# Chapter 7

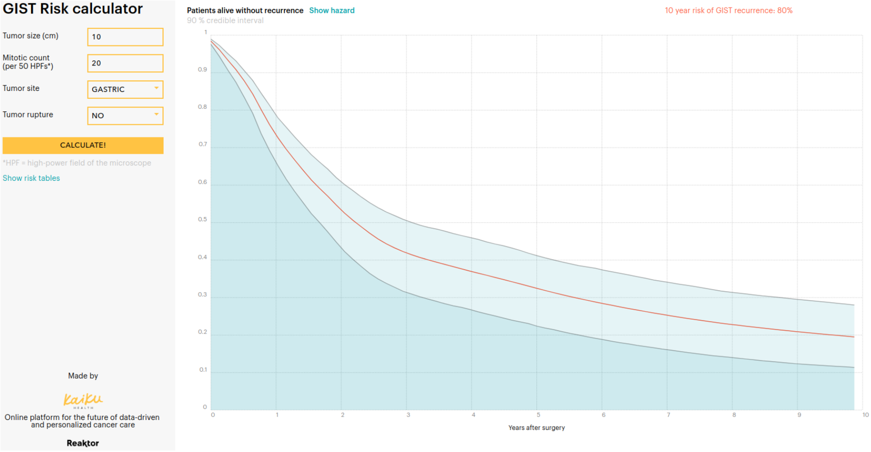
- 7.1 Measures of predictive accuracy
- 7.2 Information criteria and cross-validation
  - Instead of 7.2, read:
    Vehtari, A., Gelman, A., Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing. 27(5):1413–1432. arXiv preprint.
- 7.3 Model comparison based on predictive performance
- 7.4 Model comparison using Bayes factors
- 7.5 Continuous model expansion / sensitivity analysis
- 7.5 Example (may be skipped)

# Model assesment, selection and inference after selection

- Extra material at https://avehtari.github.io/modelselection/
  - Videos, Slides, Notebooks, References
  - The most relevant for the course is the first part of the talk “Model assessment, comparison and selection at Master class in Bayesian statistics, CIRM, Marseille”

# Predicting concrete quality

# Predicting cancer recurrence



# Predictive performance

- True predictive performance is found out by using it to make predictions and comparing predictions to true observations
  - external validation

# Predictive performance

- True predictive performance is found out by using it to make predictions and comparing predictions to true observations
  - external validation
- Expected predictive performance
  - approximates the external validation

# Predictive performance

- We need to choose the utility/cost function
- Application specific utility/cost functions are important
  - eg. money, life years, quality adjusted life years, etc.

# Predictive performance

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

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

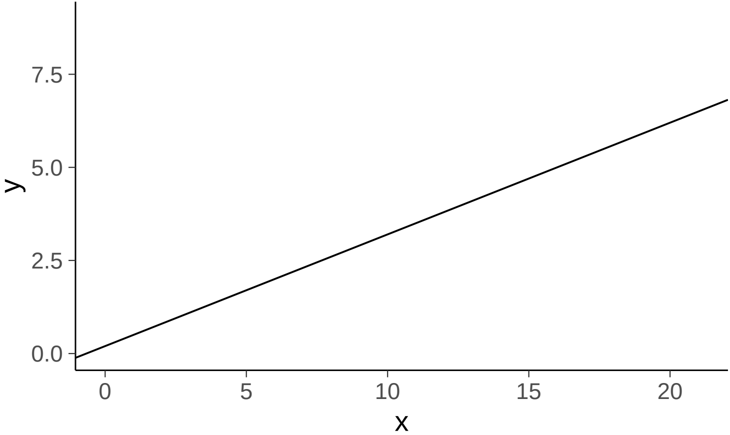
# Outline

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

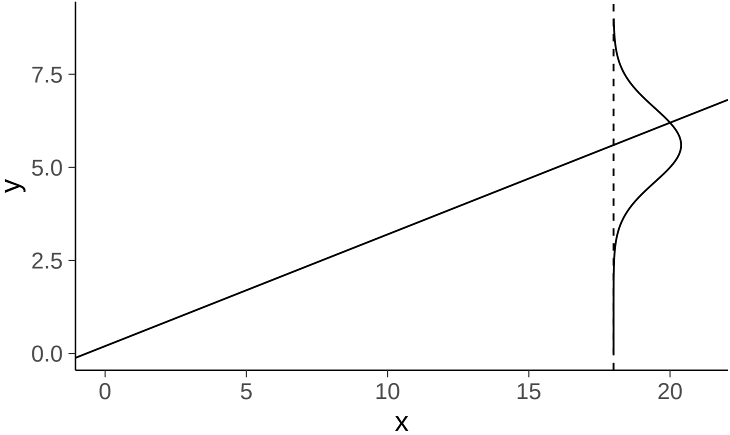
# Stan and loo package

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

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

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

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

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

elpd_diff        se
     -0.2       0.1
```

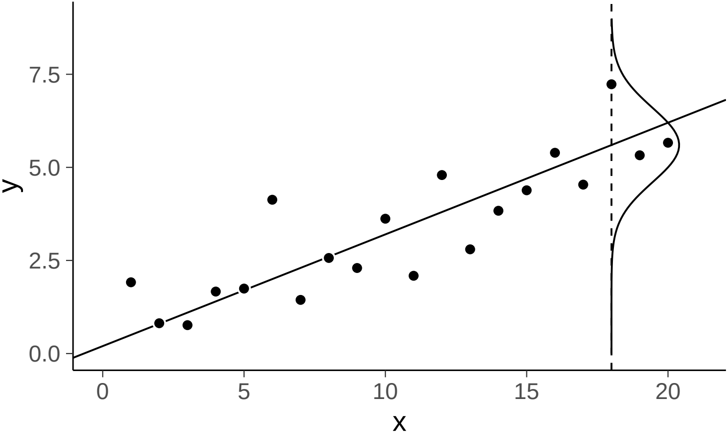
# True mean y = a + bx



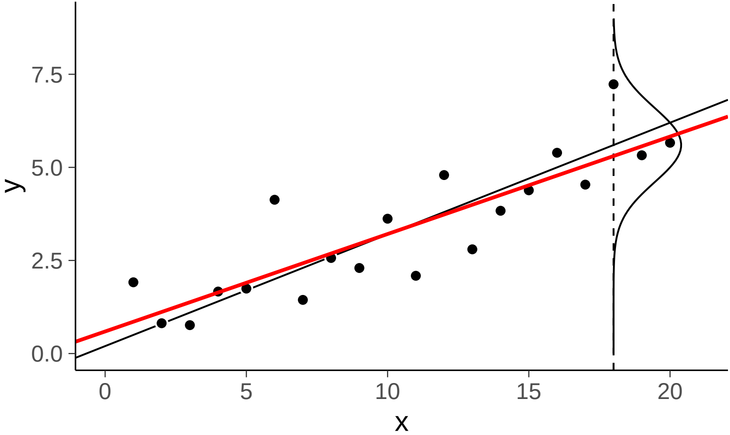
## True mean and sigma



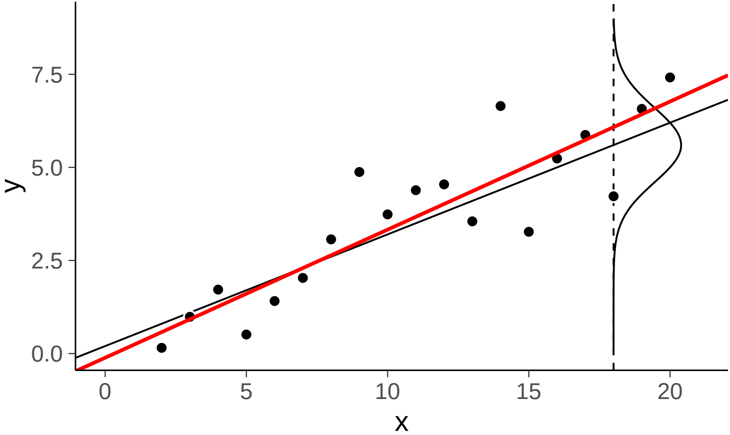
# Data



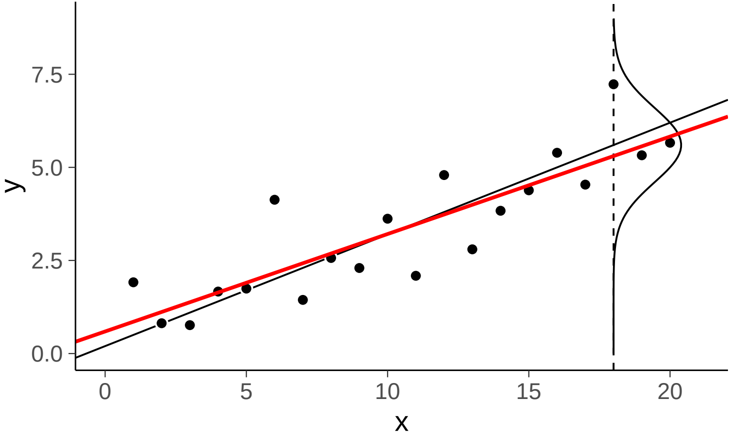
# Posterior mean



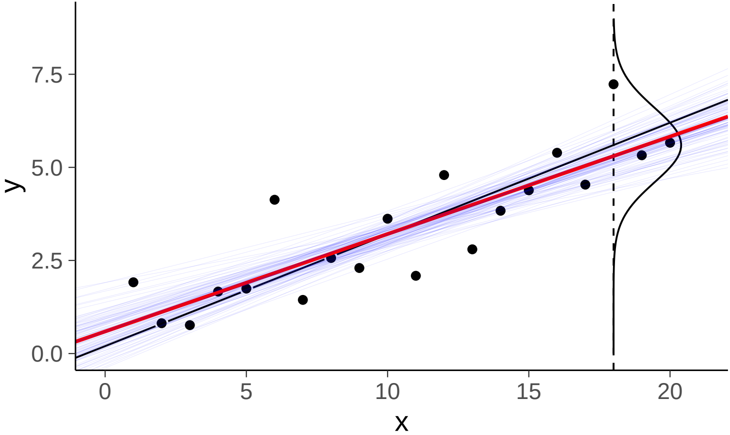
# Posterior mean, alternative data realisation



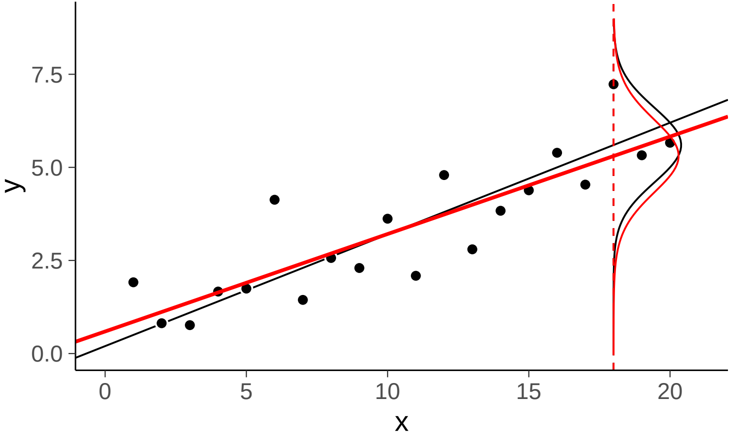
# Posterior mean



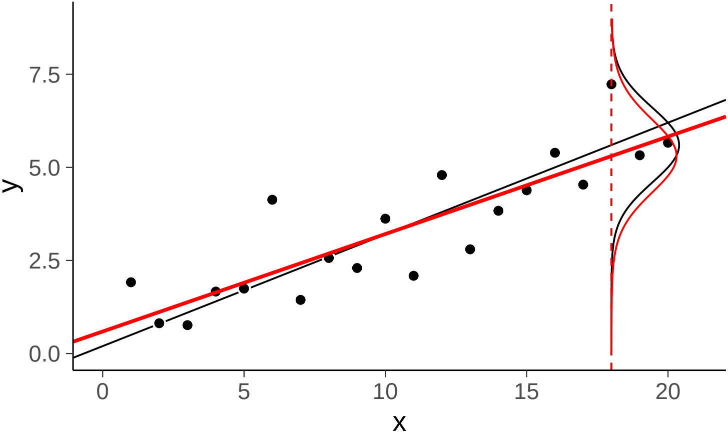
# Posterior draws



# Posterior predictive distribution

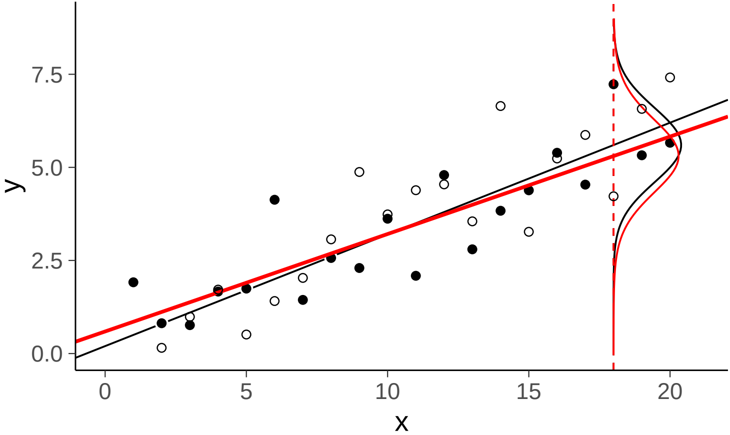


# Posterior predictive distribution

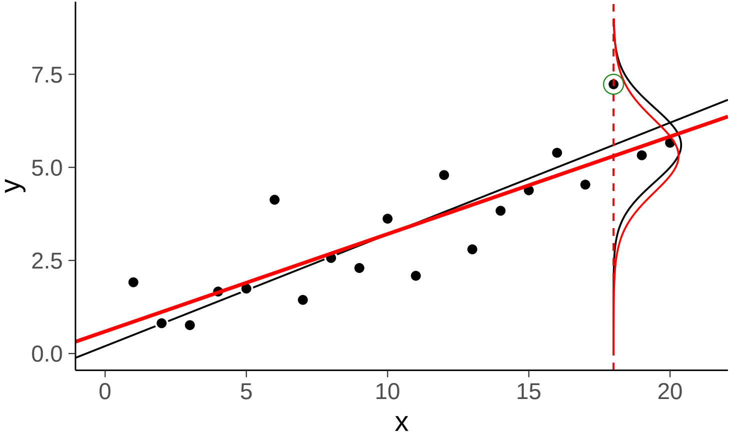



$$p(\tilde{y}|\tilde{x} = 18, x, y) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x, y)d\theta$$

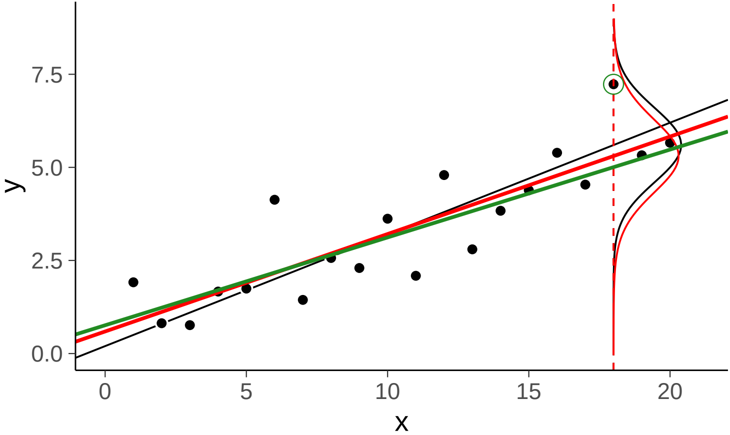
# New data



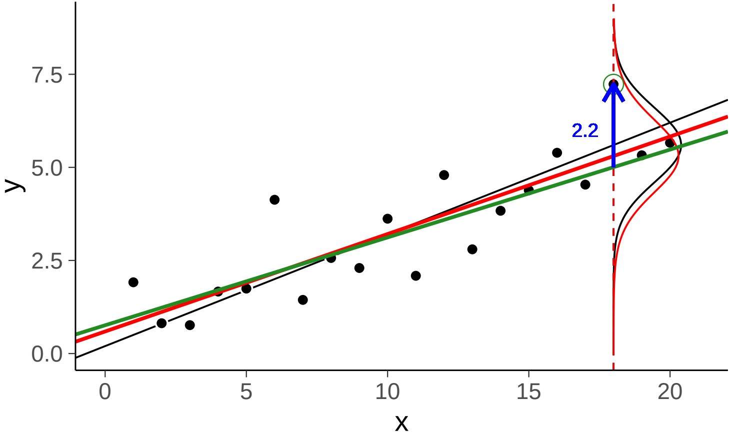
# Posterior predictive distribution

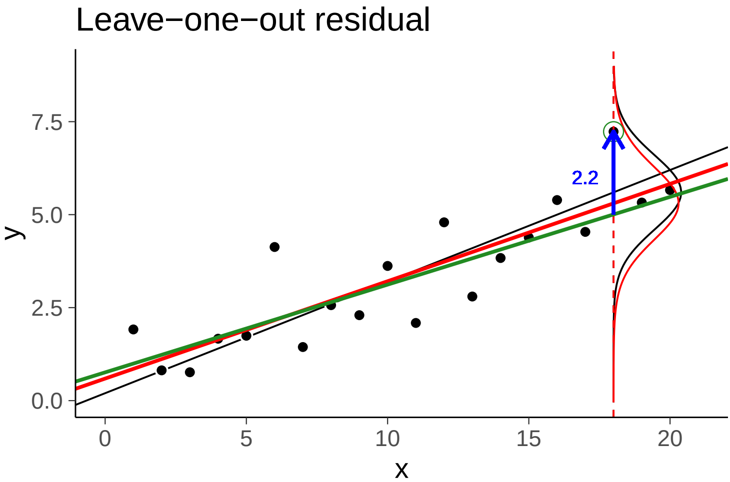


# Leave−one−out mean

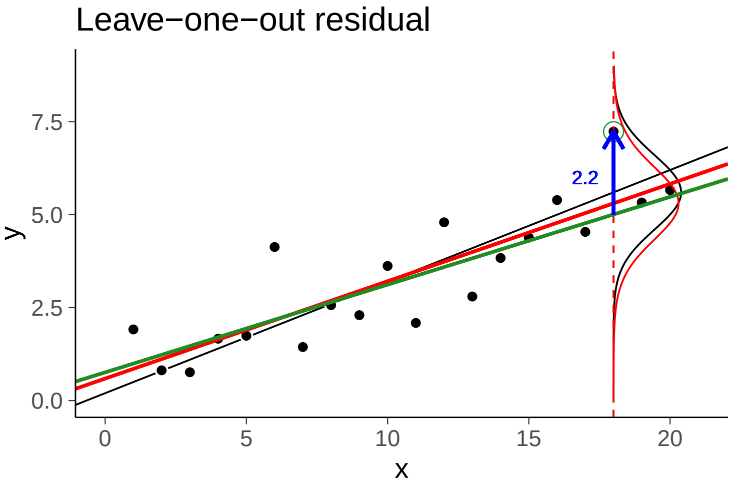


# Leave−one−out residual



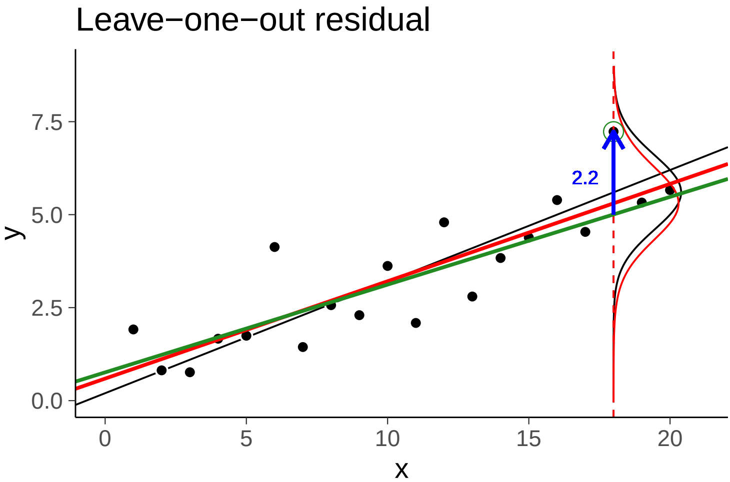



$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$




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

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

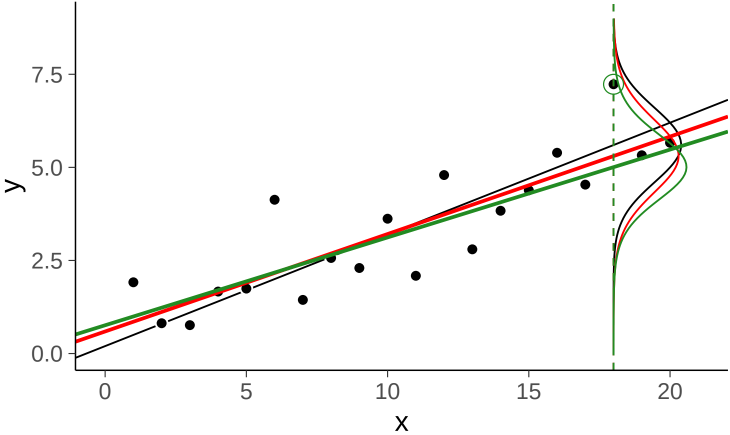



$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

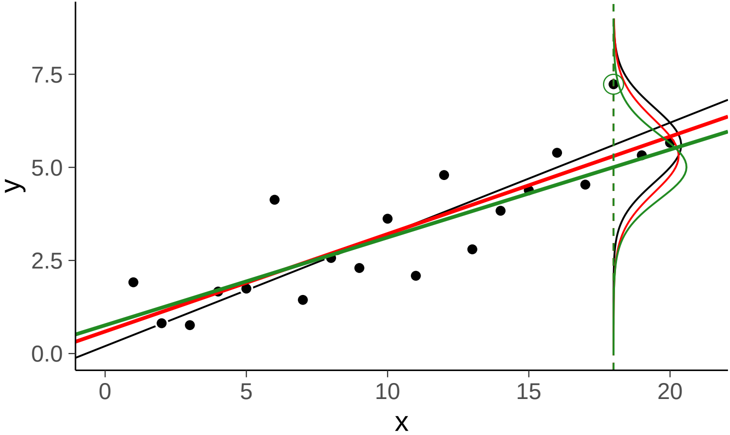
Can be use to compute, e.g., RMSE, $R^2$, 90% error

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

# Leave−one−out predictive distribution

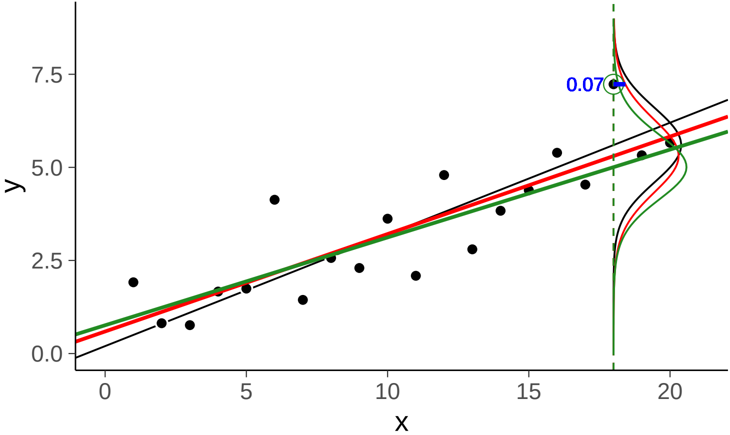


Leave−one−out predictive distribution

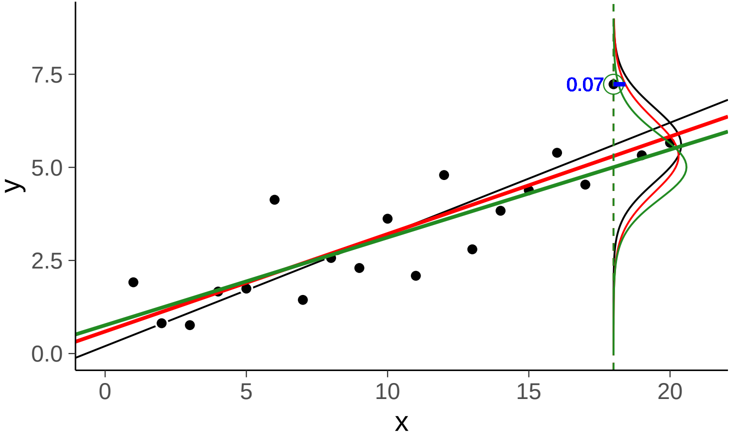



$$p(\tilde{y}|\tilde{x}=18, x_{-18}, y_{-18}) = \int p(\tilde{y}|\tilde{x}=18,\theta)p(\theta|x_{-18}, y_{-18})d\theta$$

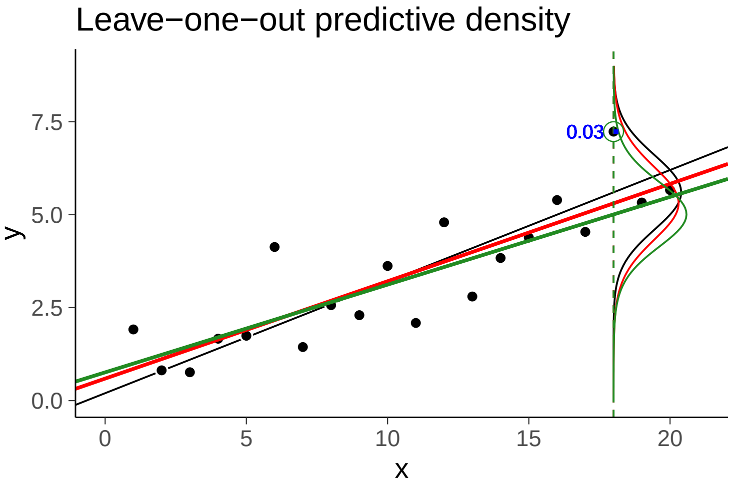
# Posterior predictive density



# Posterior predictive density

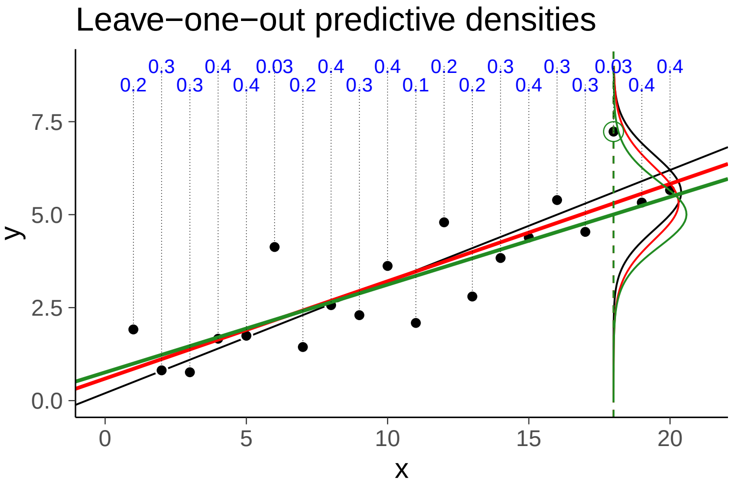



$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$



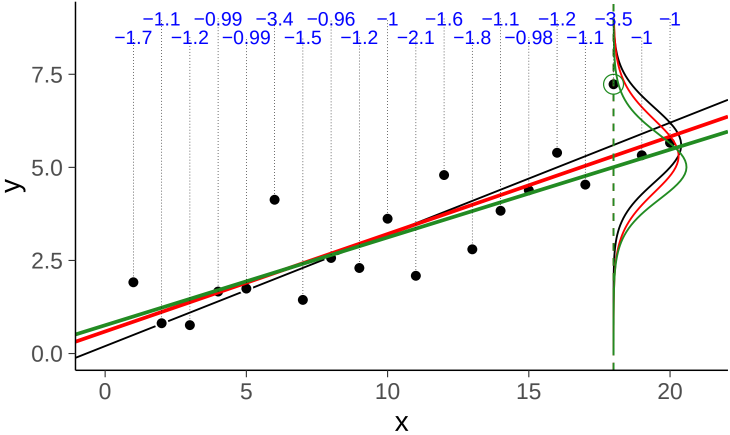

$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$

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



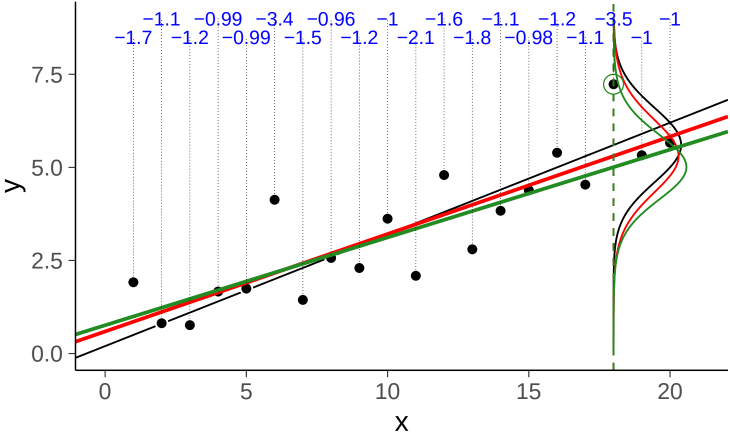

$p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

# Leave−one−out log predictive densities

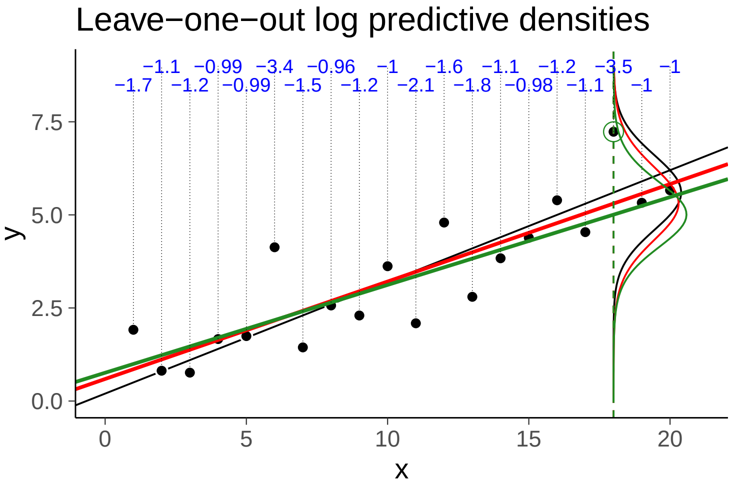



$\log p(y_i | x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

# Leave−one−out log predictive densities

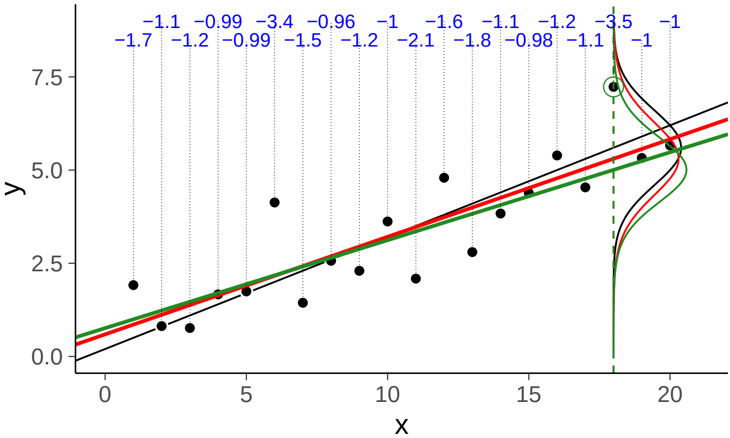



$\sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$




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

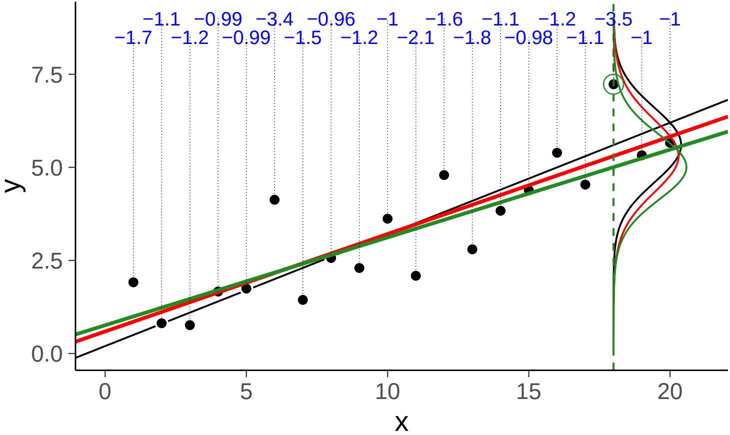
## Leave−one−out log predictive densities




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

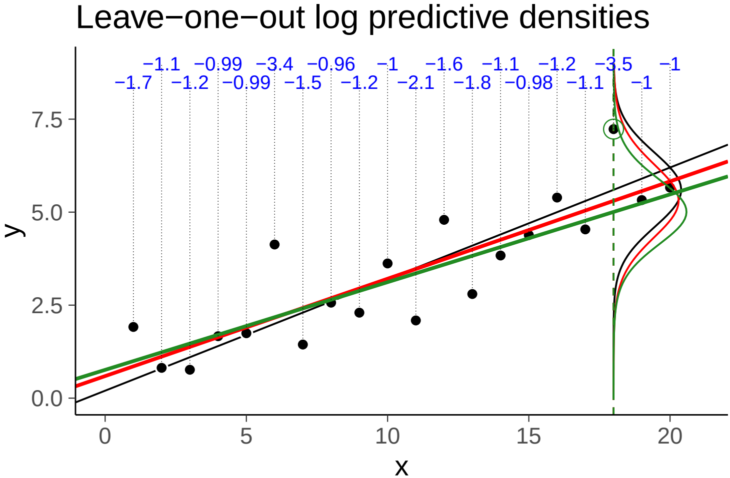
unbiased estimate of log posterior pred. density for new data

# Leave−one−out log predictive densities




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

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

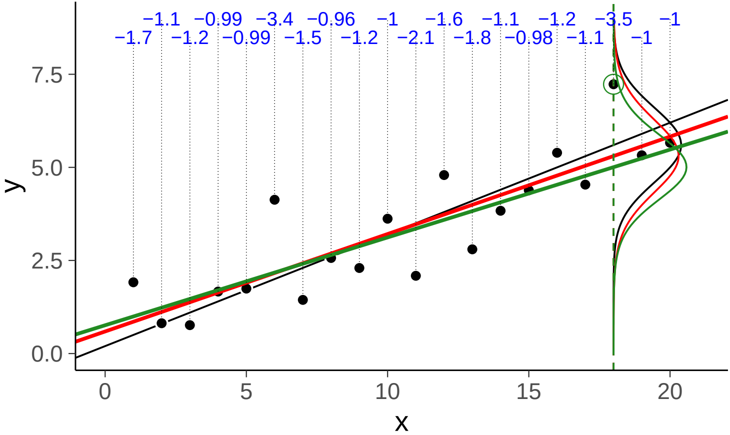



$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

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

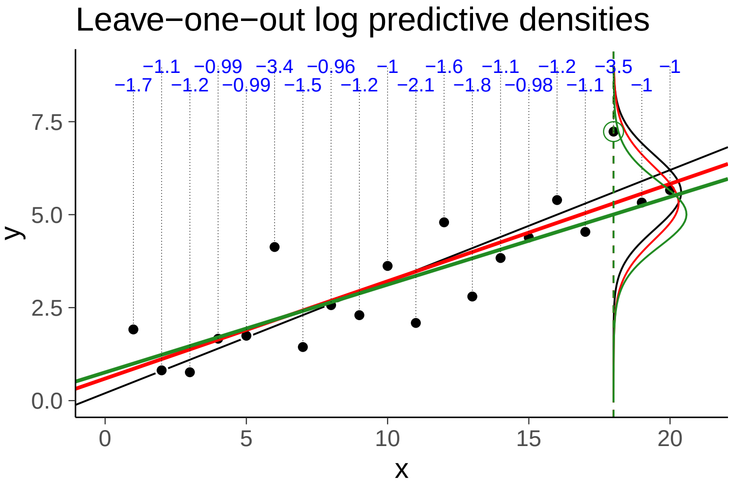
$\text{p\_loo} = \text{lpd} - \text{elpd\_loo} \approx 2.7$

# Leave−one−out log predictive densities




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

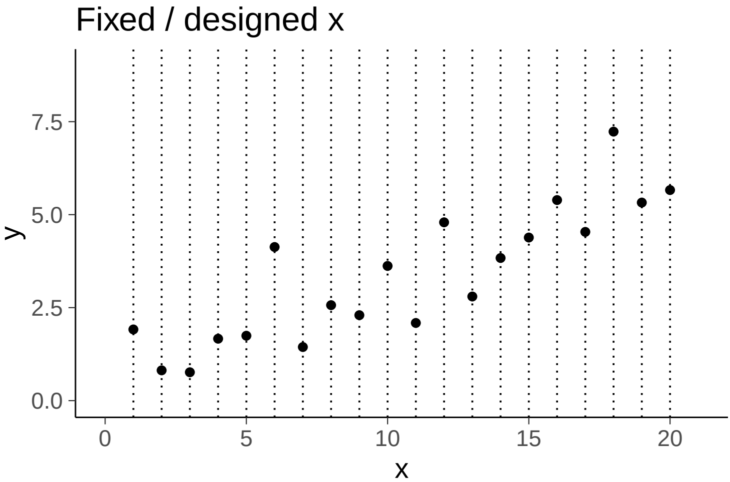
$$\text{SE} = \text{sd}(\log p(y_i|x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$$




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

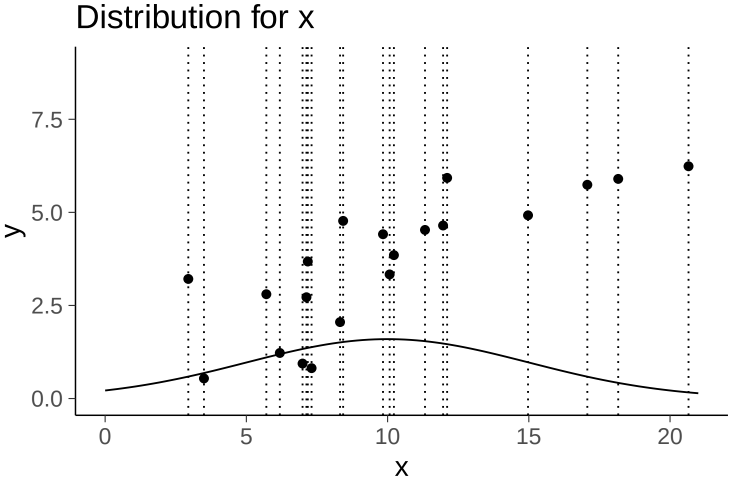
$\text{SE} = \text{sd}(\log p(y_i|x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$

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




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

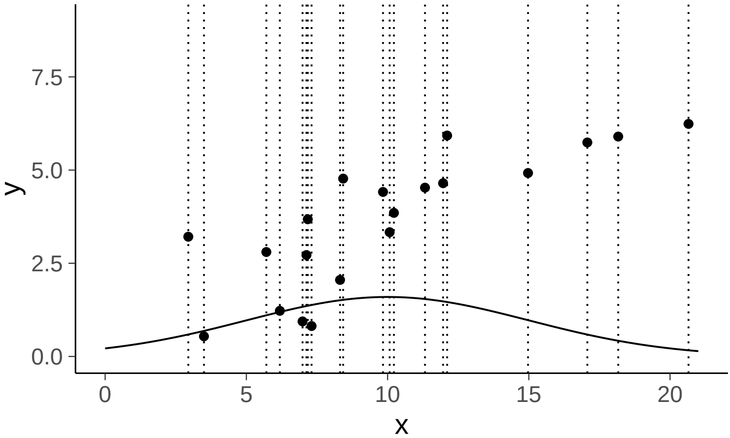
see Vehtari & Ojanen (2012) and andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/




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

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

# Distribution for x

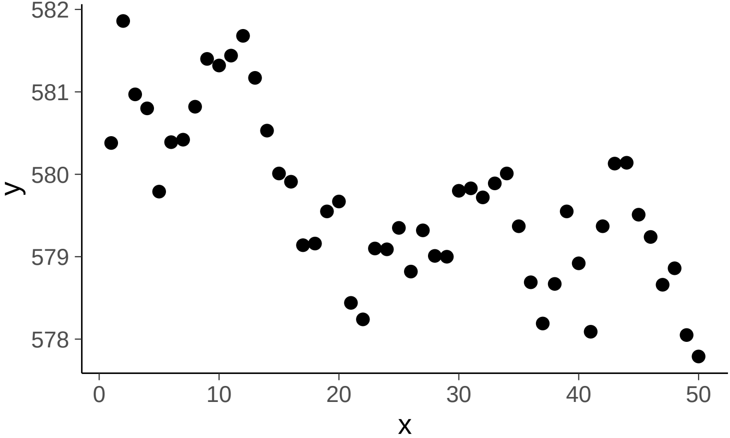



LOO is ok for random $x$. SE is uncertainty about $y|x$ and $x$.

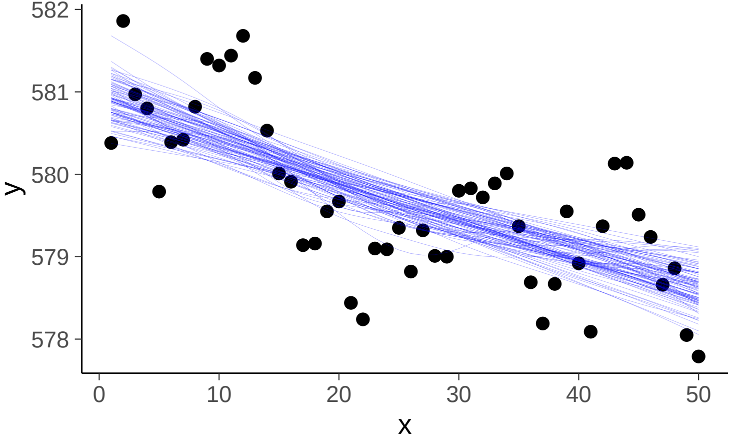
Covariate shift can be handled with importance weighting or modelling
see Vehtari & Ojanen (2012) and andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/

# loo package

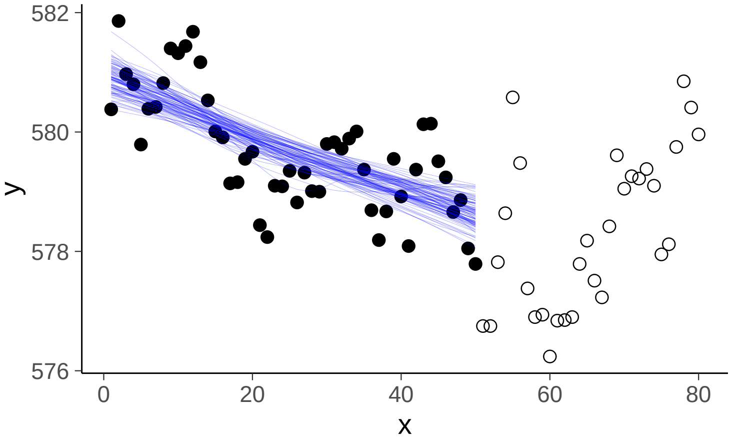
```
Computed from 4000 by 20 log-likelihood matrix

         Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0

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

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

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


582
581
580
579
578
y
0
10
20
30
40
50
x

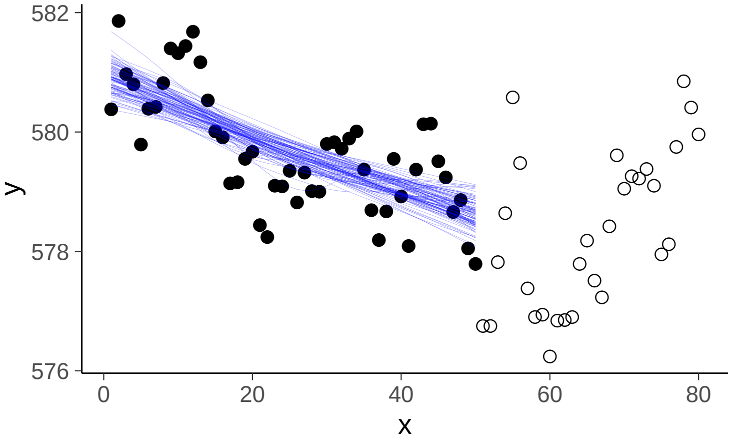
# Nonlinear model fit



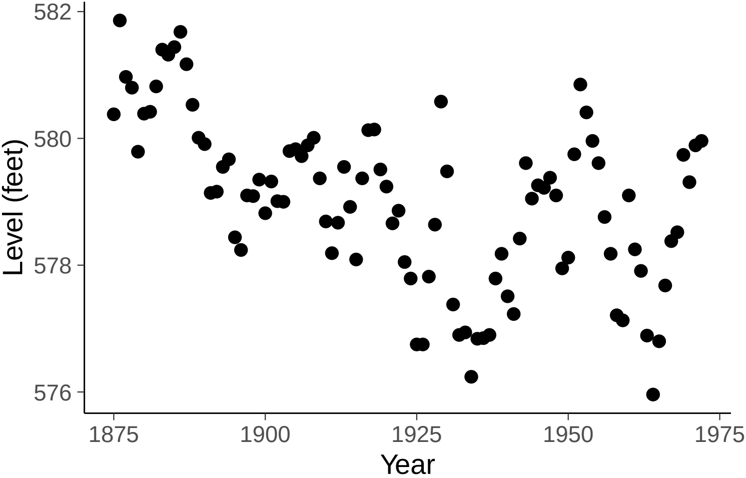
# Nonlinear model fit + new data



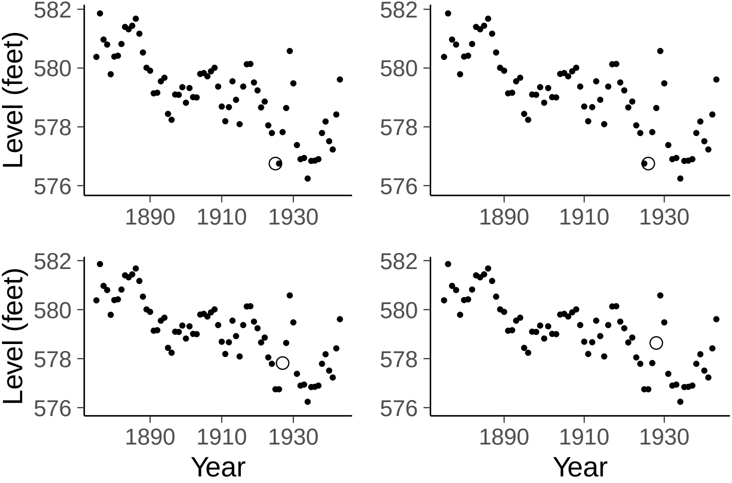
# Nonlinear model fit + new data



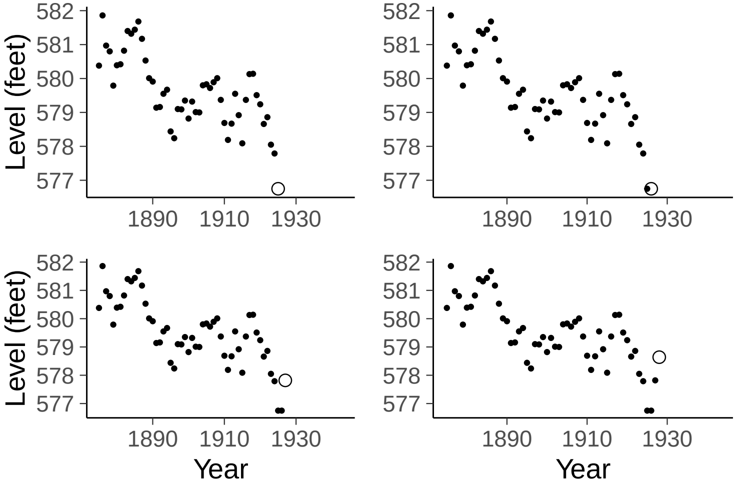

Extrapolation is more difficult



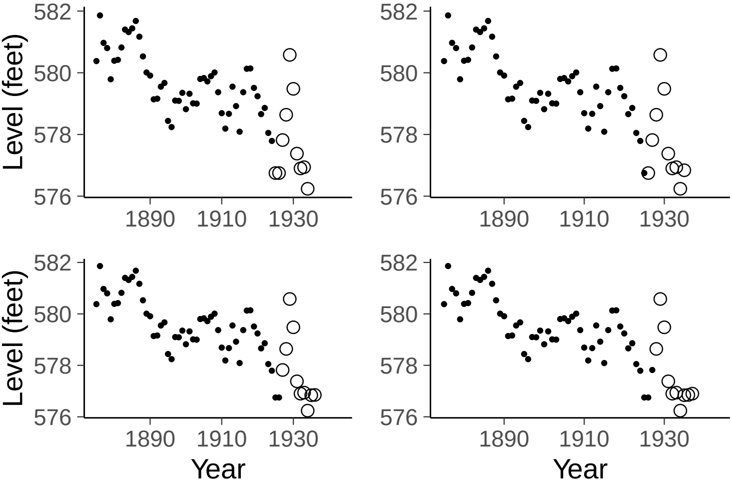

Can LOO or other cross-validation be used with time series?



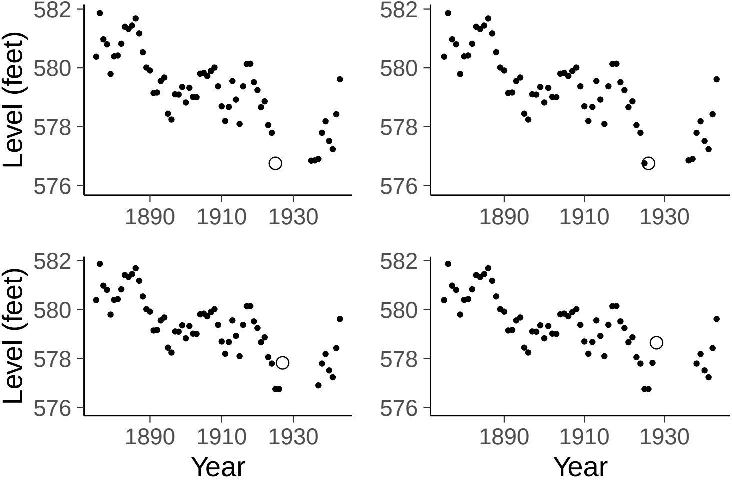

Leave-one-out cross-validation is ok for assessing conditional model



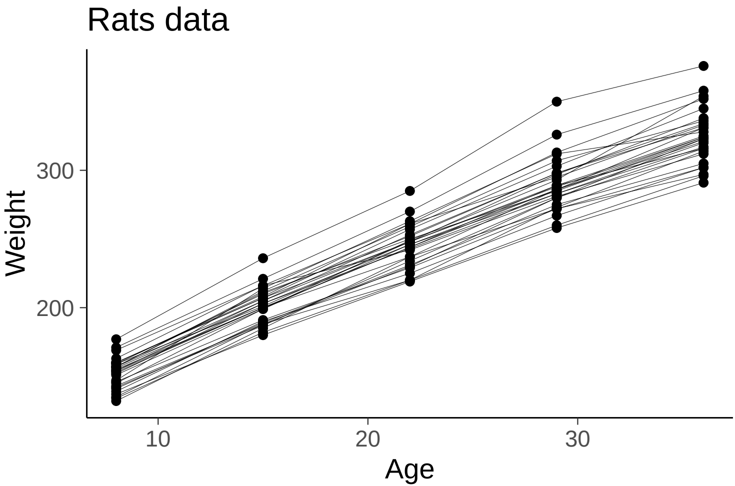

Leave-future-out cross-validation is better for predicting future




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

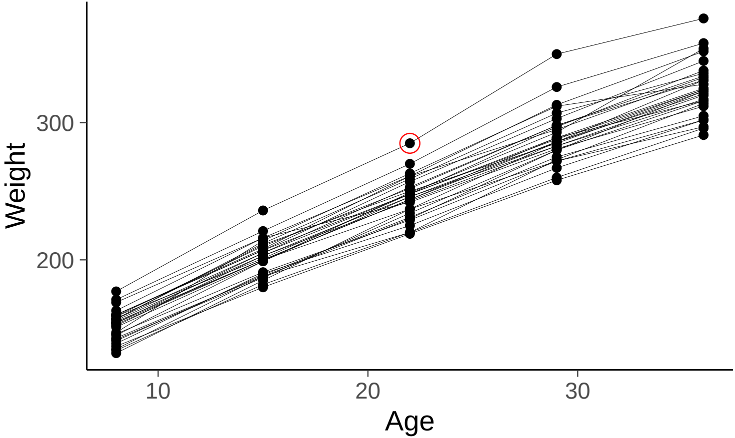



*m*-step-ahead leave-a-block-out cross-validation



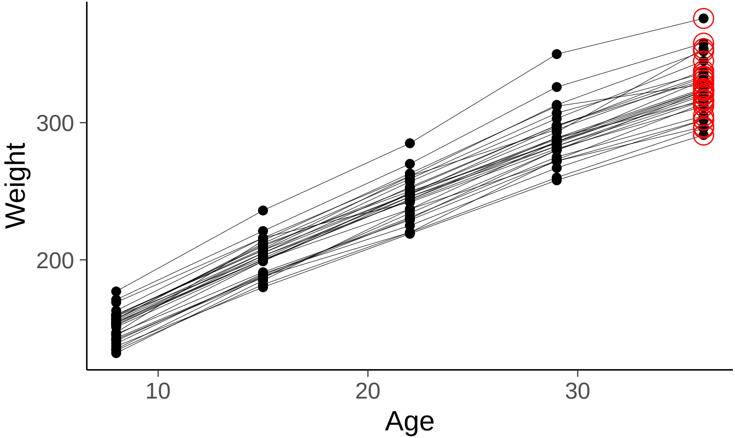

Can LOO or other cross-validation be used with hierarchical data?

# Leave−one−out?



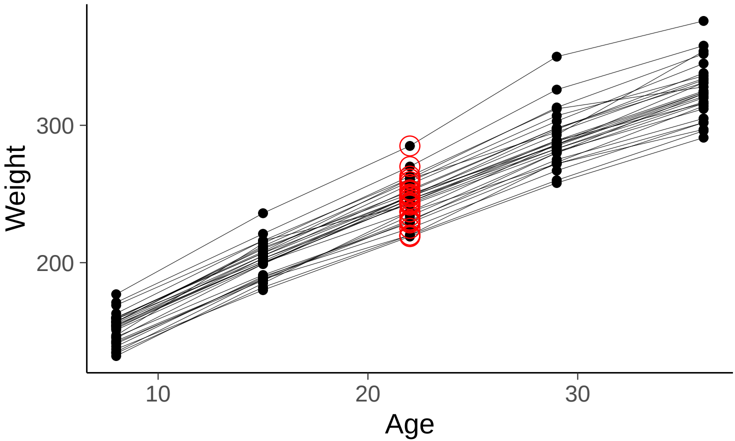

Yes!

## 1−step−ahead?

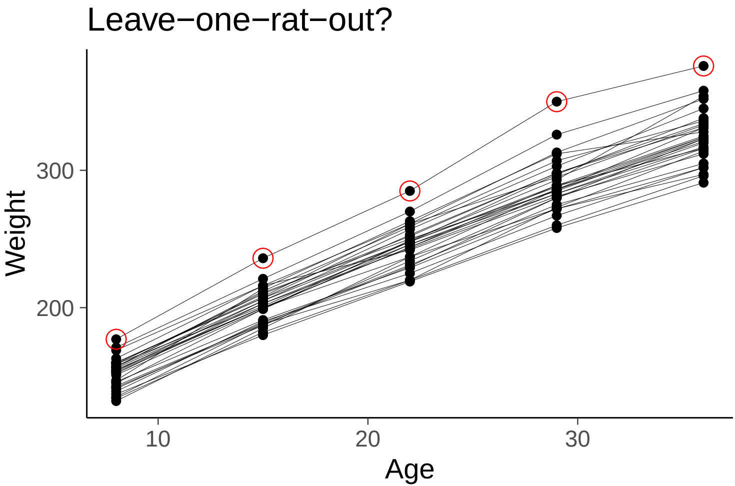



Yes!

## Leave−one−time−point−out?

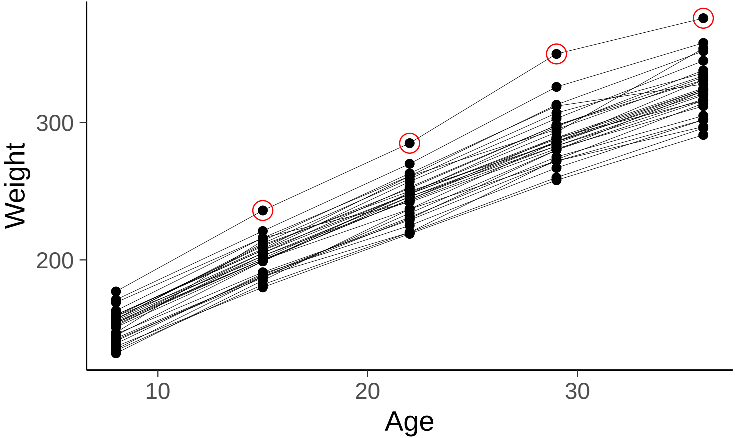



Yes!




Yes!

# Predict given initial weight?




Yes!

# Summary of data generating mechanisms and prediction tasks

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

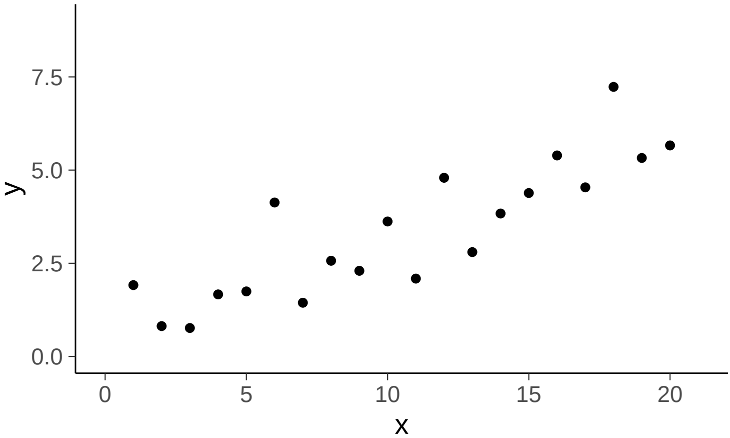
see Vehtari & Ojanen (2012) and andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/

# Fast cross-validation

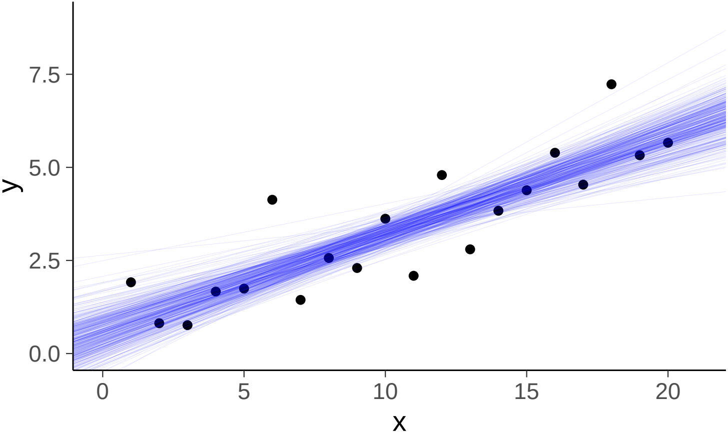
- Pareto smoothed importance sampling LOO (PSIS-LOO)
- K-fold cross-validation

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

# Data

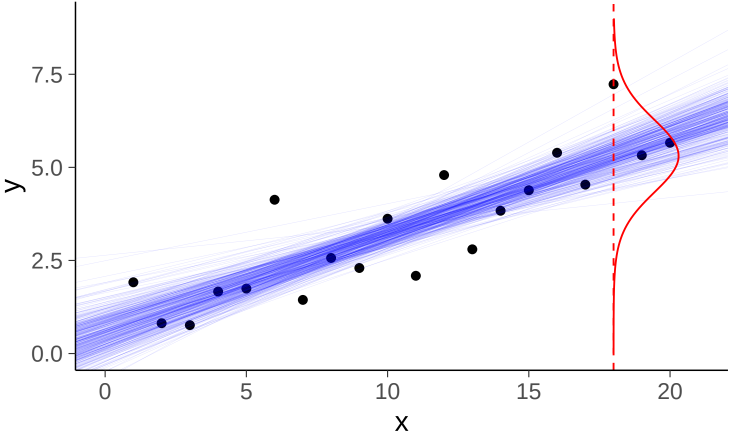


# Posterior draws



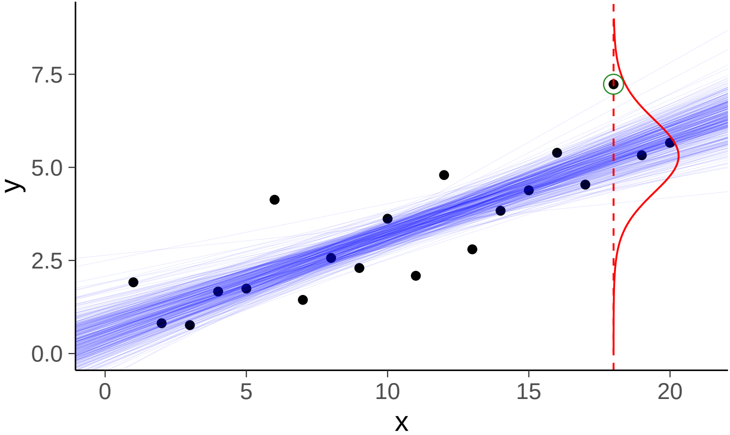

$\theta^{(s)} \sim p(\theta|x, y)$

# Posterior predictive distribution

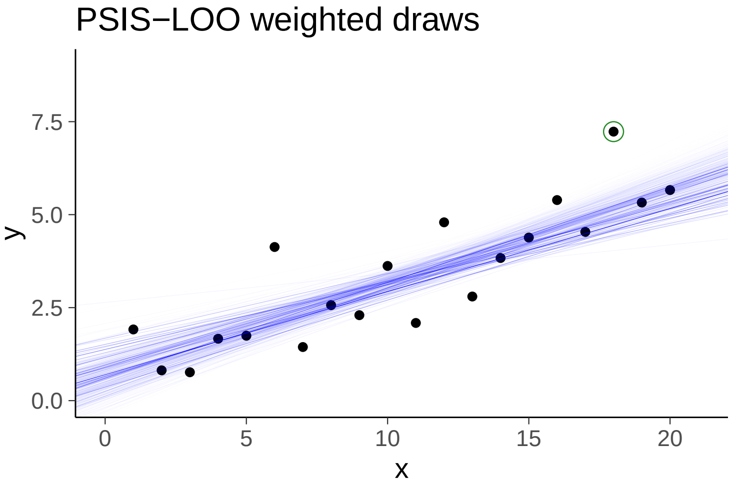



$$\theta^{(s)} \sim p(\theta|x,y), \quad p(\tilde{y}|\tilde{x},x,y) \approx \frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}|\tilde{x},\theta^{(s)})$$

# Posterior predictive distribution

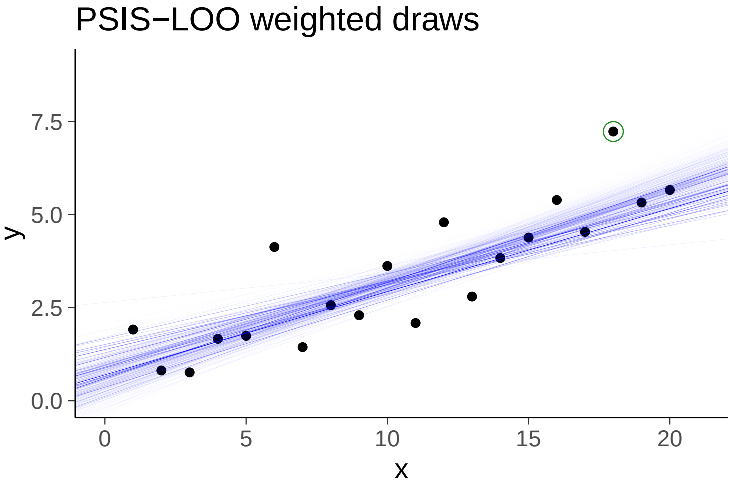



$$\theta^{(s)} \sim p(\theta|x,y), \quad p(\tilde{y}|\tilde{x},x,y) \approx \frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}|\tilde{x},\theta^{(s)})$$



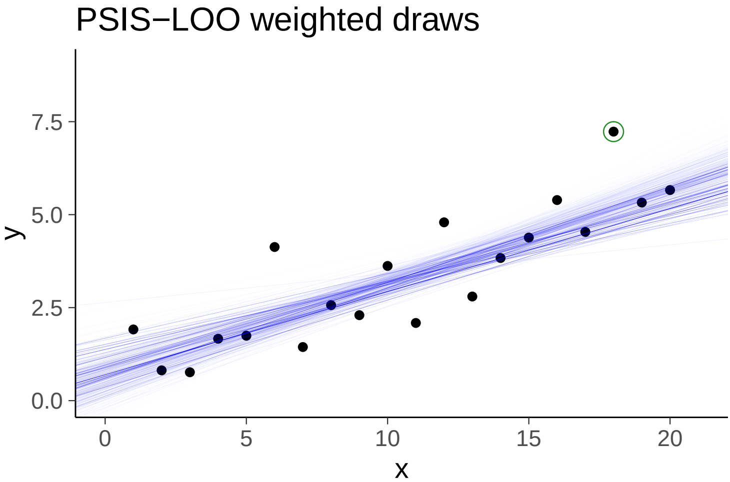

$\theta^{(s)} \sim p(\theta|x, y)$

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




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

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

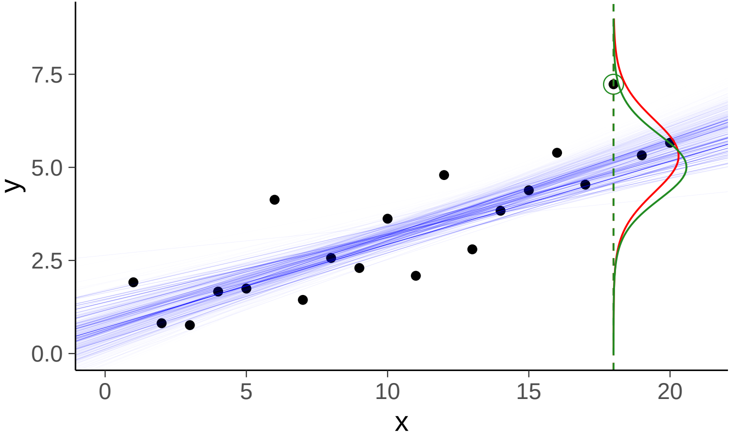



$\theta^{(s)} \sim p(\theta|x, y)$

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

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

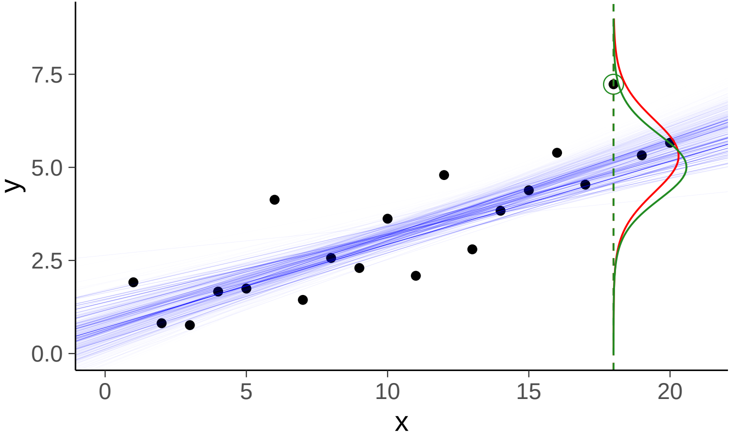
## PSIS−LOO weighted predictive distribution




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

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

# PSIS−LOO weighted predictive distribution

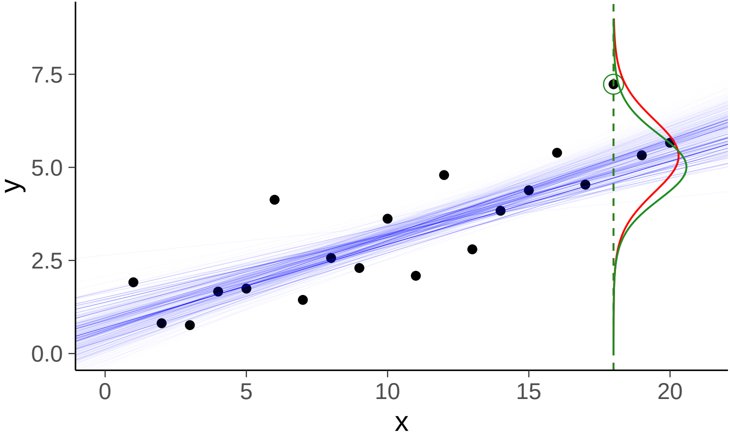



$\theta^{(s)} \sim p(\theta|x, y)$

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

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

# PSIS−LOO weighted predictive distribution

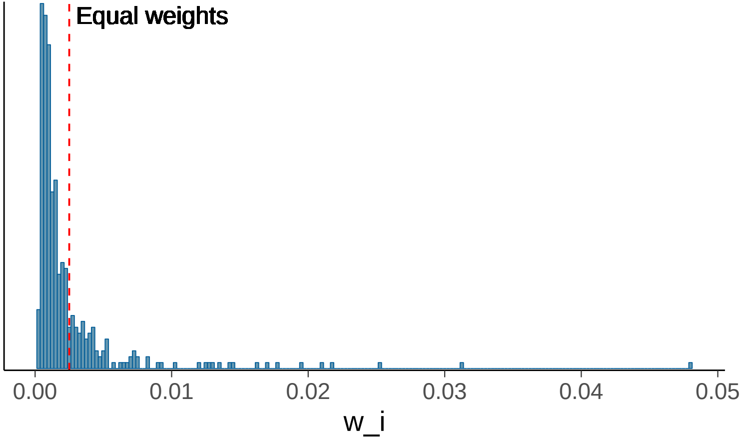



$\theta^{(s)} \sim p(\theta|x, y)$

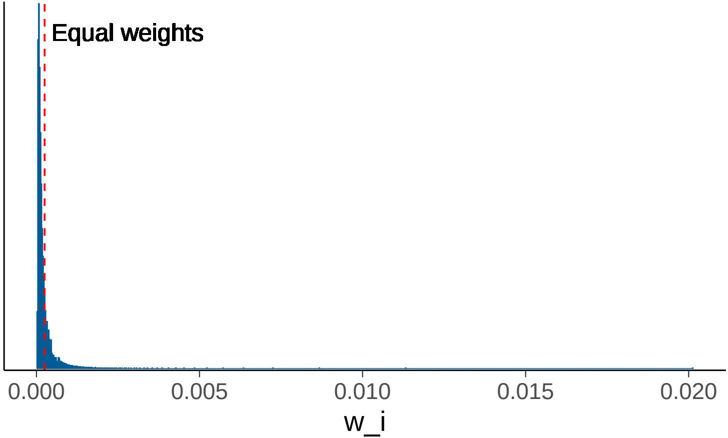
$r_i^{(s)} = p(\theta^{(s)}|x_{-i}, y_{-i})/p(\theta^{(s)}|x, y) \propto 1/p(y_i|x_i, \theta^{(s)})$

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

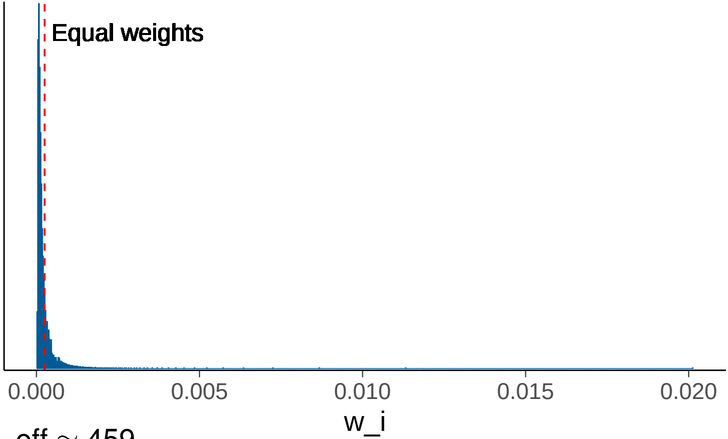
# 400 importance weights for leave−18th−out



# 4000 importance weights for leave−18th−out



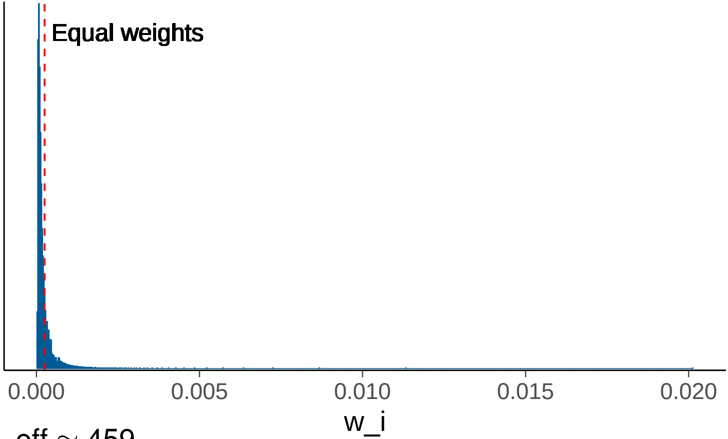
# 4000 importance weights for leave−18th−out




n_eff ≈ 459

see Vehtari, Gelman & Gabry (2017b)

## 4000 importance weights for leave−18th−out

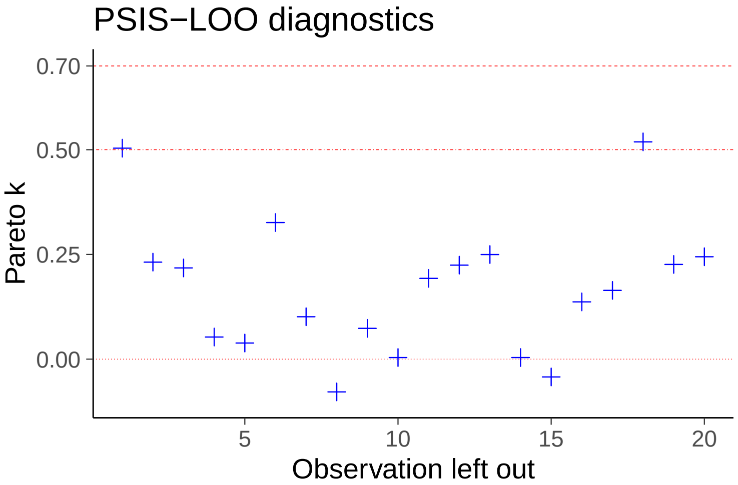


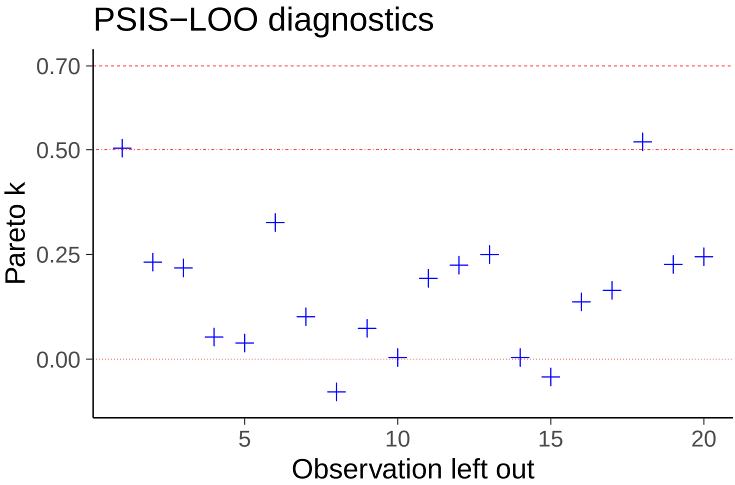

n_eff $\approx$ 459

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.

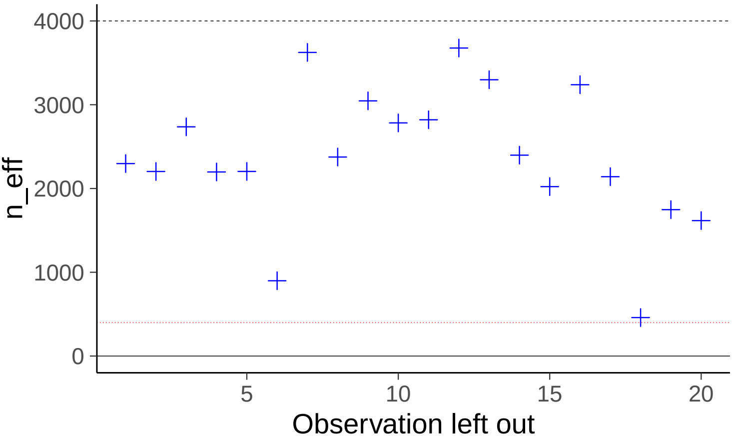
see Vehtari, Gelman & Gabry (2017b)


PSIS−LOO diagnostics
0.70
0.50
0.25
0.00
Pareto k
5
10
15
20
Observation left out




```
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>
```

## 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>
```

# `loo` package

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

         Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0

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

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

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

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

# Stan code

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

# Stan code

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

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

# Stan code

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

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

- RStanARM and BRMS compute log_lik by default

# Pareto smoothed importance sampling LOO

- PSIS-LOO for hierarchical models
  - leave-one-group out is challenging for PSIS-LOO

    see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration

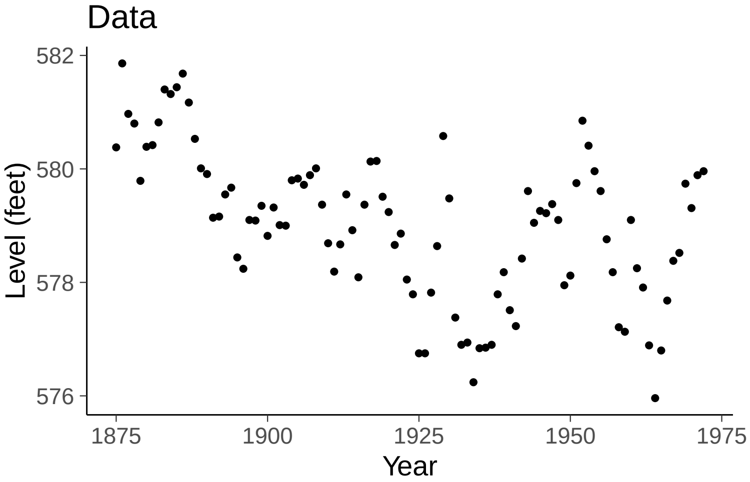
# Pareto smoothed importance sampling LOO

- PSIS-LOO for hierarchical models
    - leave-one-group out is challenging for PSIS-LOO

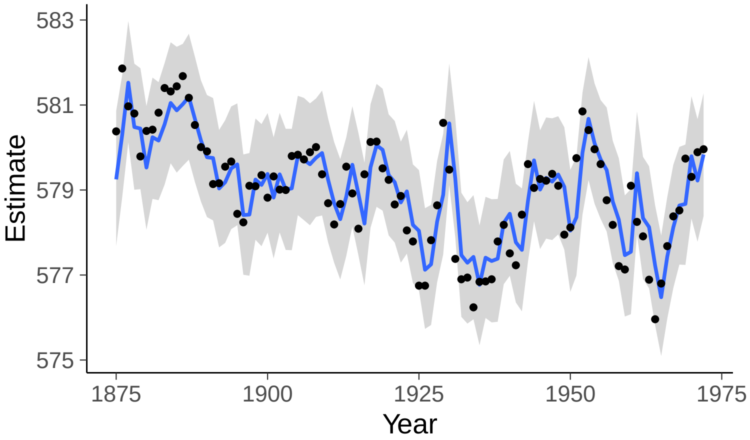
      see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration
- PSIS-LOO for non-factorizable models
    - mc-stan.org/loo/articles/loo2-non-factorizable.html

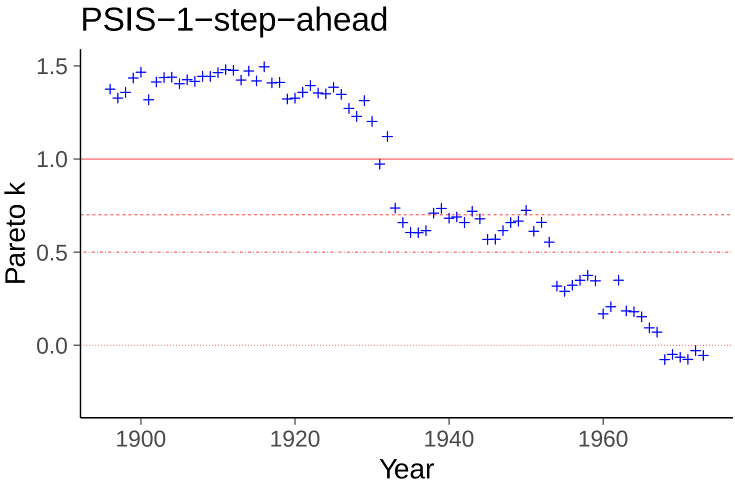
# Pareto smoothed importance sampling LOO

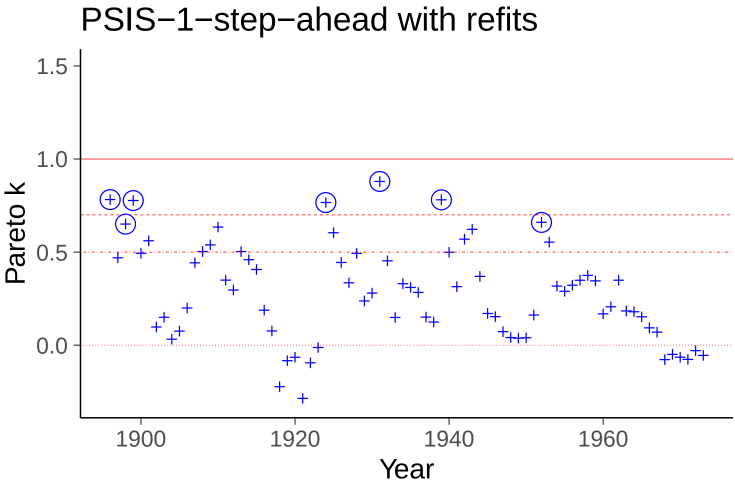
- PSIS-LOO for hierarchical models
  - leave-one-group out is challenging for PSIS-LOO
    see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration
- PSIS-LOO for non-factorizable models
  - mc-stan.org/loo/articles/loo2-non-factorizable.html
- PSIS-LOO for time series
 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


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

# AR−2 prediction with 95% interval




PSIS−1−step−ahead
Pareto k
1.5
1.0
0.5
0.0
1900
1920
1940
1960
Year

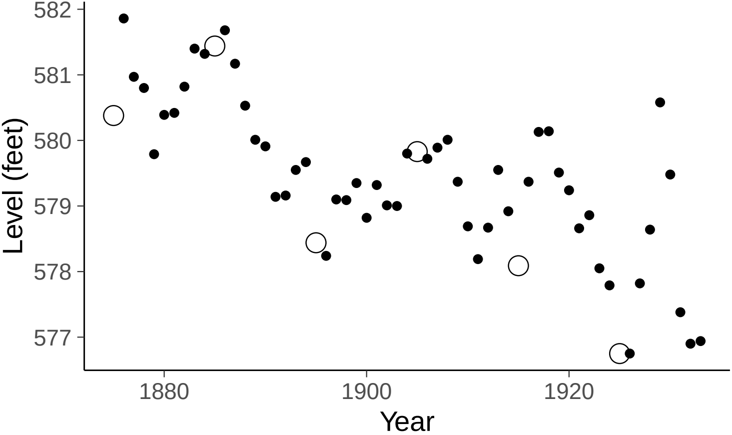



mc-stan.org/loo/articles/loo2-lfo.html

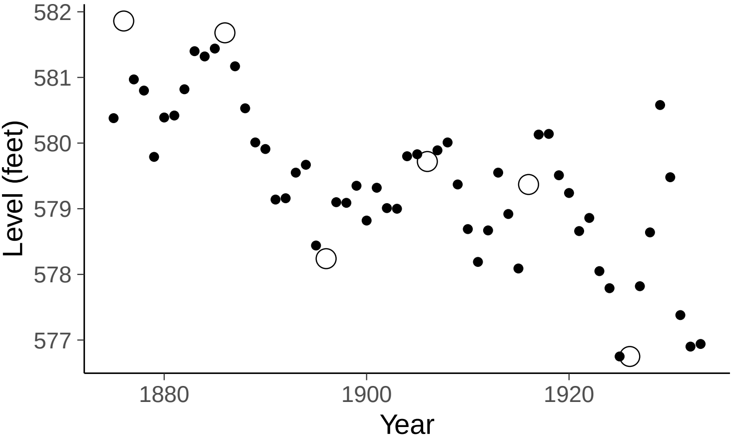
# K-fold cross-validation

- K-fold cross-validation can approximate LOO
  - all uses for LOO
- K-fold cross-validation can be used for hierarchical models
  - good for leave-one-group-out
- K-fold cross-validation can be used for time series
  - with leave-block-out

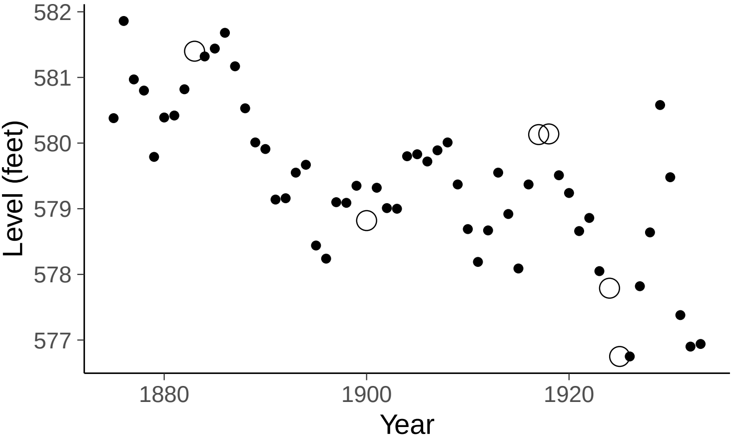
# Balance k−fold approximation of LOO



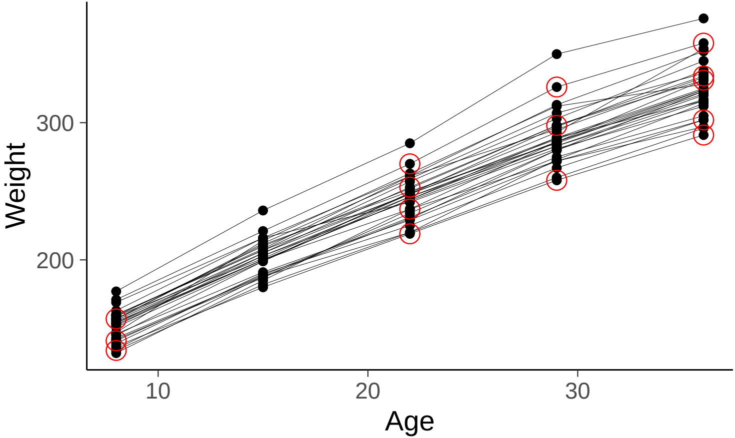
# Balance k−fold approximation of LOO

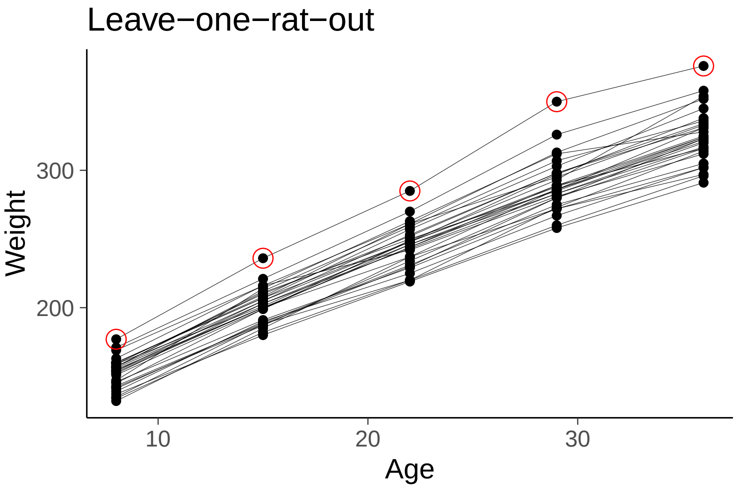


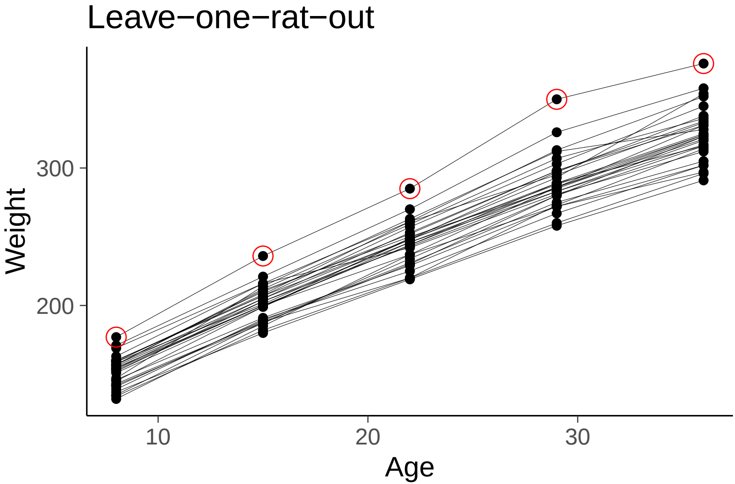
# Random k−fold approximation of LOO



# Random kfold approximation of LOO




Leave−one−rat−out
Weight
300
200
Age
10
20
30




kfold_split_random()

kfold_split_balanced()

kfold_split_stratified()

# WAIC vs PSIS-LOO

see Vehtari, Gelman & Gabry (2017a)

# WAIC vs PSIS-LOO

- WAIC has same assumptions as LOO

see Vehtari, Gelman & Gabry (2017a)

# WAIC vs PSIS-LOO

- WAIC has same assumptions as LOO
- PSIS-LOO is more accurate

see Vehtari, Gelman & Gabry (2017a)

# WAIC vs PSIS-LOO

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

see Vehtari, Gelman & Gabry (2017a)

# WAIC vs PSIS-LOO

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

see Vehtari, Gelman & Gabry (2017a)

# WAIC vs PSIS-LOO

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

see Vehtari, Gelman & Gabry (2017a)

# *IC

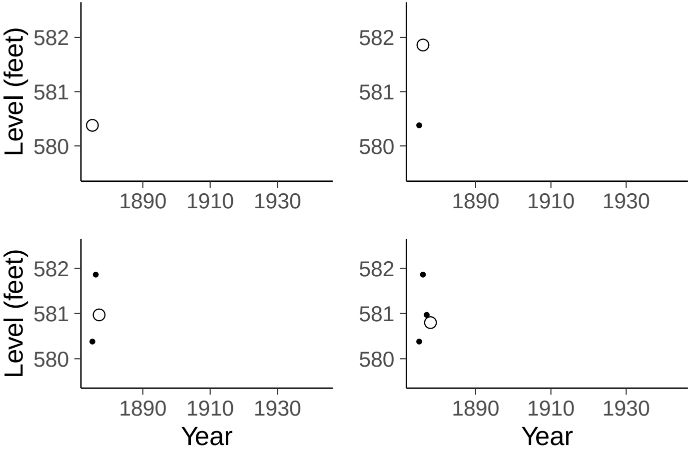
- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction
- BIC is an approximation for marginal likelihood
- TIC, NIC, RIC, PIC, BPIC, QIC, AICc, ...

# Marginal likelihood / Bayes factor

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

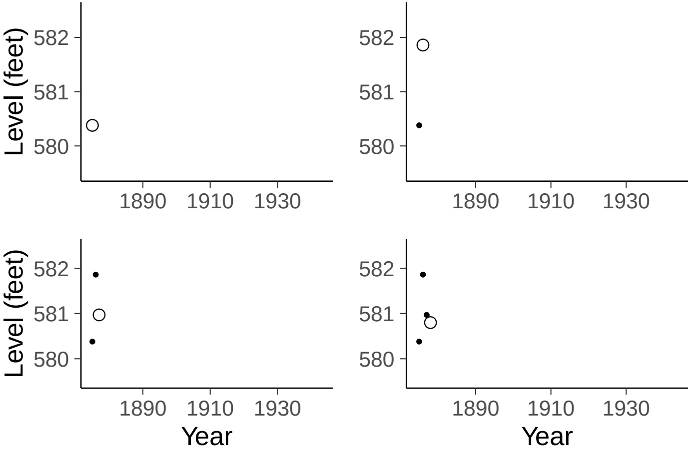
# Marginal likelihood / Bayes factor

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



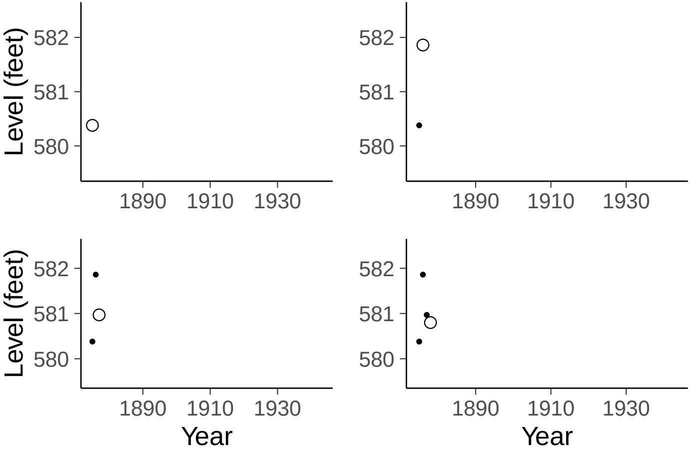
# Marginal likelihood / Bayes factor

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



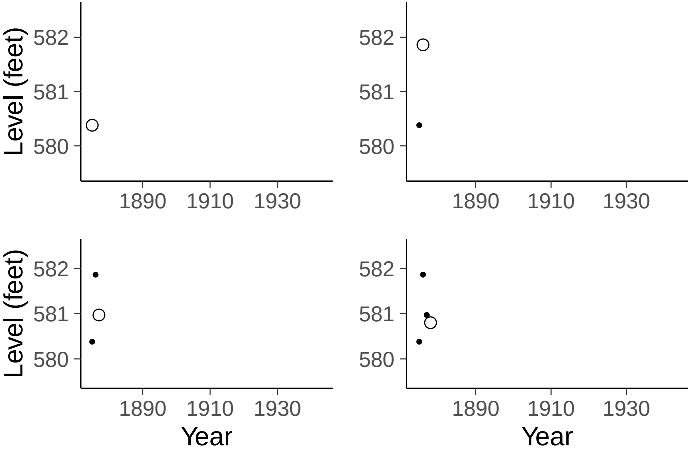
# Marginal likelihood / Bayes factor

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



# Marginal likelihood / Bayes factor

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



# Cross-validation for model assessment

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

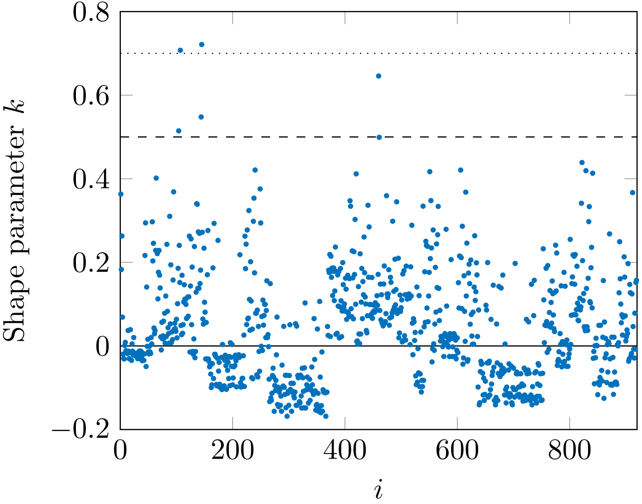
# Cross-validation for model assessment

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

see demos avehtari.github.io/modelselection/

# Radon example

PSIS-LOO diagnostics

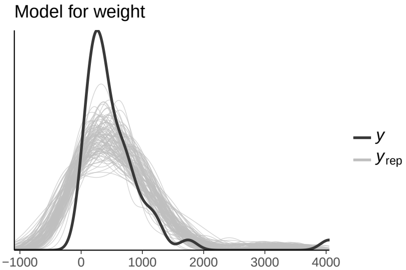


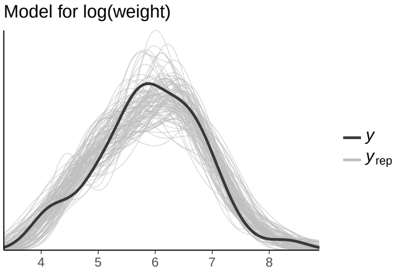

see Vehtari, Gelman & Gabry (2017a)

# Sometimes cross-validation is not needed

# Sometimes cross-validation is not needed

- Posterior predictive checking is often sufficient



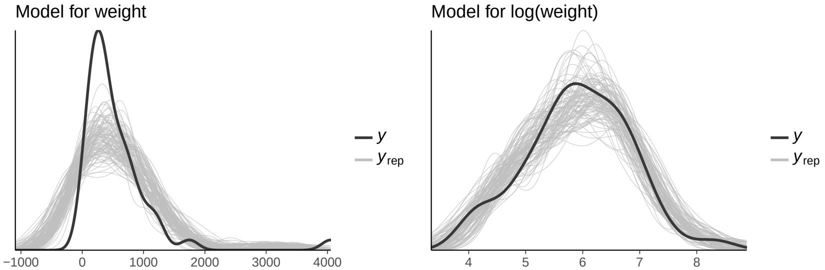




Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2019): Regression and Other Stories, Chapter 11.

# Sometimes cross-validation is not needed

- Posterior predictive checking is often sufficient




Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2019): Regression and Other Stories, Chapter 11.

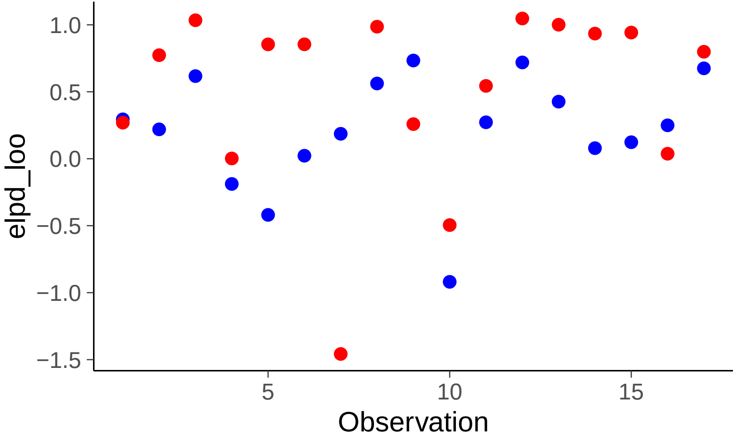
- BDA3, Chapter 6
- Gabry, Simpson, Vehtari, Betancourt, Gelman (2018). Visualization in Bayesian workflow. JRSS A, preprint arXiv:1709.01449
- mc-stan.org/bayesplot/articles/graphical-ppcs.html
- betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

# Model comparison

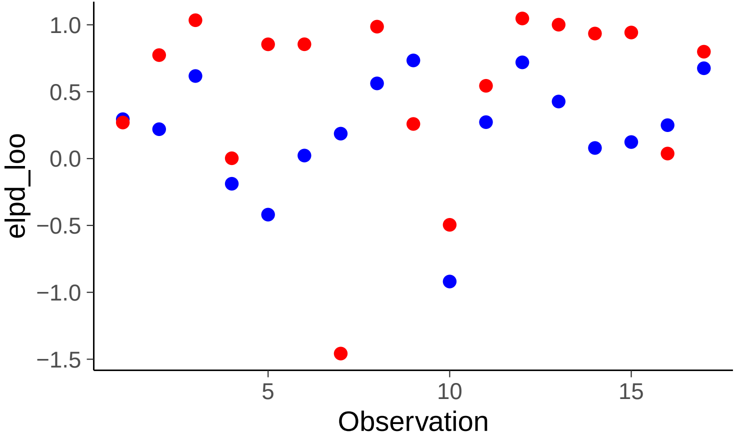
- “A popular hypothesis has it that primates with larger brains produce more energetic milk, so that brains can grow quickly” (from Statistical Rethinking)
  - Model 1: formula = kcal.per.g $\sim$ neocortex
  - Model 2: formula = kcal.per.g $\sim$ neocortex + log(mass)

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

# Pointwise comparison LOO models: Model 1

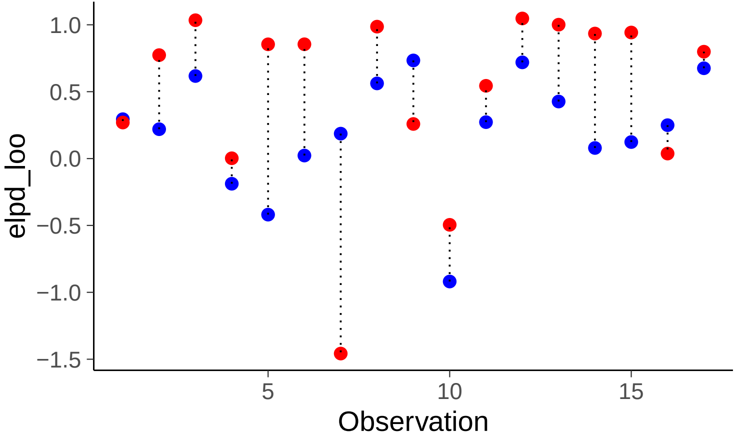


## Pointwise comparison LOO models: Model 1




Model 1 elpd_loo $\approx$ 3.7, SE=1.8
Model 2 elpd_loo $\approx$ 8.4, SE=2.8

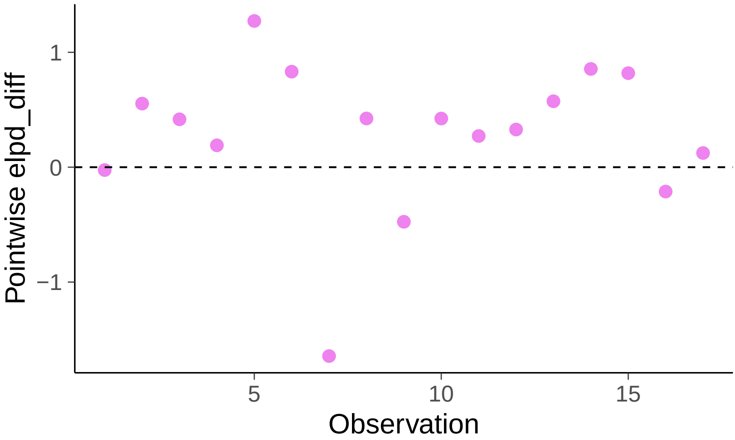
## Pointwise comparison LOO models: Model 1




Model 1 elpd_loo ≈ 3.7, SE=1.8

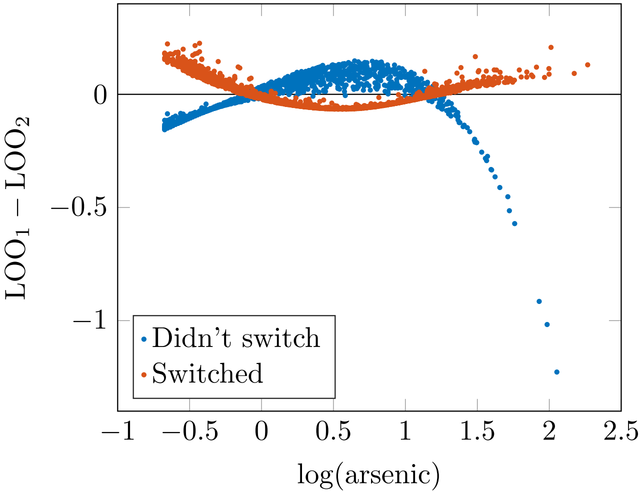
Model 2 elpd_loo ≈ 8.4, SE=2.8

# Pointwise comparison LOO models




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

elpd_diff        se
      4.7       2.7
```

# Arsenic well example – Model comparison

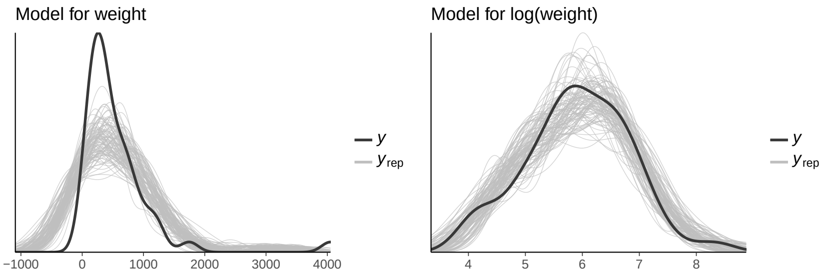



An estimated difference in $\mathrm{elpd}_{\mathrm{loo}}$ of 16.4 with SE of 4.4.

see Vehtari, Gelman & Gabry (2017a)

# Sometimes cross-validation is not needed

- Posterior predictive checking is often sufficient




Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2019): Regression and Other Stories, Chapter 11.

- BDA3, Chapter 6
- Gabry, Simpson, Vehtari, Betancourt, Gelman (2018). Visualization in Bayesian workflow. JRSS A, preprint arXiv:1709.01449
- mc-stan.org/bayesplot/articles/graphical-ppcs.html
- betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

# Sometimes cross-validation is not needed

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

# Sometimes cross-validation is not needed

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

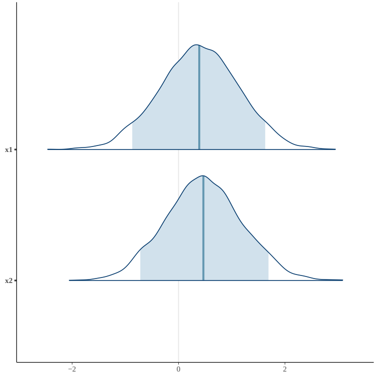
# Sometimes cross-validation is not needed

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

# Sometimes cross-validation is not needed

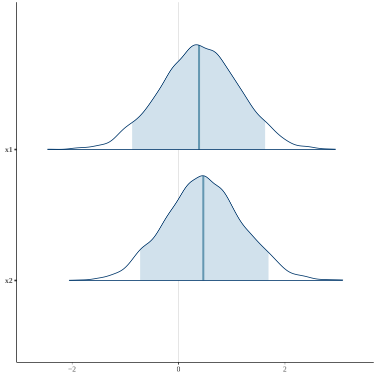
- For some very simple cases you may assume that true model is included in the list of models considered ($M$-closed)
  - see predictive model selection in $M$-closed case by San Martini and Spezzaferri (1984)
  - but you should not force your design of experiment or analysis to stay in the simplified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

# Sometimes predictive model comparison can be useful

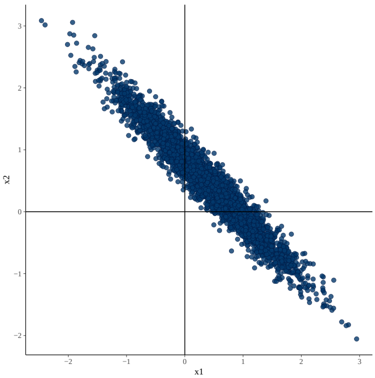



Marginal posterior intervals

# Sometimes predictive model comparison can be useful



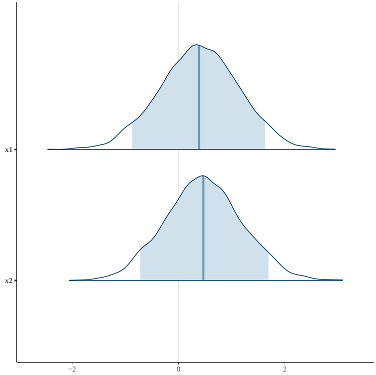

Marginal posterior intervals



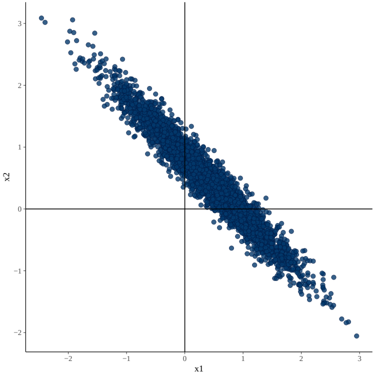

Joint posterior density

rstanarm + bayesplot

# Sometimes predictive model comparison can be useful




Marginal posterior intervals




Joint posterior density

rstanarm + bayesplot

see also Collinear demo

# What if one is not clearly better than others?

# What if one is not clearly better than others?

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

# What if one is not clearly better than others?

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

# What if one is not clearly better than others?

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

# What if one is not clearly better than others?

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

# Model averaging

- Prefer continuous model expansion

# Model averaging

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

# Model averaging

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

# Cross-validation and model selection

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

# Cross-validation and model selection

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

# Cross-validation and model selection

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

# Selection induced bias and overfitting

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

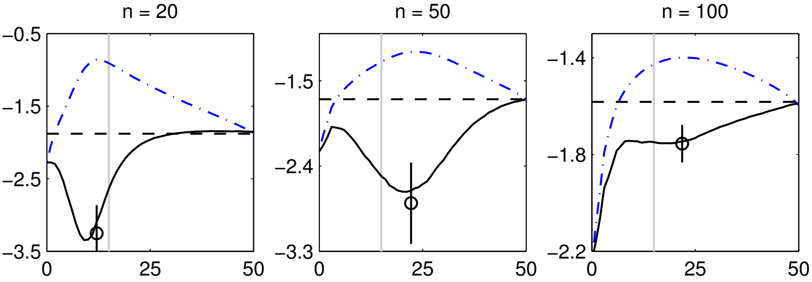
# Selection induced bias and overfitting

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

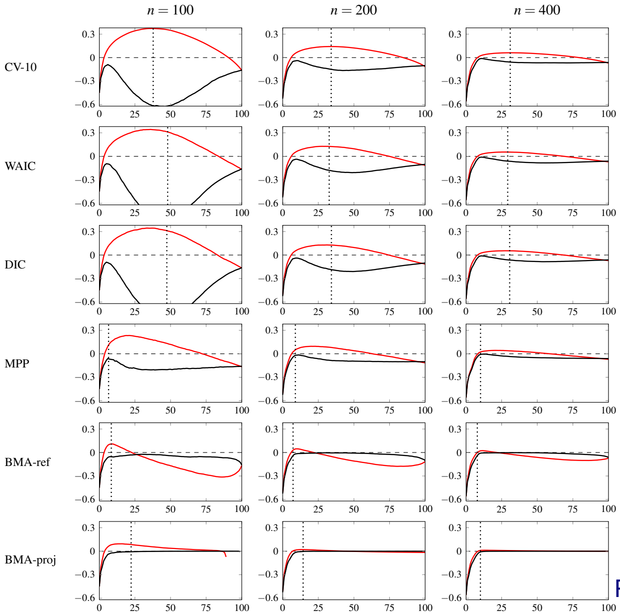
# Selection induced bias and overfitting

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

# Selection induced bias in variable selection

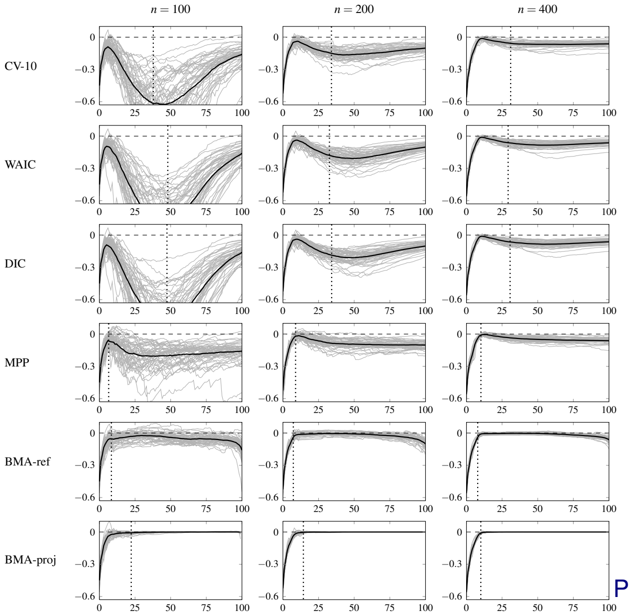


# Selection induced bias in variable selection



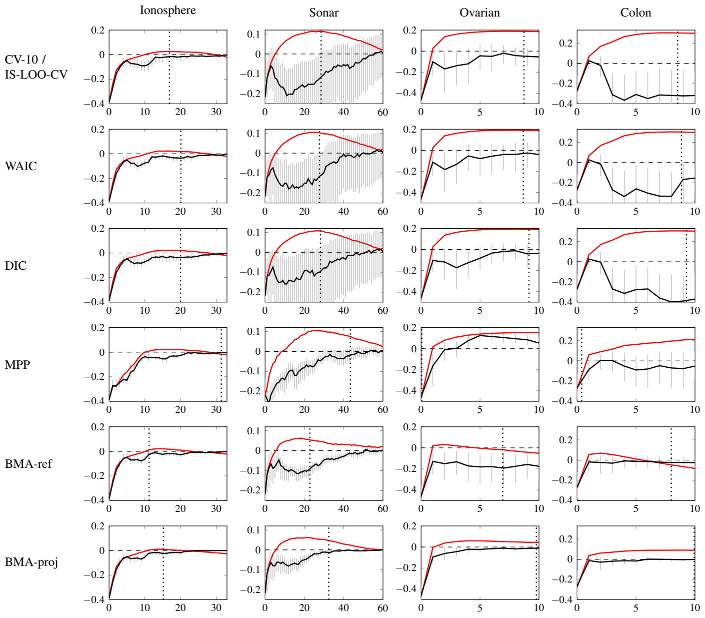

Piironen & Vehtari (2017)

# Selection induced bias in variable selection




Piironen & Vehtari (2017)

# Selection induced bias in variable selection




Piironen & Vehtari (2017)

# Take-home messages

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

# Take-home messages

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

# Take-home messages

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

# Take-home messages

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

# Take-home messages

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