pareto k estimates are ok (k < 0.7).
See help('pareto-k-diagnostic') for details.

Model comparison:
(negative 'elpd_diff' favors 1st model, positive favors 2nd)

elpd_diff        se
     -0.2       0.1
```



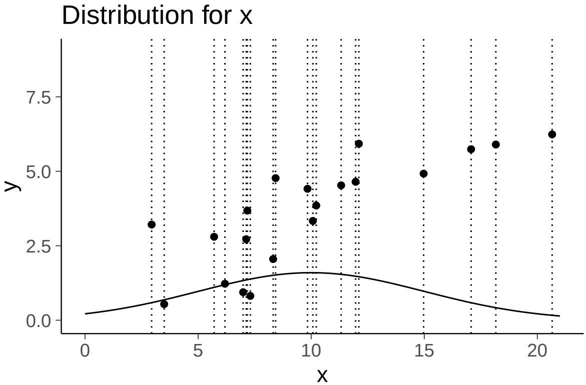
LOO is ok for fixed / designed x. SE is uncertainty about $y|x$.

see Vehtari & Ojanen (2012) and `andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/`



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

see Vehtari & Ojanen (2012) and `andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/`



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

Covariate shift can be handled with importance weighting or modelling

see Vehtari & Ojanen (2012) and `andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/`

Stan and loo package

```
Computed from 4000 by 20 log-likelihood matrix

         Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0
------
Monte Carlo SE of elpd_loo is 0.1.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     18    90.0%   899
 (0.5, 0.7]   (ok)        2    10.0%   459
   (0.7, 1]   (bad)       0     0.0%   <NA>
   (1, Inf)   (very bad)  0     0.0%   <NA>

All Pareto k estimates are ok (k < 0.7).
See help('pareto-k-diagnostic') for details.

Model comparison:
(negative 'elpd_diff' favors 1st model, positive favors 2nd)

elpd_diff        se
     -0.2       0.1
```

loo package

```
Computed from 4000 by 20 log-likelihood matrix

         Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0
------
Monte Carlo SE of elpd_loo is 0.1.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     18    90.0%   899
 (0.5, 0.7]   (ok)        2    10.0%   459
   (0.7, 1]   (bad)       0     0.0%   <NA>
   (1, Inf)   (very bad)  0     0.0%   <NA>

All Pareto k estimates are ok (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

Section 3

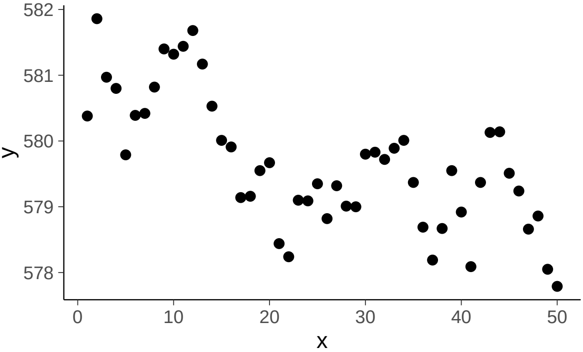
Information criteria

Section 4

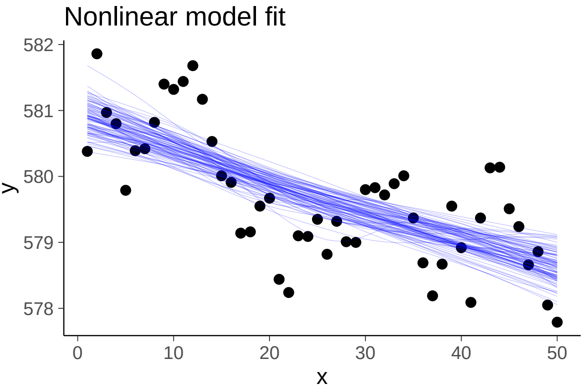
To move up

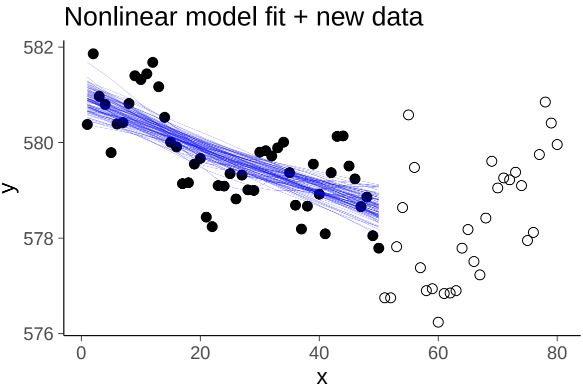
UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up



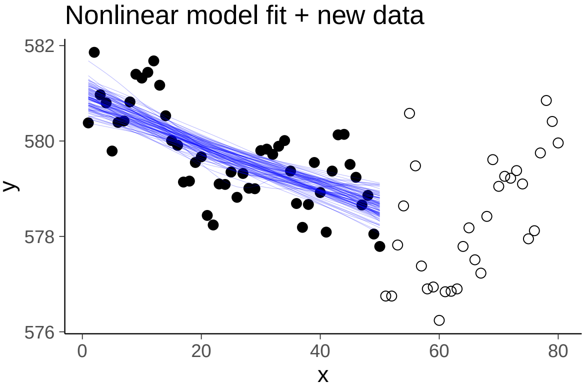
UPPSALA UNIVERSITET



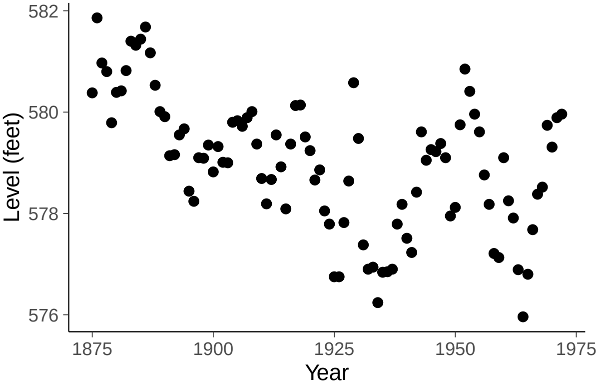

Nonlinear model fit + new data
582
580
578
576
y
0
20
40
60
80
x

UPPSALA UNIVERSITET

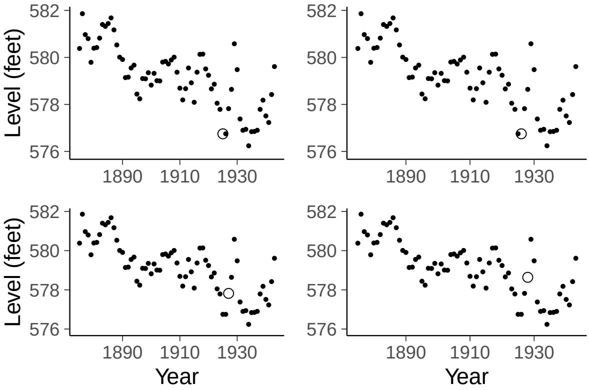
- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and loo
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up



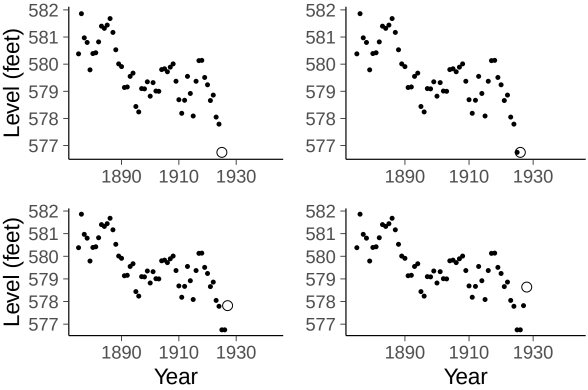
Extrapolation is more difficult



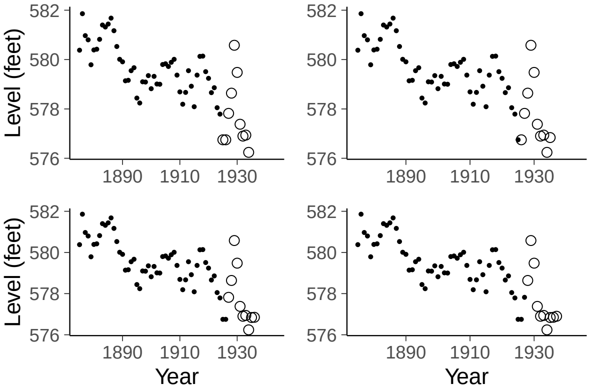
Can LOO or other cross-validation be used with time series?



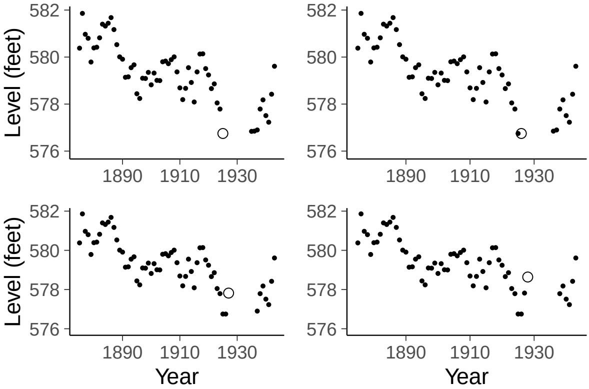
Leave-one-out cross-validation is ok for assessing conditional model



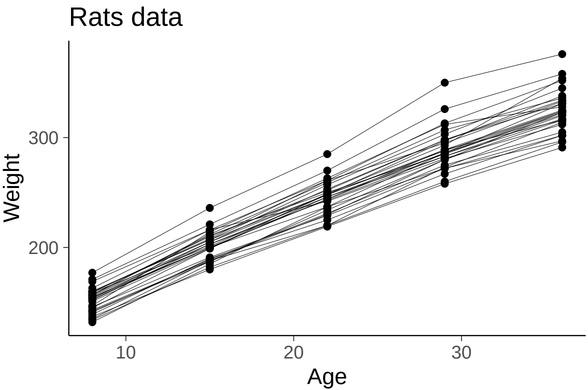
Leave-future-out cross-validation is better for predicting future



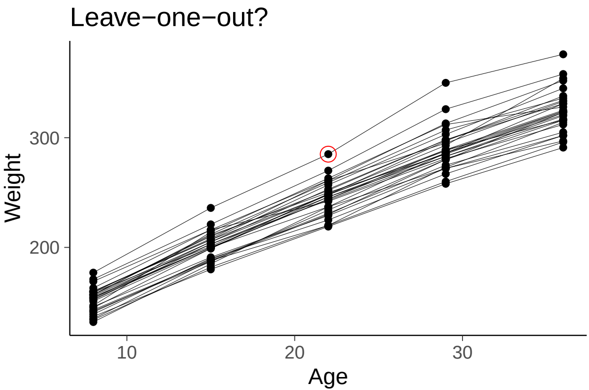
m-step-ahead cross-validation is better for predicting further future



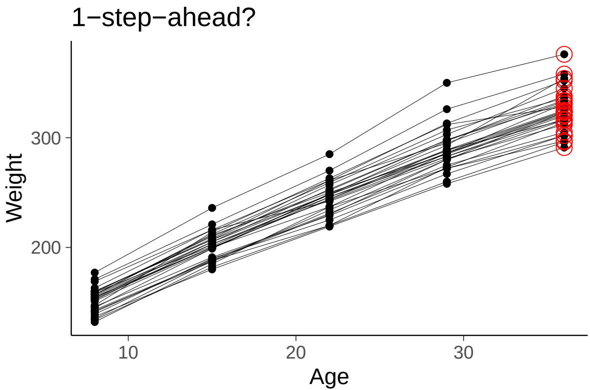
m-step-ahead leave-a-block-out cross-validation



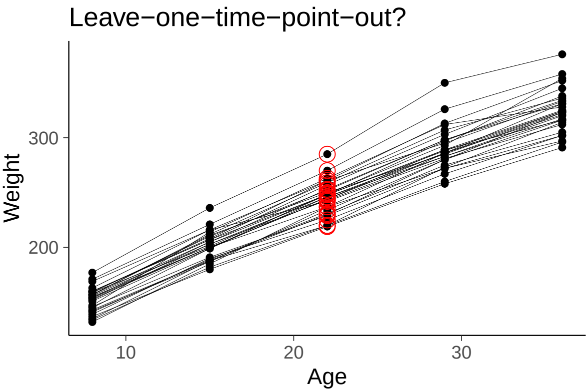
Can LOO or other cross-validation be used with hierarchical data?



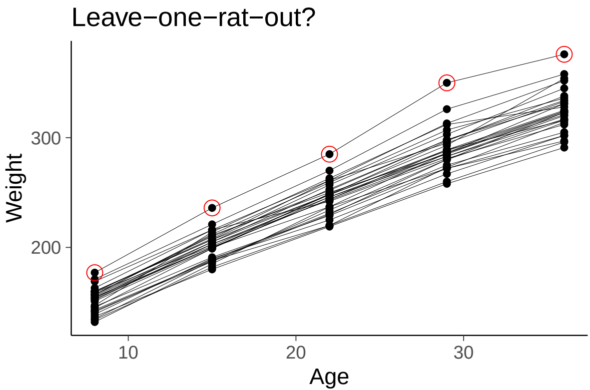
Yes!



Yes!

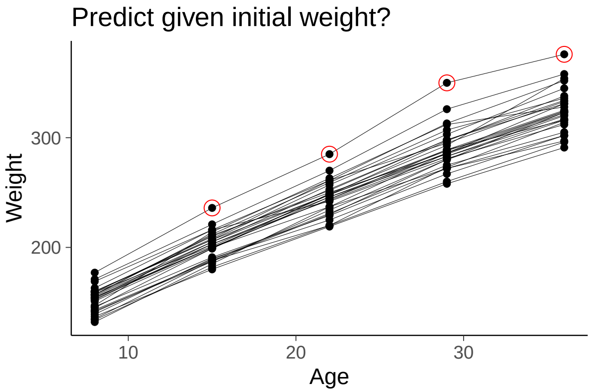


Yes!



Yes!

UPPSALA UNIVERSITET



Yes!

Summary of data generating mechanisms and prediction tasks

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

see Vehtari & Ojanen (2012) and `andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/`

UPPSALA UNIVERSITET

Fast cross-validation

1. Pareto smoothed importance sampling LOO (PSIS-LOO)
2. K-fold cross-validation

see Vehtari, Gelman & Gabry (2017a) and `mc-stan.org/loo/`

loo package

```
 Computed from 4000 by 20 log-likelihood matrix

         Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0

------
Monte Carlo SE of elpd_loo is 0.1.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     18    90.0%   899
 (0.5, 0.7]   (ok)        2    10.0%   459
   (0.7, 1]   (bad)       0     0.0%   <NA>
   (1, Inf)   (very bad)  0     0.0%   <NA>

All Pareto k estimates are ok (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

see more in Vehtari, Gelman & Gabry (2017b)

Stan code

$$\log(r_i^{(s)}) = \log(1/p(y_i|x_i, \theta^{(s)})) = -\text{log\_lik}[i]$$

Stan code

$$\log(r_i^{(s)}) = \log(1/p(y_i|x_i,\theta^{(s)})) = -\mathrm{log\_lik}[i]$$

```
...
model {
  alpha ~ normal(pmualpha, psalpha);
  beta ~ normal(pmubeta, psbeta);
  y ~ normal(mu, sigma);
}
generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = normal_lpdf(y[i] | mu[i], sigma);
}
```

Stan code

$$\log(r_i^{(s)}) = \log(1/p(y_i|x_i,\theta^{(s)})) = -\text{log\_lik}[i]$$

```
...
model {
  alpha ~ normal(pmualpha, psalpha);
  beta ~ normal(pmubeta, psbeta);
  y ~ normal(mu, sigma);
}
generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = normal_lpdf(y[i] | mu[i], sigma);
}
```

1. RStanARM and BRMS compute log_lik by default

Pareto smoothed importance sampling LOO

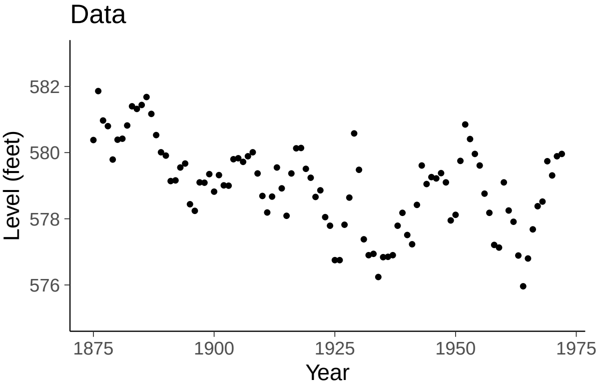
1. PSIS-LOO for hierarchical models
   1.1 leave-one-group out is challenging for PSIS-LOO
   see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration

Pareto smoothed importance sampling LOO

1. PSIS-LOO for hierarchical models
   1.1 leave-one-group out is challenging for PSIS-LOO
   see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration
2. PSIS-LOO for non-factorizable models
   2.1 `mc-stan.org/loo/articles/loo2-non-factorizable.html`

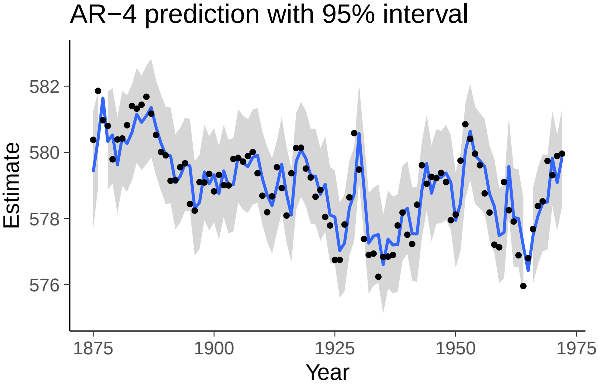
Pareto smoothed importance sampling LOO

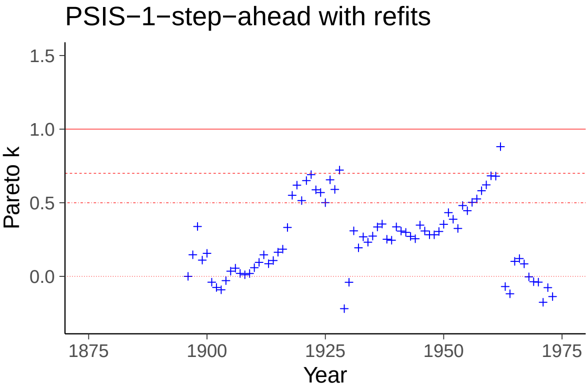
1. PSIS-LOO for hierarchical models
   1.1 leave-one-group out is challenging for PSIS-LOO
   see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration
2. PSIS-LOO for non-factorizable models
   2.1 `mc-stan.org/loo/articles/loo2-non-factorizable.html`
3. PSIS-LOO for time series
   3.1 Approximate leave-future-out cross-validation
   `mc-stan.org/loo/articles/loo2-lfo.html`


Data
582
580
578
576
Level (feet)
1875
1900
1925
1950
1975
Year

UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up





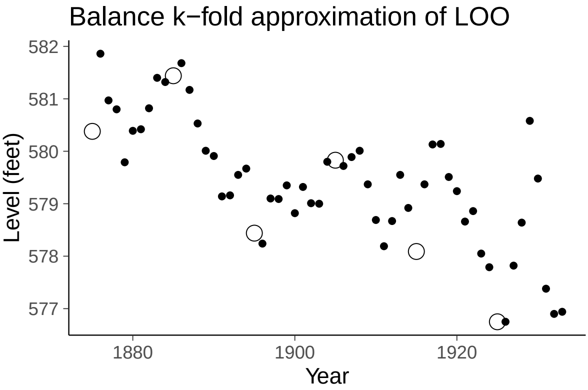
mc-stan.org/loo/articles/loo2-lfo.html

K-fold cross-validation

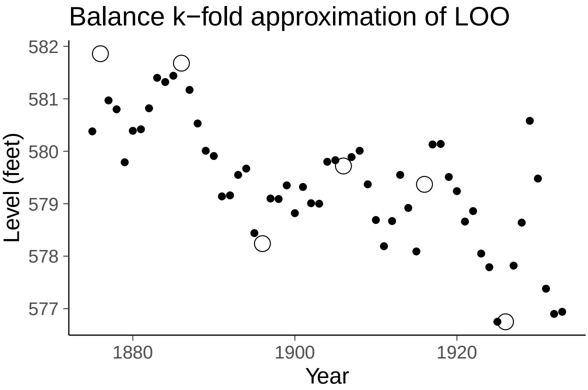
1. K-fold cross-validation can approximate LOO
   1.1 all uses for LOO
2. K-fold cross-validation can be used for hierarchical models
   2.1 good for leave-one-group-out
3. K-fold cross-validation can be used for time series
   3.1 with leave-block-out

UPPSALA UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up

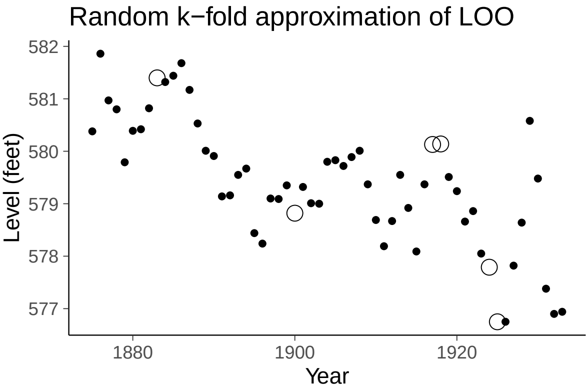


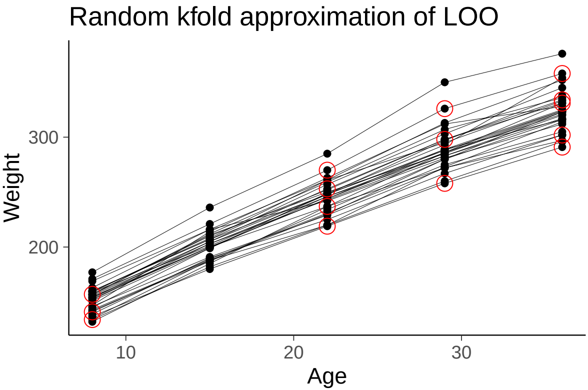
UPPSALA UNIVERSITET



UPPSALA UNIVERSITET

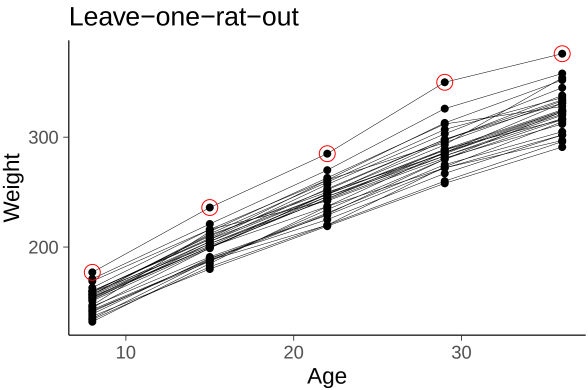
- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up

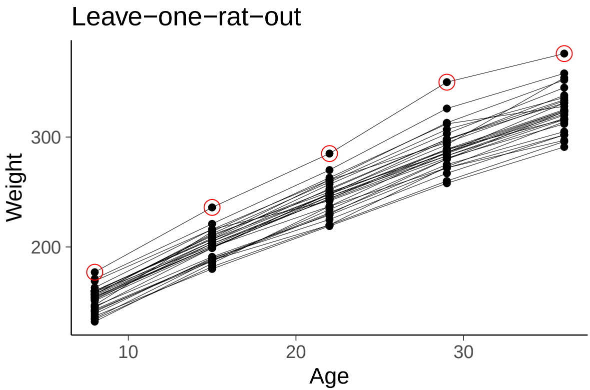



Random kfold approximation of LOO
Weight
300
200
10
20
30
Age

UPPSALA
UNIVERSITET

- Model assessment and selection
  - Measures of predictive accuracy
- Cross-validation
  - When is LOO applicable
  - PSIS-LOO and `loo`
  - Comparison and selection
  - Additional reading
- Information criteria
- To move up





kfold_split_random()

kfold_split_balanced()

kfold_split_stratified()

WAIC vs PSIS-LOO

see Vehtari, Gelman & Gabry (2017a)

WAIC vs PSIS-LOO

1. WAIC has same assumptions as LOO

see Vehtari, Gelman & Gabry (2017a)

WAIC vs PSIS-LOO

1. WAIC has same assumptions as LOO
2. PSIS-LOO is more accurate

see Vehtari, Gelman & Gabry (2017a)

WAIC vs PSIS-LOO

1. WAIC has same assumptions as LOO
2. PSIS-LOO is more accurate
3. PSIS-LOO has much better diagnostics

see Vehtari, Gelman & Gabry (2017a)

WAIC vs PSIS-LOO

1. WAIC has same assumptions as LOO
2. PSIS-LOO is more accurate
3. PSIS-LOO has much better diagnostics
4. LOO makes the prediction assumption more clear, which helps if K-fold-CV is needed instead

see Vehtari, Gelman & Gabry (2017a)

WAIC vs PSIS-LOO

1. WAIC has same assumptions as LOO
2. PSIS-LOO is more accurate
3. PSIS-LOO has much better diagnostics
4. LOO makes the prediction assumption more clear, which helps if K-fold-CV is needed instead
5. Multiplying by -2 doesn't give any benefit (Watanabe didn't multiply by -2)

see Vehtari, Gelman & Gabry (2017a)

*IC

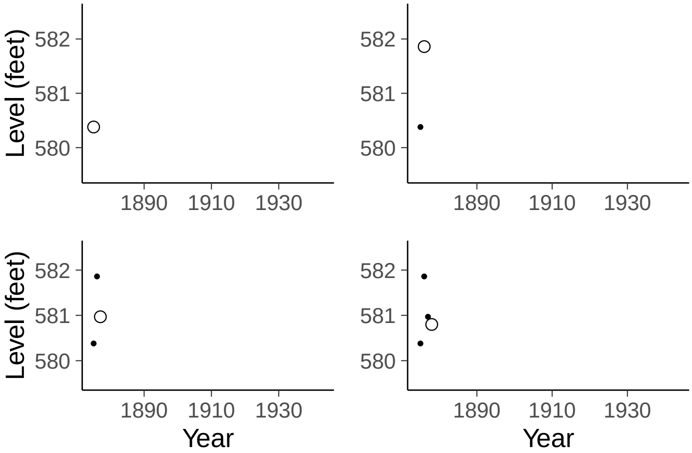
1. AIC uses maximum likelihood estimate for prediction
2. DIC uses posterior mean for prediction
3. BIC is an approximation for marginal likelihood
4. TIC, NIC, RIC, PIC, BPIC, QIC, AICc, ...

Marginal likelihood / Bayes factor

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

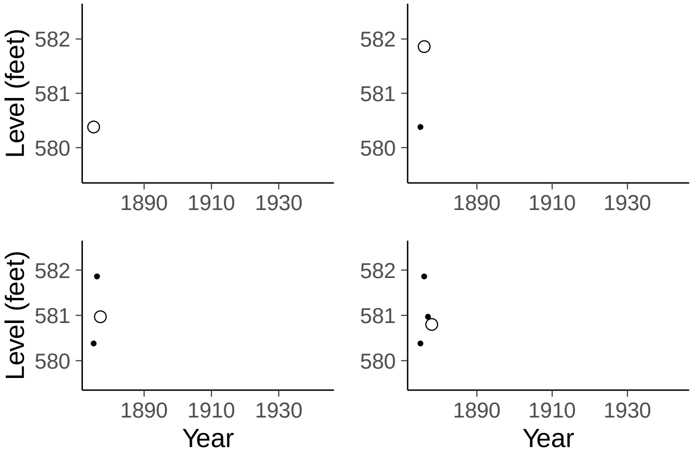
Marginal likelihood / Bayes factor

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



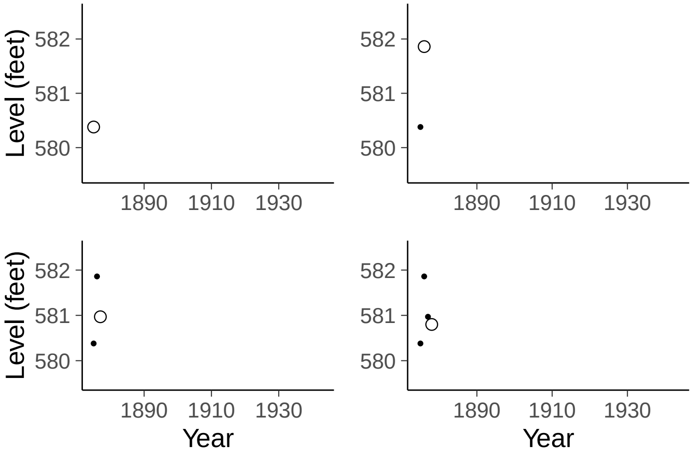
Marginal likelihood / Bayes factor

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



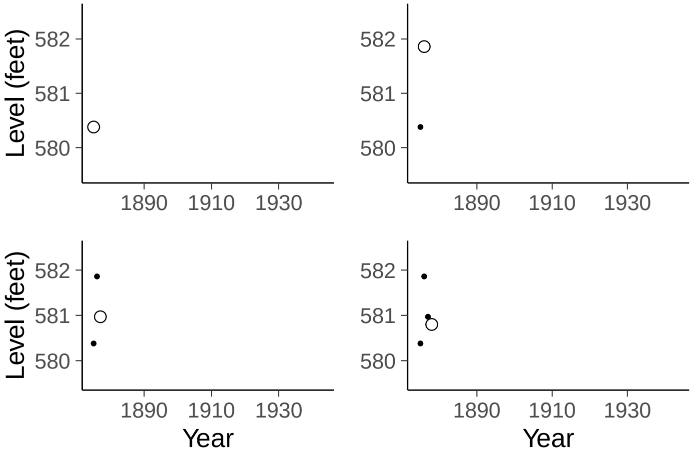
Marginal likelihood / Bayes factor

1. 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

1. 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



Cross-validation for model assessment

1. CV is good for model assessment when application specific utility/cost functions are used
   1.1 e.g. 90% absolute error

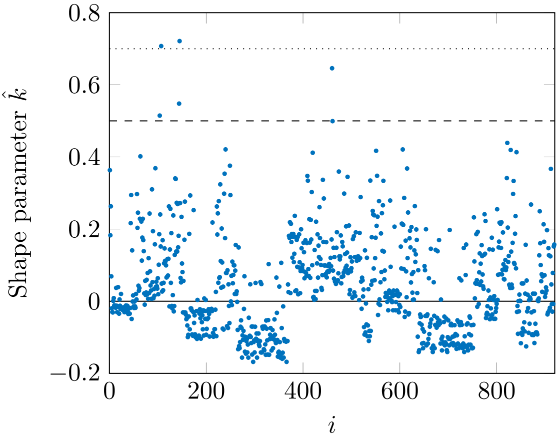
Cross-validation for model assessment

1. CV is good for model assessment when application specific utility/cost functions are used
   1.1 e.g. 90% absolute error
2. Also useful in model checking in similar way as posterior predictive checking (PPC)
   2.1 model misspecification diagnostics
   (e.g. Pareto-k and p_loo)
   2.2 checking calibration of leave-one-out predictive posteriors
   (ppc_loo_pit in bayesplot)

see demos `avehtari.github.io/modelselection/`

Radon example

PSIS-LOO diagnostics

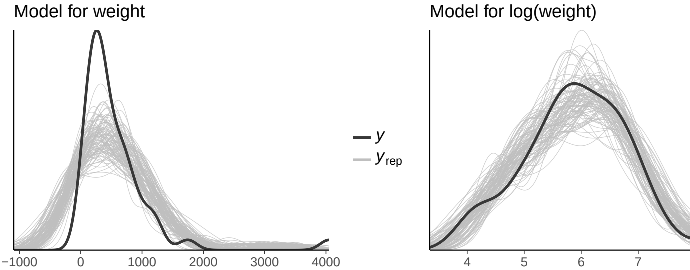


see Vehtari, Gelman & Gabry (2017a)

Sometimes cross-validation is not needed

Sometimes cross-validation is not needed

1. Posterior predictive checking is often sufficient

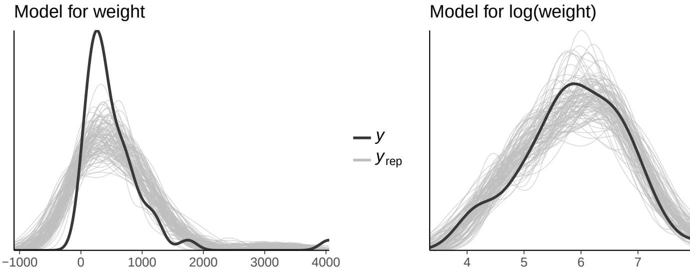


Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2020): Regression and Other Stories, Chapter 11.

Sometimes cross-validation is not needed

1. Posterior predictive checking is often sufficient



Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2020): Regression and Other Stories, Chapter 11.

1. BDA3, Chapter 6
2. Gabry, Simpson, Vehtari, Betancourt, Gelman (2019). Visualization in Bayesian workflow. JRSS A, https://doi.org/10.1111/rssa.12378
3. mc-stan.org/bayesplot/articles/graphical-ppcs.html
4. betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

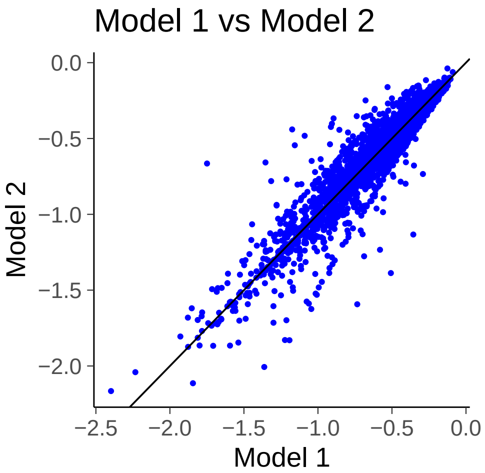
UPPSALA UNIVERSITET

Arsenic well example – Model comparison

1. Probability of switching well with high arsenic level in rural Bangladesh
   1.1 Model 1 covariates: log(arsenic) and distance
   1.2 Model 2 covariates: log(arsenic), distance and education level

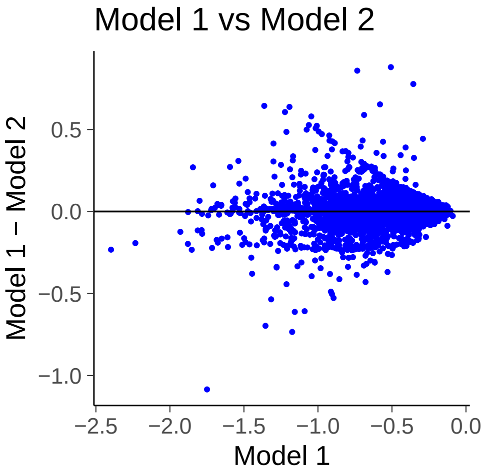
Gelman, Hill & Vehtari (2020): Regression and Other Stories, Chapter 13.

Arsenic well example – Model comparison



Model 1 elpd_loo ≈ -1952, SE=16
Model 2 elpd_loo ≈ -1938, SE=17

Arsenic well example – Model comparison




```
> loo_compare(model1, model2)
        elpd_diff  se_diff
model2     0.0       0.0
model1   -14.4       6.1
```

see Vehtari, Gelman & Gabry (2017a)

UPPSALA UNIVERSITET

Arsenic well example – Model comparison

```
> loo_compare(model1, model2)
       elpd_diff  se_diff
model2    0.0       0.0
model1  -14.4       6.1
```

`se_diff` and normal approximation for the uncertainty in the difference is good only if models are well specified and the number of observations is relatively big (more details in a forthcoming article).

UPPSALA UNIVERSITET

Sometimes cross-validation is not needed

1. For some very simple cases you may assume that true model is included in the list of models considered (M-closed)

Sometimes cross-validation is not needed

1. For some very simple cases you may assume that true model is included in the list of models considered (M-closed)
   1.1 see predictive model selection in M-closed case by San Martini and Spezzaferri (1984)

Sometimes cross-validation is not needed

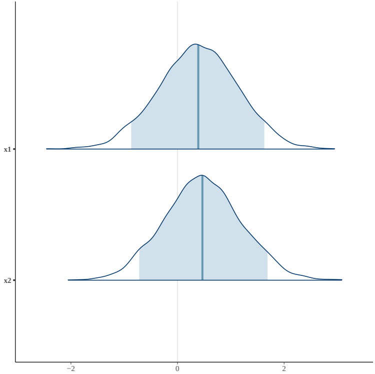
1. For some very simple cases you may assume that true model is included in the list of models considered (M-closed)
   1.1 see predictive model selection in M-closed case by San Martini and Spezzaferri (1984)
   1.2 but you should not force your design of experiment or analysis to stay in the simplified world

UPPSALA UNIVERSITET

Sometimes cross-validation is not needed

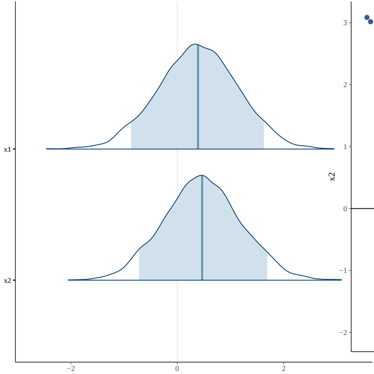
1. For some very simple cases you may assume that true model is included in the list of models considered (M-closed)
   1.1 see predictive model selection in M-closed case by San Martini and Spezzaferri (1984)
   1.2 but you should not force your design of experiment or analysis to stay in the simplified world
2. In nested case, often easier and more accurate to analyse posterior distribution of more complex model directly
   `avehtari.github.io/modelselection/betablockers.html`

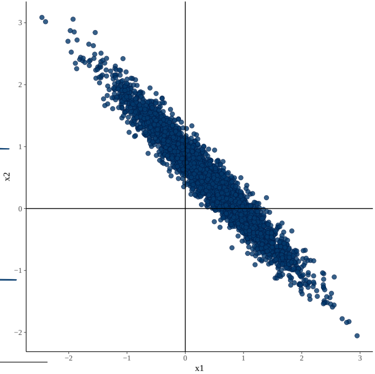
Sometimes predictive model comparison can be useful



Marginal posterior intervals

Sometimes predictive model comparison can be useful

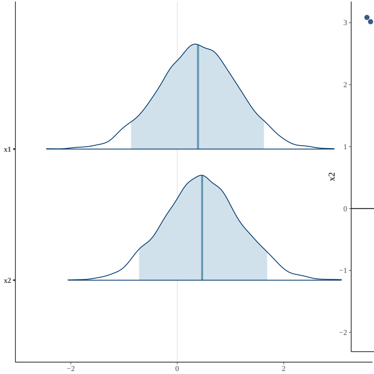




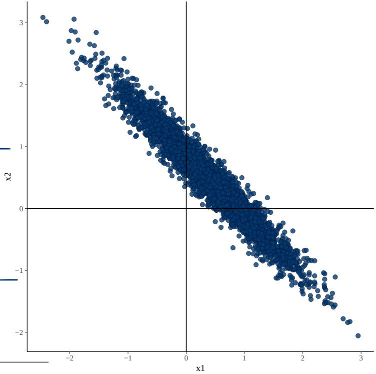
Marginal posterior intervals   Joint posterior density

rstanarm + bayesplot

Sometimes predictive model comparison can be useful



Marginal posterior intervals



Joint posterior density

rstanarm + bayesplot

see also Collinear demo

What if one is not clearly better than others?

What if one is not clearly better than others?

1. Continuous expansion including all models?
   1.1 and then analyse the posterior distribution directly
   avehtari.github.io/modelselection/betablockers.html
   1.2 sparse priors like regularized horseshoe prior instead of variable selection
   video, refs and demos at
   avehtari.github.io/modelselection/

What if one is not clearly better than others?

1. Continuous expansion including all models?
   1.1 and then analyse the posterior distribution directly
   `avehtari.github.io/modelselection/betablockers.html`
   1.2 sparse priors like regularized horseshoe prior instead of variable selection
   video, refs and demos at
   `avehtari.github.io/modelselection/`
2. Model averaging with BMA or Bayesian stacking?
   `mc-stan.org/loo/articles/loo2-example.html`

What if one is not clearly better than others?

1. Continuous expansion including all models?
   1.1 and then analyse the posterior distribution directly
   `avehtari.github.io/modelselection/betablockers.html`
   1.2 sparse priors like regularized horseshoe prior instead of variable selection
   video, refs and demos at
   `avehtari.github.io/modelselection/`
2. Model averaging with BMA or Bayesian stacking?
   `mc-stan.org/loo/articles/loo2-example.html`
3. In a nested case choose simpler if assuming some cost for extra parts?
   `andrewgelman.com/2018/07/26/parsimonious-principle-vs-integration-uncertainties/`

What if one is not clearly better than others?

1. Continuous expansion including all models?
   1.1 and then analyse the posterior distribution directly
   avehtari.github.io/modelselection/betablockers.html
   1.2 sparse priors like regularized horseshoe prior instead of variable selection
   video, refs and demos at
   avehtari.github.io/modelselection/
2. Model averaging with BMA or Bayesian stacking?
   mc-stan.org/loo/articles/loo2-example.html
3. In a nested case choose simpler if assuming some cost for extra parts?
   andrewgelman.com/2018/07/26/parsimonious-principle-vs-integration-uncertainties/
4. In a nested case choose more complex if you want to take into account all the uncertainties.
   andrewgelman.com/2018/07/26/parsimonious-principle-vs-integration-uncertainties/

UPPSALA UNIVERSITET

Model averaging

1. Prefer continuous model expansion

Model averaging

1. Prefer continuous model expansion
2. If needed integrate over the model space = model averaging

Model averaging

1. Prefer continuous model expansion
2. If needed integrate over the model space = model averaging
3. Bayesian stacking may work better than BMA
   3.1 Yao, Vehtari, Simpson, & Gelman (2018)

Cross-validation and model selection

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

Cross-validation and model selection

1. Cross-validation can be used for model selection if
   1.1 small number of models
   1.2 the difference between models is clear
2. Do not use cross-validation to choose from a large set of models
   2.1 selection process leads to overfitting

Cross-validation and model selection

1. Cross-validation can be used for model selection if
   1.1 small number of models
   1.2 the difference between models is clear
2. Do not use cross-validation to choose from a large set of models
   2.1 selection process leads to overfitting
3. Overfitting in selection process is not unique for cross-validation

Selection induced bias and overfitting

- 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)

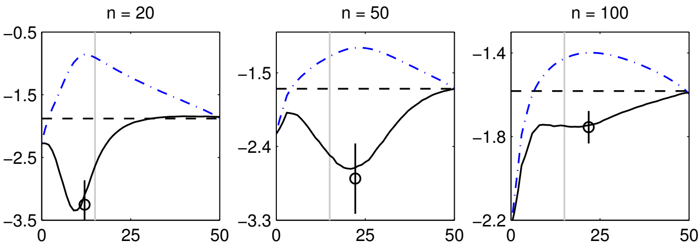
Selection induced bias and overfitting

- 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

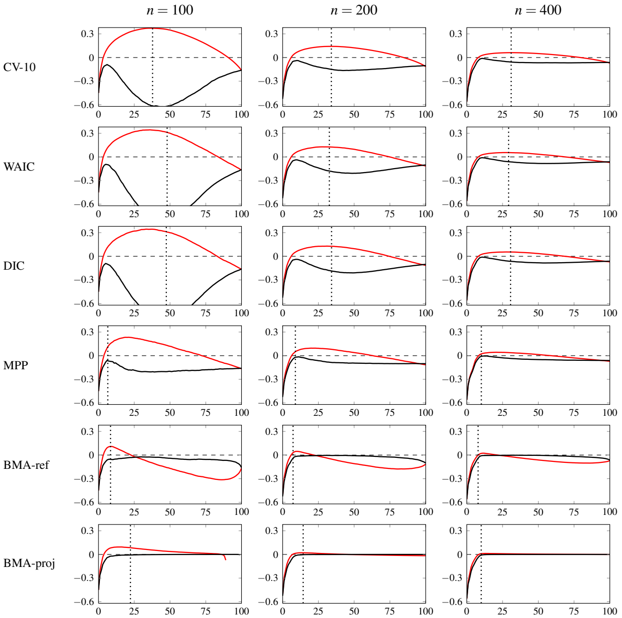
Selection induced bias and overfitting

- 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

Selection induced bias in variable selection



Selection induced bias in variable selection



Piironen & V

UPPSALA UNIVERSITET

Take-home messages

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

UPPSALA UNIVERSITET

Take-home messages

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

Take-home messages

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

UPPSALA UNIVERSITET

Take-home messages

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

Take-home messages

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