



UPPSALA
UNIVERSITET

• Introduction

Bayesian Statistics and Data Analysis

Lecture 8b

Måns Magnusson

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



UPPSALA
UNIVERSITET

● Introduction

Section 1

Introduction



Model assessment, selection and inference after selection

- Introduction

1. Extra material at

<https://avehtari.github.io/modelselection/>

- 1.1 Videos, Slides, Notebooks, References

- 1.2 The most relevant for the course is the first part of the talk "Model assessment, comparison and selection at Master class in Bayesian statistics, CIRM, Marseille"



UPPSALA
UNIVERSITET

Predicting concrete quality

- Introduction



UPPSALA
UNIVERSITET

● Introduction

GIST Risk calculator

Tumor size (cm)

Mitotic count (per 50 HPFs*)

Tumor site

Tumor rupture

CALCULATE!

*HPF = high-power field of the microscope

[Show risk tables](#)

Made by

kariku
HEALTH

Online platform for the future of data-driven
and personalized cancer care

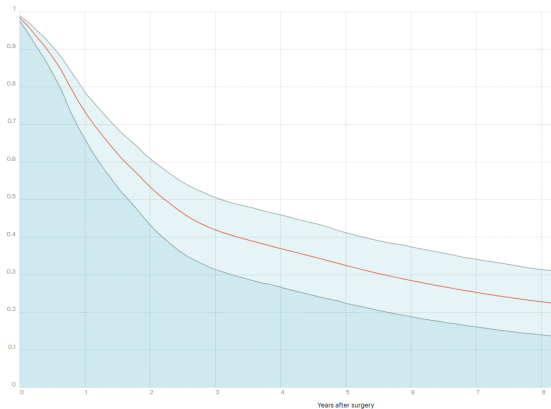
Reaktor

Predicting cancer recurrence

Patients alive without recurrence [Show hazard](#)

90 % credible interval

10 year risk of GIST re





- Introduction

- 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



- Introduction

- We need to choose the utility/cost function
- Application specific utility/cost functions are important
 - eg. money, life years, quality adjusted life years, etc.
- If are interested overall in the goodness of the predictive distribution, or we don't know (yet) the application specific utility, then good information theoretically justified choice is log-score

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



Outline

1. What is cross-validation
 - 1.1 Leave-one-out cross-validation (elpd_loo, p_loo)
 - 1.2 Uncertainty in LOO (SE)
2. When is cross-validation applicable?
 - 2.1 data generating mechanisms and prediction tasks
 - 2.2 leave-many-out cross-validation
3. Fast cross-validation
 - 3.1 PSIS and diagnostics in loo package (Pareto k, n_eff, Monte Carlo SE)
 - 3.2 K-fold cross-validation
4. Related methods (WAIC, *IC, BF)
5. Model comparison and selection (elpd_diff, se)
6. Model averaging with Bayesian stacking



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.

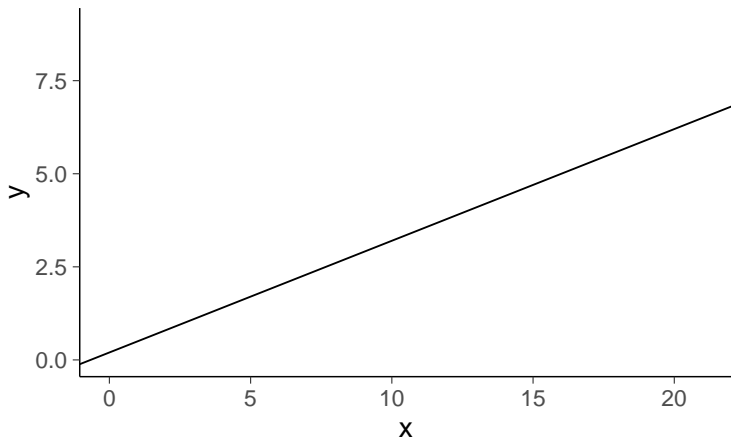
Model comparison:

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

elpd_diff	se
-0.2	0.1



True mean $y = a + bx$

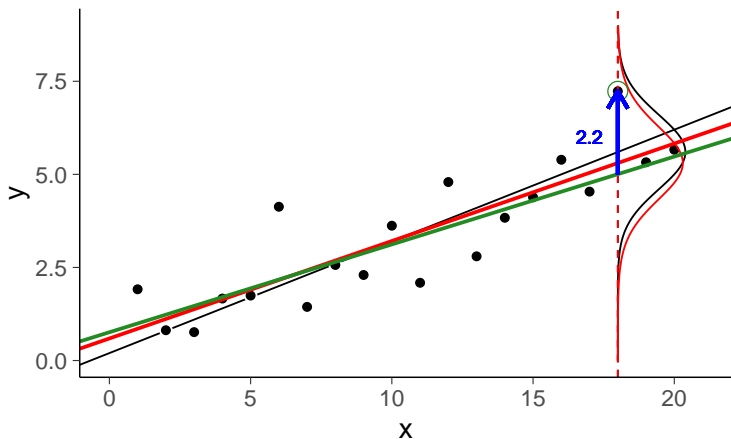


True mean and sigma





Leave-one-out residual



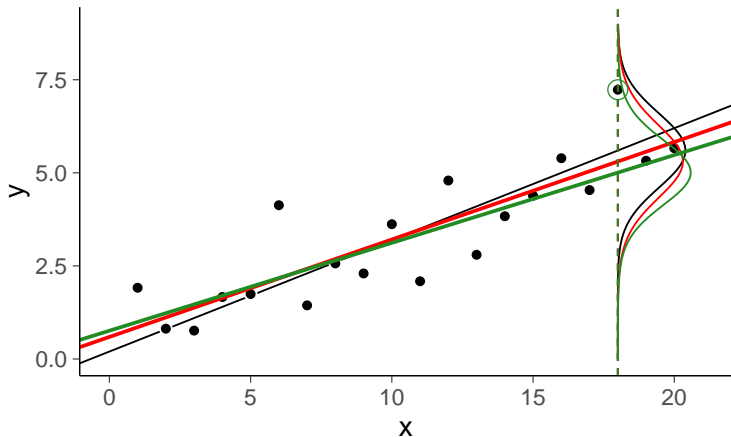
$$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$$

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



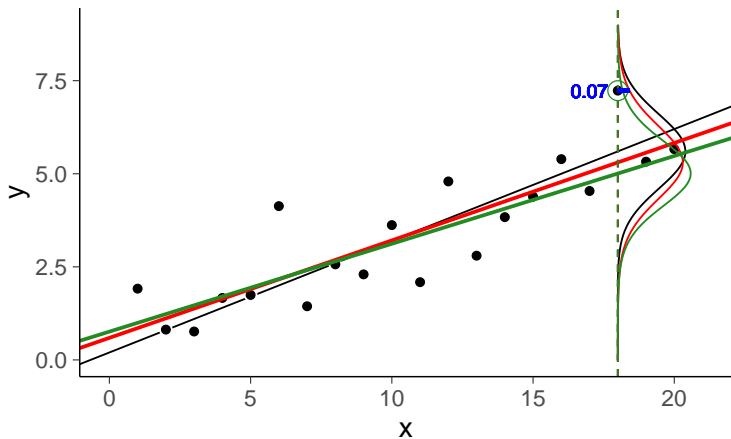
$$p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18}) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x_{-18}, y_{-18})d\theta$$



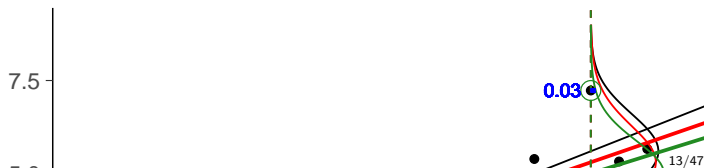
UPPSALA
UNIVERSITET

● Introduction

Posterior predictive density



Leave-one-out predictive density

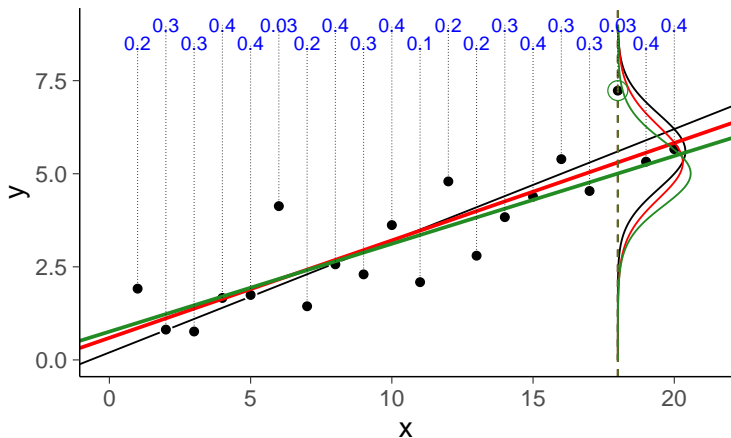




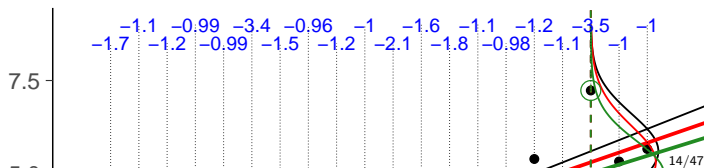
UPPSALA
UNIVERSITET

● Introduction

Leave-one-out predictive densities



Leave-one-out log predictive densities

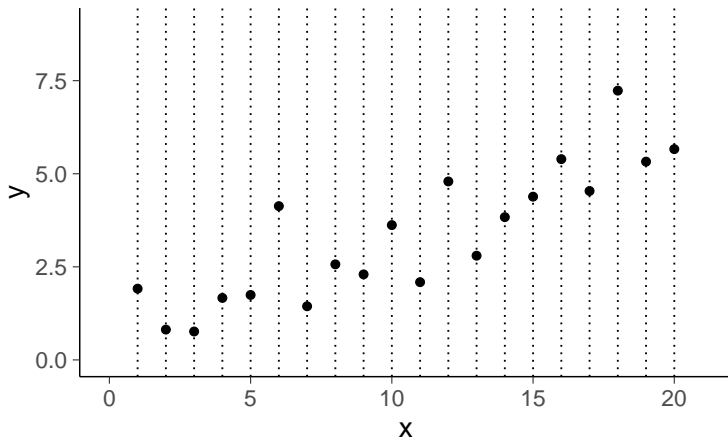




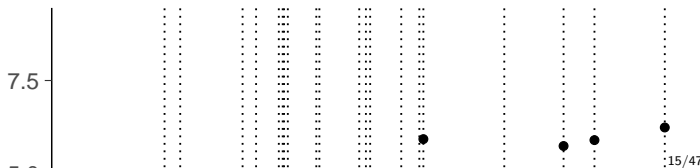
UPPSALA
UNIVERSITET

• Introduction

Fixed / designed x



Distribution for x





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:

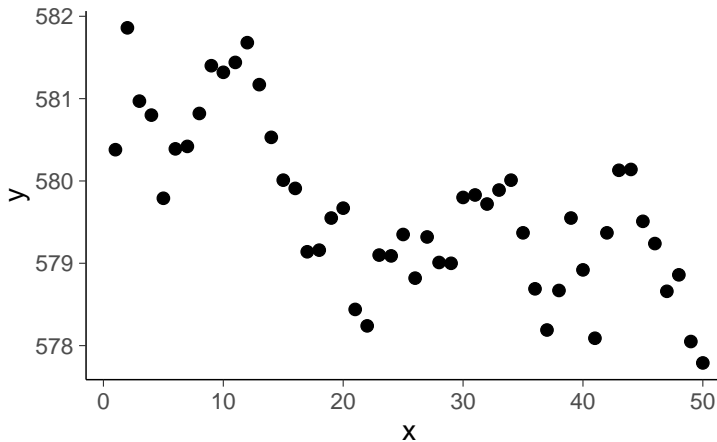
		Count	Pct.	Min.	n_eff
(-Inf, 0.5]	(good)	18	90.0%	899	
(0.5, 0.7]	(ok)	2	10.0%	459	
(0.7, 1]	(bad)	0	0.0%	<NA>	
(1, Inf)	(very bad)	0	0.0%	<NA>	

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

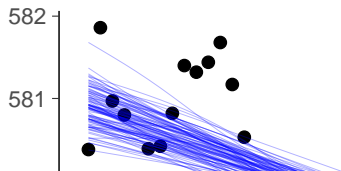


UPPSALA
UNIVERSITET

• Introduction



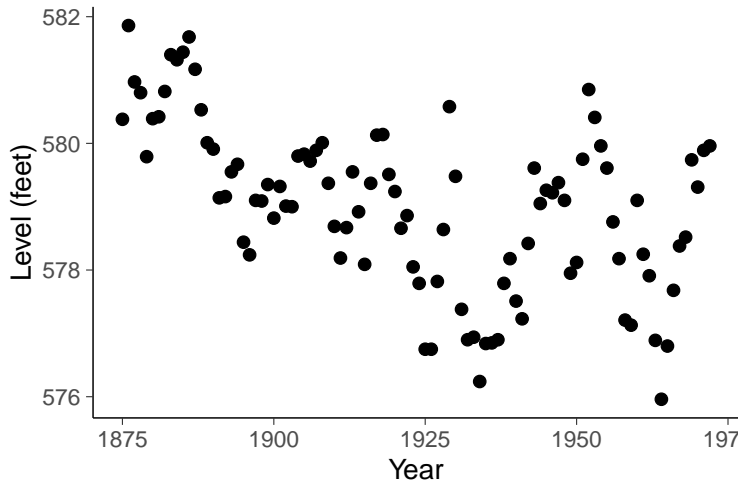
Nonlinear model fit





UPPSALA
UNIVERSITET

• Introduction

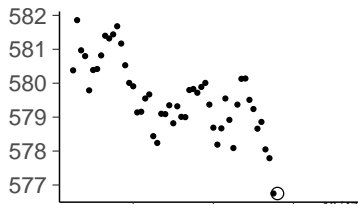
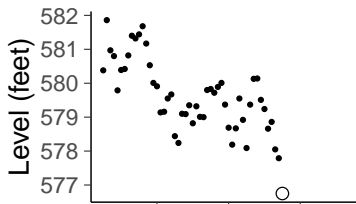
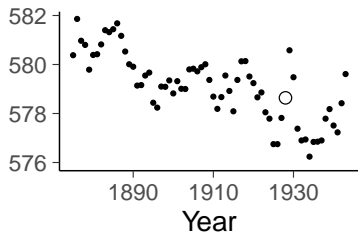
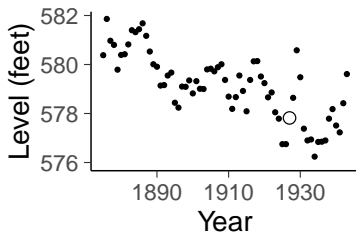
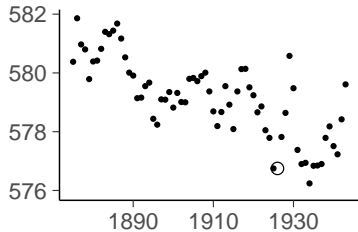
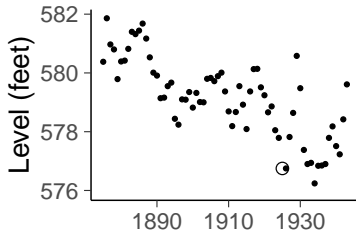


Can LOO or other cross-validation be used with time series?



UPPSALA
UNIVERSITET

● Introduction

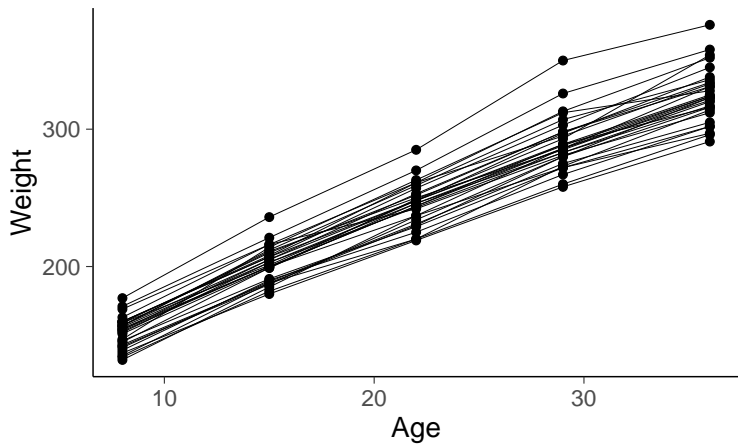




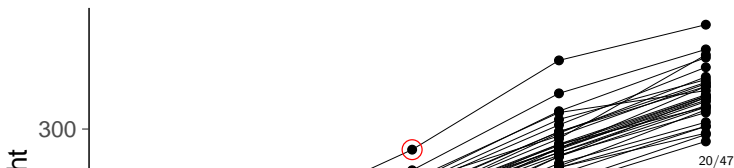
UPPSALA
UNIVERSITET

● Introduction

Rats data



Leave-one-out?





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

● Introduction

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 \(2017b\)](#)



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

- Introduction

```
...  
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

● Introduction

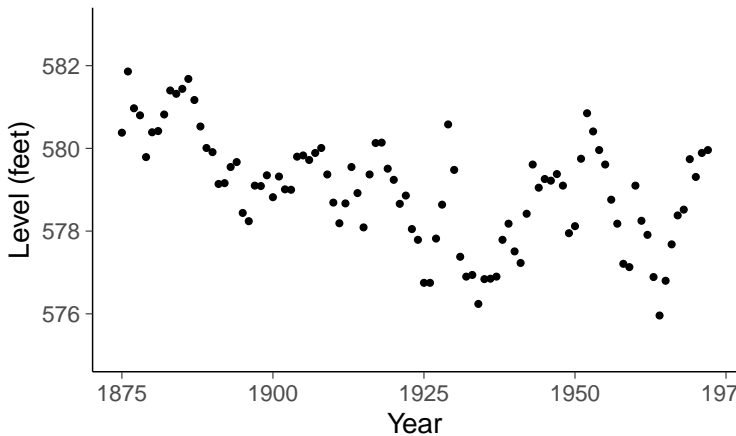
1. PSIS-LOO for hierarchical models
 - 1.1 leave-one-group out is challenging for PSIS-LOO
see Merkel, Furr and Rabe-Hesketh (2018) for an
approach using quadrature integration
2. PSIS-LOO for non-factorizable models
 - 2.1 mc-stan.org/loo/articles/loo2-non-factorizable.html
3. PSIS-LOO for time series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



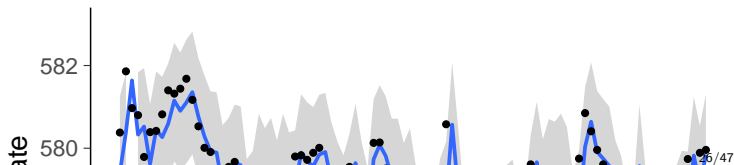
UPPSALA
UNIVERSITET

● Introduction

Data



AR-4 prediction with 95% interval





● Introduction

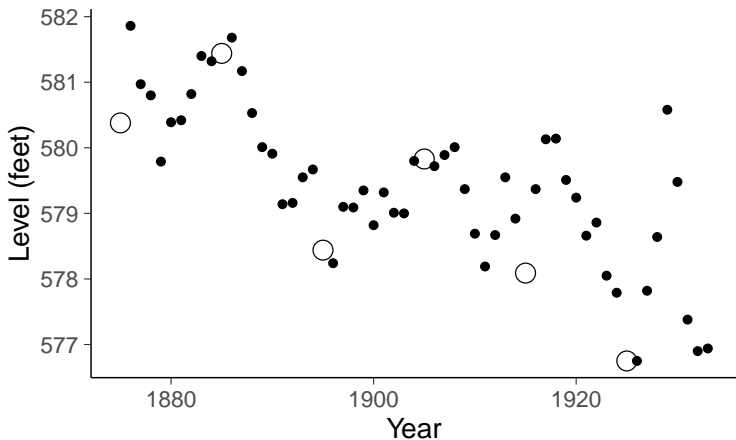
1. K-fold cross-validation can approximate LOO
 - 1.1 all uses for LOO
2. K-fold cross-validation can be used for hierarchical models
 - 2.1 good for leave-one-group-out
3. K-fold cross-validation can be used for time series
 - 3.1 with leave-block-out



UPPSALA
UNIVERSITET

● Introduction

Balance k-fold approximation of LOO



Balance k-fold approximation of LOO





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 \(2017a\)](#)



UPPSALA
UNIVERSITET

*IC

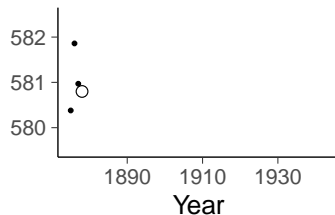
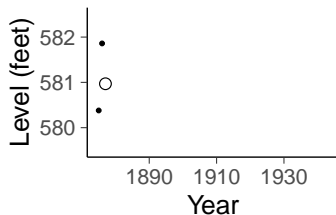
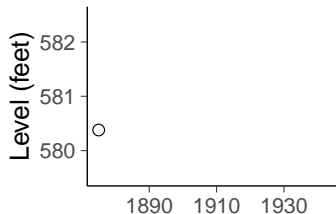
● Introduction

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
 - which makes it very sensitive to prior and
 - unstable in case of misspecified models also asymptotically





Cross-validation for model assessment

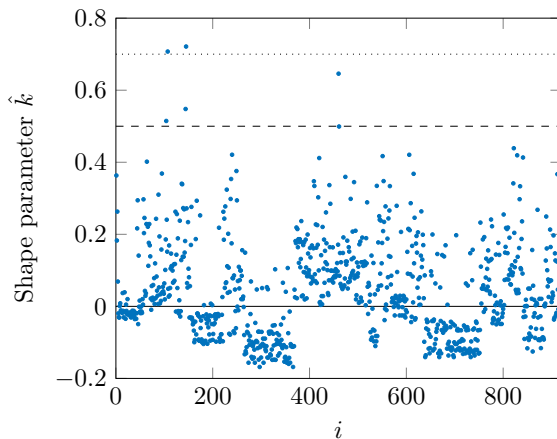
1. CV is good for model assessment when application specific utility/cost functions are used
 - 1.1 e.g. 90% absolute error
2. Also useful in model checking in similar way as posterior predictive checking (PPC)
 - 2.1 model misspecification diagnostics (e.g. Pareto- k and p_{loo})
 - 2.2 checking calibration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

see demos avehtari.github.io/modelselection/



Radon example

PSIS-LOO diagnostics

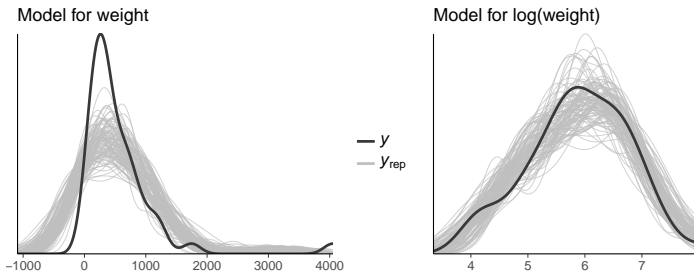


see [Vehtari, Gelman & Gabry \(2017a\)](#)



Sometimes cross-validation is not needed

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



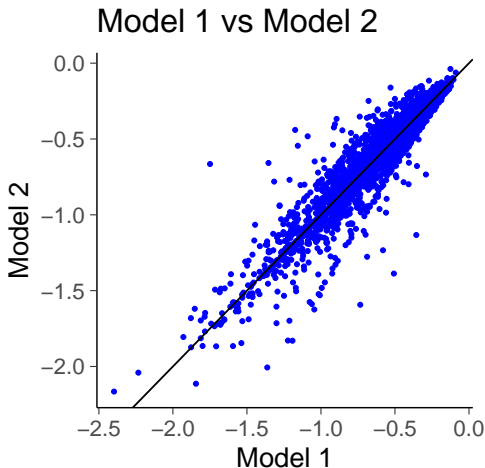
Arsenic well example – Model comparison

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

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



Arsenic well example – Model comparison



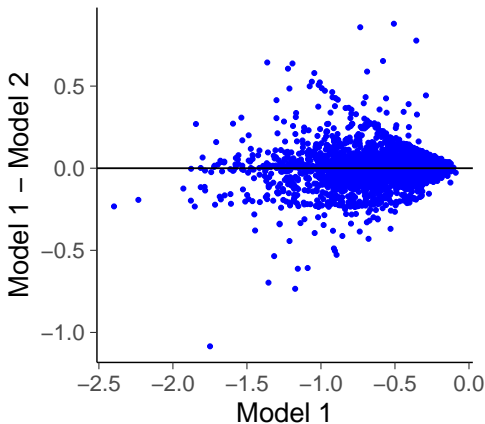
Model 1 $\text{elpd}_{\text{loo}} \approx -1952$, $\text{SE}=16$

Model 2 $\text{elpd}_{\text{loo}} \approx -1938$, $\text{SE}=17$



Arsenic well example – Model comparison

Model 1 vs Model 2



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

see [Vehtari, Gelman & Gabry](#)

(2017a)



- Introduction

```
> 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).

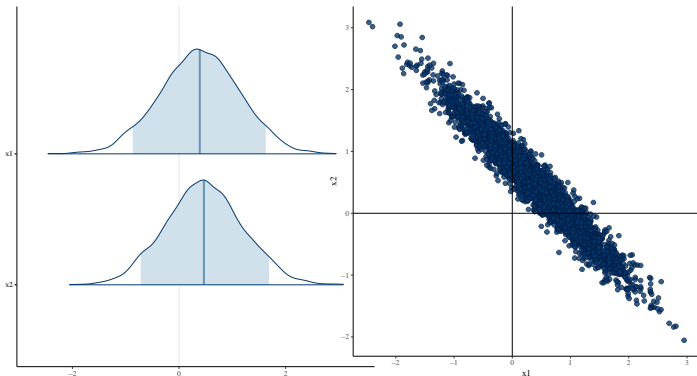


Sometimes cross-validation is not needed

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

Joint posterior density

`rstanarm` + `bayesplot`

see also [Collinear demo](#)



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/



● Introduction

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)



● Introduction

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

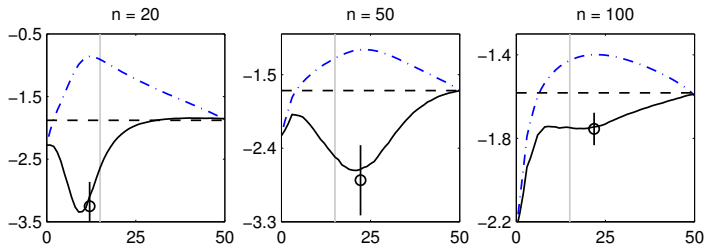


• Introduction

- 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

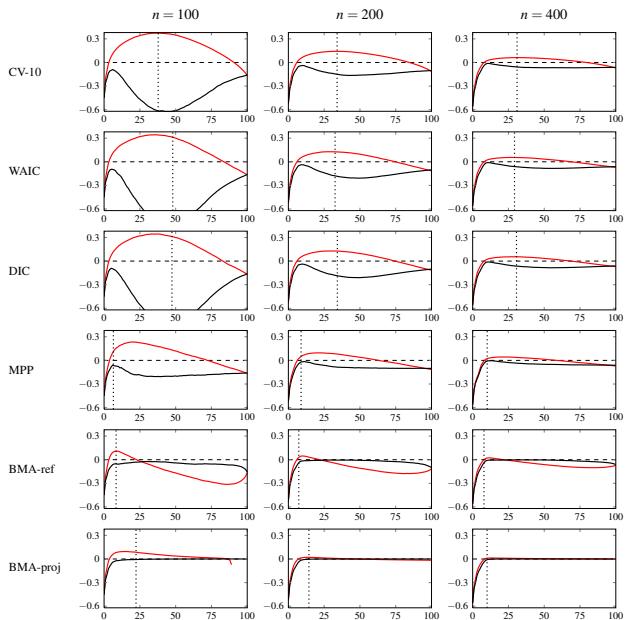


Selection induced bias in variable selection





Selection induced bias in variable selection





Take-home messages

● Introduction

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