



UPPSALA
UNIVERSITET

Bayesian Statistics and Data Analysis

Lecture 8b

Måns Magnusson

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA
UNIVERSITET

- **Model assessment and selection**
 - Measures of predictive accuracy
 - Model selection
- **Cross-validation**
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- **Information criteria**
- **Model averaging and Stacking**
- **Summary**

Section 1

Model assessment and selection



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Modeling complex phenomena with models that are simplified

All models are wrong... but some are useful.



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- 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



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

- 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



UPPSALA
UNIVERSITET

Goal of model evaluation

- **Model assessment and selection**

- Measures of predictive accuracy
- Model selection

- **Cross-validation**

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- **Information criteria**

- **Model averaging and Stacking**

- **Summary**

- **Model choice is a (model-)decision-theoretic problem**



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

- Model choice is a (model-)decision-theoretic problem
- Evaluate the **utility** of a model M for new **unseen data** \tilde{y} :

$$U = \int u(\tilde{y}) p_{\text{true}}(\tilde{y}) d\tilde{y},$$

where \tilde{y} is an unseen observation generated from the true data generating process $p_{\text{text}}(\tilde{y})$, and y are observed data and $u(\tilde{y})$ is a utility function.



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

- Model choice is a (model-)decision-theoretic problem
- Evaluate the **utility** of a model M for new **unseen data** \tilde{y} :

$$U = \int u(\tilde{y}) p_{\text{true}}(\tilde{y}) d\tilde{y},$$

where \tilde{y} is an unseen observation generated from the true data generating process $p_{\text{text}}(\tilde{y})$, and y are observed data and $u(\tilde{y})$ is a utility function.

- The expectation is with respect to p_{true} (f in BDA3)



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

- Model choice is a (model-)decision-theoretic problem
- Evaluate the **utility** of a model M for new **unseen data** \tilde{y} :

$$U = \int u(\tilde{y}) p_{\text{true}}(\tilde{y}) d\tilde{y},$$

where \tilde{y} is an unseen observation generated from the true data generating process $p_{\text{text}}(\tilde{y})$, and y are observed data and $u(\tilde{y})$ is a utility function.

- The expectation is with respect to p_{true} (f in BDA3)
- Choose the model function to **maximize our utility**



UPPSALA
UNIVERSITET

Model choice utility

- **Model assessment and selection**

- Measures of predictive accuracy
- Model selection

- **Cross-validation**

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- **Information criteria**

- **Model averaging and Stacking**

- **Summary**

- Application specific utility/cost functions are important
 - eg. money, life years, quality adjusted life years, etc.



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- 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 for model M

$$\log p_M(y^{\text{rep}}|y)$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- 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 for model M

$$\log p_M(y^{\text{rep}}|y)$$

- We want the "best" model to explain the data



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Subsection 1

Measures of predictive accuracy



- Point residuals

$$e_i = y_i - E(\tilde{y}_i|y),$$

where

$$E(\tilde{y}|y) = \int \tilde{y} p(\tilde{y}|y) d\tilde{y},$$

i.e. the **expected predicted value**

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Point residuals

$$e_i = y_i - E(\tilde{y}_i|y),$$

where

$$E(\tilde{y}|y) = \int \tilde{y} p(\tilde{y}|y) d\tilde{y},$$

i.e. the **expected predicted value**

- Mean squared (prediction) error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_i^n [y_i - E(\tilde{y}_i|y)]^2.$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA
UNIVERSITET

Probabilistic predictions

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- The log score (a local and proper scoring rule)*

$$\log p(y|\theta)$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- The log score (a local and proper scoring rule)*

$$\log p(y|\theta)$$

- The log predictive density (lpd)

$$\begin{aligned}\text{lpd} &= \log p(y|y) \\ &= \log \int p(y|\theta)p(\theta|y)d\theta\end{aligned}$$



- The lpd is usually approximated with the log **point** predictive density (lppd or just lpd)

$$\begin{aligned}\text{lppd} &= \sum_i^n \log p(y_i|y) \\ &\approx \log p(y|y)\end{aligned}$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- The lpd is usually approximated with the log **point** predictive density (lppd or just lpd)

$$\begin{aligned}\text{lppd} &= \sum_i^n \log p(y_i|y) \\ &\approx \log p(y|y)\end{aligned}$$

- Estimation using MCMC

$$\text{lppd} = \sum_i^n \log \left(\frac{1}{S} \sum_s^S p(y_i|\theta_s) \right)$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Subsection 2

Model selection



- Evaluate how model M **generalizes to unseen data** \tilde{y} (the *expected log predictive density*):

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

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

- $\log p_M(\tilde{y}|y)$ is the log score (the utility of the model)

- Model assessment and selection
 - Measures of predictive accuracy
 - **Model selection**
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Evaluate how model M **generalizes to unseen data** \tilde{y} (the *expected log predictive density*):

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

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

- $\log p_M(\tilde{y}|y)$ is the log score (the utility of the model)
- The expectation is with respect to p_{true}



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Evaluate how model M **generalizes to unseen data** \tilde{y} (the *expected log predictive density*):

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

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

- $\log p_M(\tilde{y}|y)$ is the log score (the utility of the model)
- The expectation is with respect to p_{true}
- p_{true} is (almost always) **unknown**



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Evaluate how model M **generalizes to unseen data** \tilde{y} (the *expected log predictive density*):

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

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

- $\log p_M(\tilde{y}|y)$ is the log score (the utility of the model)
- The expectation is with respect to p_{true}
- p_{true} is (almost always) **unknown**
- The utility function is the log scoring rule.



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Section 2

Cross-validation



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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- 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 $p_{\text{true}}(\tilde{y}_i)$, and y are observed data.



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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- 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 $p_{\text{true}}(\tilde{y}_i)$, and y are observed data.

- Can we approximate $p_{\text{true}}(\tilde{y}_i)$?



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

- Approximate $p_{\text{true}}(\tilde{y}_i)$ with data y

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



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

- Approximate $p_{\text{true}}(\tilde{y}_i)$ with data y
- 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**

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- **Cross-validation**
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



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

- Approximate $p_{\text{true}}(\tilde{y}_i)$ with data y
- 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**

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

- **Analogy:** Monte Carlo approximation using our data
- Similar to **jack-knife resampling**

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- **Cross-validation**
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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

- Approximate $p_{\text{true}}(\tilde{y}_i)$ with data y
- 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**

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

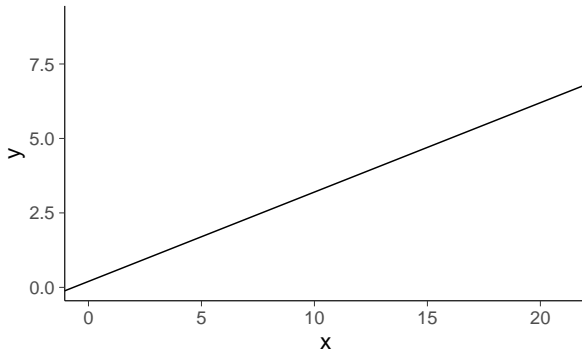
- **Analogy:** Monte Carlo approximation using our data
- Similar to **jack-knife resampling**
- The elpd , lpd and efficient number of parameters (p_{loo})

$$\text{elpd}_{\text{loo}} = \text{lpd} - p_{\text{loo}}$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

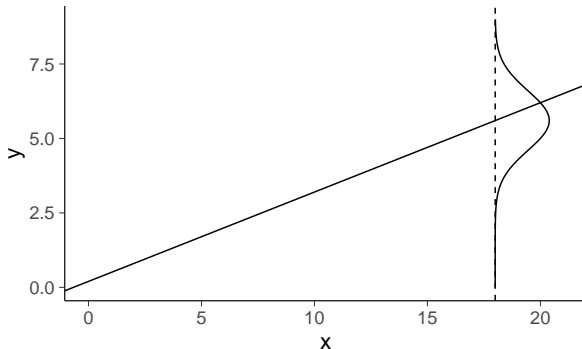
True mean $y = a + bx$





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

True mean and sigma





UPPSALA UNIVERSITET

- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

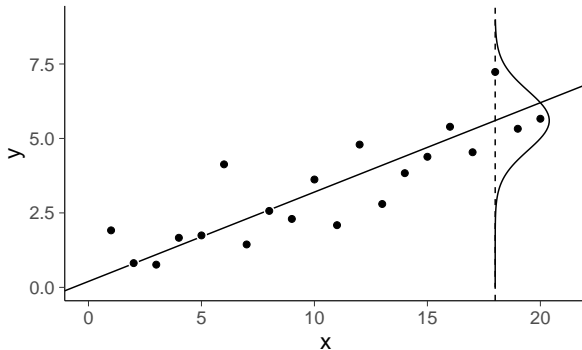
- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

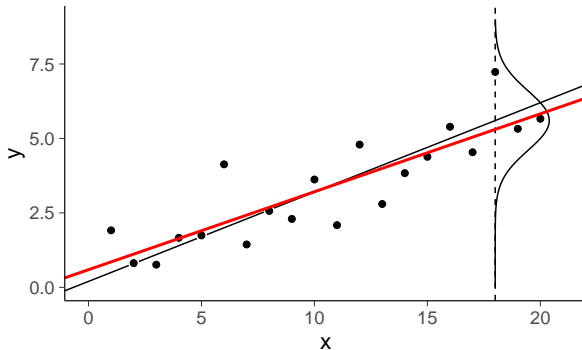
Data





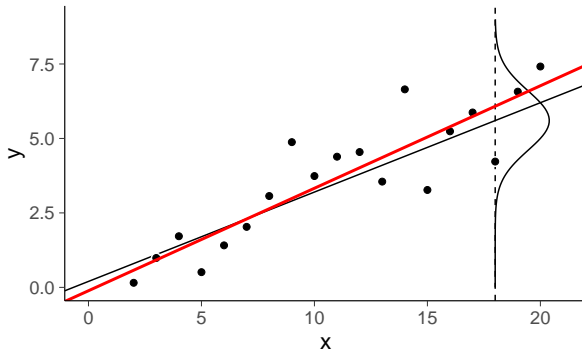
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Posterior mean





Posterior mean, alternative data realisation

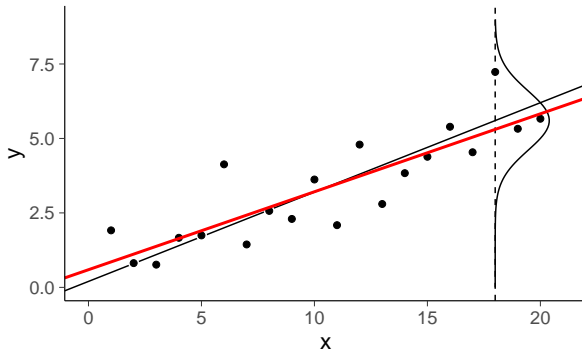


- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

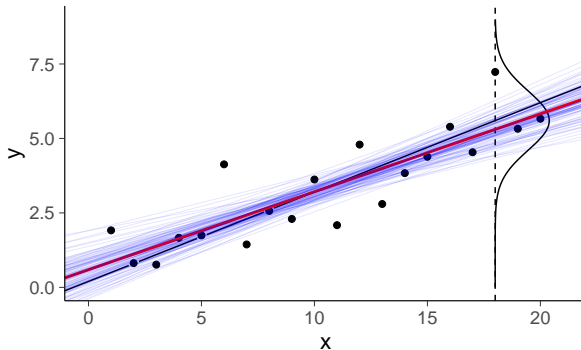
Posterior mean





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

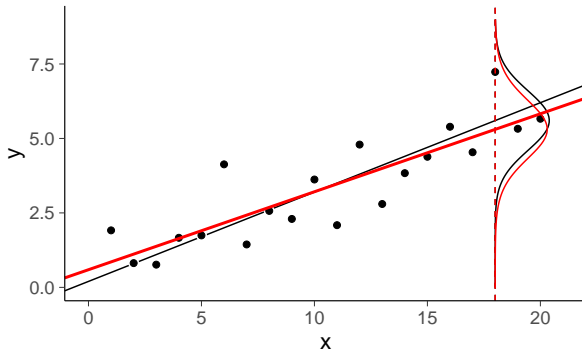
Posterior draws





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

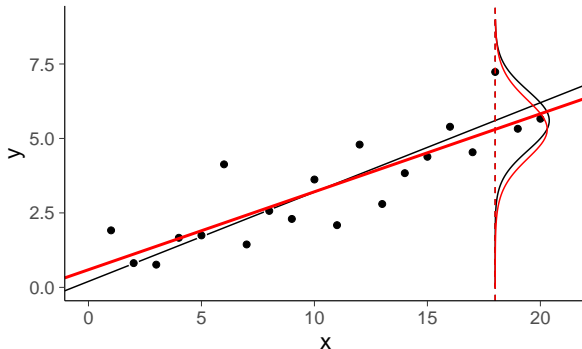
Posterior predictive distribution





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

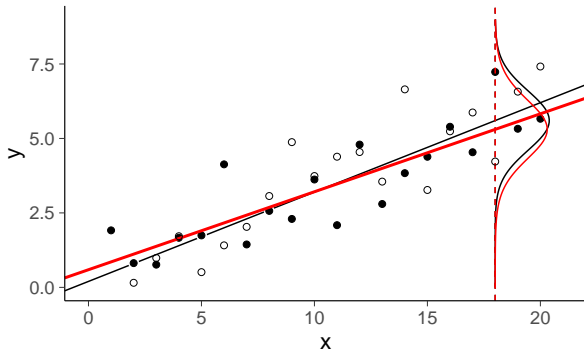
Posterior predictive distribution



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



New data

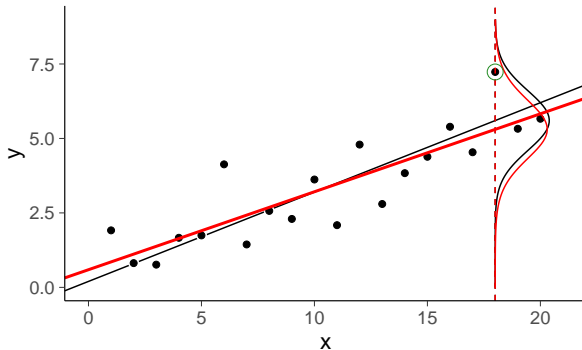


- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Posterior predictive distribution





- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

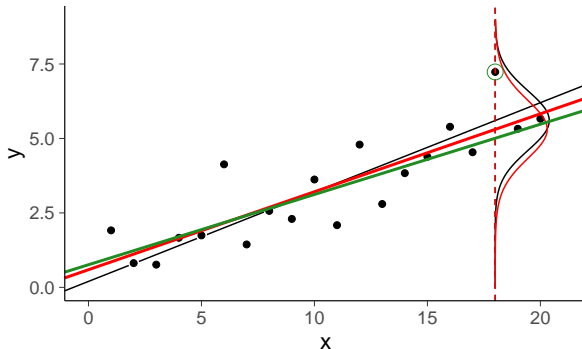
- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

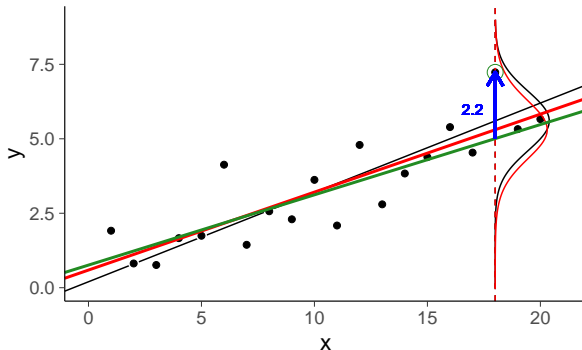
- Summary

Leave-one-out mean





Leave-one-out residual



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

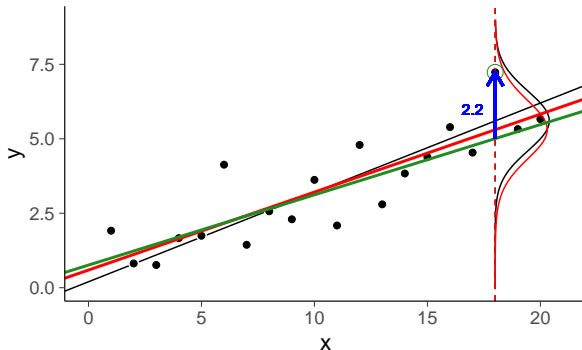
- Information criteria

- Model averaging and Stacking

- Summary



Leave-one-out residual



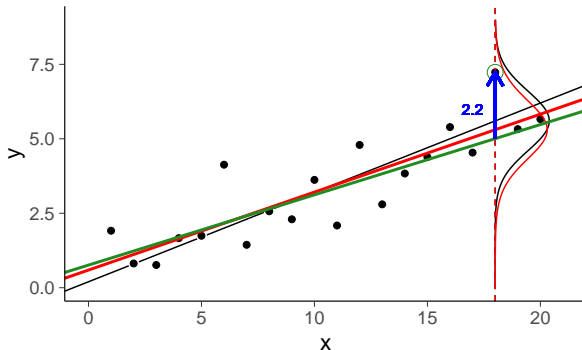
$$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

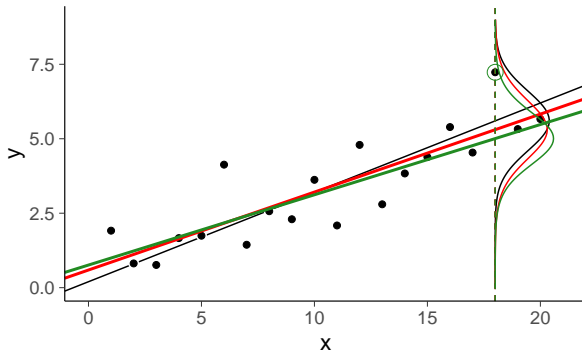
- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

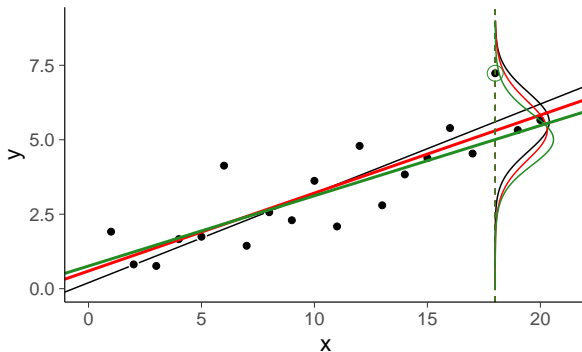
- Summary

Leave-one-out predictive distribution





Leave-one-out predictive distribution



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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

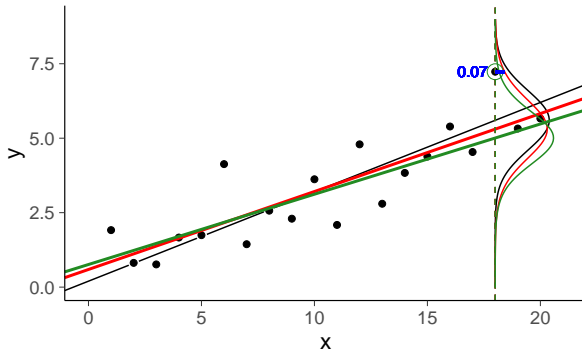
- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

Posterior predictive density





- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

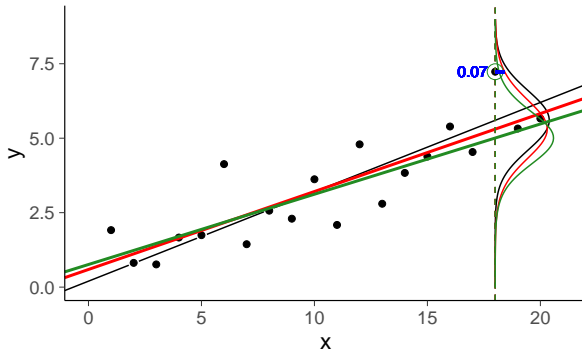
- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

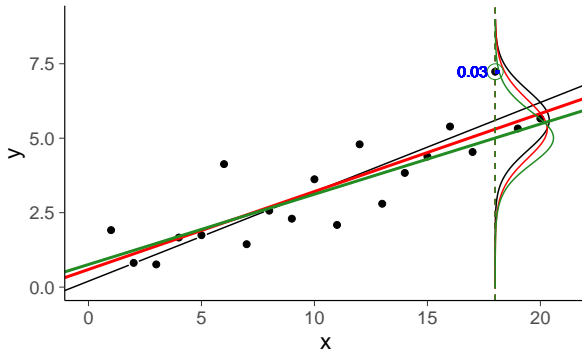
Posterior predictive density



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



Leave-one-out predictive density



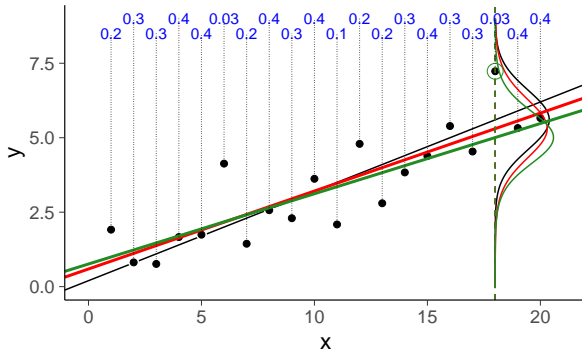
$$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$$

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-out predictive densities

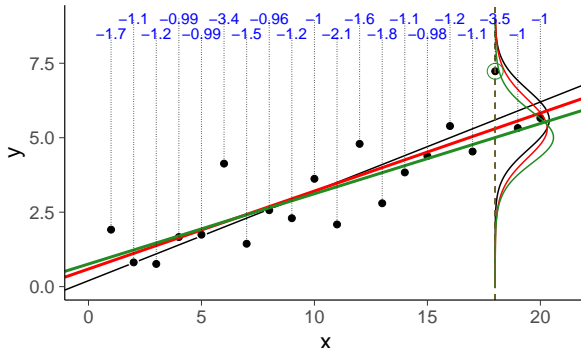


$$p(y_i | x_i, x_{-i}, y_{-i}), \quad i = 1, \dots, 20$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-out log predictive densities

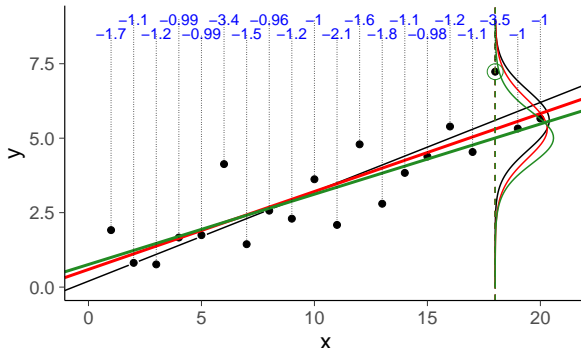


$$\log p(y_i | x_i, x_{-i}, y_{-i}), \quad i = 1, \dots, 20$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-out log predictive densities

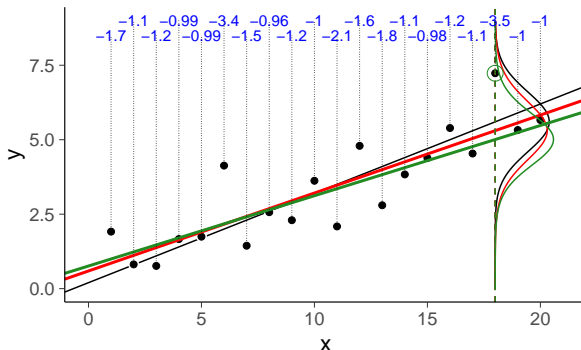


$$\sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-out log predictive densities

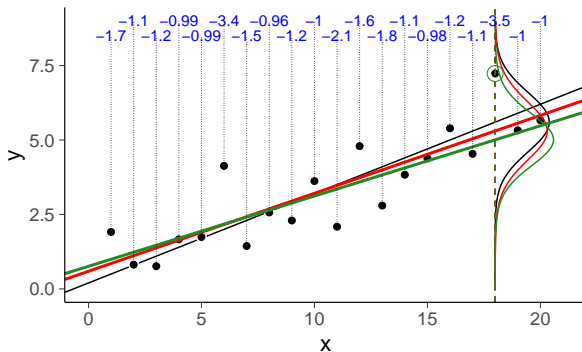


$$\text{elpd_loo} = \sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-out log predictive densities



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

unbiased estimate of log posterior pred. density for new data

- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

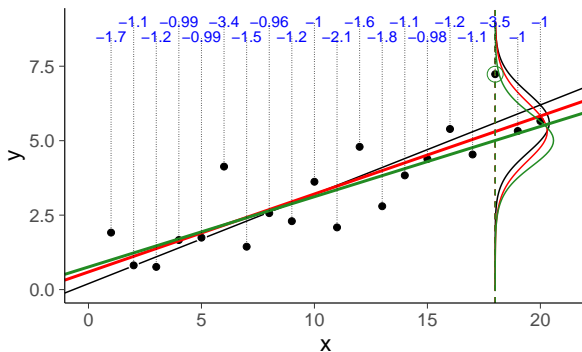
- Information criteria

- Model averaging and Stacking

- Summary



Leave-one-out log predictive densities



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

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

- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

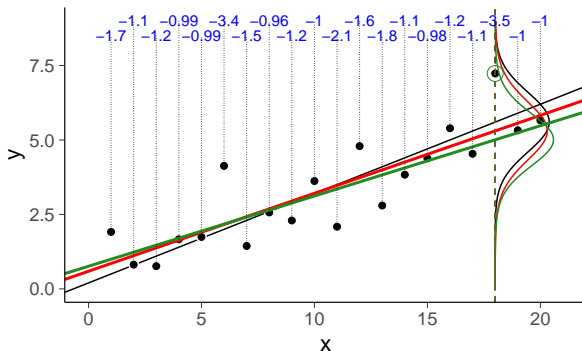
- Information criteria

- Model averaging and Stacking

- Summary



Leave-one-out log predictive densities



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

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

$$\text{p_loo} = \text{lpd} - \text{elpd_loo} \approx 2.7$$

- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

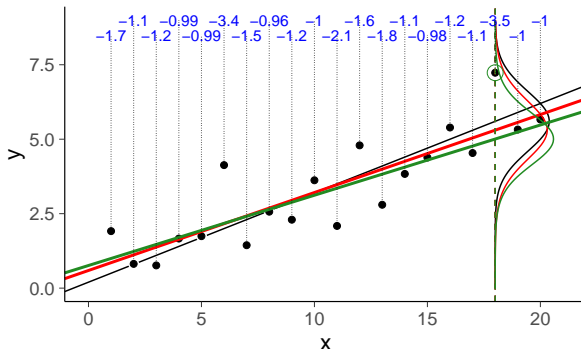
- Information criteria

- Model averaging and Stacking

- Summary



Leave-one-out log predictive densities



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

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

- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary



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 assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Subsection 1

When is LOO applicable



UPPSALA
UNIVERSITET

Pro and cons with LOO-CV

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- + Intuitive
- + Robust
- + Good theoretical properties



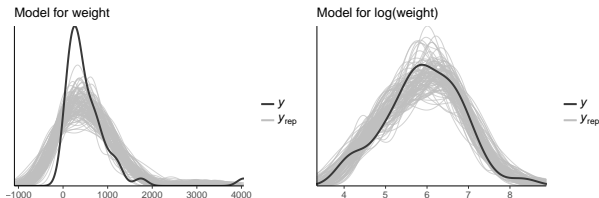
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- + Intuitive
- + Robust
- + Good theoretical properties
 - Can be costly (naive LOO-CV mean n posterior computations)



Sometimes cross-validation is not needed

- Posterior predictive checking can be sufficient



Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2020, Ch. 11)

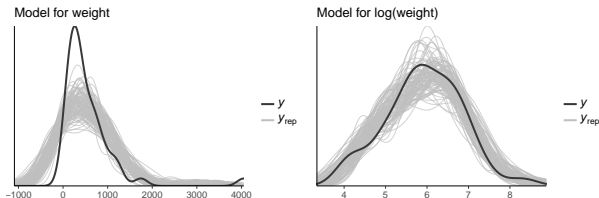
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Sometimes cross-validation is not needed

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Posterior predictive checking can be sufficient



Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2020, Ch. 11)

- In nested case, often easier and more accurate to analyse posterior distribution of more complex model directly



Data generating mechanisms and prediction tasks

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

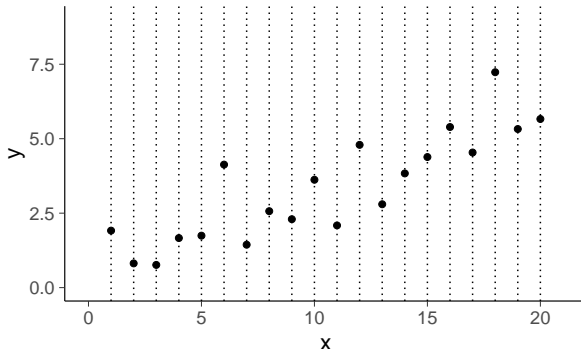
1. You have to make some assumptions on data generating mechanism p_{true}

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

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



Fixed / designed x

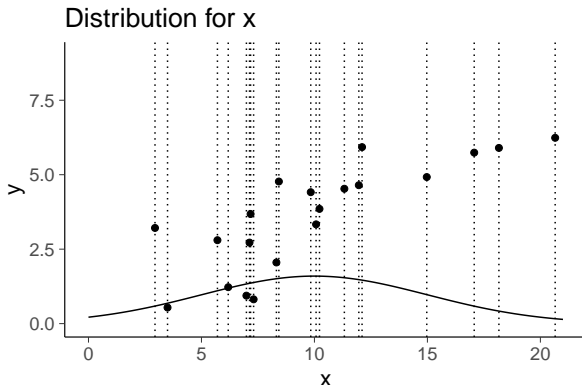


LOO is ok for fixed / designed x : $p_{\text{true}}(y|x)$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



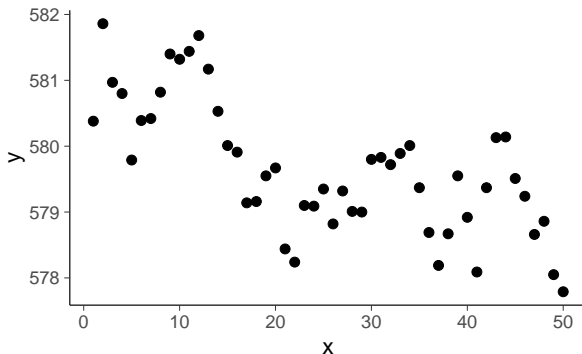
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



LOO is ok for random x . $p_{\text{true}}(y, x)$

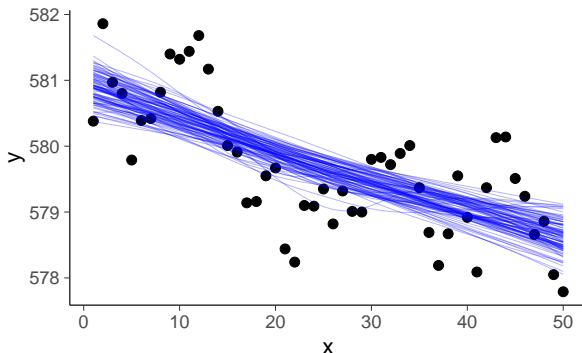


- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary





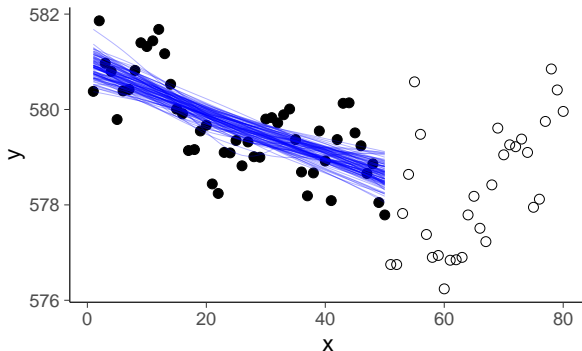
Nonlinear model fit



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - *When is LOO applicable*
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Nonlinear model fit + new data



Extrapolation is more difficult

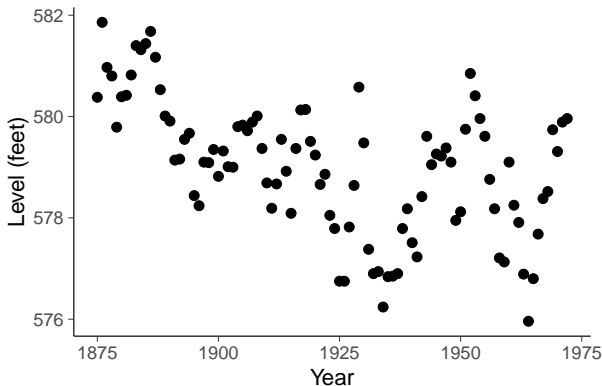
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

LOO for time series data



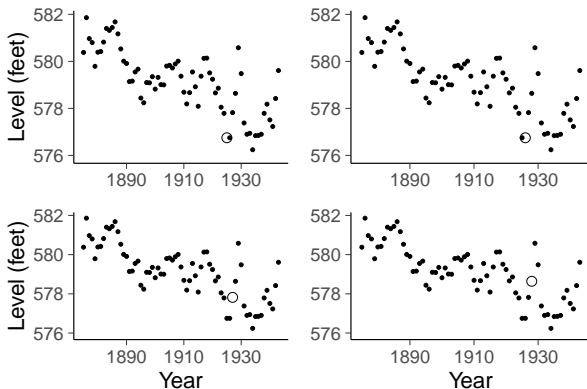
p_{true} for time series?



UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

LOO for time series

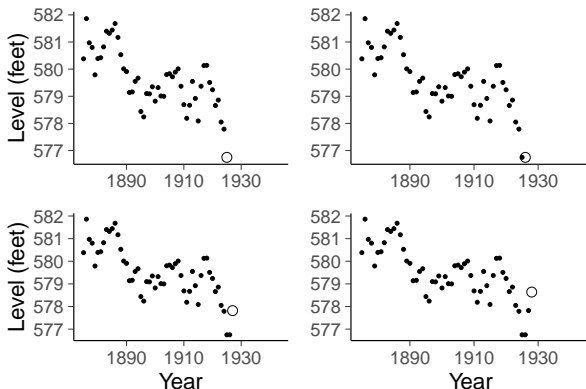


Leave-one-out cross-validation is ok for assessing conditional model



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

LOO for time series

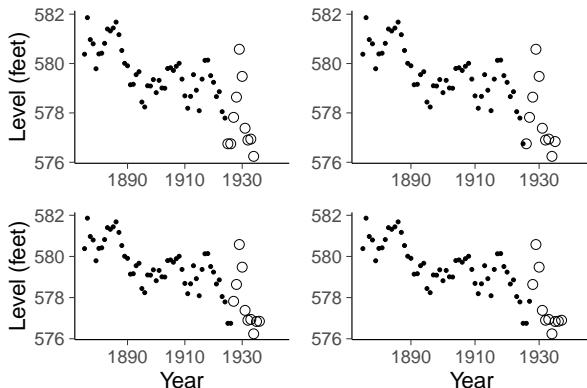


Leave-future-out cross-validation is better for predicting future



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

LOO for time series

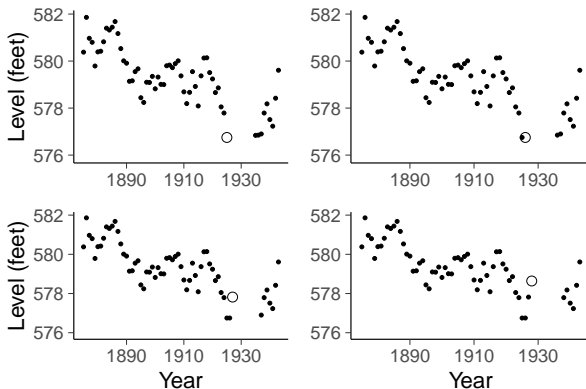


m-step-ahead cross-validation is better for predicting further future



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

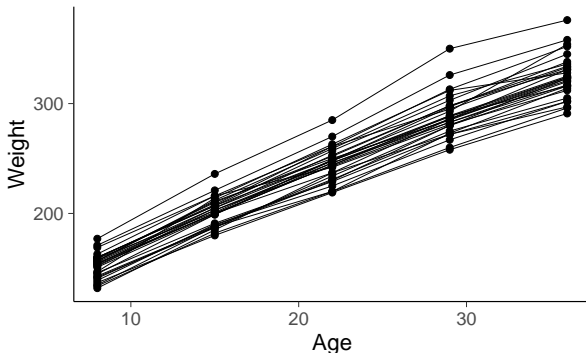
LOO for time series



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



Rats data

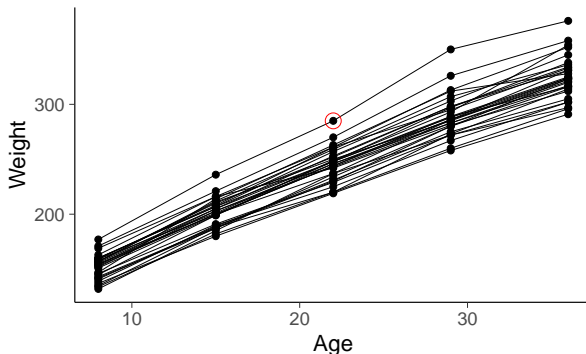


Can LOO or other cross-validation be used with hierarchical data?

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-out?

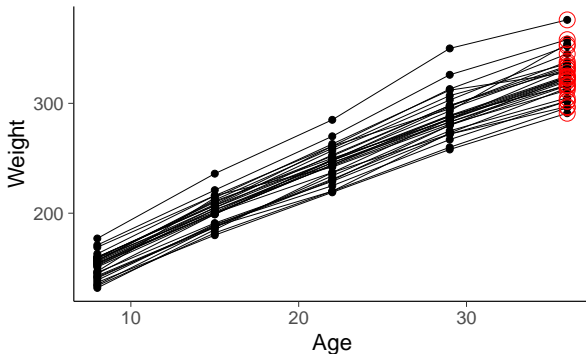


Yes!

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



1-step-ahead?

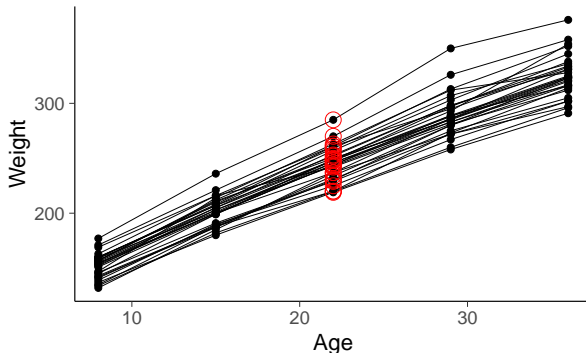


Yes!

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Leave-one-time-point-out?



Yes!

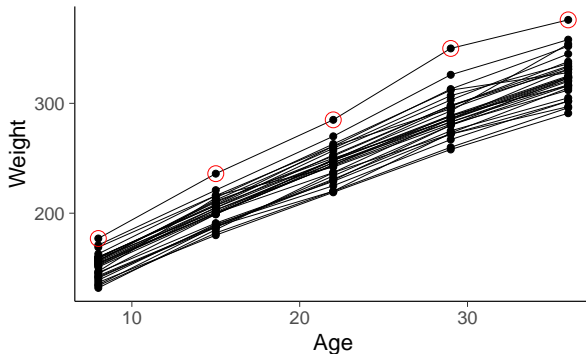
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

LOO for hierarchical data

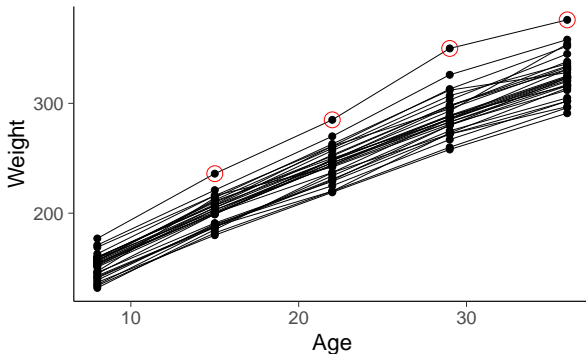
Leave-one-rat-out?



Yes!



Predict given initial weight?



Yes!

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA
UNIVERSITET

Fast cross-validation

1. Pareto smoothed importance sampling LOO (PSIS-LOO)

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA UNIVERSITET

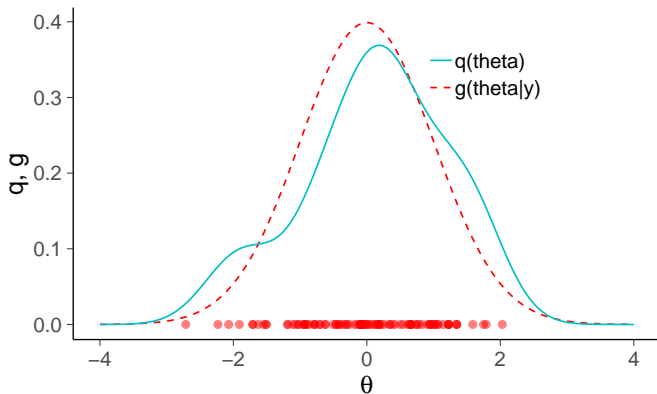
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Subsection 2

PSIS LOO-CV



Target, proposal, and draws



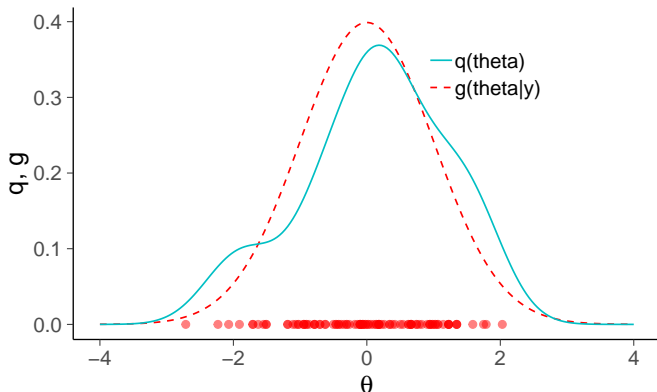
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Recap: Importance sampling

Target, proposal, and draws



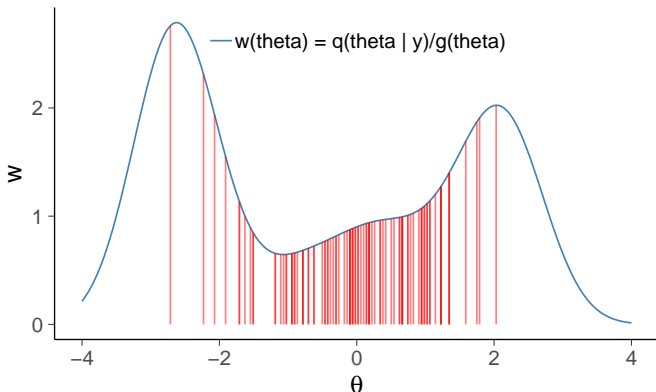
$$E[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}, \quad \text{where} \quad w_s = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Recap: Importance sampling

Draws and importance weights



$$E[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}, \quad \text{where } w_s = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Let $f = p(\tilde{y}_i|y_{-i})$, $g = p(\theta|y)$, and $q = p(\theta|y_{-i})$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Let $f = p(\tilde{y}_i|y_{-i})$, $g = p(\theta|y)$, and $q = p(\theta|y_{-i})$
- Having samples θ^s from $p(\theta^s|y)$

$$p(\tilde{y}_i|y_{-i}) \approx \frac{\sum_{s=1}^S p(\tilde{y}_i|\theta^s) w_i^s}{\sum_{s=1}^S w_i^s},$$

where w_i^s are importance weights and

$$w_i^s = \frac{p(\theta^s|y_{-i})}{p(\theta^s|y)}.$$



- We get a nice solution

$$w_i^s = \frac{p(\theta^s | y_{-i})}{p(\theta^s | y)} = \frac{\frac{p(\theta^s) \prod_{j \neq i} p(y_j | \theta^s)}{p(y_{-i})}}{\frac{p(\theta^s) \prod_j p(y_j | \theta^s)}{p(y)}} \propto \frac{1}{p(y_i | \theta^s)}.$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- We get a nice solution

$$w_i^s = \frac{p(\theta^s | y_{-i})}{p(\theta^s | y)} = \frac{\frac{p(\theta^s) \prod_{j \neq i} p(y_j | \theta^s)}{p(y_{-i})}}{\frac{p(\theta^s) \prod_j p(y_j | \theta^s)}{p(y)}} \propto \frac{1}{p(y_i | \theta^s)}.$$

- If evaluated with $\tilde{y}_i = y_i$

$$p(\tilde{y}_i | y_{-i}) \approx \frac{\sum_{s=1}^S p(\tilde{y}_i | \theta^s) w_i^s}{\sum_{s=1}^S w_i^s} = \frac{1}{\sum_{s=1}^S \frac{1}{p(y_i | \theta^s)}},$$

and

$$\log w_i^{(s)} = -\log p(y_i | \theta^{(s)})$$

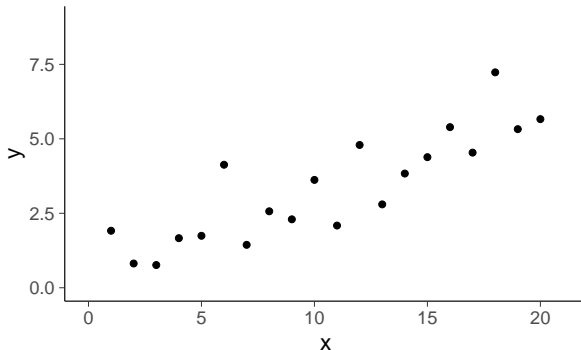
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

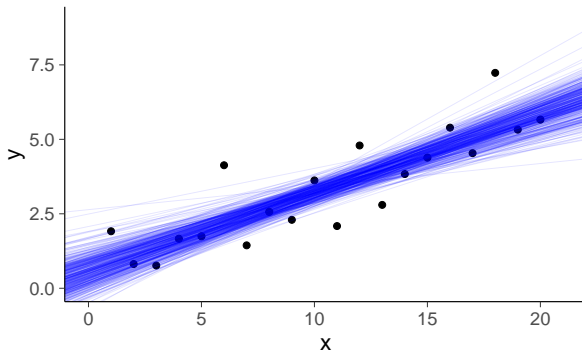
Data





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Posterior draws

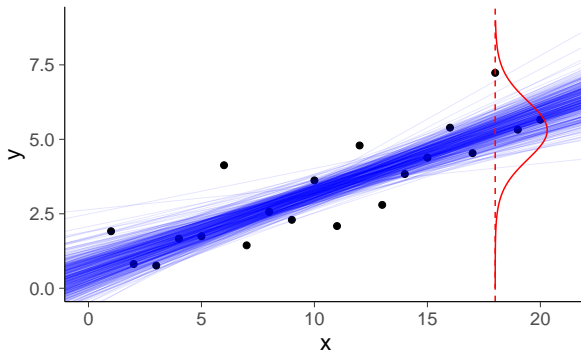


$$\theta^{(s)} \sim p(\theta|x, y)$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Posterior predictive distribution

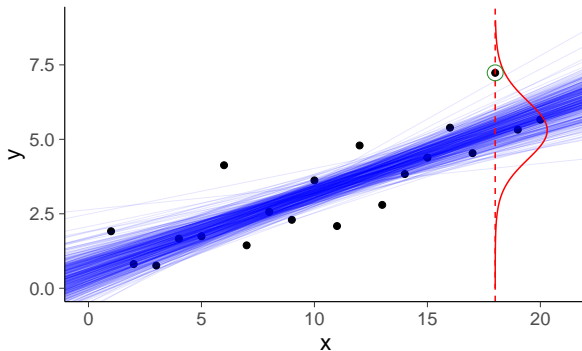


$$\theta^{(s)} \sim p(\theta|x, y), \quad p(\tilde{y}|\tilde{x}, x, y) \approx \frac{1}{S} \sum_{s=1}^S p(\tilde{y}|\tilde{x}, \theta^{(s)})$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Posterior predictive distribution

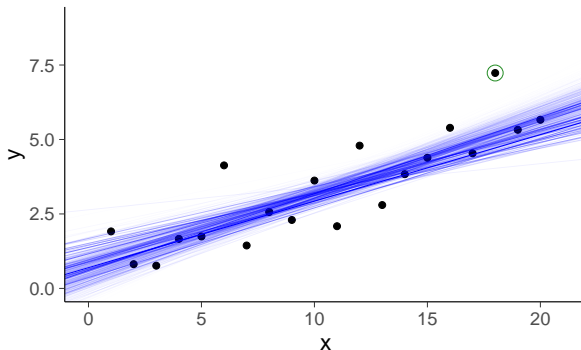


$$\theta^{(s)} \sim p(\theta|x, y), \quad p(\tilde{y}|\tilde{x}, x, y) \approx \frac{1}{S} \sum_{s=1}^S p(\tilde{y}|\tilde{x}, \theta^{(s)})$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

PSIS-LOO weighted draws



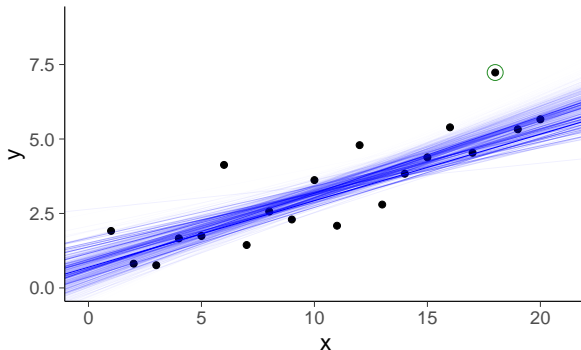
$$\theta^{(s)} \sim p(\theta|x, y)$$

$$w_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y)$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

PSIS-LOO weighted draws



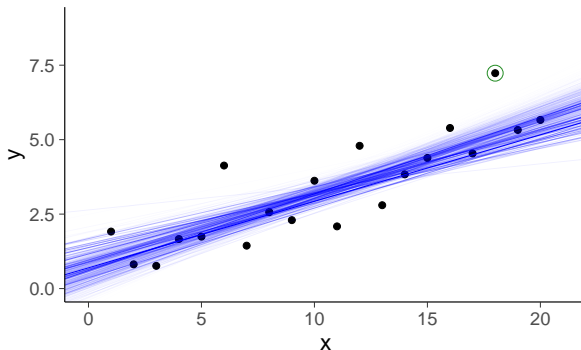
$$\theta^{(s)} \sim p(\theta|x, y)$$

$$w_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y) \propto 1/p(y_i|x_i, \theta^{(s)})$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

PSIS-LOO weighted draws



$$\theta^{(s)} \sim p(\theta|x, y)$$

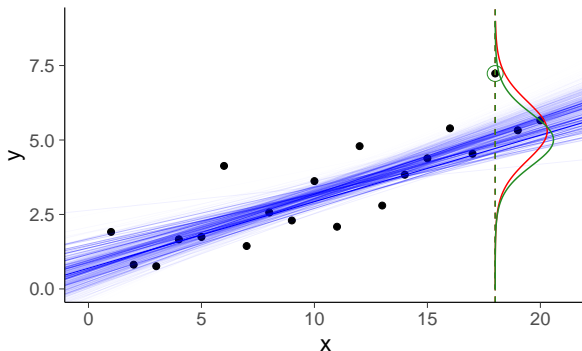
$$w_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y) \propto 1/p(y_i|x_i, \theta^{(s)})$$

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

PSIS-LOO weighted predictive distribution

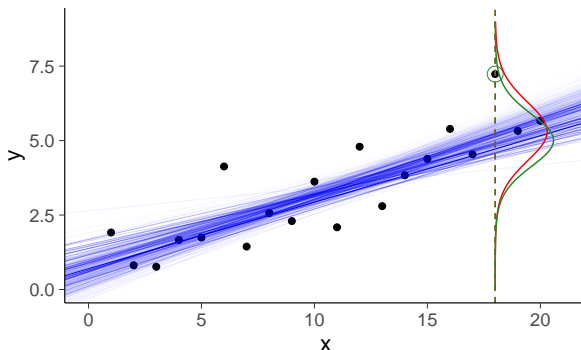


$$\theta^{(s)} \sim p(\theta|x, y)$$

$$w_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y) \propto 1/p(y_i|x_i, \theta^{(s)})$$



PSIS-LOO weighted predictive distribution



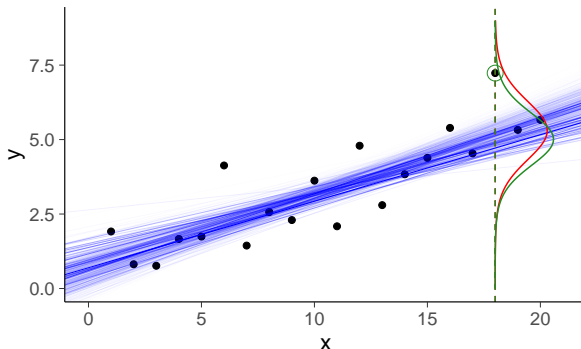
$$\theta^{(s)} \sim p(\theta|x, y)$$

$$w_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y) \propto 1/p(y_i|x_i, \theta^{(s)})$$

$$p(y_i|x_i, x_{-i}, y_{-i}) \approx \sum_{s=1}^S [w_i^{(s)} p(y_i|x_i, \theta^{(s)})]$$



PSIS-LOO weighted predictive distribution



$$\theta^{(s)} \sim p(\theta|x, y)$$

$$w_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y) \propto 1/p(y_i|x_i, \theta^{(s)})$$

$$p(y_i|x_i, x_{-i}, y_{-i}) \approx \sum_{s=1}^S [w_i^{(s)} p(y_i|x_i, \theta^{(s)})]$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Recap: Pareto smoothed importance sampling

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
 - Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
 - Information criteria
 - Model averaging and Stacking
 - Summary
- Pareto-Smoothed Importance sampling smooth the weights according to a Generalized Pareto(k) distribution
 - Pareto- k diagnostic estimate the number of existing moments ($\lfloor 1/k \rfloor$)



Recap: Pareto smoothed importance sampling

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
 - Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
 - Information criteria
 - Model averaging and Stacking
 - Summary
- Pareto-Smoothed Importance sampling smooth the weights according to a Generalized Pareto(k) distribution
 - Pareto- k diagnostic estimate the number of existing moments ($\lfloor 1/k \rfloor$)
 - Finite variance and central limit theorem for $k < 1/2$



Recap: Pareto smoothed importance sampling

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Pareto-Smoothed Importance sampling smooth the weights according to a Generalized Pareto(k) distribution
- Pareto- k diagnostic estimate the number of existing moments ($\lfloor 1/k \rfloor$)
- Finite variance and central limit theorem for $k < 1/2$
- Finite mean and generalized central limit theorem for $k < 1$, but pre-asymptotic constant grows impractically large for $k > 0.7$



Recap: Pareto smoothed importance sampling

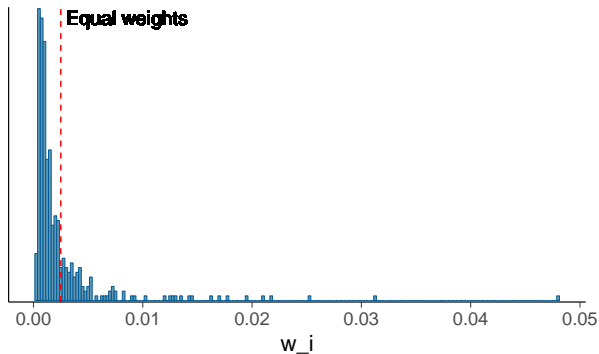
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- Pareto-Smoothed Importance sampling smooth the weights according to a Generalized Pareto(k) distribution
- Pareto- k diagnostic estimate the number of existing moments ($\lfloor 1/k \rfloor$)
- Finite variance and central limit theorem for $k < 1/2$
- Finite mean and generalized central limit theorem for $k < 1$, but pre-asymptotic constant grows impractically large for $k > 0.7$
- Hence $k > 0.7$ can be used to indicate what observation that can "hold-out" using IS



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

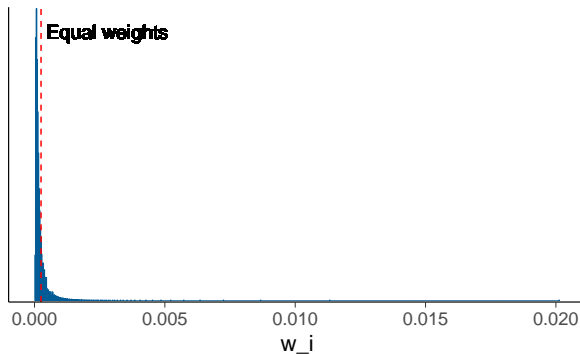
400 importance weights for leave-18th-out





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

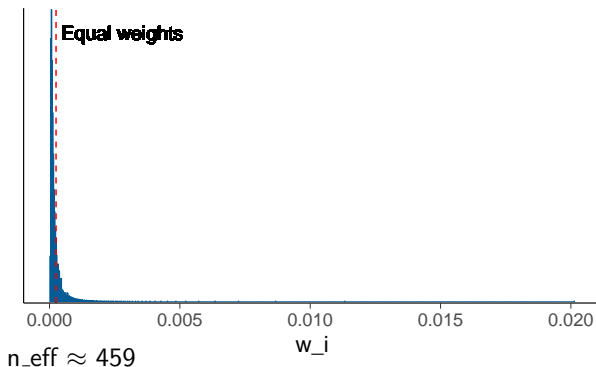
4000 importance weights for leave-18th-out





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

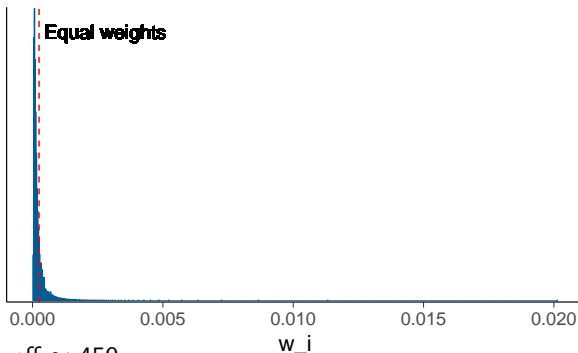
4000 importance weights for leave-18th-out





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

4000 importance weights for leave-18th-out



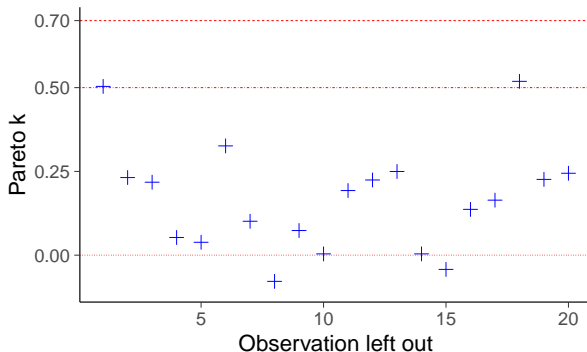
$n_{\text{eff}} \approx 459$

Pareto: $\hat{k} \approx 0.52$ Pareto \hat{k} estimates the tail shape which determines the convergence rate of PSIS. Less than 0.7 is ok.



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

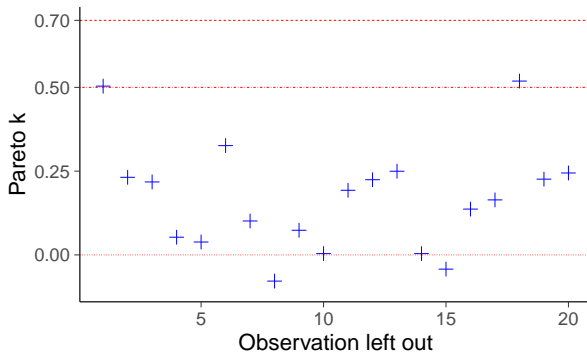
PSIS-LOO diagnostics





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

PSIS-LOO diagnostics

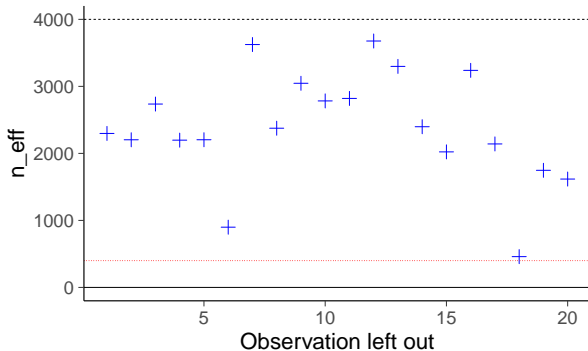


Pareto k diagnostic values:

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



PSIS-LOO diagnostics



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>

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA UNIVERSITET

Stan code

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Stan code

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

```
...
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);
}
```



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 assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - **PSIS LOO-CV**
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. PSIS-LOO for hierarchical models

1.1 leave-one-group out is challenging for PSIS-LOO



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. PSIS-LOO for hierarchical models
 - 1.1 leave-one-group out is challenging for PSIS-LOO
2. PSIS-LOO for non-factorizable models

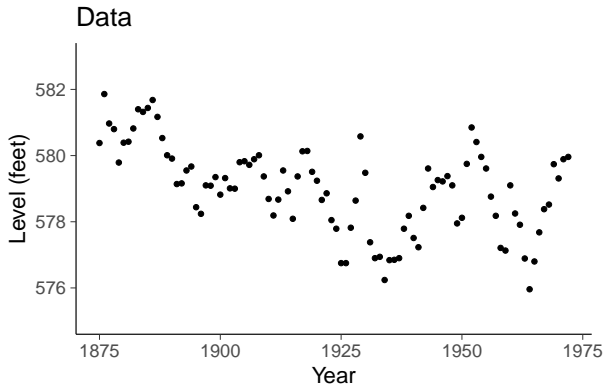


- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. PSIS-LOO for hierarchical models
 - 1.1 leave-one-group out is challenging for PSIS-LOO
2. PSIS-LOO for non-factorizable models
3. PSIS-LOO for time series
 - 3.1 Approximate leave-future-out cross-validation

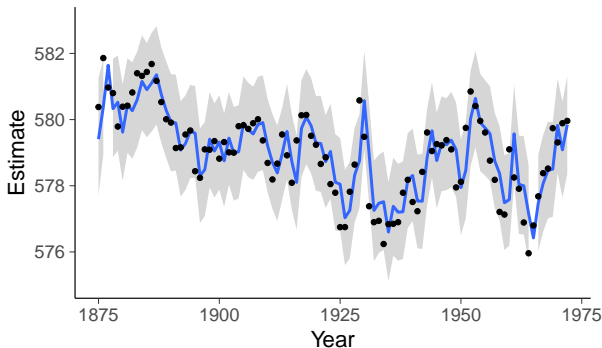


- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary





AR-4 prediction with 95% interval



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

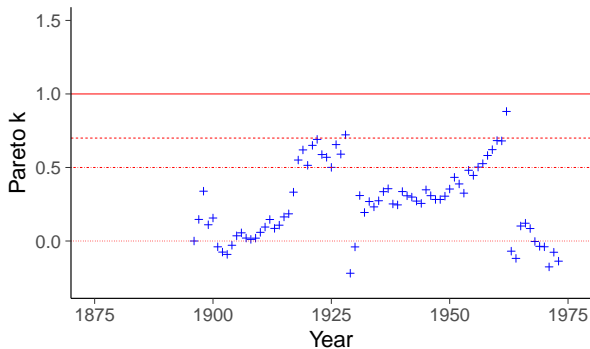


UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

PSIS-LOO for time series

PSIS-1-step-ahead with refits



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



UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Subsection 3

K-fold CV



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- We can approximate $elpd_{loo}$ using K-fold CV



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- We can approximate $elpd_{loo}$ using K-fold CV
 1. for k in K
 - 1.1 hold $1/K$ observations out (i.e. $y_{-y(k)}$)
 - 1.2 compute $p(\theta|y_{-y(k)})$
 - 1.3 compute lpd for $y_{-y(k)}$
 2. sum the lpd over K folds



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

- We can approximate $elpd_{loo}$ using K-fold CV
 1. for k in K
 - 1.1 hold $1/K$ observations out (i.e. $y_{-y(k)}$)
 - 1.2 compute $p(\theta|y_{-y(k)})$
 - 1.3 compute lpd for $y_{-y(k)}$
 2. sum the lpd over K folds
- We only need to estimate K models



K-fold cross-validation

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

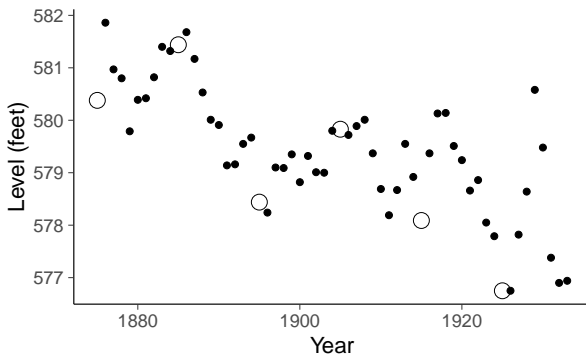
1. K-fold cross-validation can approximate LOO
 - 1.1 can be used all uses for LOO
2. K-fold cross-validation can be used for hierarchical models
 - 2.1 good for leave-one-group-out
3. K-fold cross-validation can be used for time series
 - 3.1 with leave-block-out



UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Balance k-fold approximation of LOO

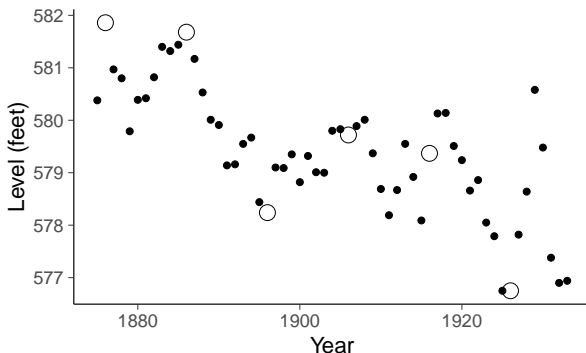




UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Balance k-fold approximation of LOO

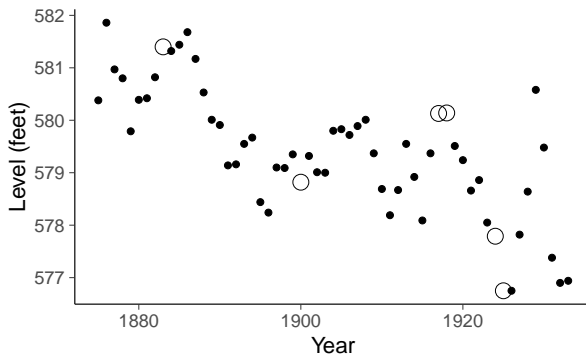




UPPSALA UNIVERSITET

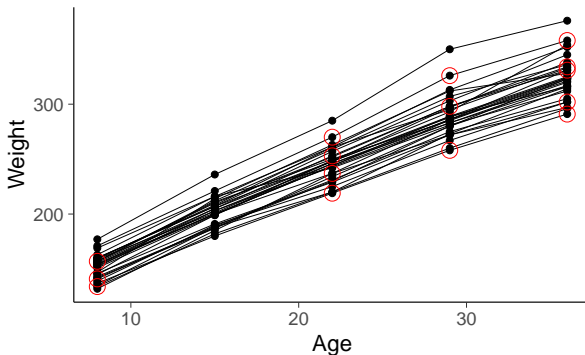
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Random k-fold approximation of LOO





Random kfold approximation of LOO

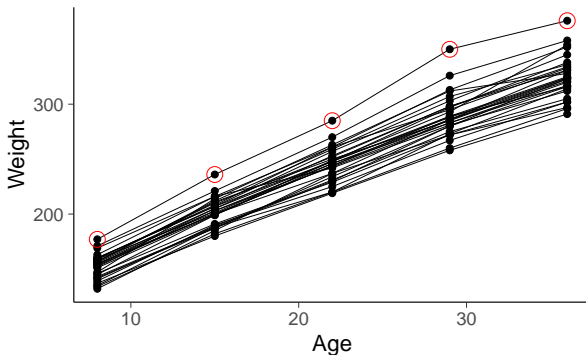


- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA UNIVERSITET

Leave-one-rat-out



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. CV is good for model assessment when application specific utility/cost functions are used



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - **K-fold CV**
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. CV is good for model assessment when application specific utility/cost functions are used
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`)



UPPSALA UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Subsection 4

Model Comparison



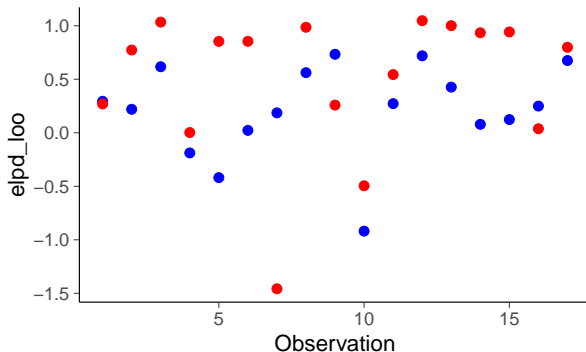
- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. “A popular hypothesis has it that primates with larger brains produce more energetic milk, so that brains can grow quickly” (from Statistical Rethinking)
 - 1.1 Model 1: formula = kcal.per.g \sim neocortex
 - 1.2 Model 2: formula = kcal.per.g \sim neocortex + log(mass)



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

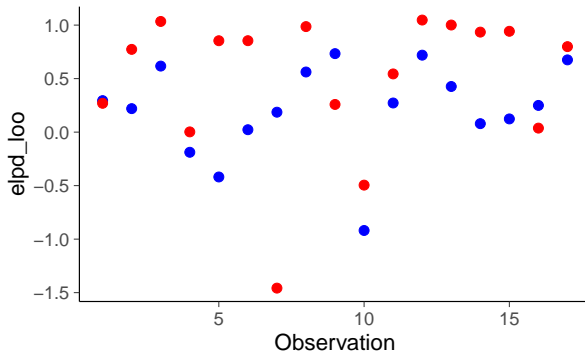
Pointwise comparison LOO models: Model 1





- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Pointwise comparison LOO models: Model 1



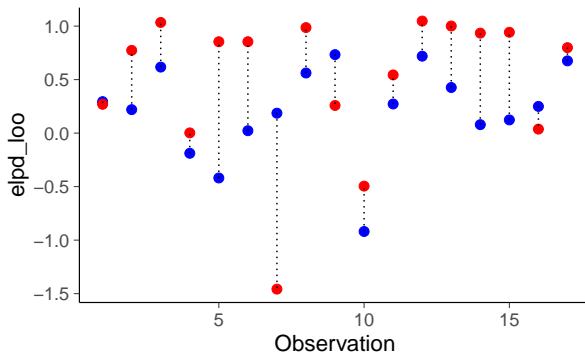
Model 1 $\text{elpd_loo} \approx 3.7$, $\text{SE}=1.8$

Model 2 $\text{elpd_loo} \approx 8.4$, $\text{SE}=2.8$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Pointwise comparison LOO models: Model 1

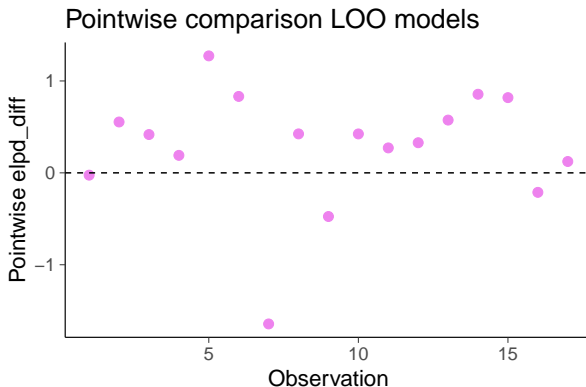


Model 1 $\text{elpd_loo} \approx 3.7$, $\text{SE}=1.8$

Model 2 $\text{elpd_loo} \approx 8.4$, $\text{SE}=2.8$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



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

elpd_diff	se
4.7	2.7



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Arsenic well example – Model comparison

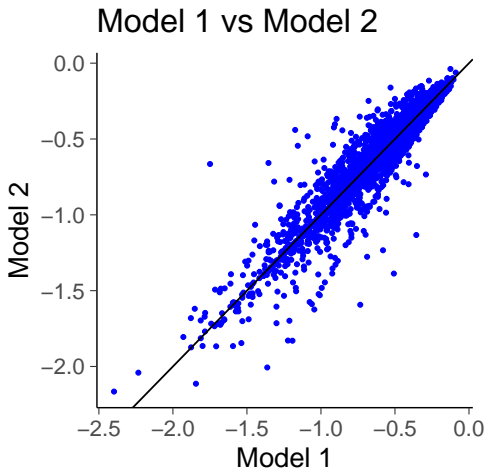
1. Probability of switching well with high arsenic level in rural Bangladesh
 - 1.1 Model 1 covariates: $\log(\text{arsenic})$ and distance
 - 1.2 Model 2 covariates: $\log(\text{arsenic})$, distance and education level

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Arsenic well example – Model comparison

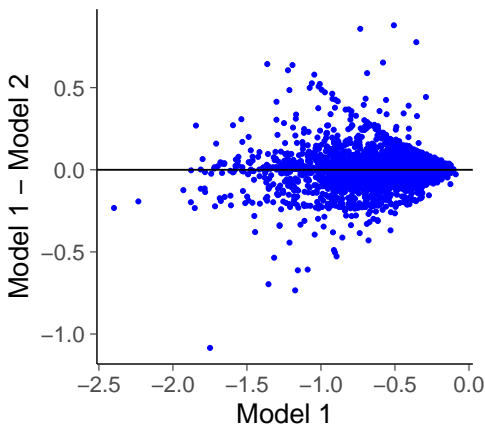


Model 1 $\text{elpd}_{\text{loo}} \approx -1952$, $\text{SE}=16$

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



Model 1 vs Model 2



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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



Arsenic well example – Model comparison

- Model assessment and selection

- Measures of predictive accuracy
- Model selection

- Cross-validation

- When is LOO applicable
- PSIS LOO-CV
- K-fold CV
- Model Comparison

- Information criteria

- Model averaging and Stacking

- Summary

```
> 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.



UPPSALA
UNIVERSITET

What if one is not clearly better than others?

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



What if one is not clearly better than others?

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Continuous expansion including all models?

1.1 and then analyse the posterior distribution directly

1.2 sparse priors like regularized horseshoe prior instead of variable selection



What if one is not clearly better than others?

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Continuous expansion including all models?
 - 1.1 and then analyse the posterior distribution directly
 - 1.2 sparse priors like regularized horseshoe prior instead of variable selection
2. Model averaging with BMA or Bayesian stacking?



What if one is not clearly better than others?

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Continuous expansion including all models?
 - 1.1 and then analyse the posterior distribution directly
 - 1.2 sparse priors like regularized horseshoe prior instead of variable selection
2. Model averaging with BMA or Bayesian stacking?
3. In a nested case choose simpler if assuming some cost for extra parts?



What if one is not clearly better than others?

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Continuous expansion including all models?
 - 1.1 and then analyse the posterior distribution directly
 - 1.2 sparse priors like regularized horseshoe prior instead of variable selection
2. Model averaging with BMA or Bayesian stacking?
3. In a nested case choose simpler if assuming some cost for extra parts?
4. In a nested case choose more complex if you want to take into account all the uncertainties.



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
 - Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
 - Information criteria
 - Model averaging and Stacking
 - Summary
- 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



Selection induced bias and overfitting

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
 - Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
 - Information criteria
 - Model averaging and Stacking
 - Summary
- 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
 - 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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
 - Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
 - Information criteria
 - Model averaging and Stacking
 - Summary
- 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
 - 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



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Section 3

Information criteria



1. Remember the relationship between the elpd and the lpd

$$\text{elpd}_{\text{loo}} = \sum_i \log p(y_i|y) - p_{\text{loo}}$$

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Remember the relationship between the elpd and the lpd

$$\text{elpd}_{\text{loo}} = \sum_i \log p(y_i | y) - p_{\text{loo}}$$

2. Akaike information criteria

$$\text{elpd}_{\text{AIC}} = \sum_i \log p(y_i | \hat{\theta}_{\text{MLE}}) - k$$

3. AIC

$$\text{AIC} = -2\text{elpd}_{\text{AIC}}$$



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Deviance information criteria

$$\text{elpd}_{\text{AIC}} = \sum_i \log p(y_i | \hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}}$$

2. We estimate the number of effeicient parameters with

3. DIC

$$\text{DIC} = -2\text{elpd}_{\text{DIC}}$$



UPPSALA
UNIVERSITET

WAIC vs PSIS-LOO

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary



UPPSALA
UNIVERSITET

WAIC vs PSIS-LOO

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. WAIC has same assumptions as LOO



UPPSALA
UNIVERSITET

WAIC vs PSIS-LOO

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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)



- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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, AIC_c, ...



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Section 4

Model averaging and Stacking



UPPSALA
UNIVERSITET

Special lecture!

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

1. Special lecture on model averaging and stacking next week!



UPPSALA
UNIVERSITET

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

Section 5

Summary



Take-home messages

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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

- Model assessment and selection
 - Measures of predictive accuracy
 - Model selection
- Cross-validation
 - When is LOO applicable
 - PSIS LOO-CV
 - K-fold CV
 - Model Comparison
- Information criteria
- Model averaging and Stacking
- Summary

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