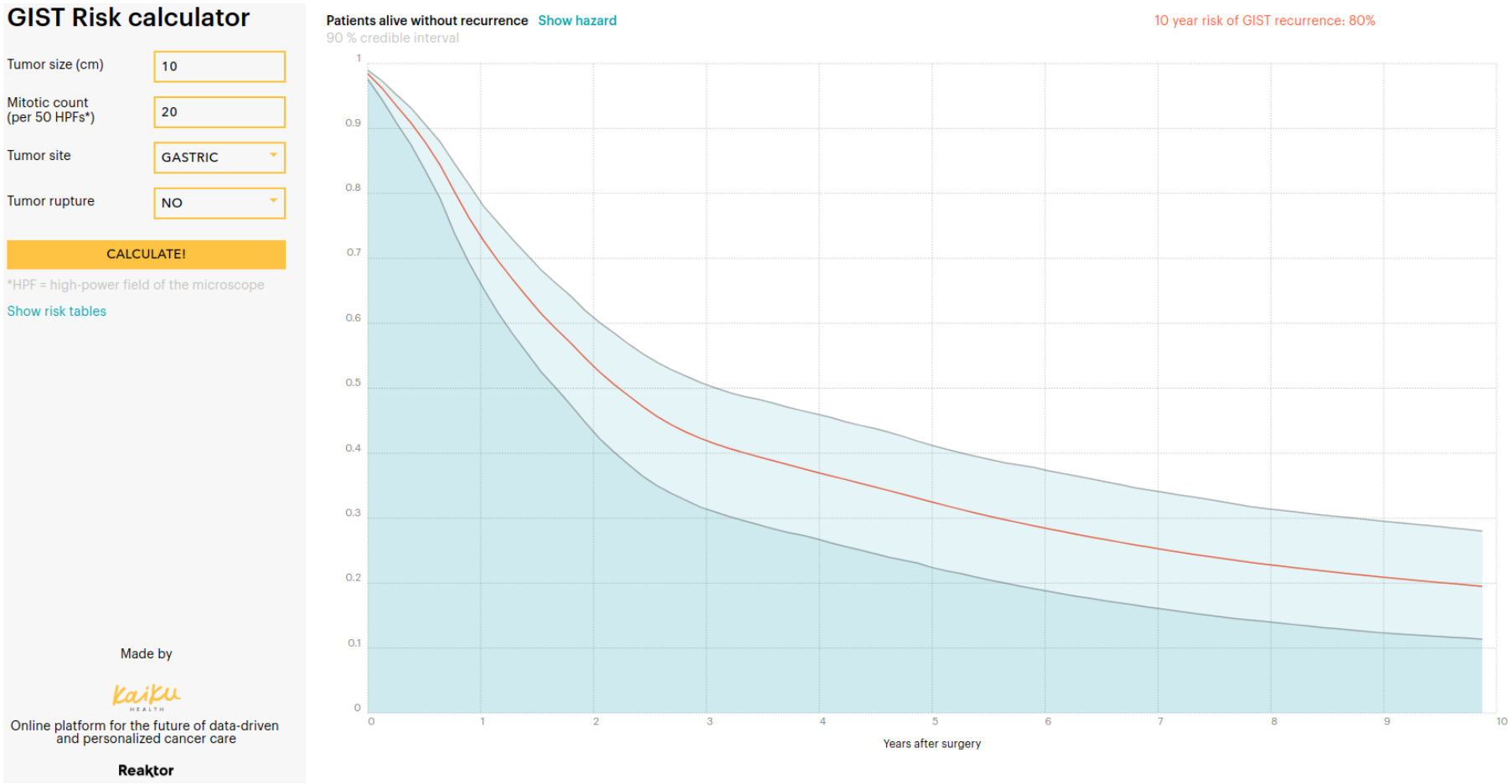
Model assessment and selection

Aki Vehtari, Aalto University

Predicting concrete quality

Predicting cancer recurrence

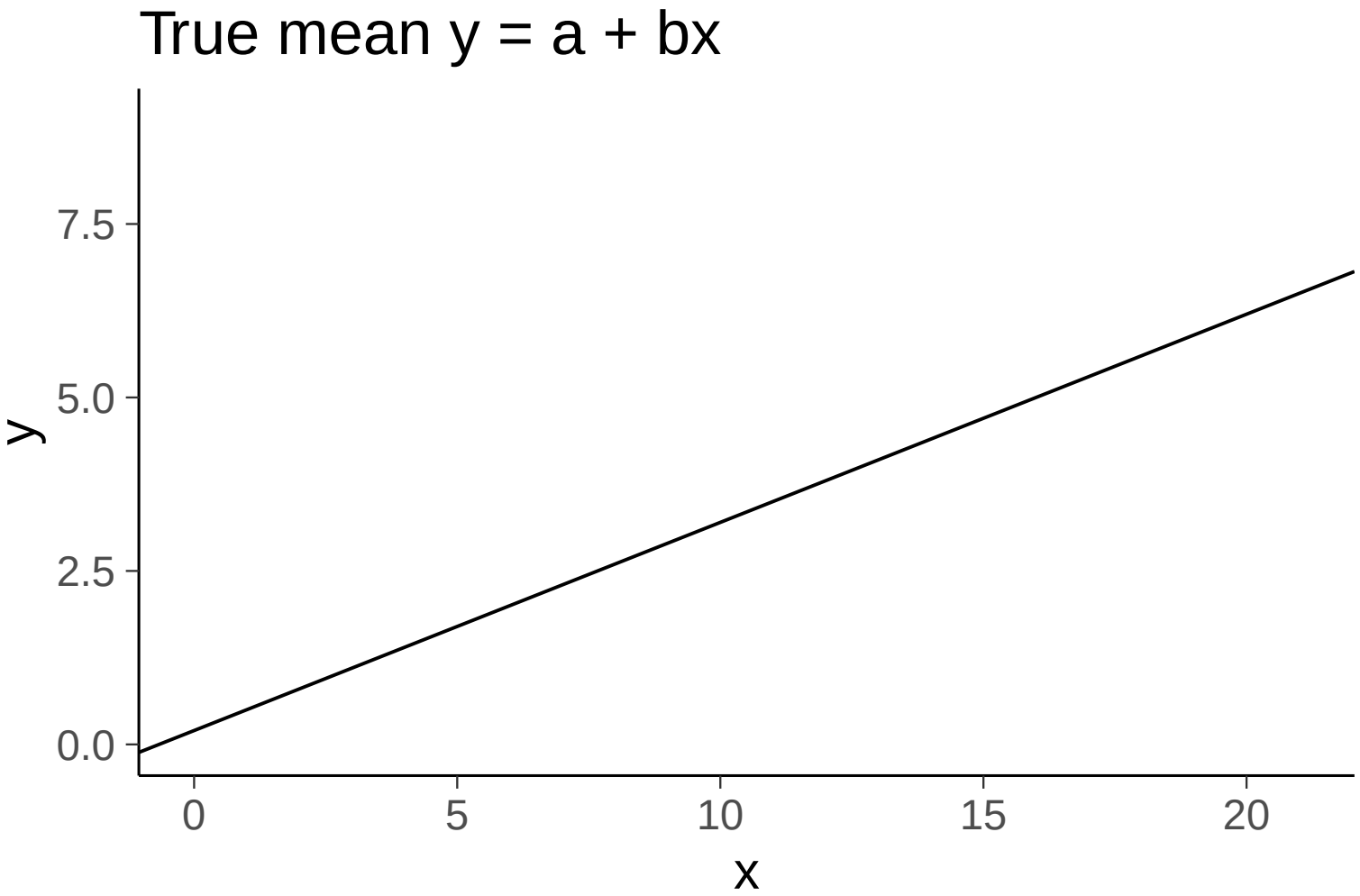


loo package

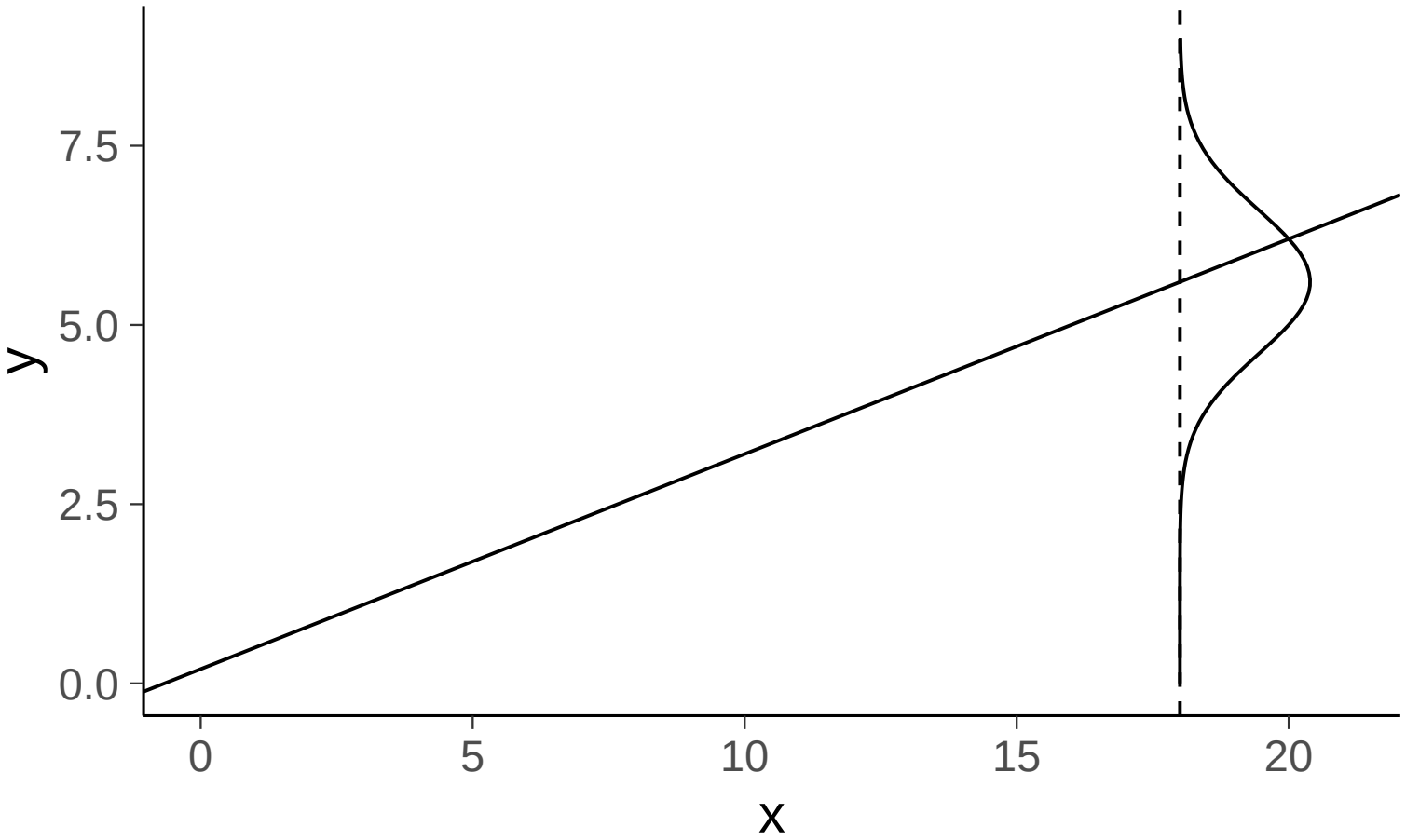
```
 Computed from 4000 by 20 log-likelihood matrix

          Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0
looic        58.9 6.7
------

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
```

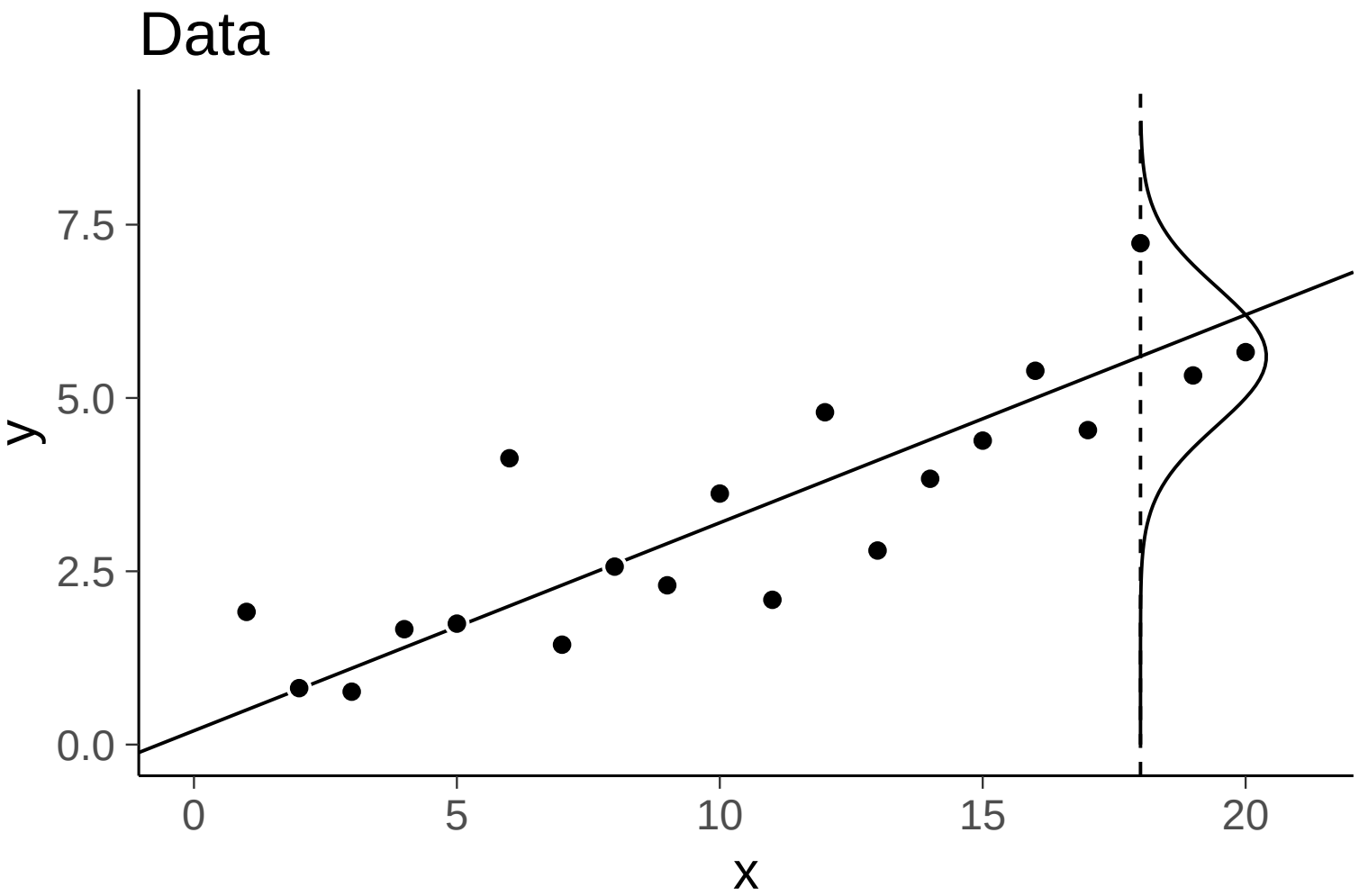
Outline

- 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 (stacking, loo weights)

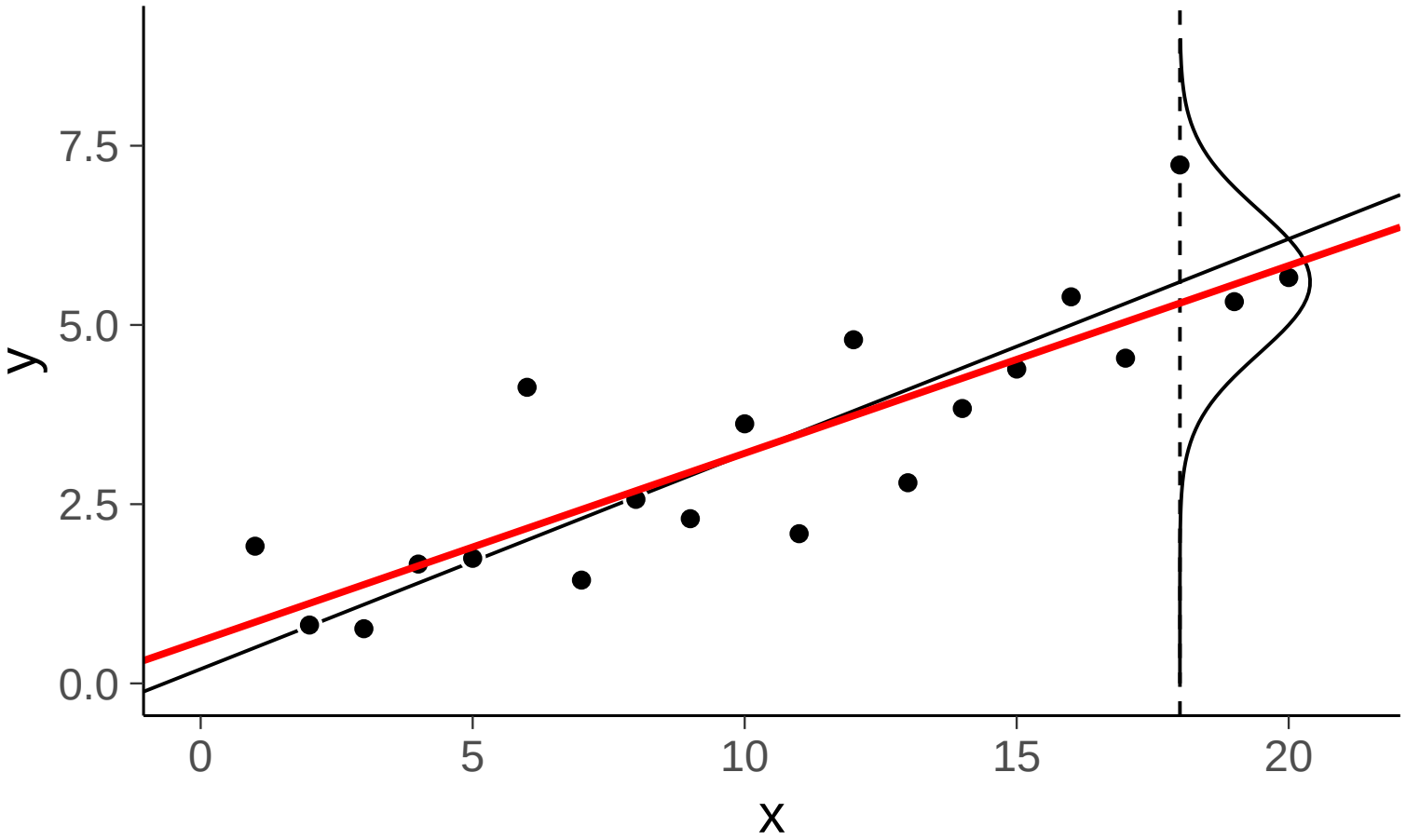

True mean y = a + bx
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

True mean and sigma

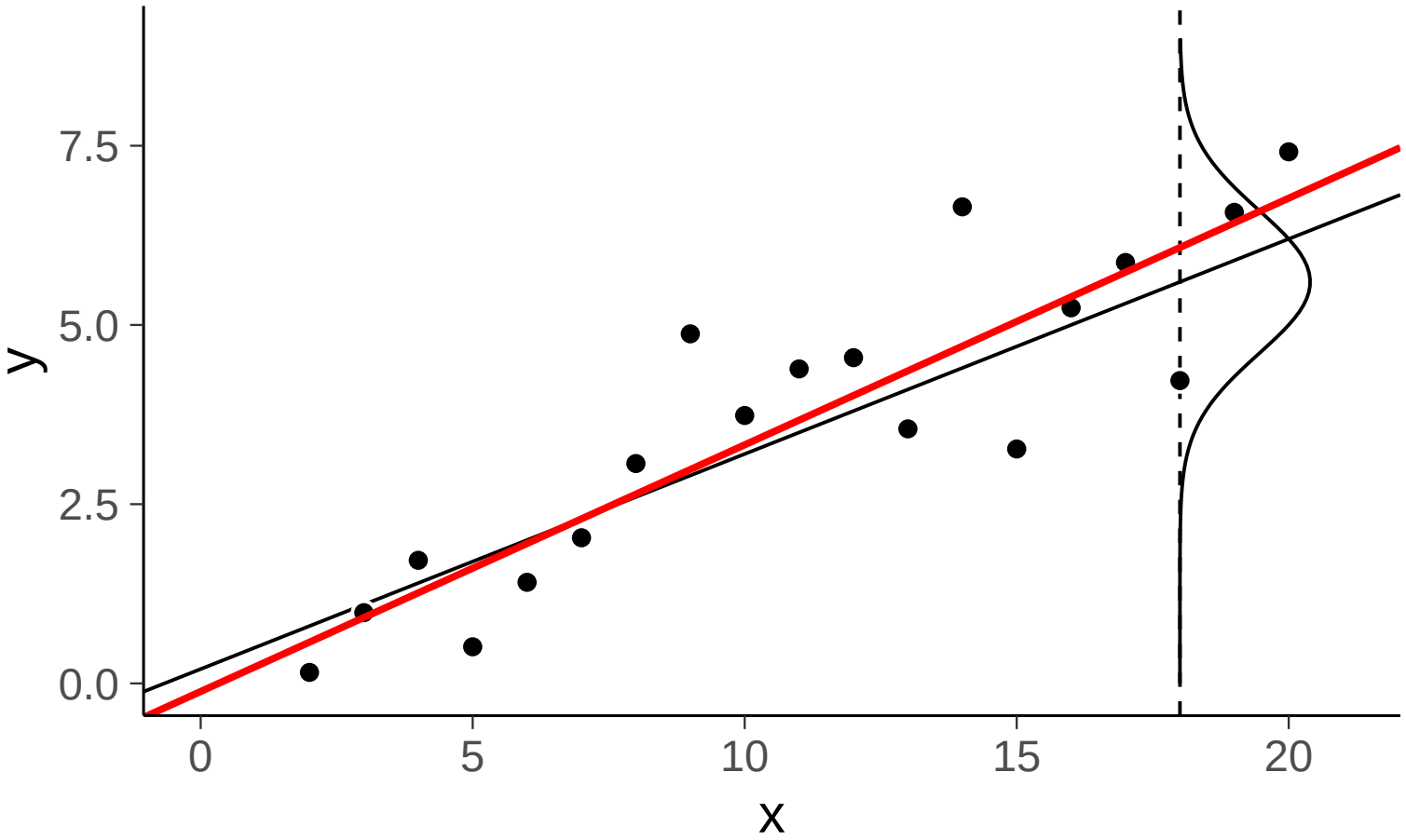



Data
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

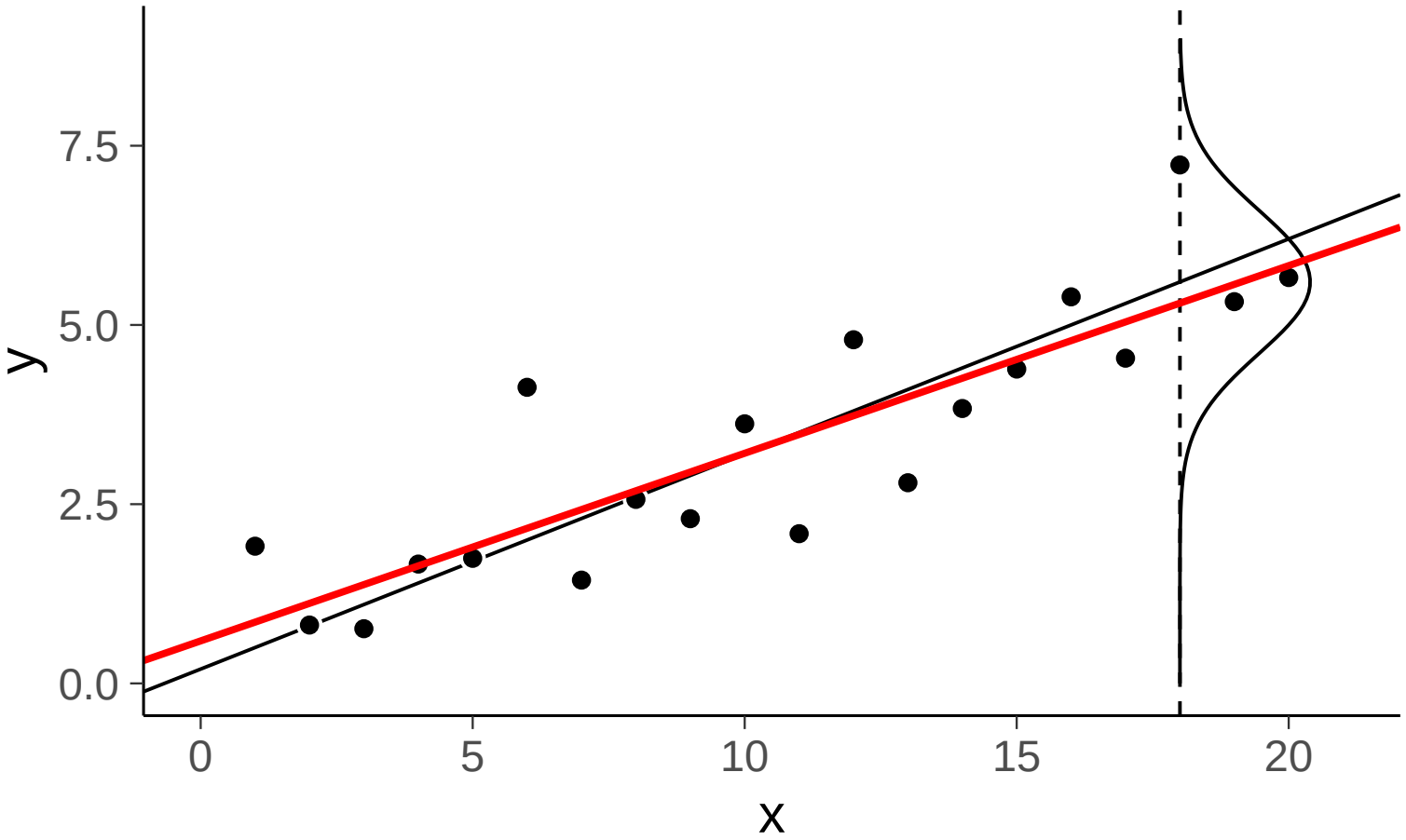
Posterior mean



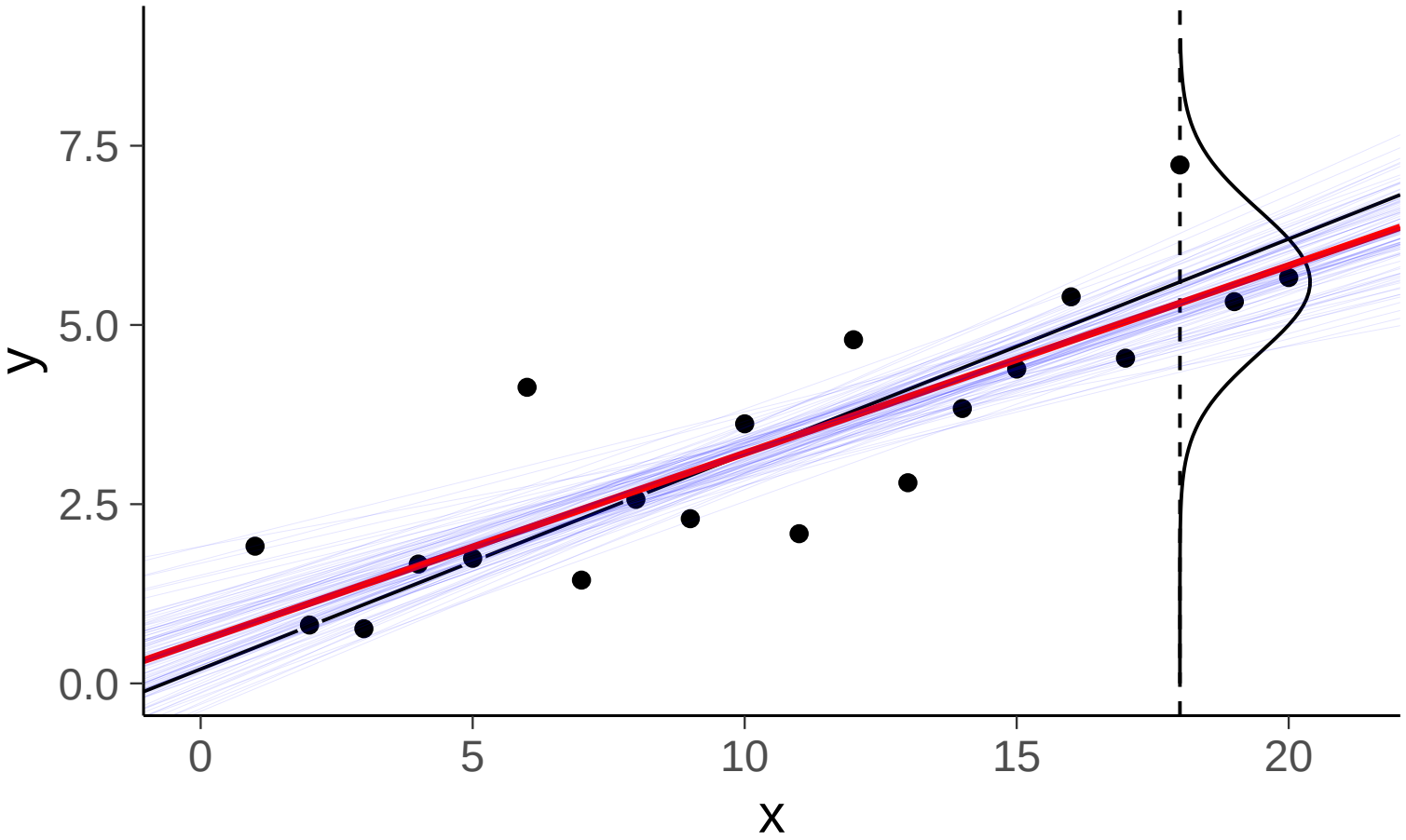
Posterior mean, alternative data realisation



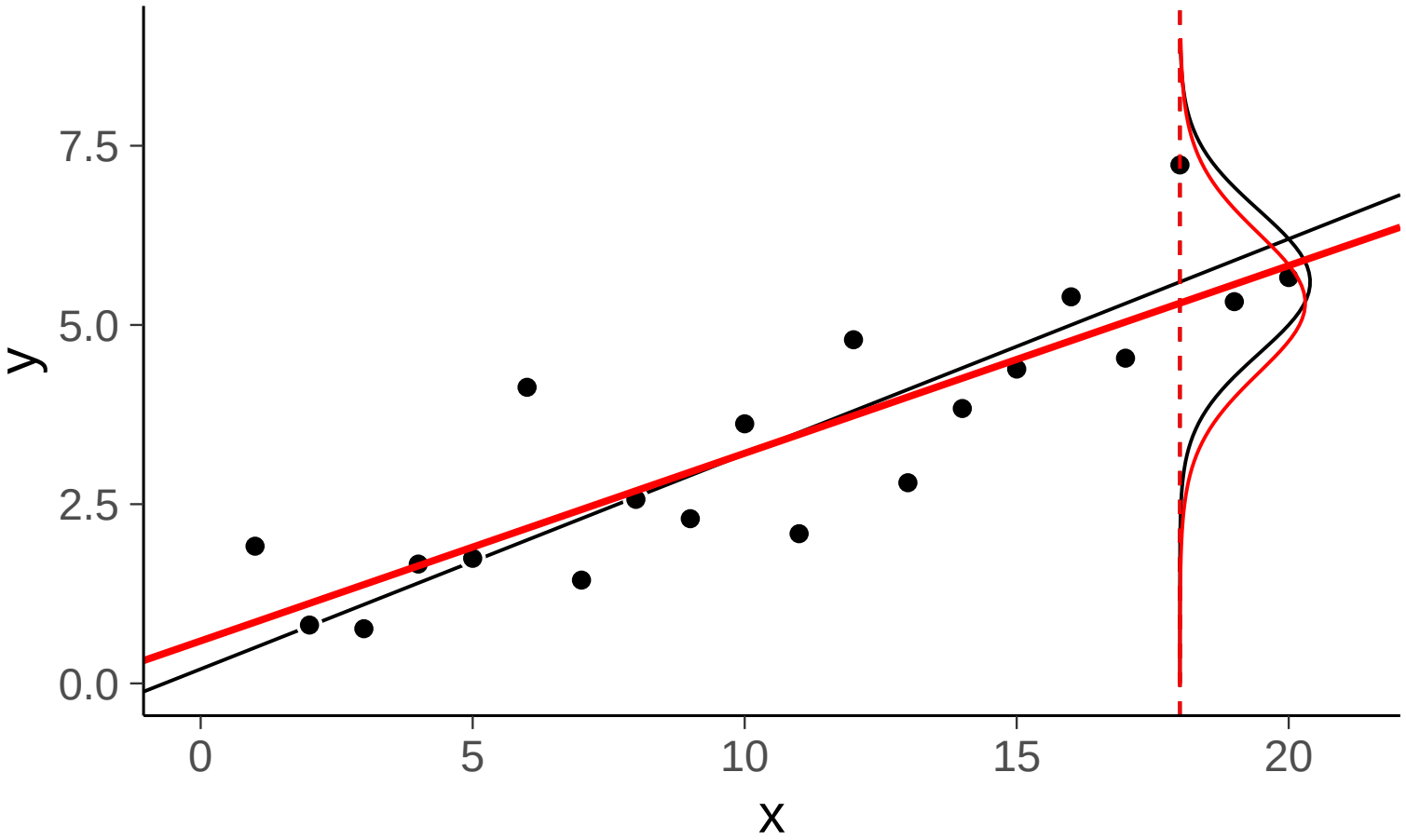
Posterior mean



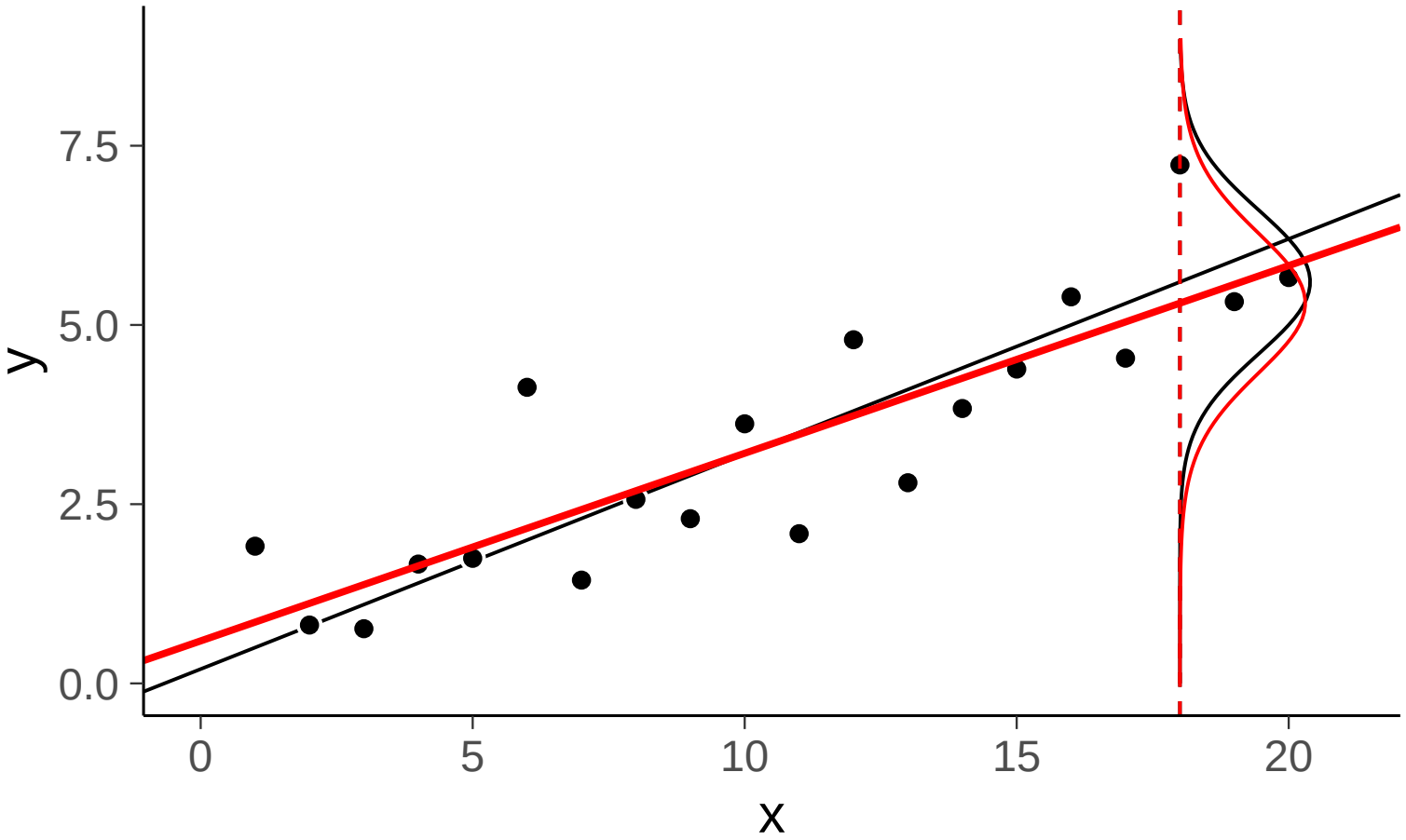
Posterior draws



Posterior predictive distribution

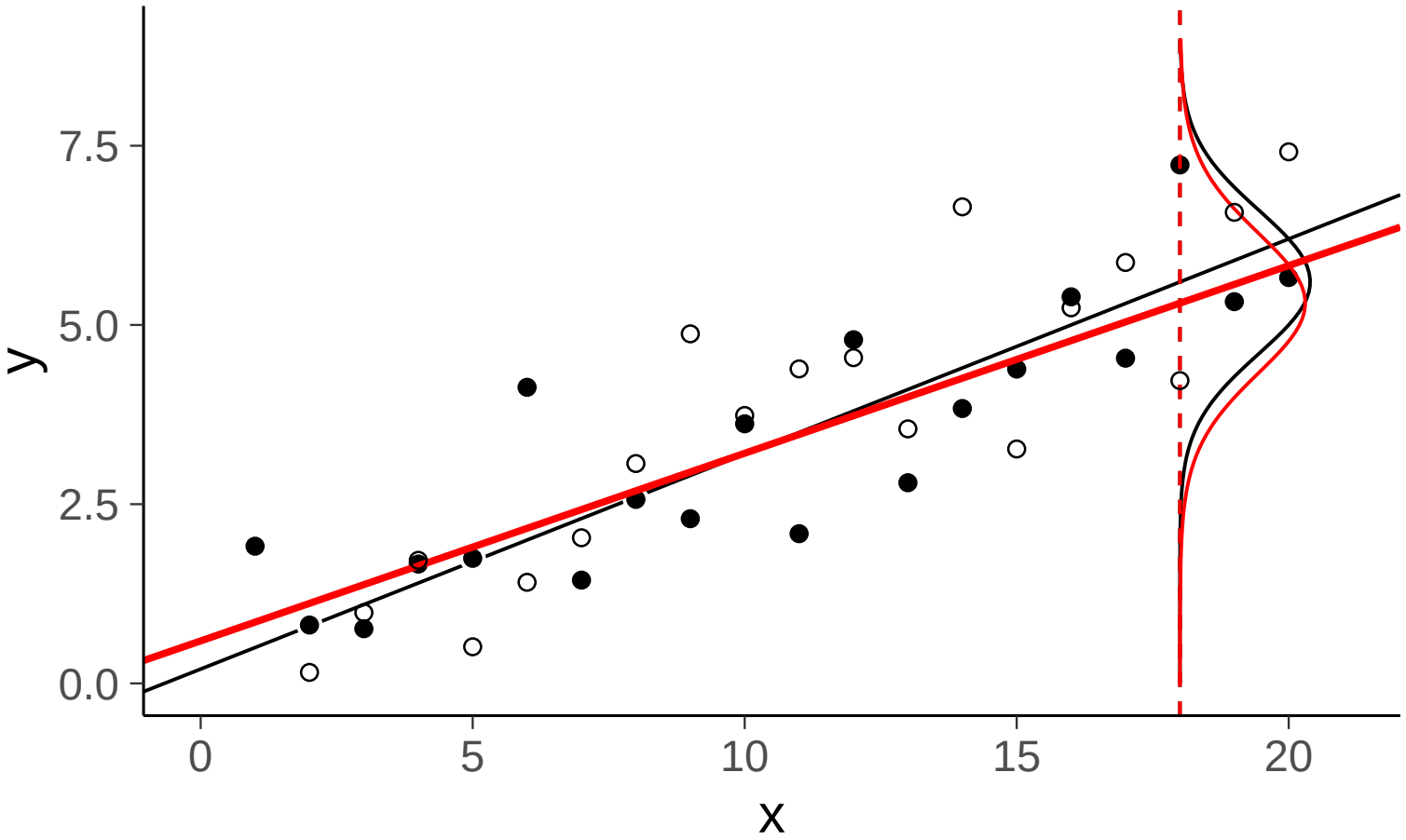


Posterior predictive distribution

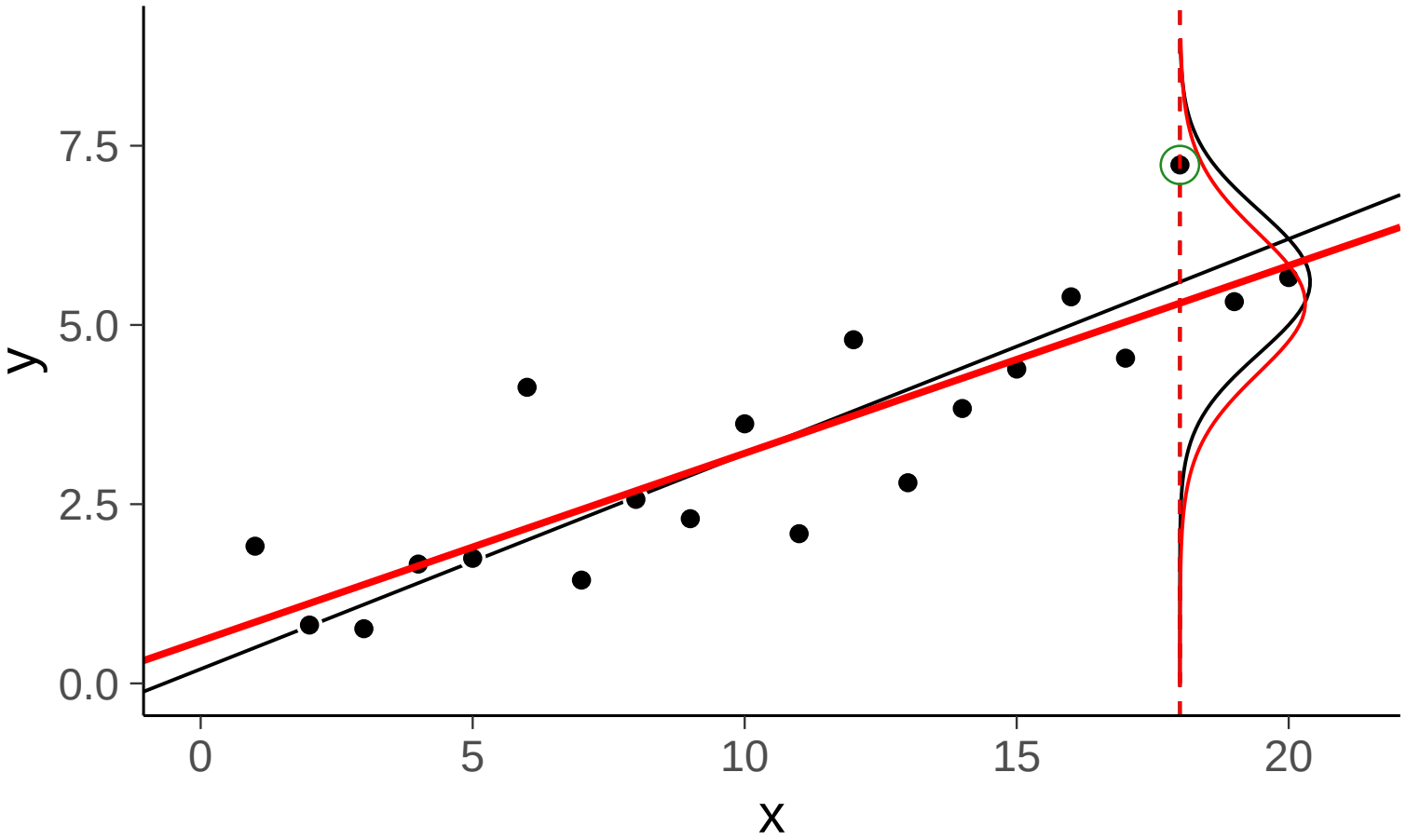


$$p(\tilde{y}|\tilde{x} = 18, x, y) = \int p(\tilde{y}|\tilde{x} = 18, \theta) p(\theta|x, y) d\theta$$

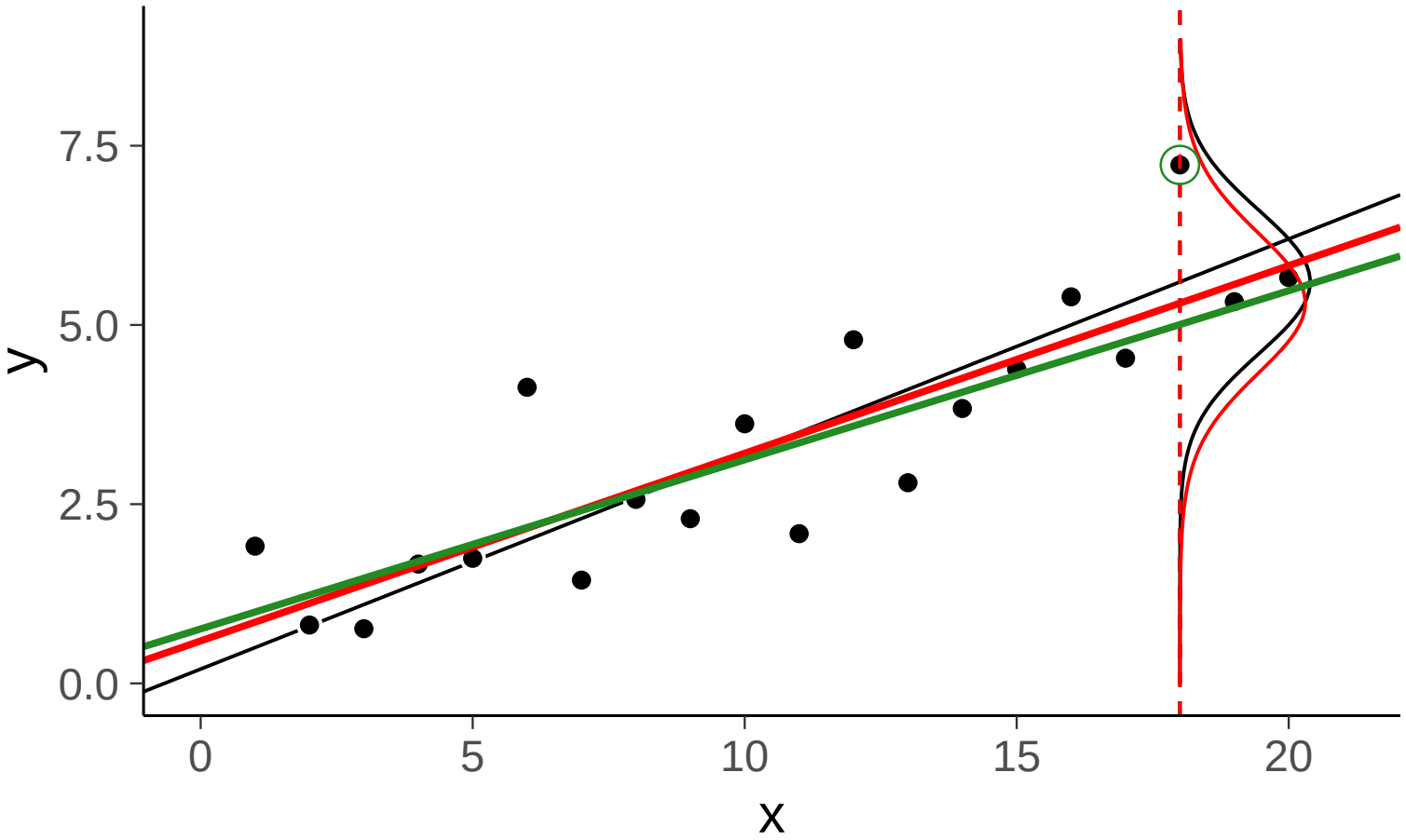
New data



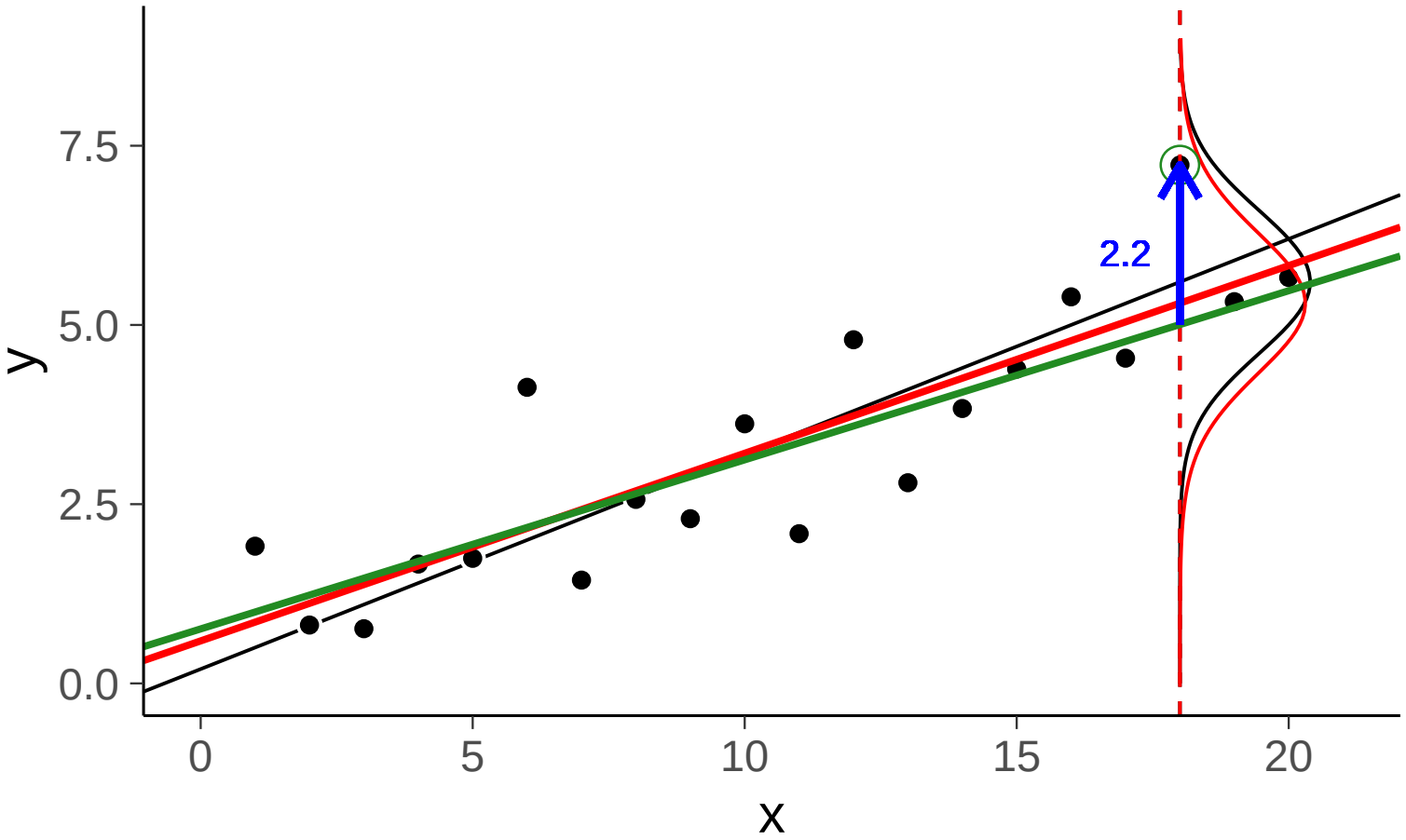
Posterior predictive distribution



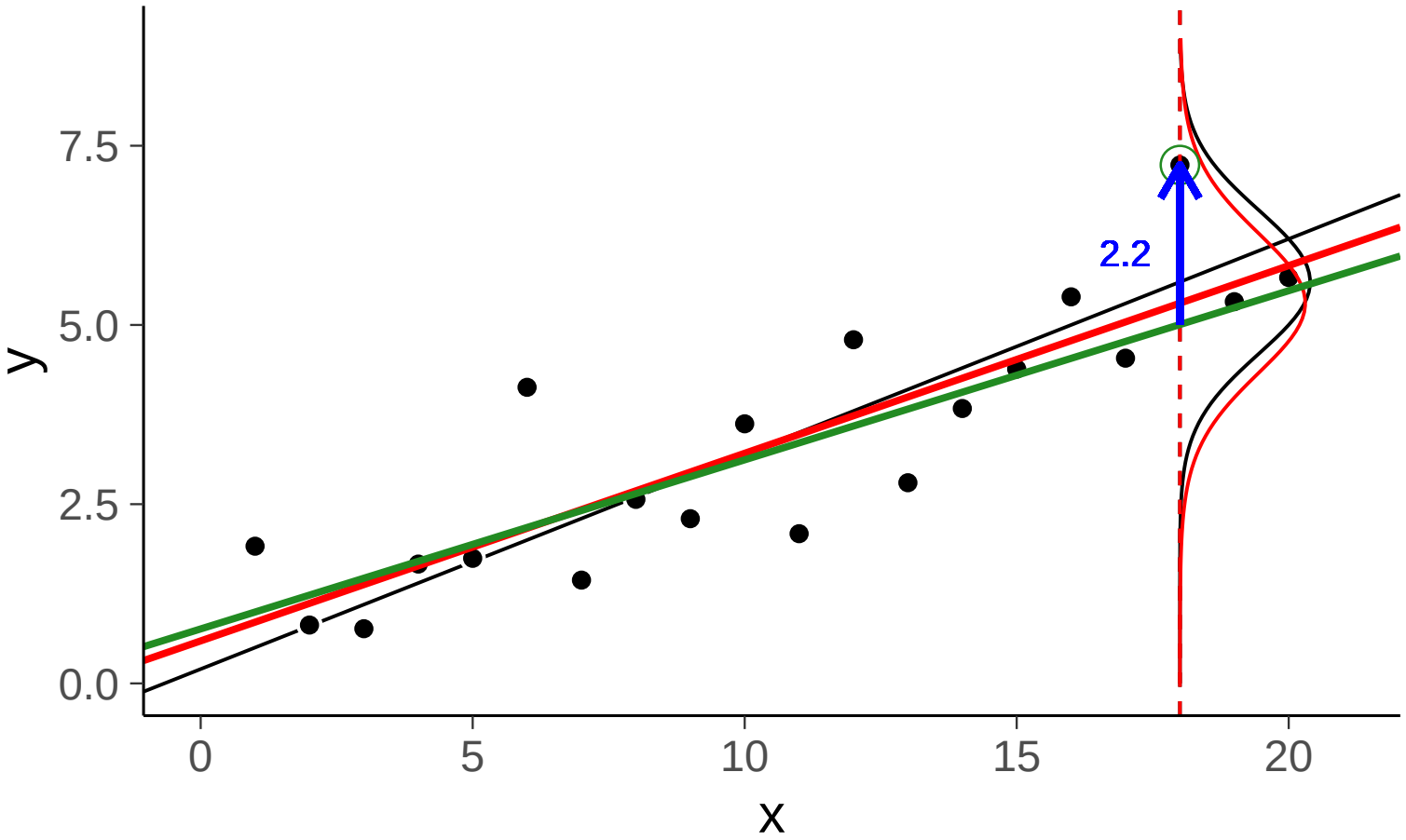
Leave−one−out mean



Leave−one−out residual

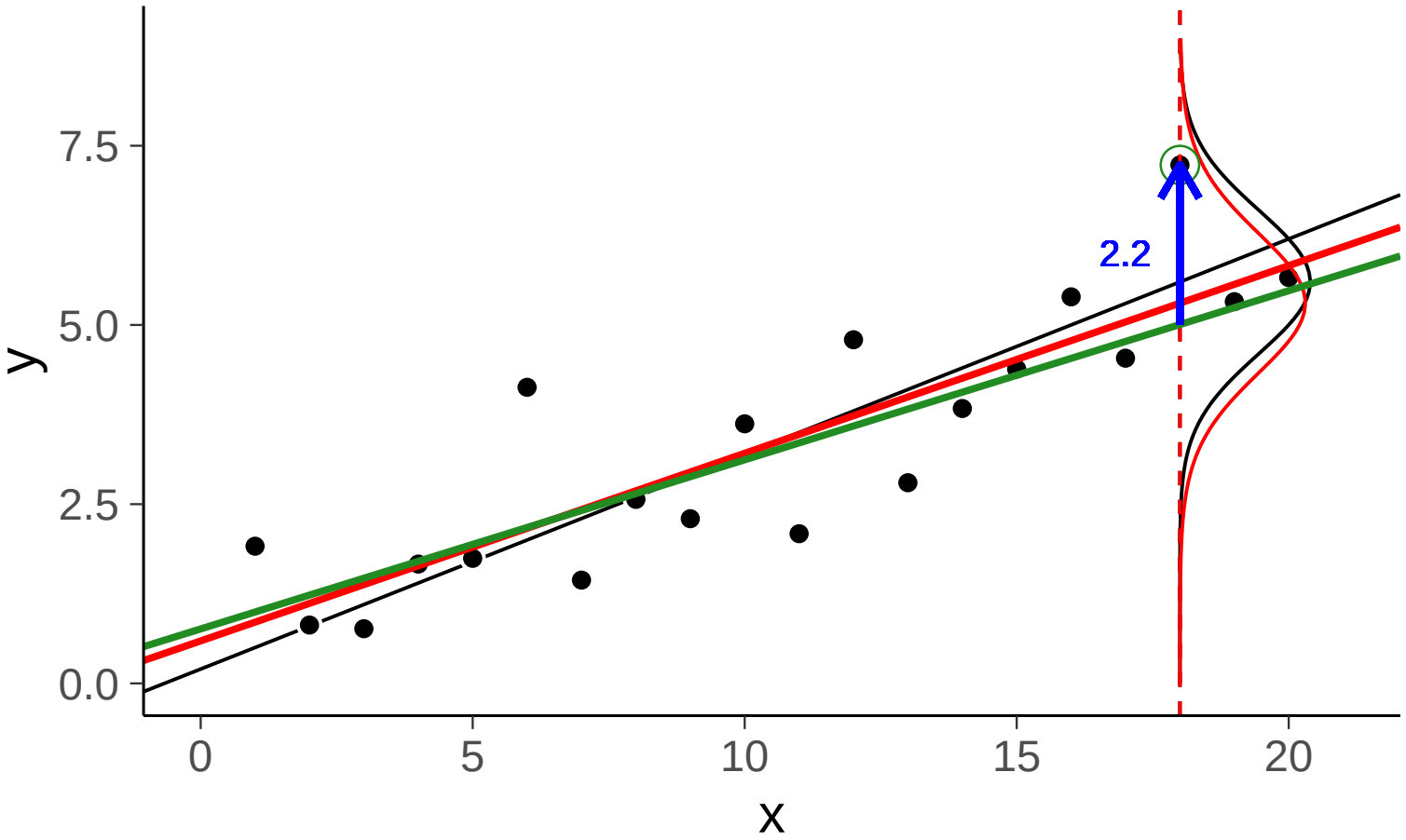


Leave−one−out residual



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

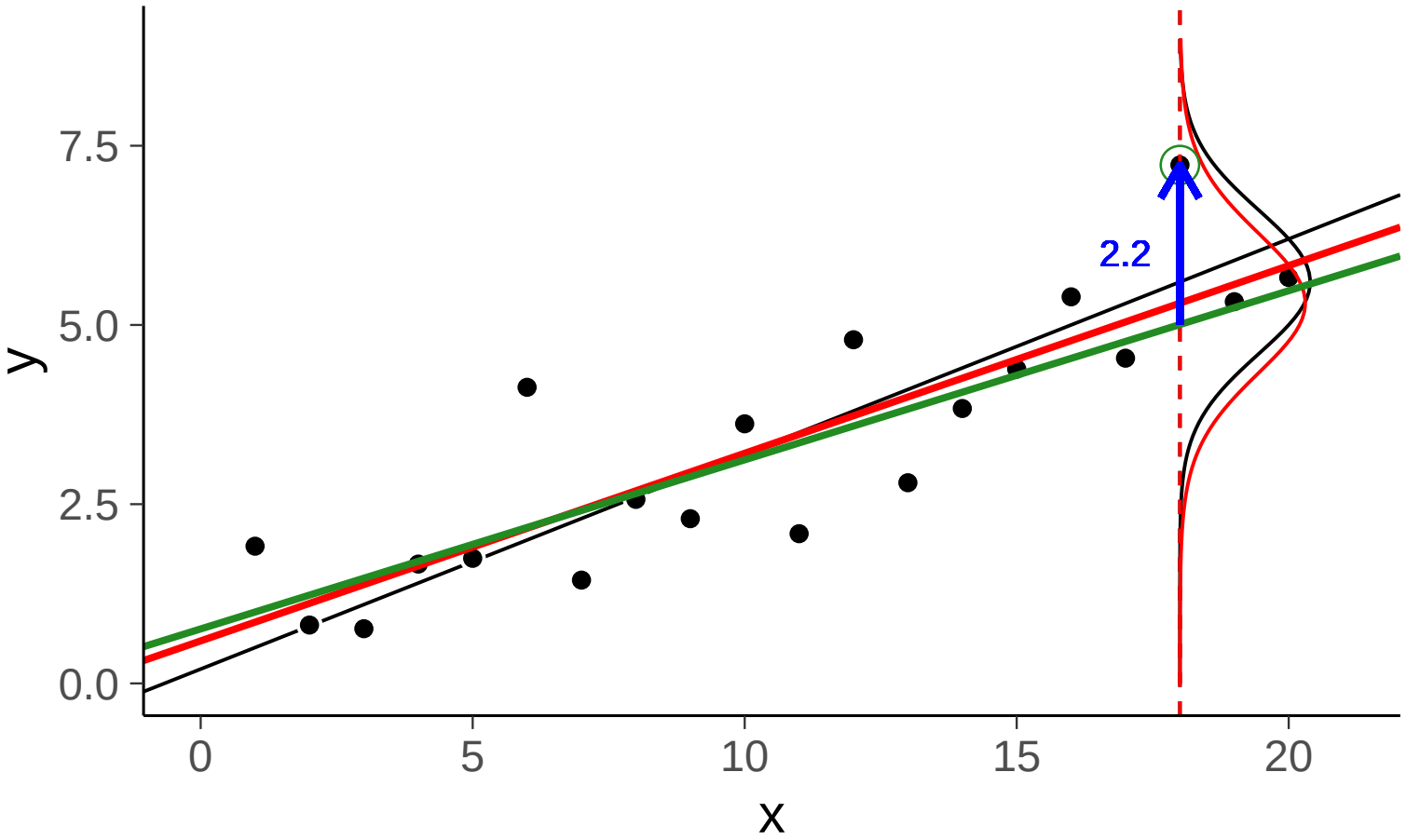
Leave−one−out residual



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

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

Leave−one−out residual

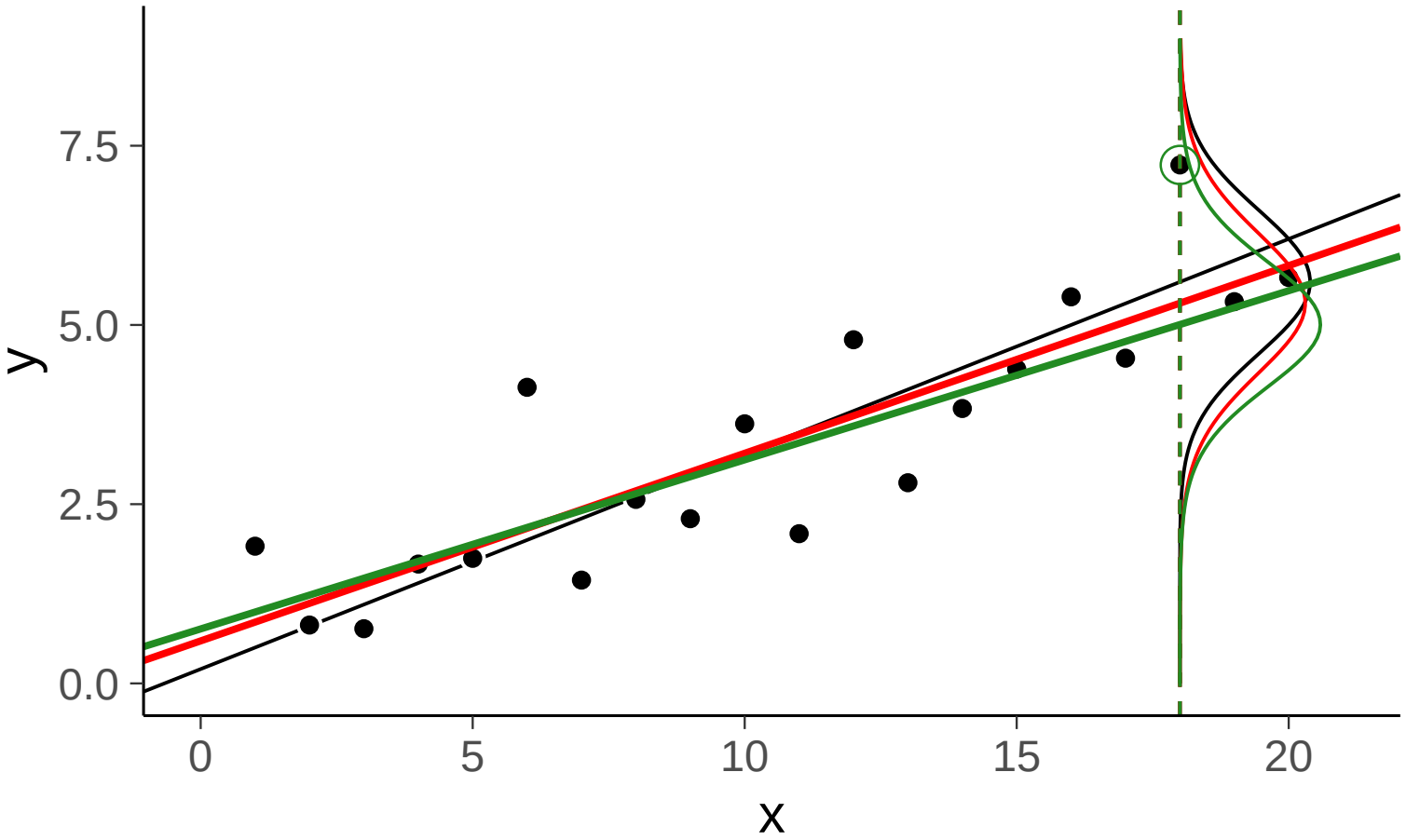


$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

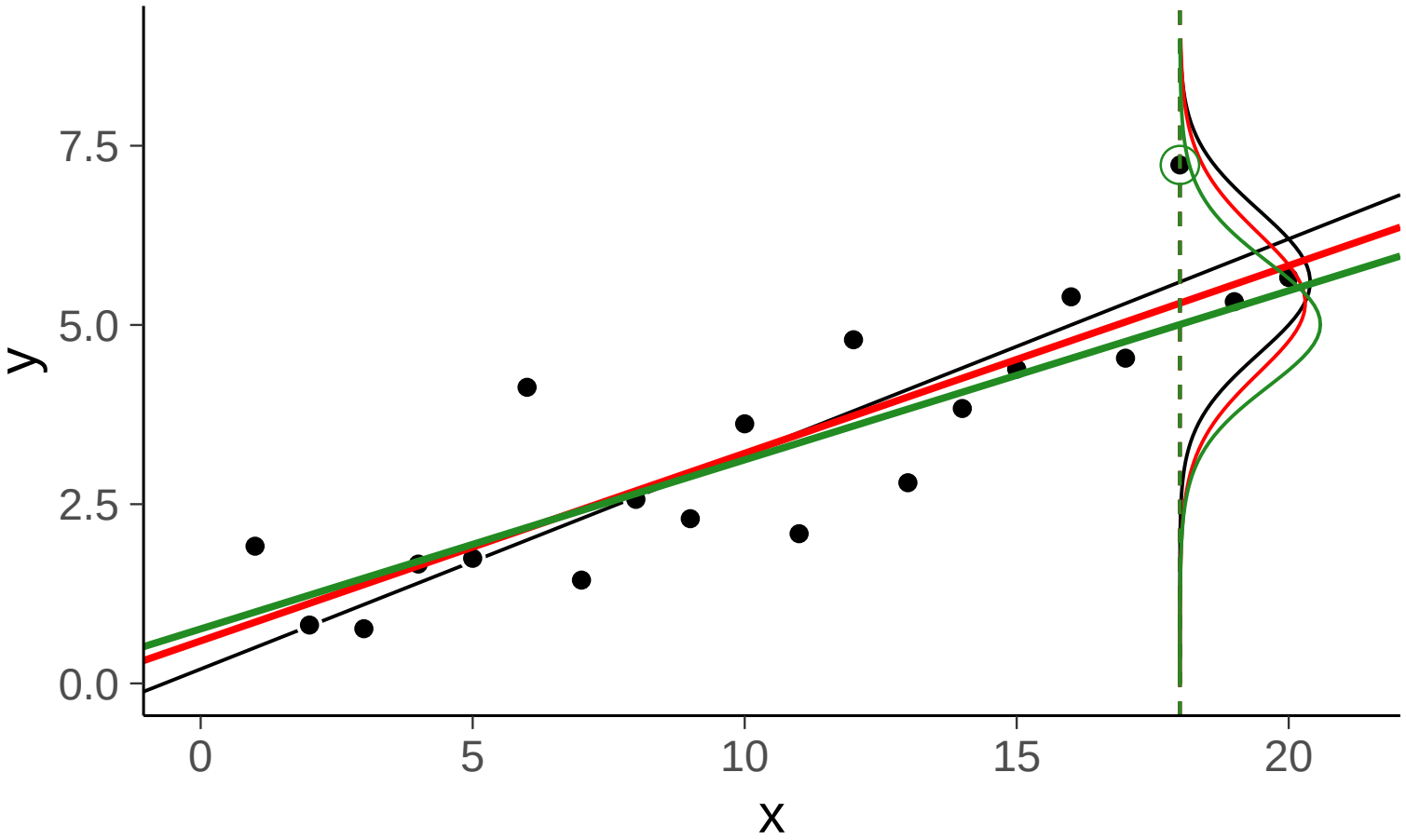
Can be use to compute, e.g., RMSE, R^2, 90% error

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

Leave−one−out predictive distribution

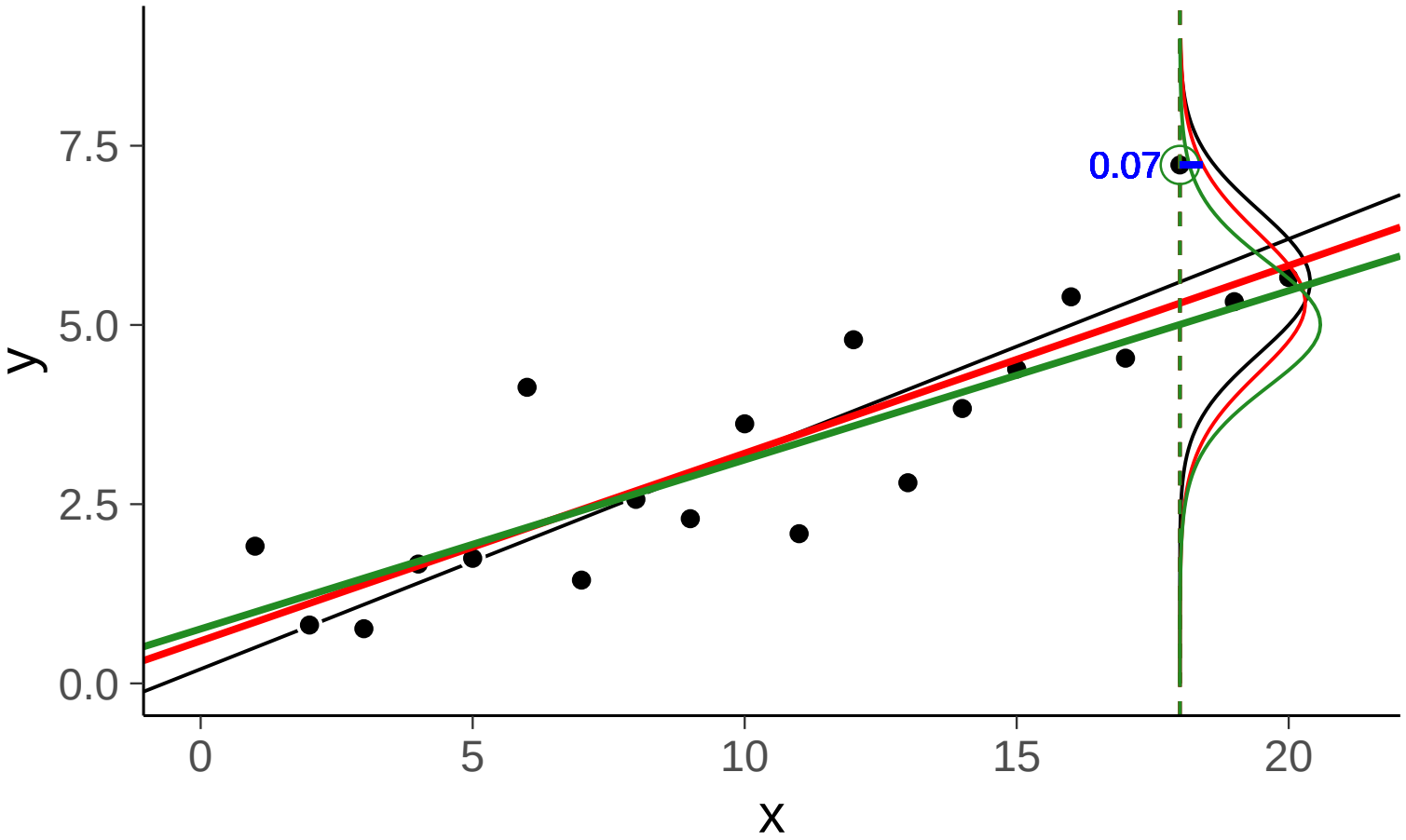


Leave−one−out predictive distribution

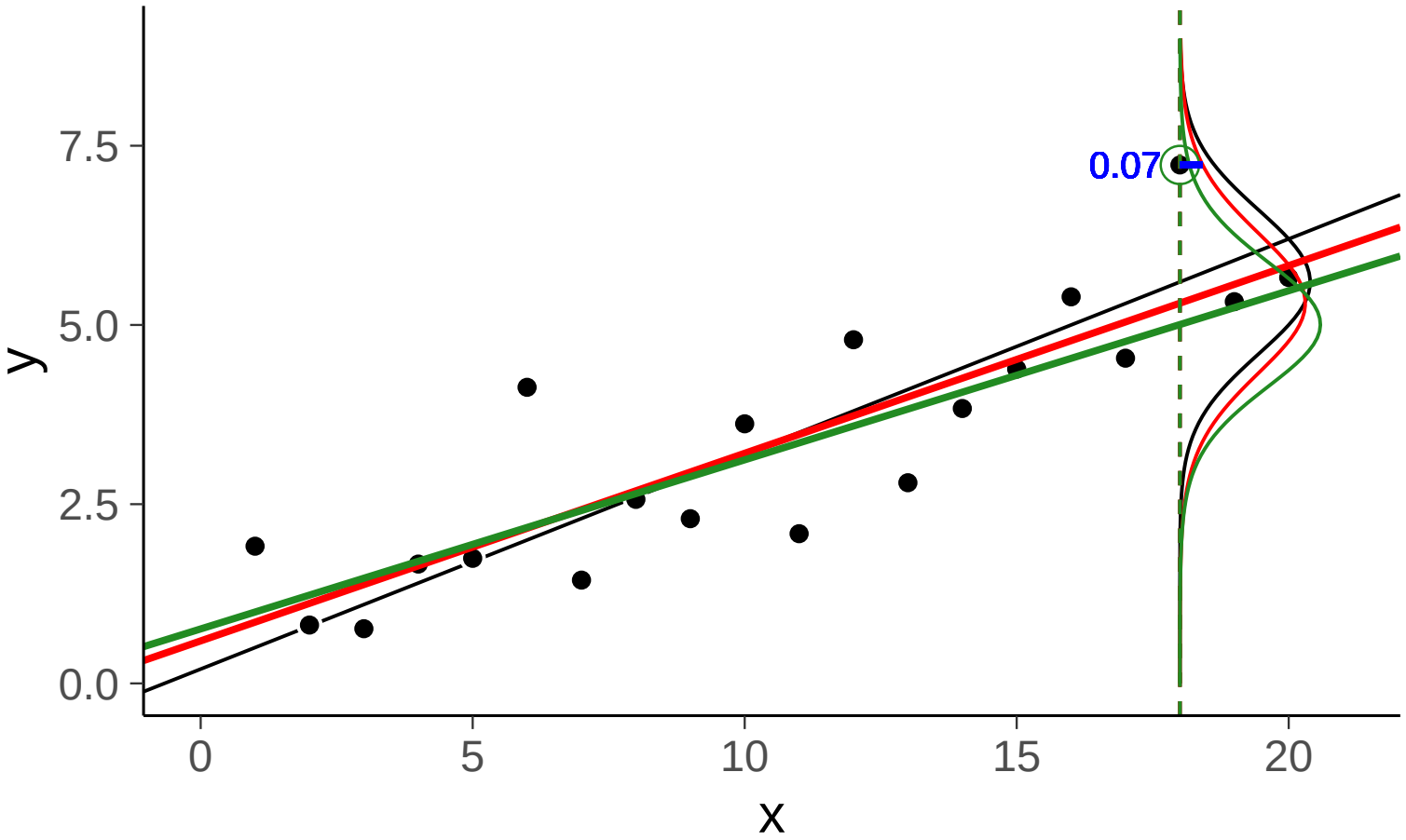


$$p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18}) = \int p(\tilde{y}|\tilde{x} = 18, \theta) p(\theta|x_{-18}, y_{-18}) d\theta$$

Posterior predictive density

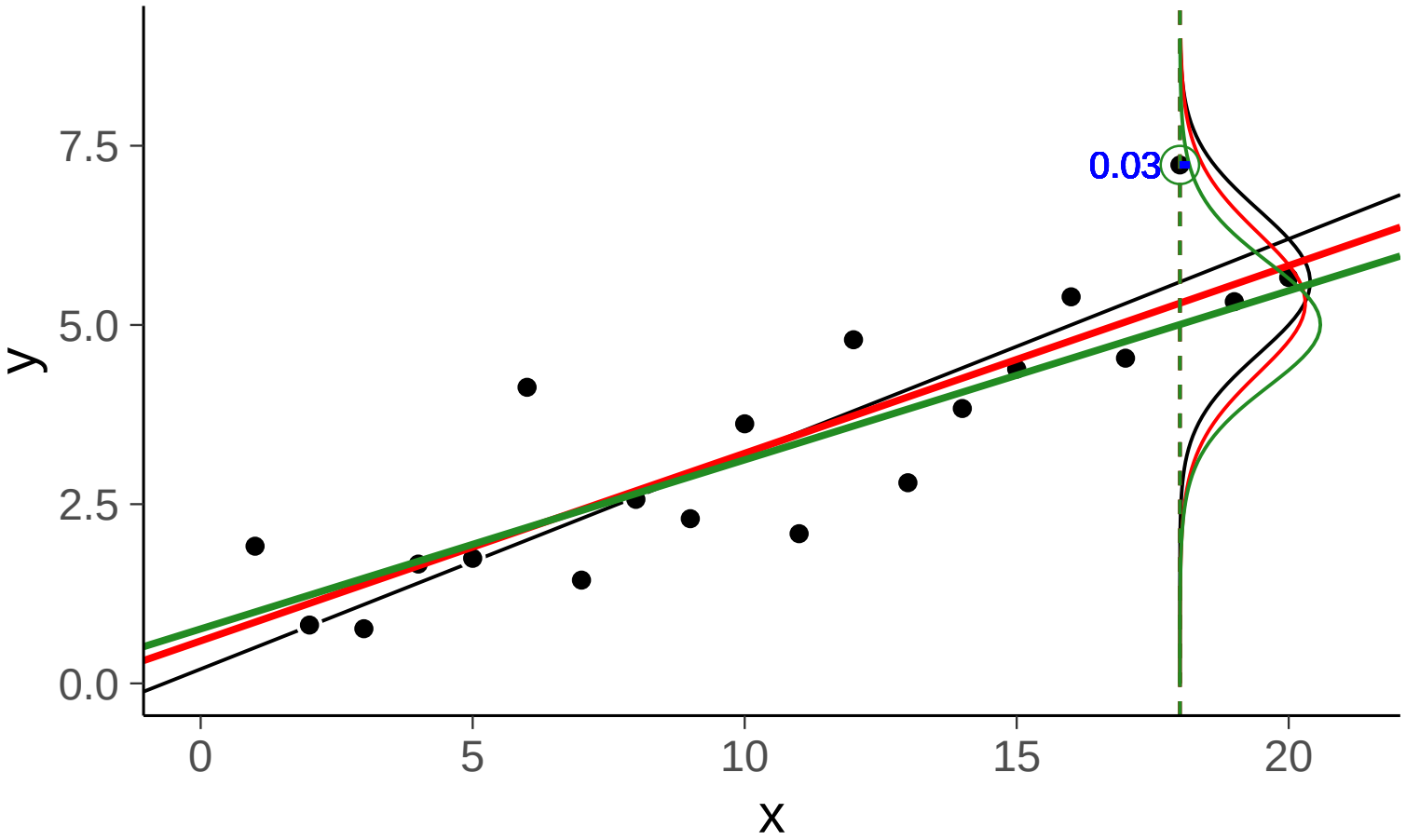


Posterior predictive density



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

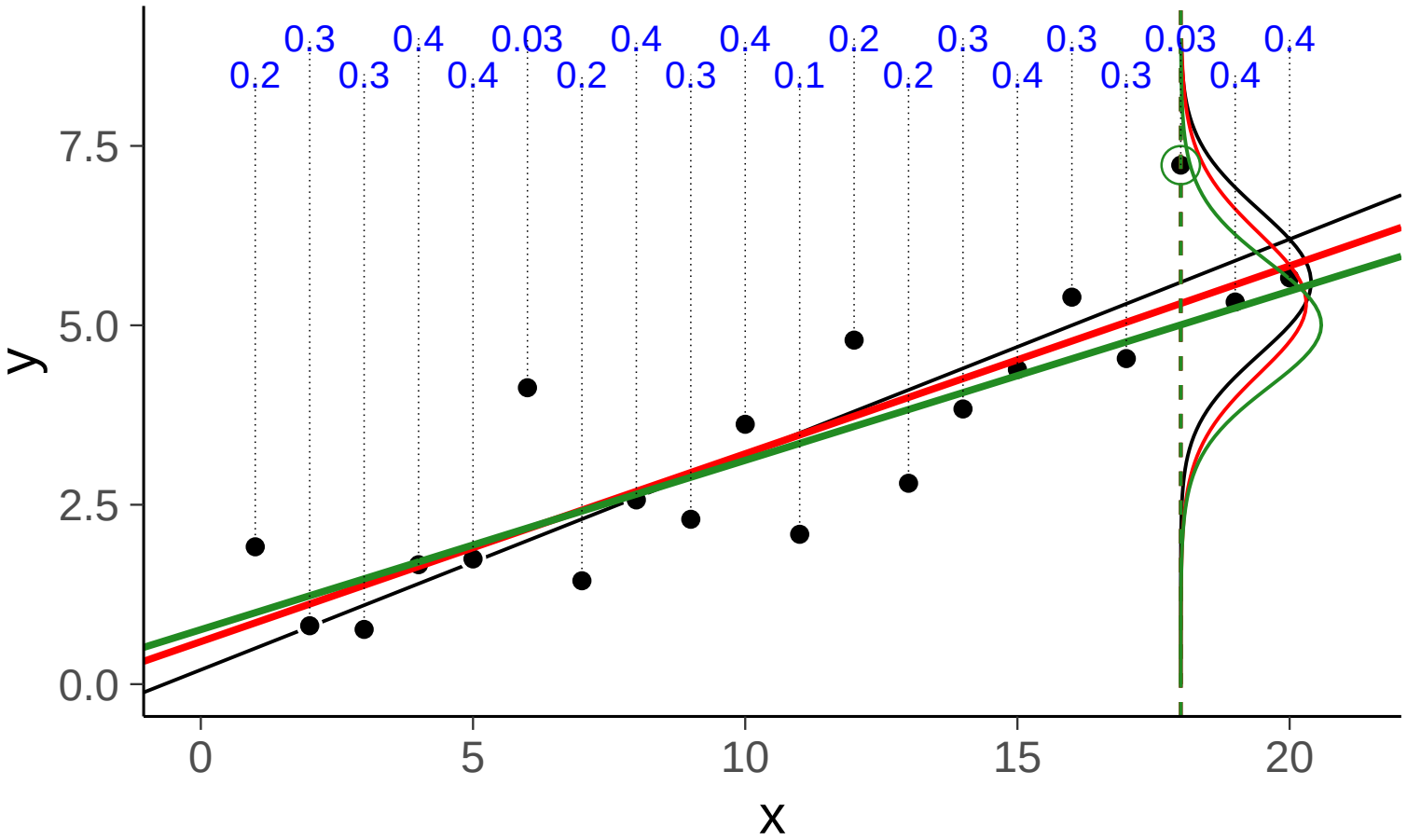
Leave−one−out predictive density



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

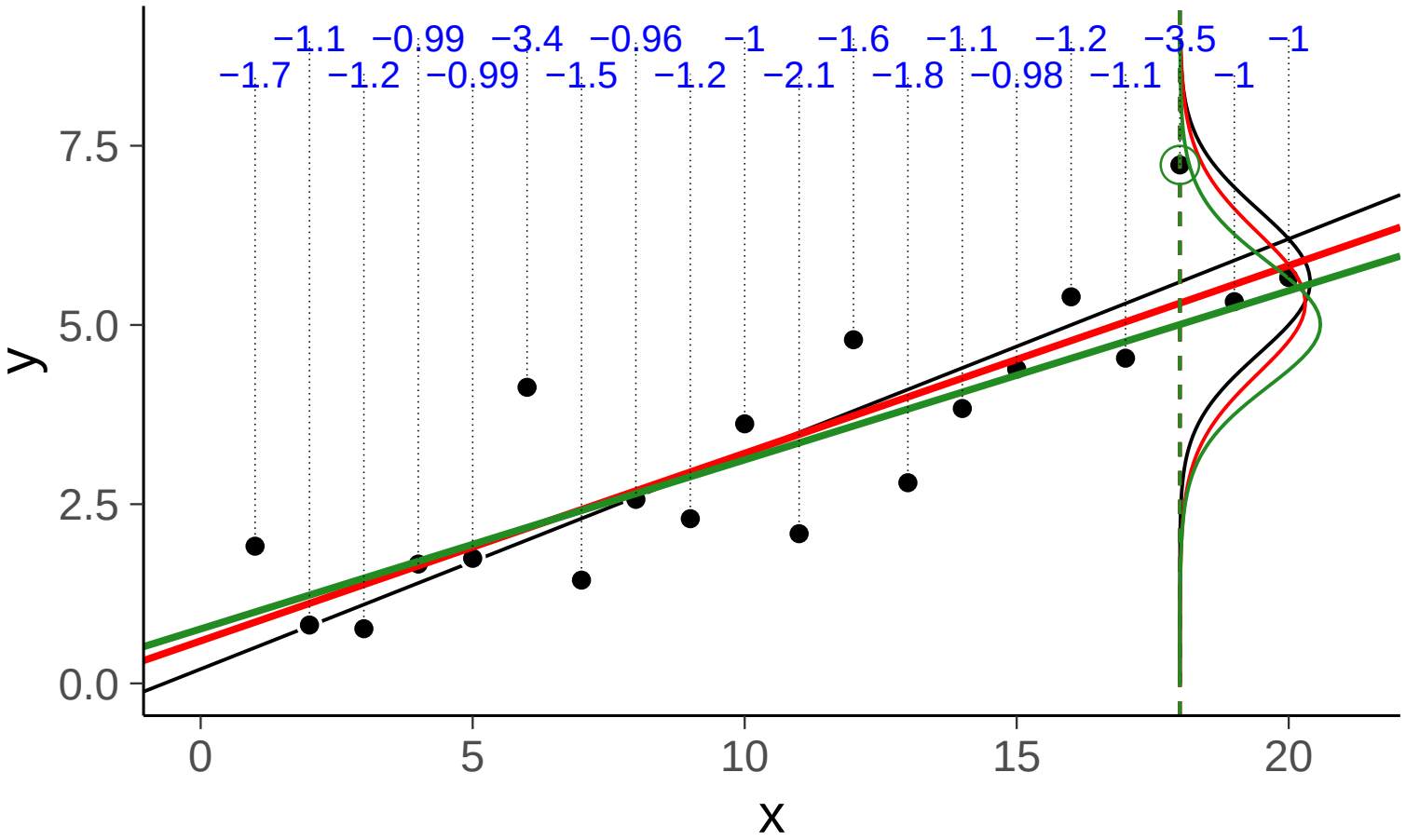
$p(\tilde{y} = y_{18} | \tilde{x} = 18, x_{-18}, y_{-18}) \approx 0.03$

Leave−one−out predictive densities



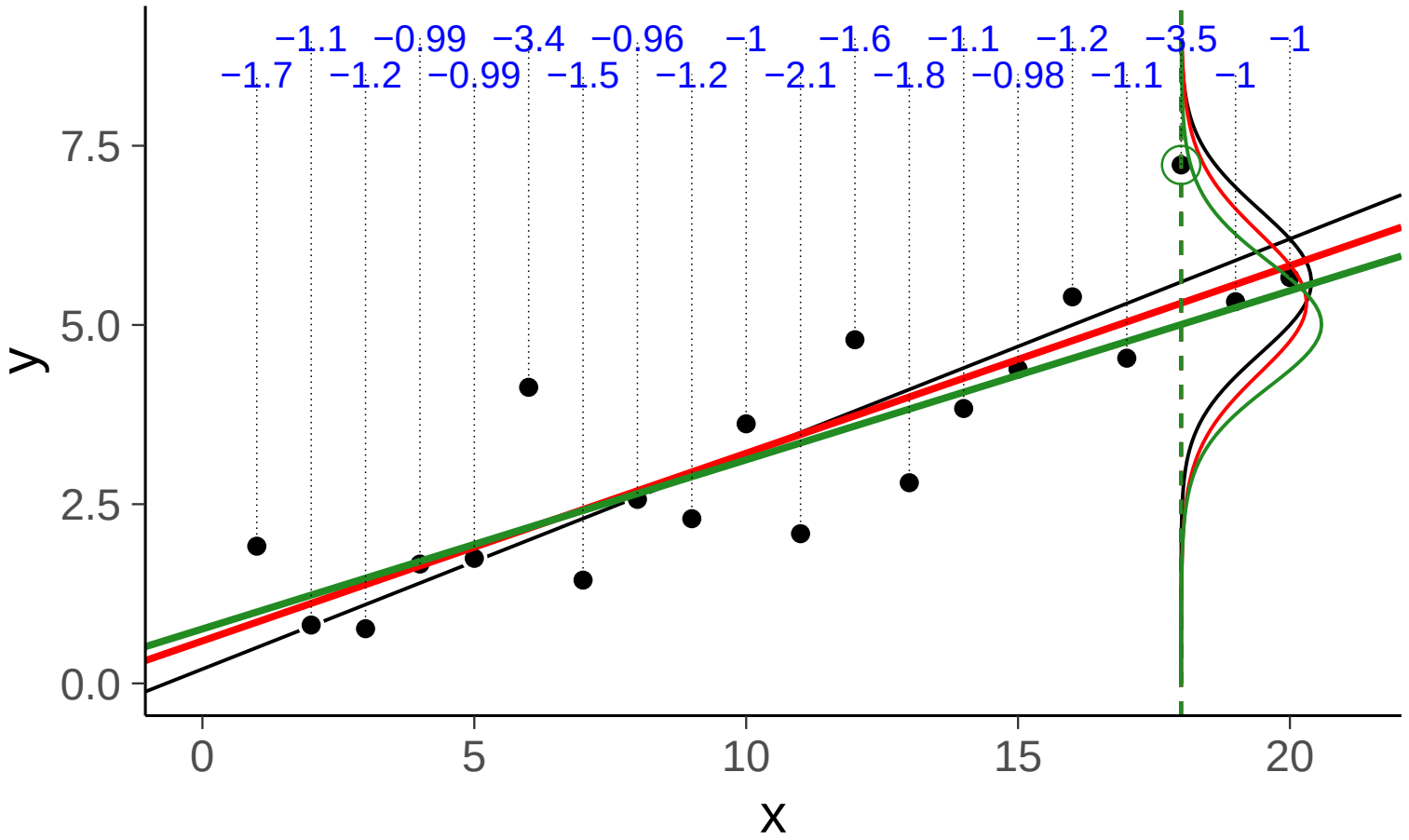
$p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

Leave−one−out log predictive densities



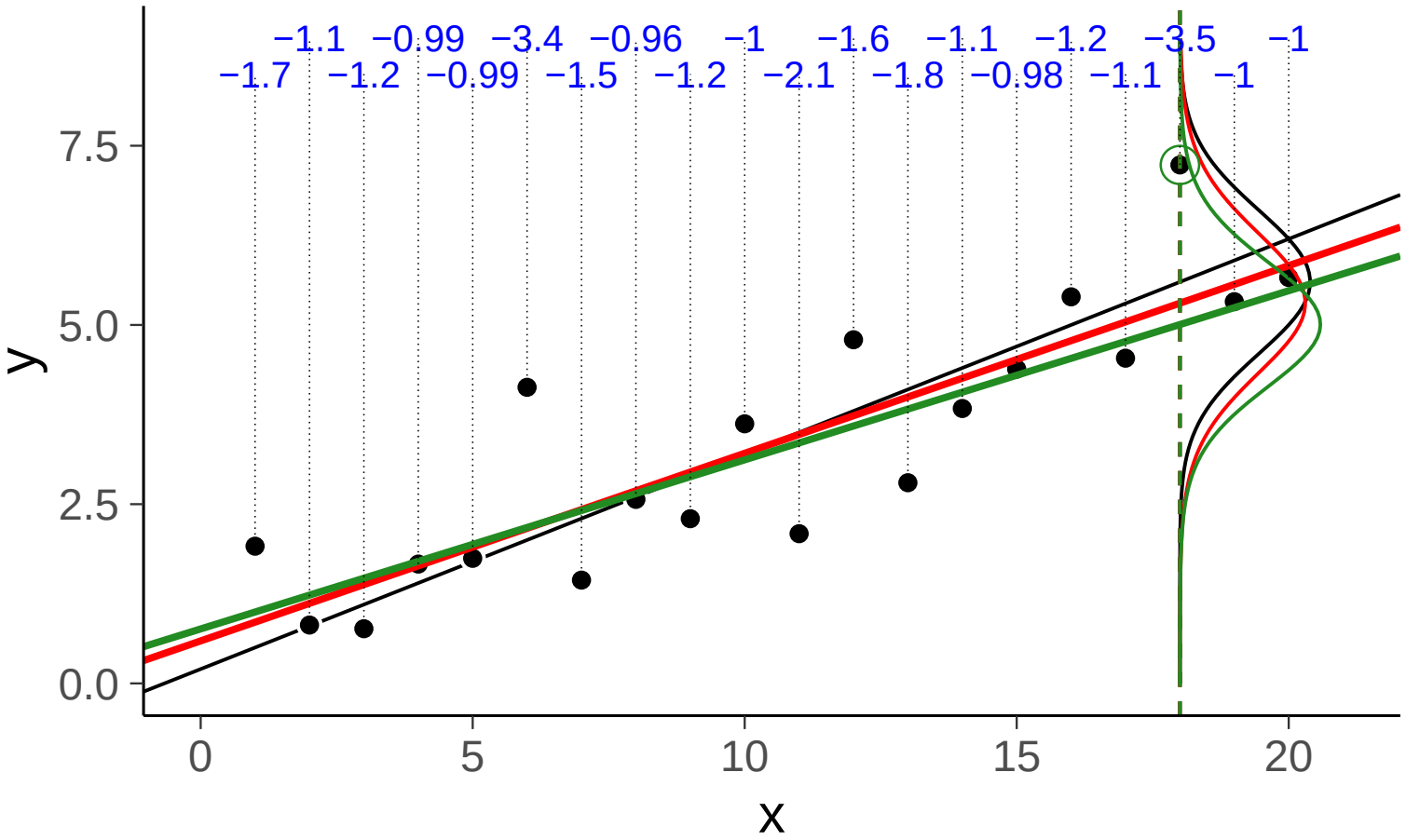
$$\log p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$$

Leave−one−out log predictive densities



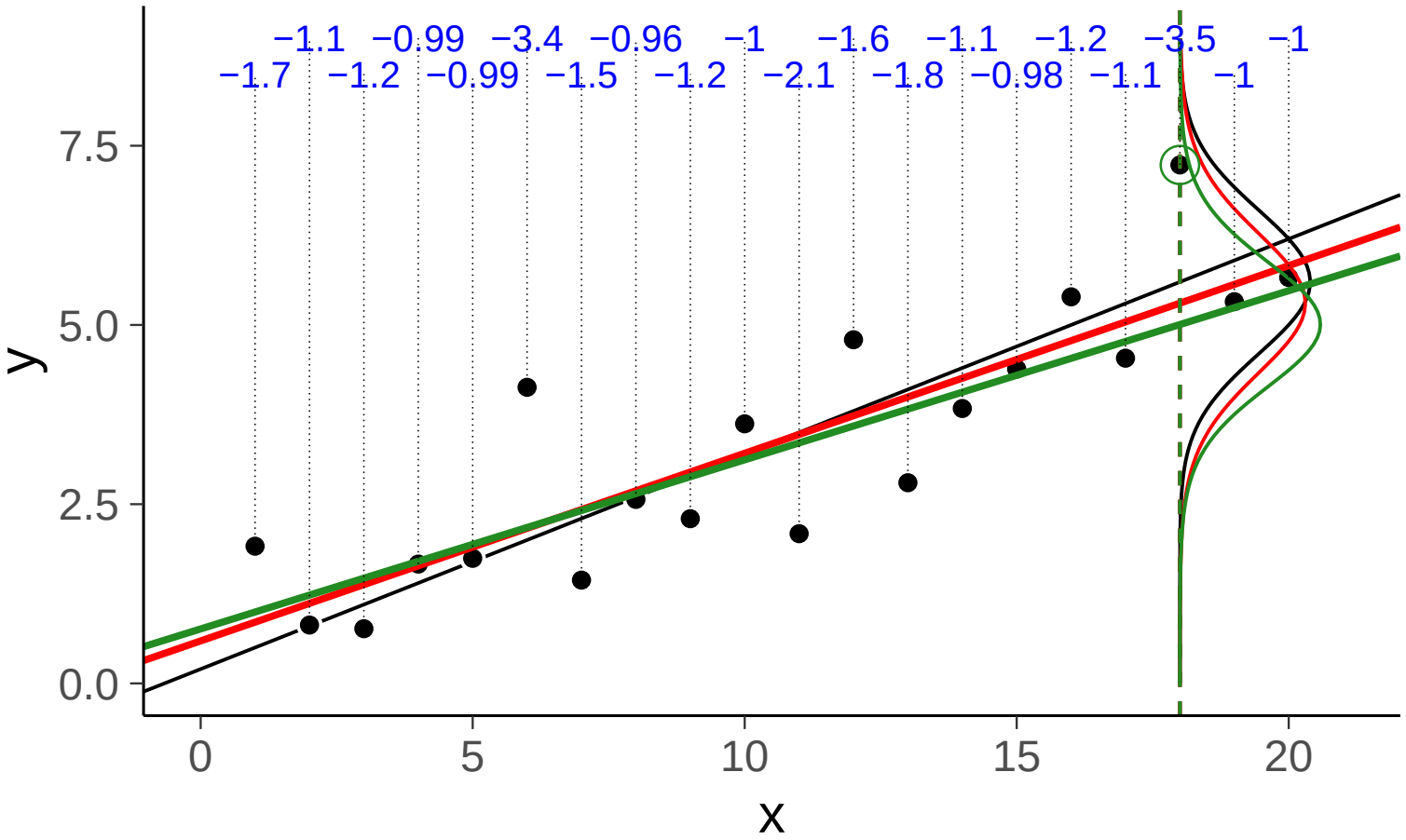
$$\sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$$

Leave−one−out log predictive densities



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

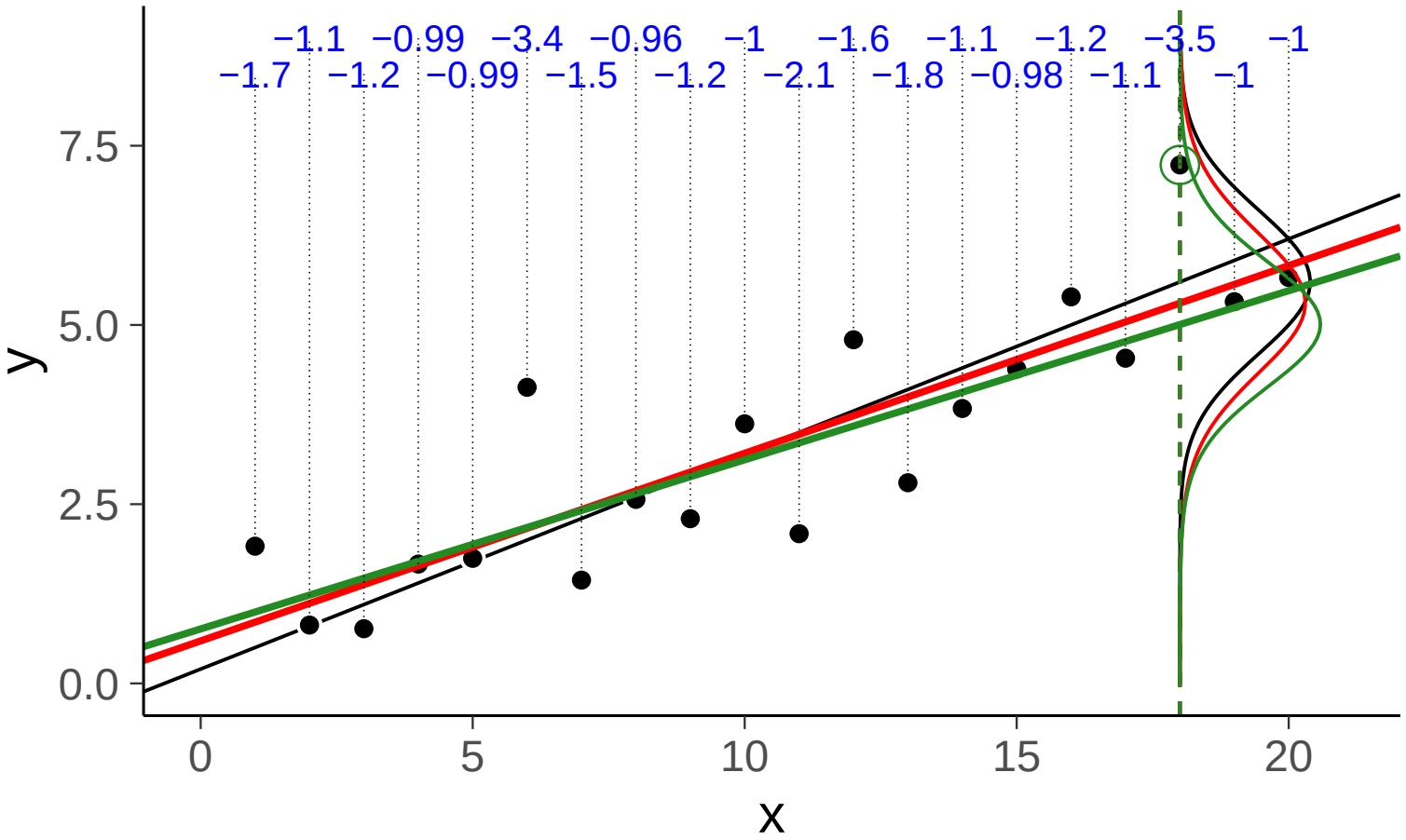
Leave−one−out log predictive densities



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

unbiased estimate of log posterior pred. density for new data

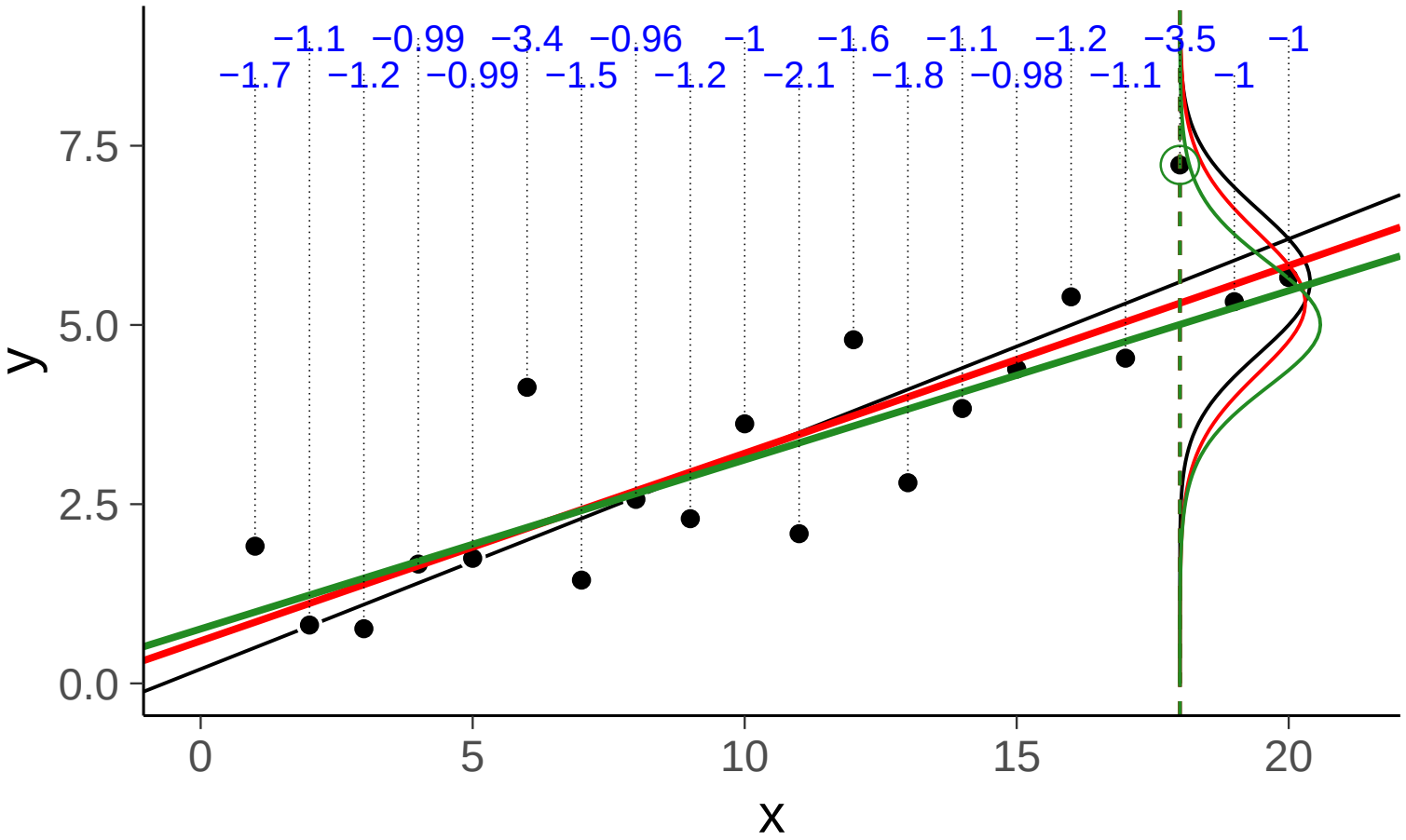
Leave−one−out log predictive densities



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

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

Leave−one−out log predictive densities

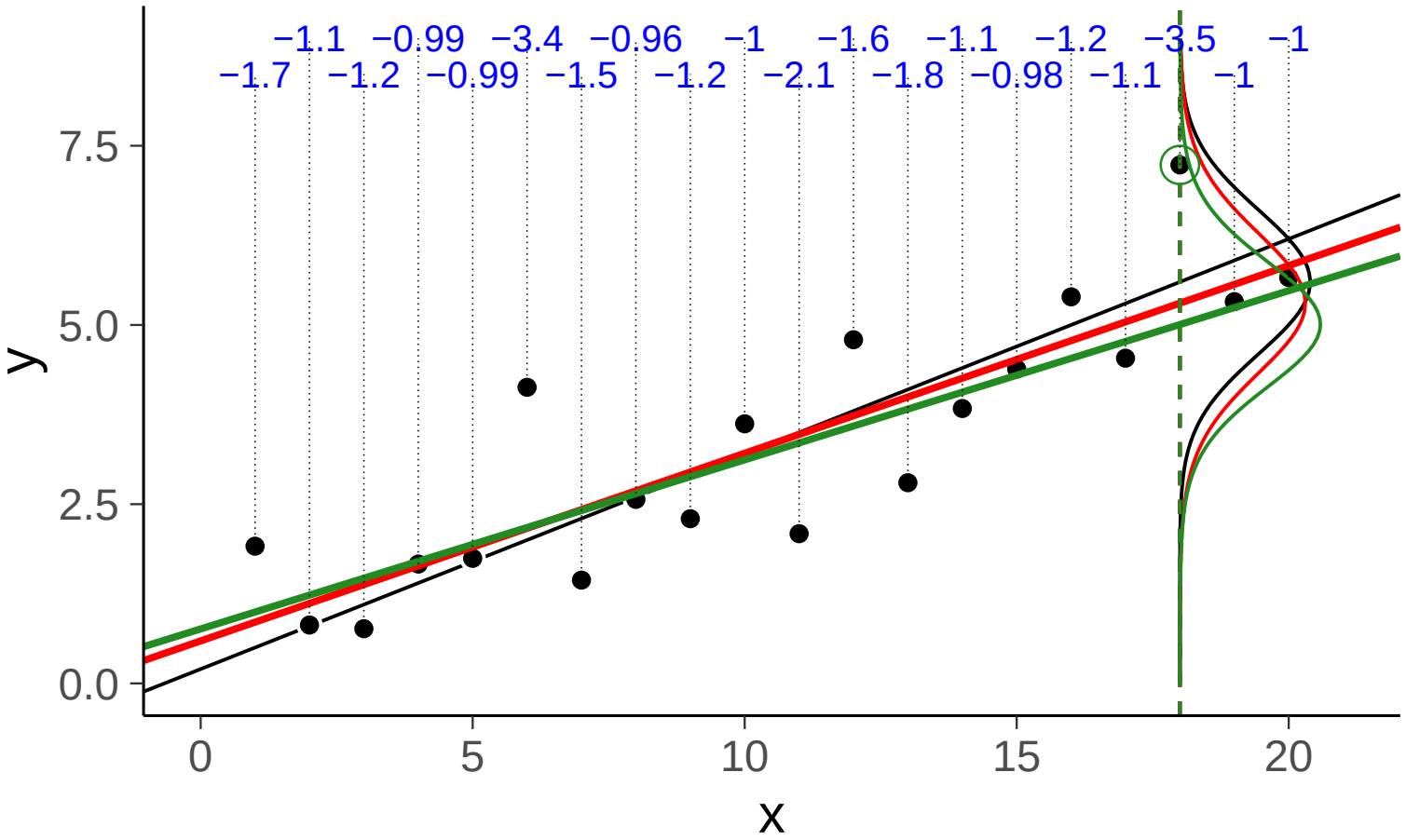


$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$$

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

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

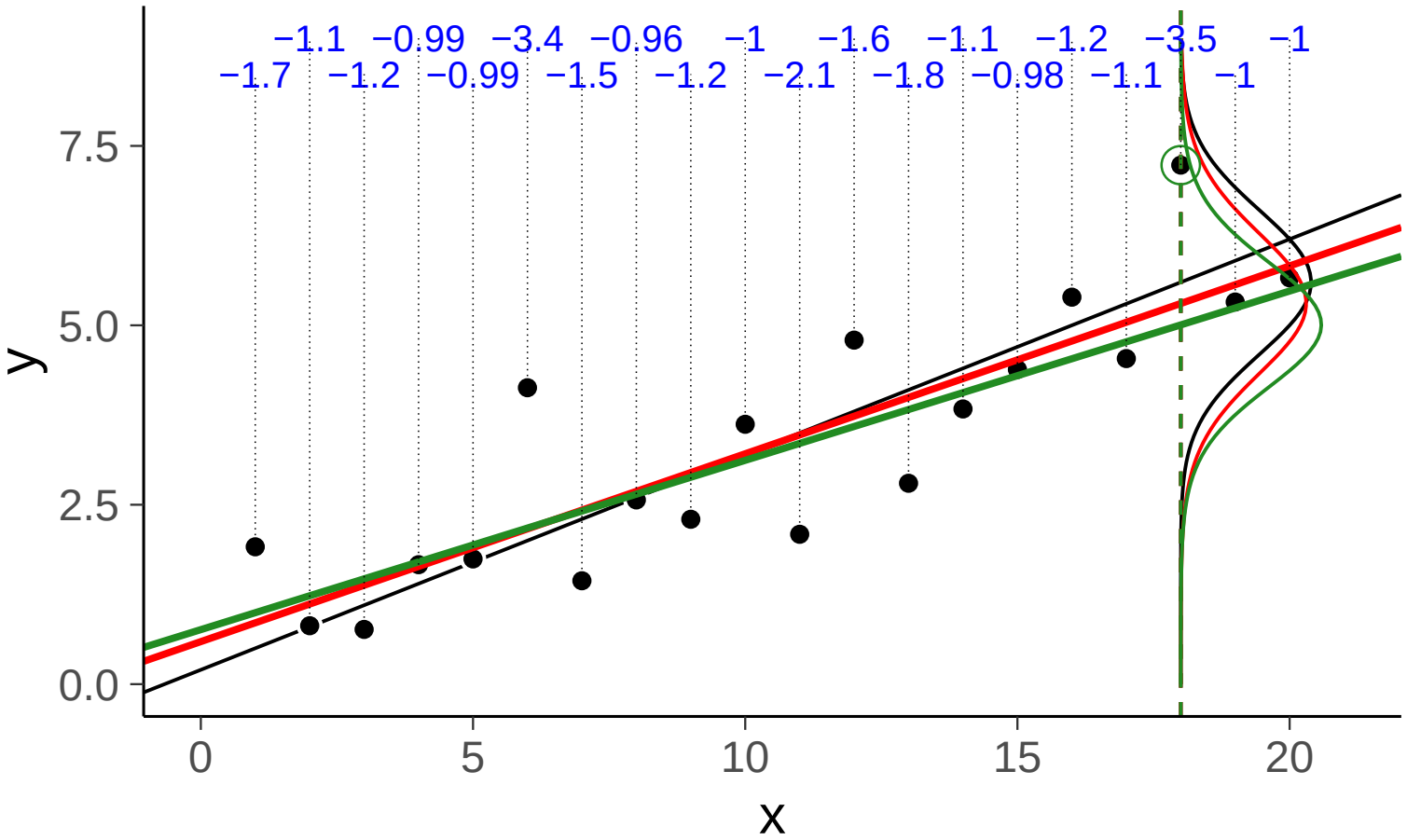
Leave−one−out log predictive densities



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

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

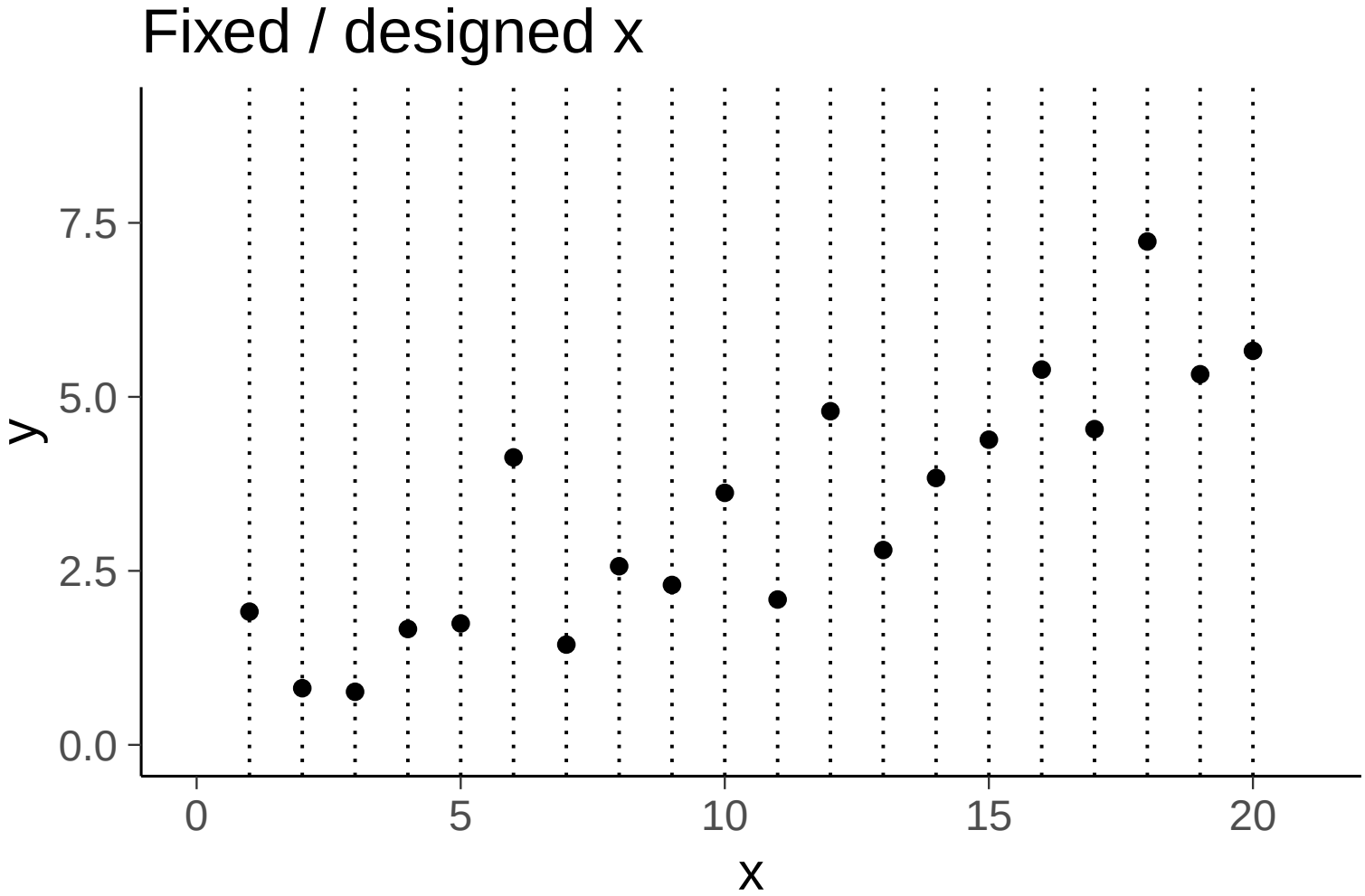
Leave−one−out log predictive densities



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

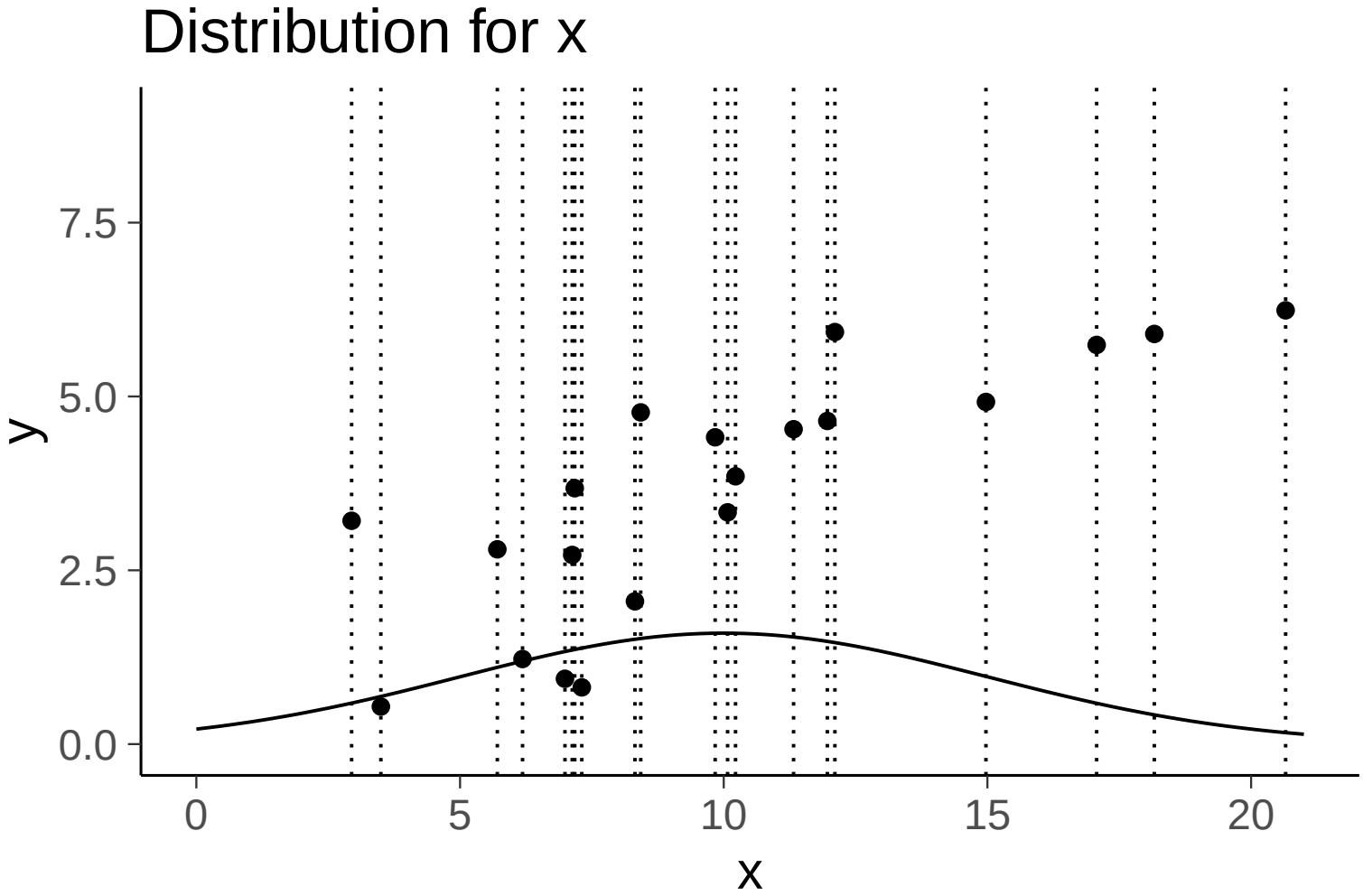
$$\text{SE} = \text{sd}(\log p(y_i|x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$$

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



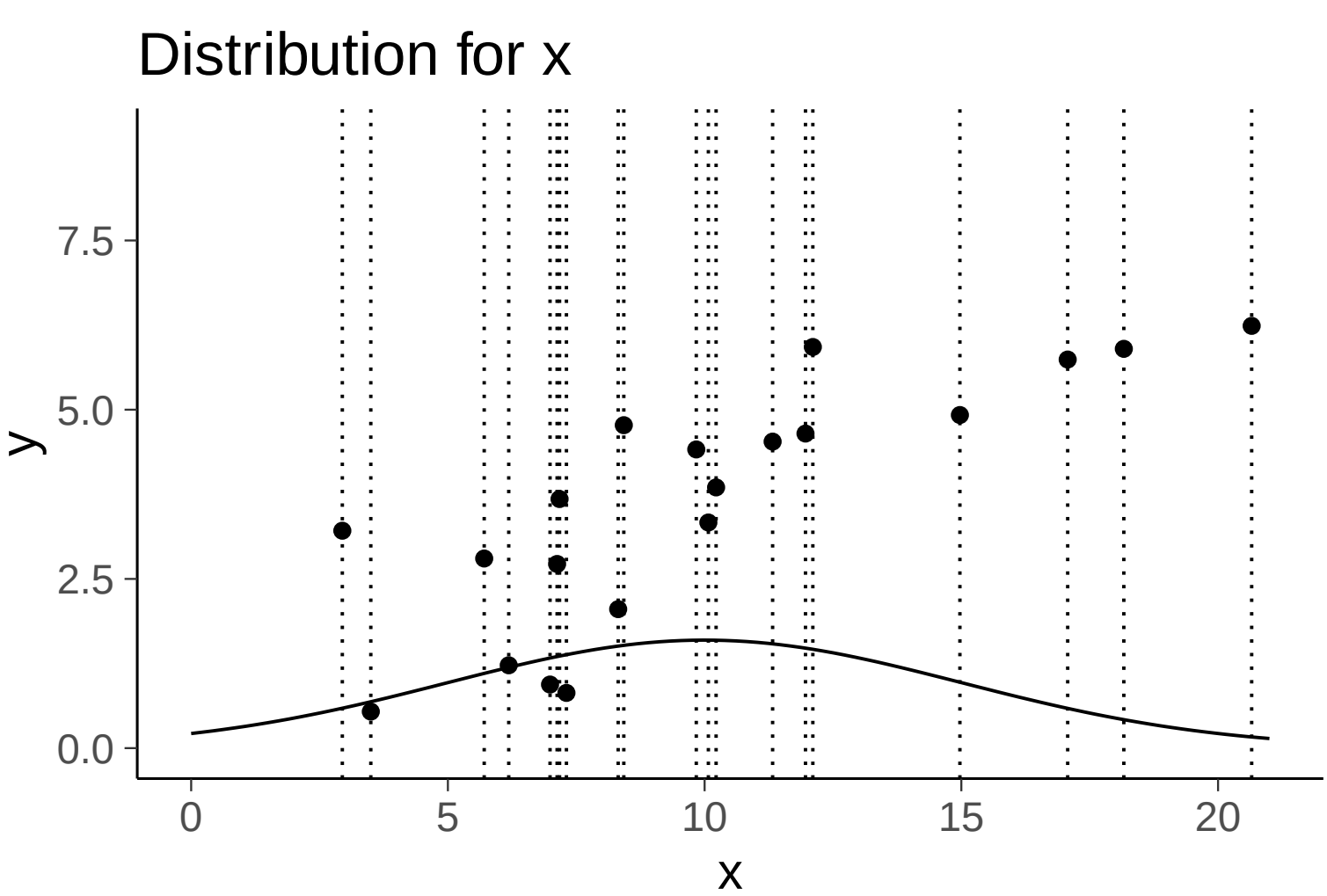
LOO is ok for fixed / designed x. SE is uncertainty about $y|x$.

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/



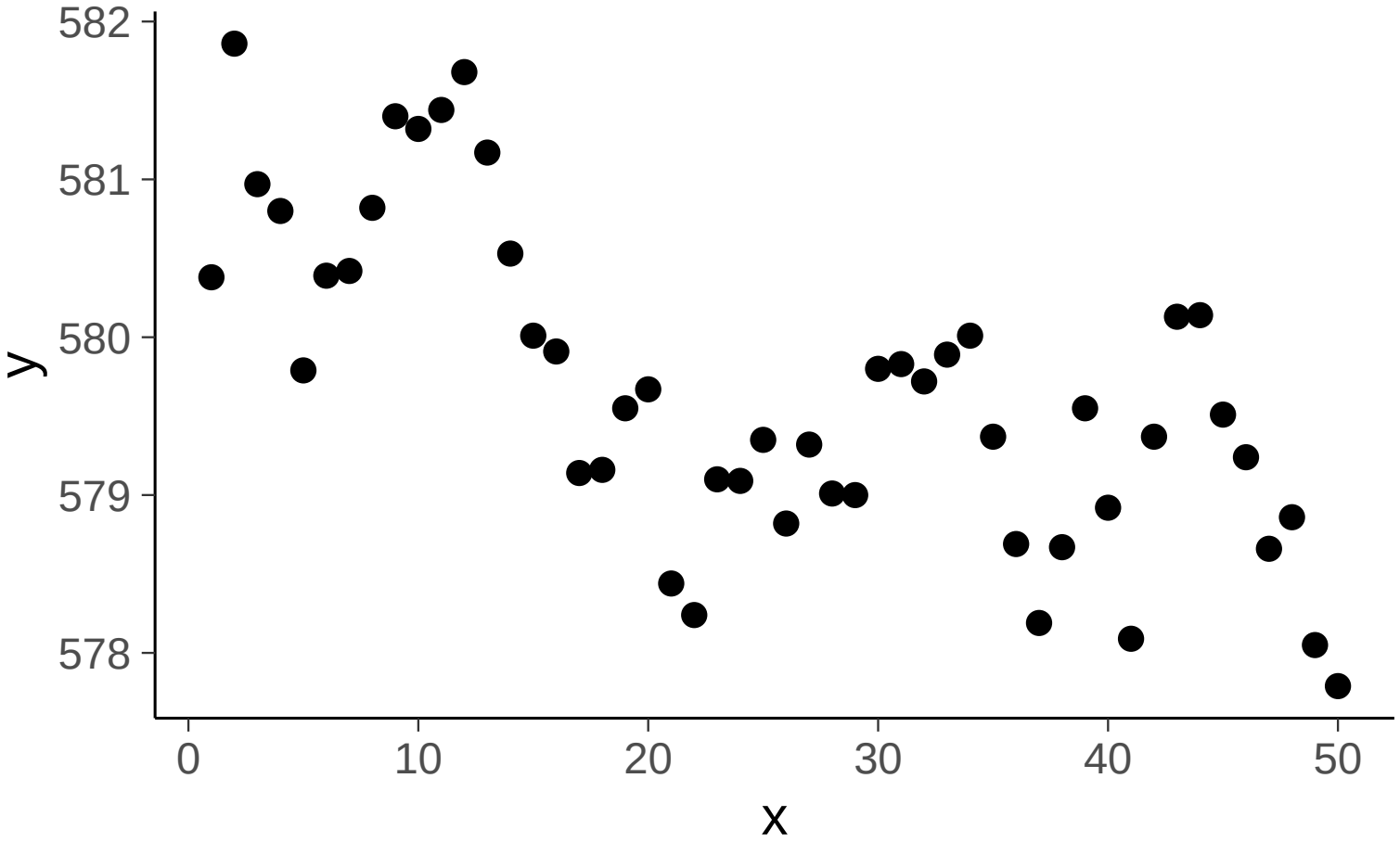
LOO is ok for random x. SE is uncertainty about $y|x$ and x.

Covariate shift can be handled with importance weighting or modelling

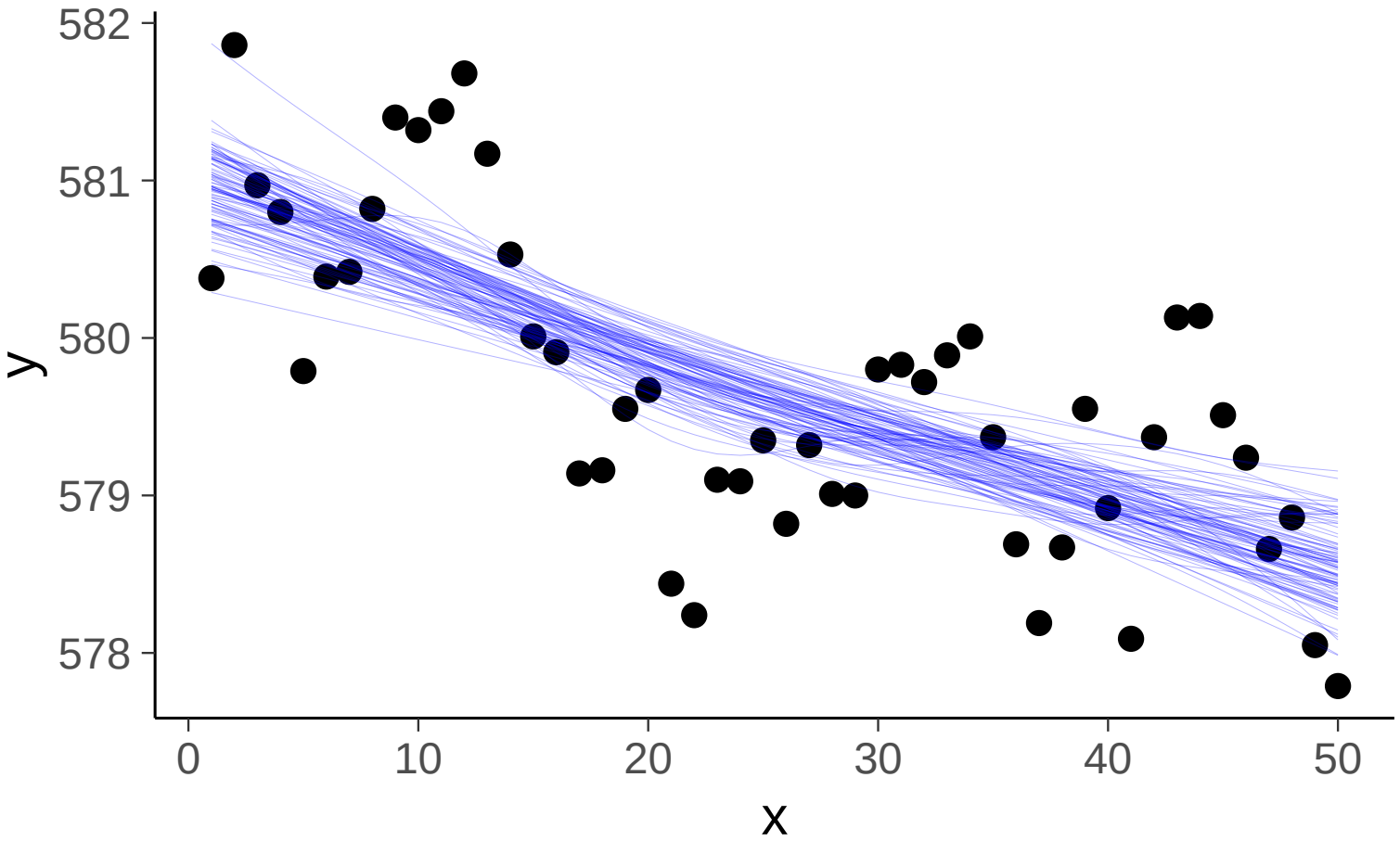
see Vehtari & Ojanen (2012) and
andrewgelman.com/2018/08/03/loo-cross-validation-approaches-valid/

loo package

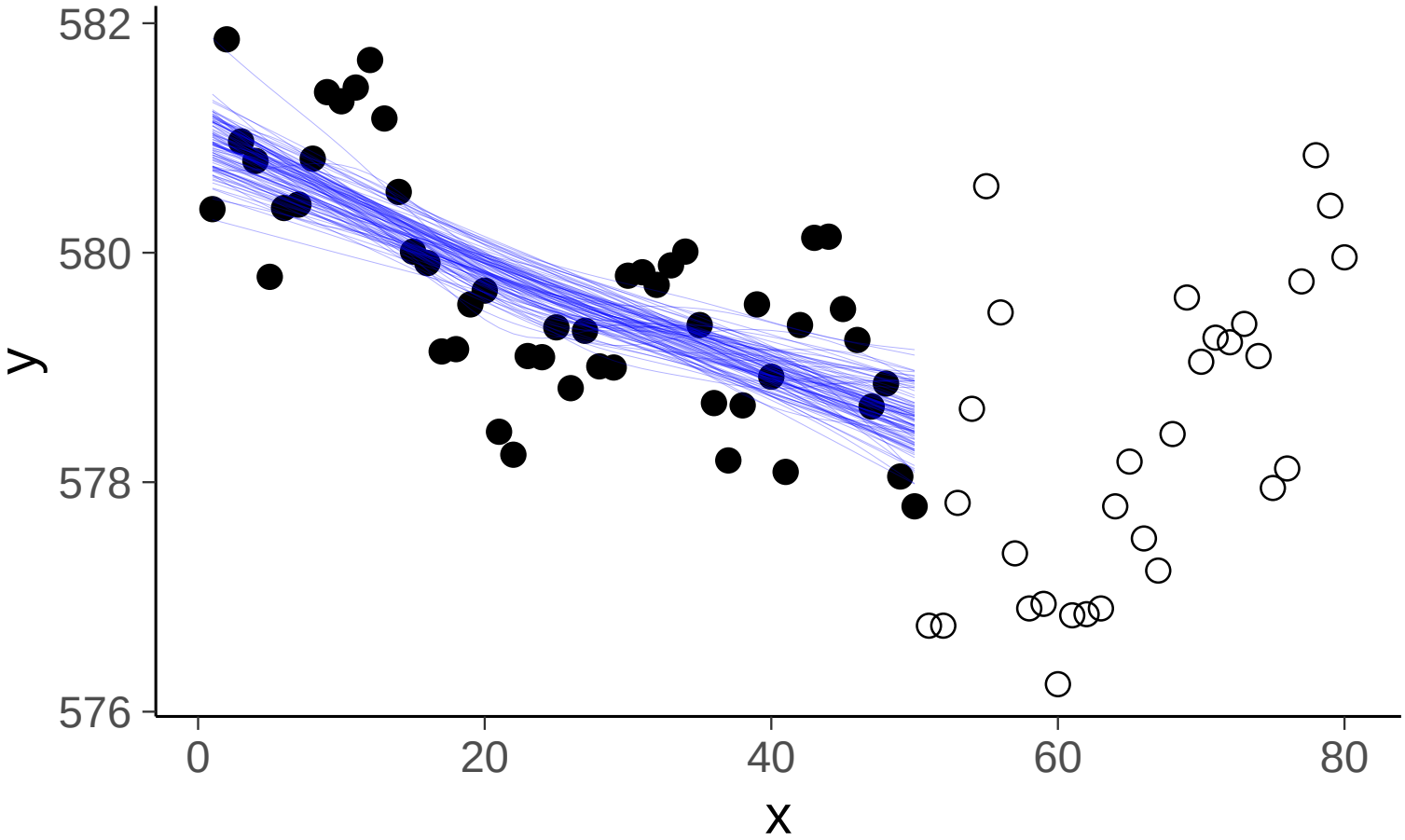
```
 Computed from 4000 by 20 log-likelihood matrix

          Estimate  SE
elpd_loo    -29.5 3.3
p_loo         2.7 1.0
looic        58.9 6.7

------

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

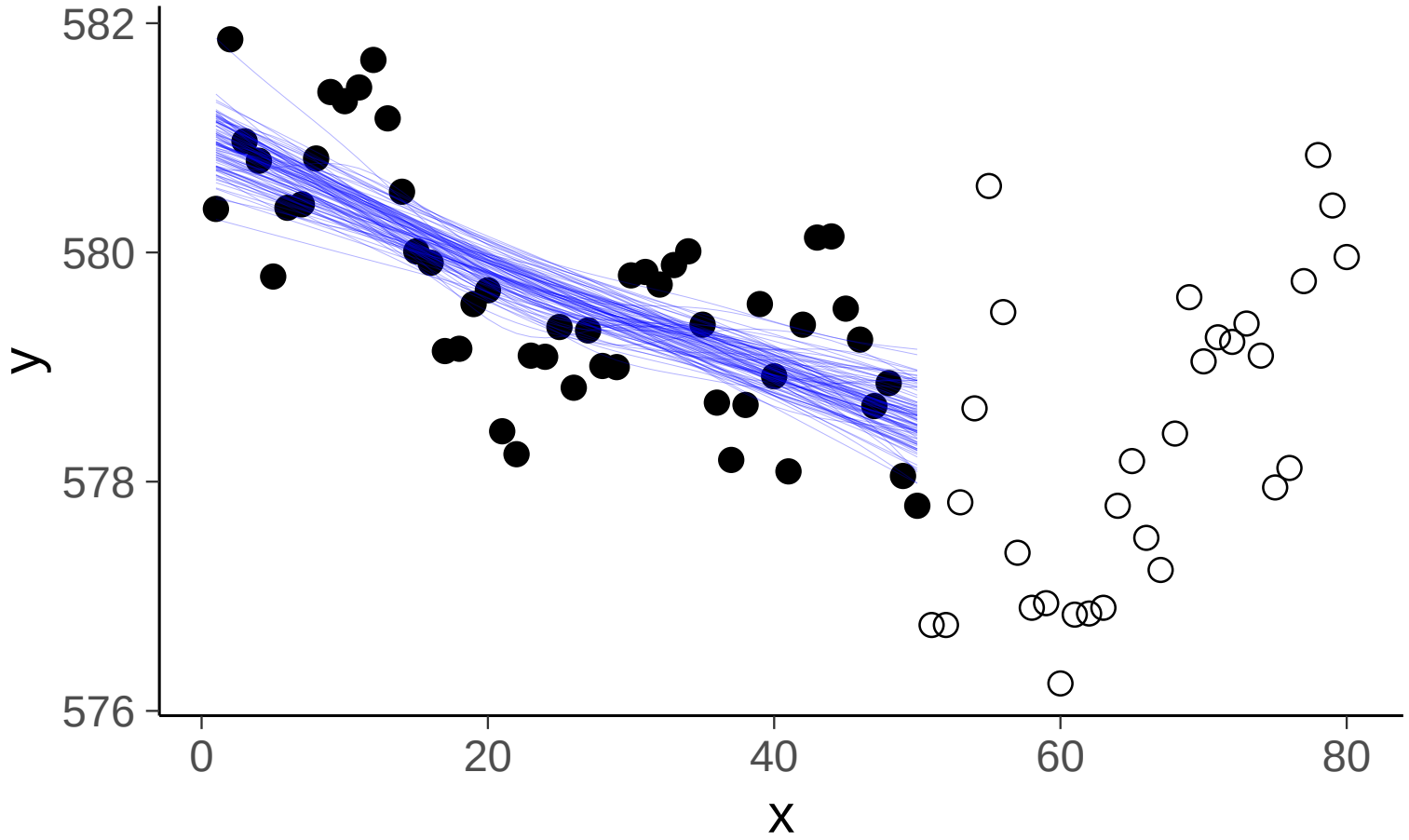
Nonlinear model fit



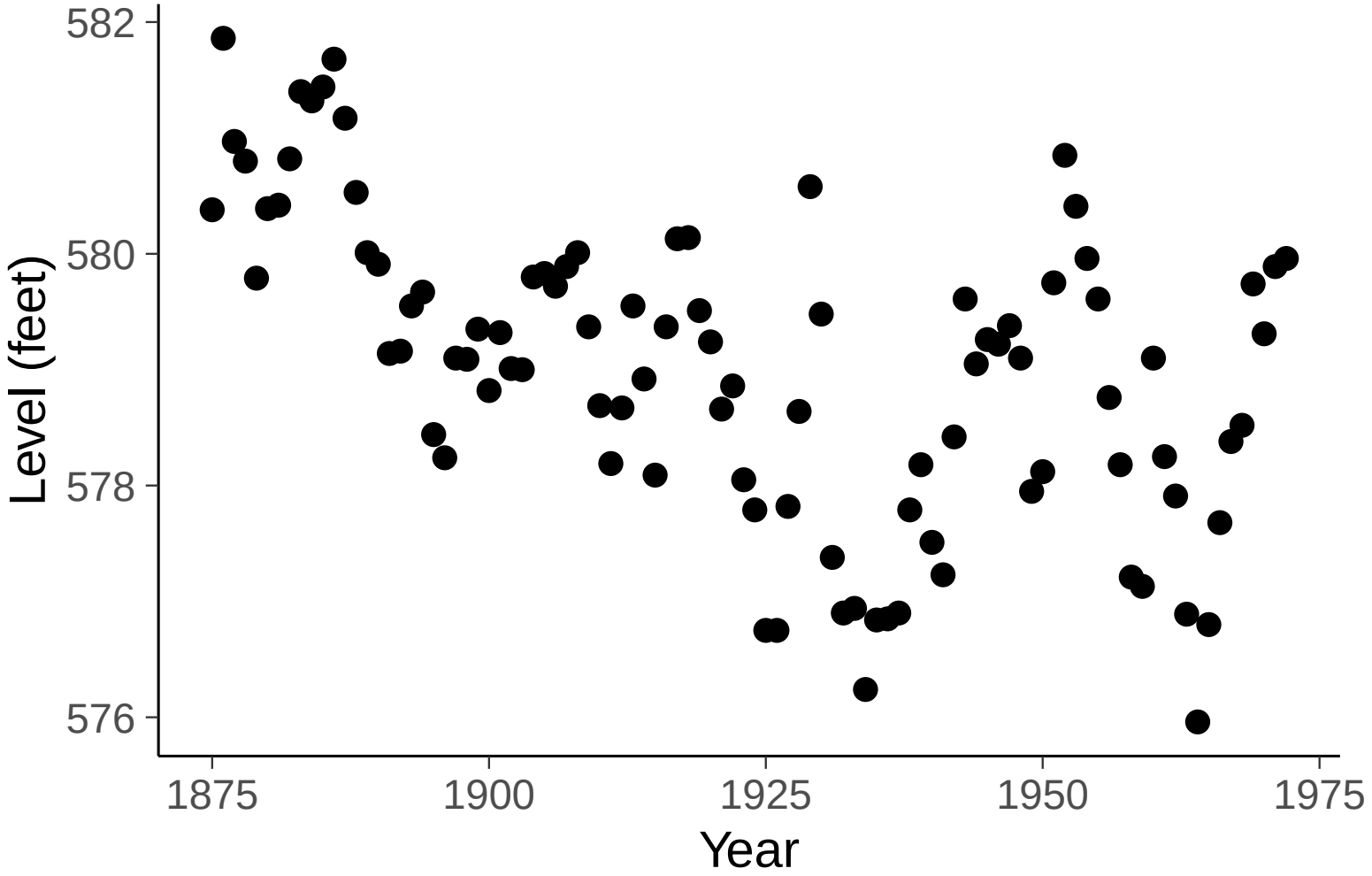
Nonlinear model fit + new data



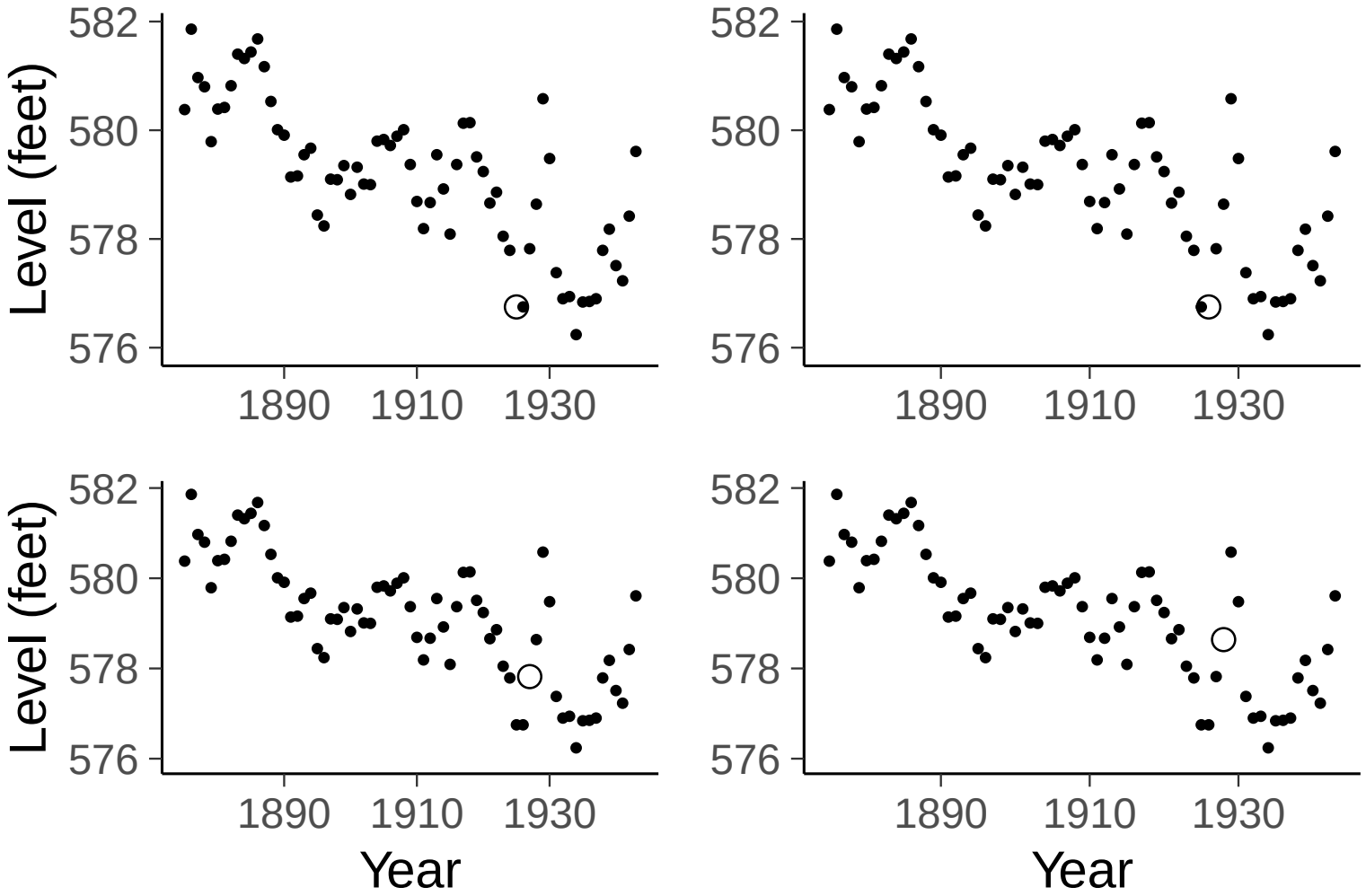
Nonlinear model fit + new data



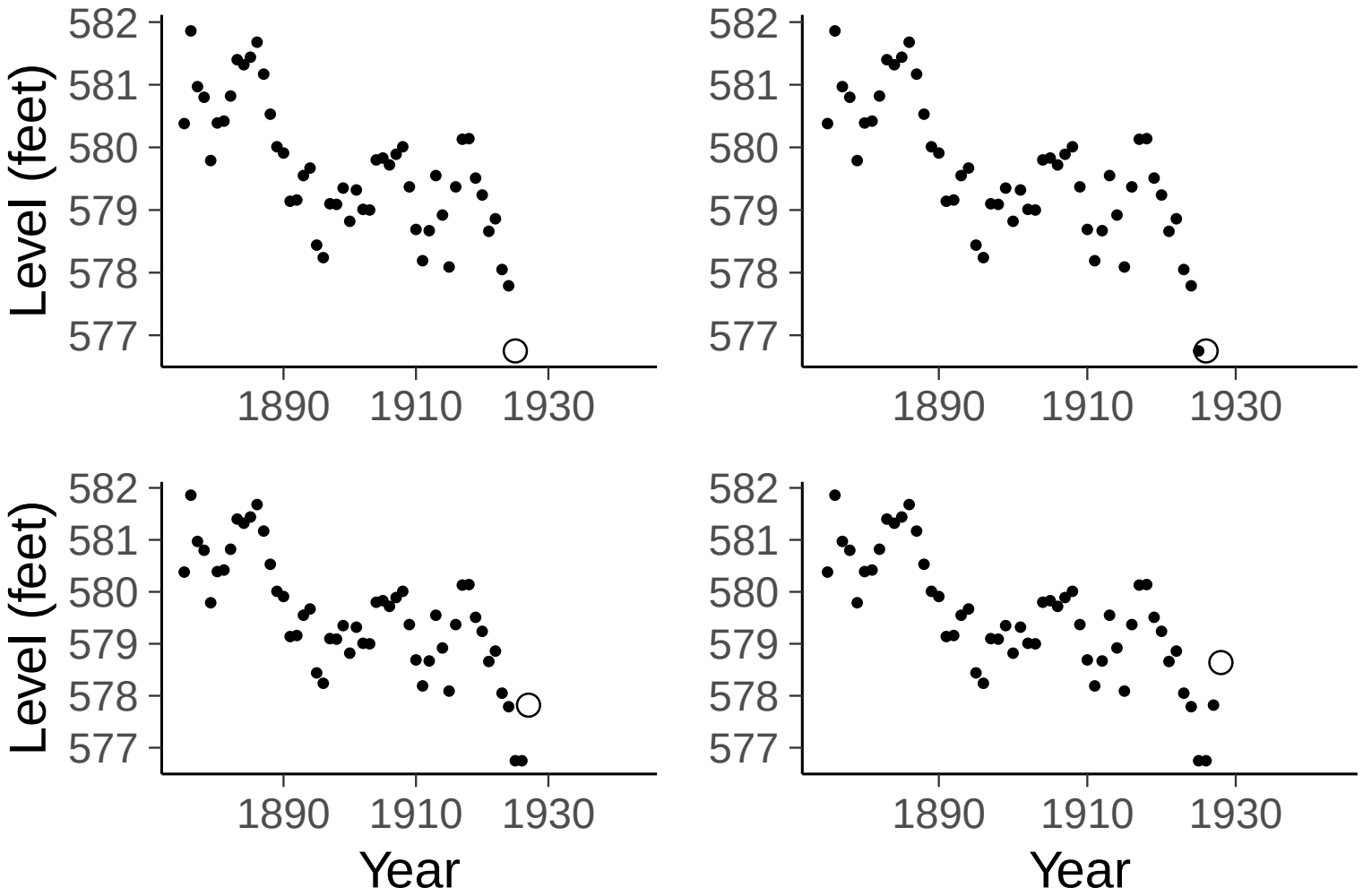
Extrapolation is more difficult



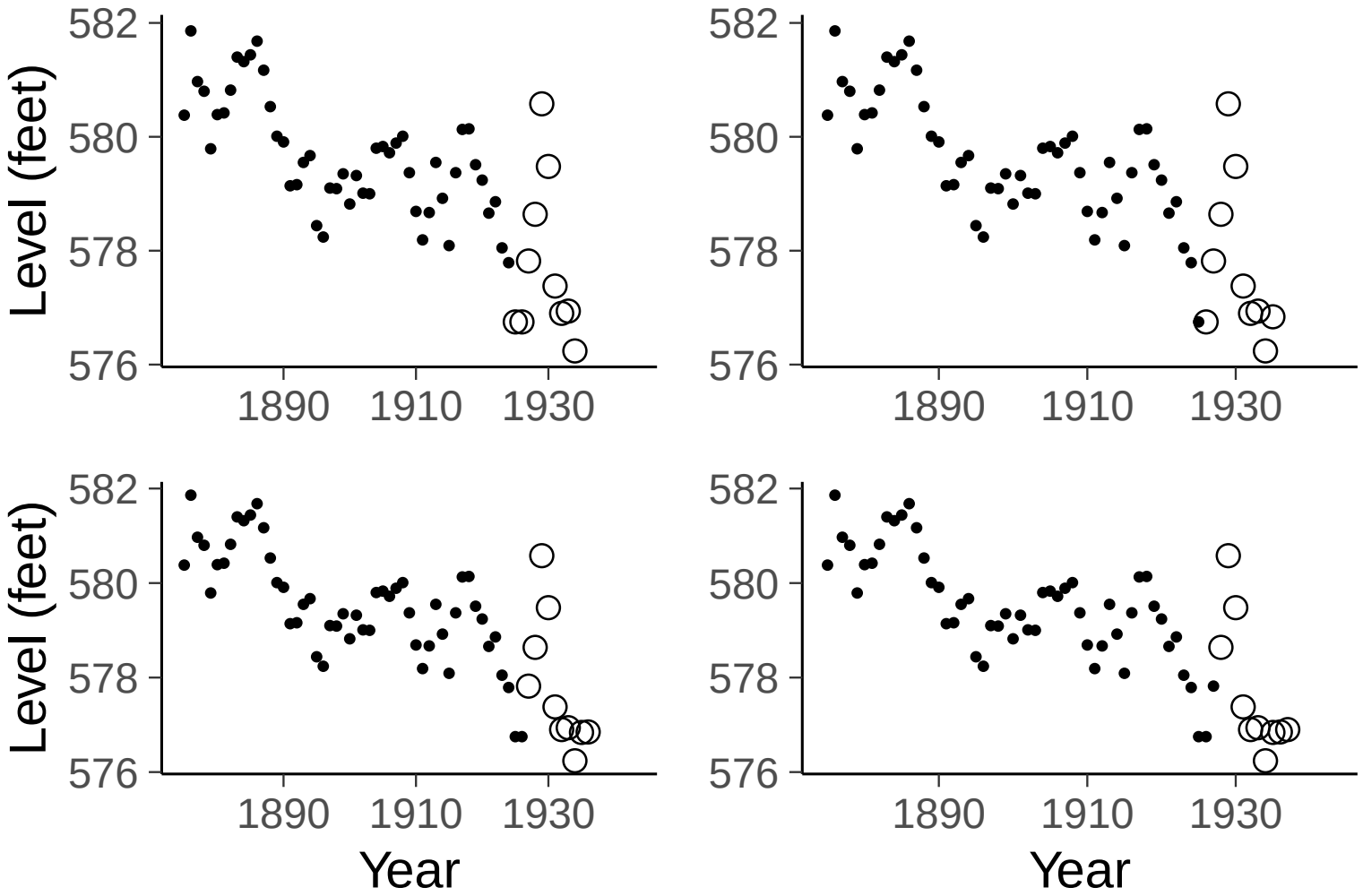
Can LOO or other cross-validation be used with time series?



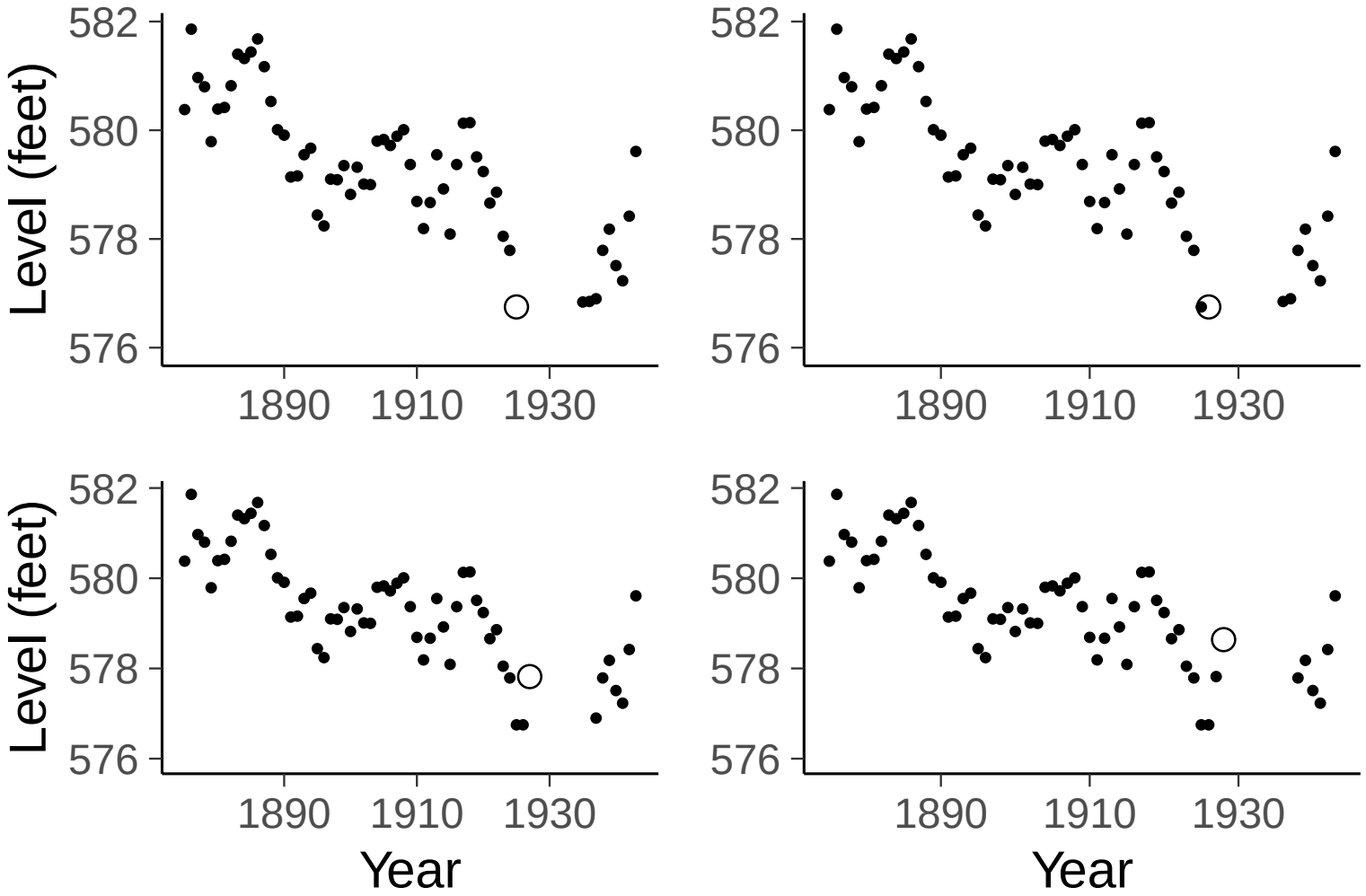
Leave-one-out cross-validation is ok for assessing conditional model



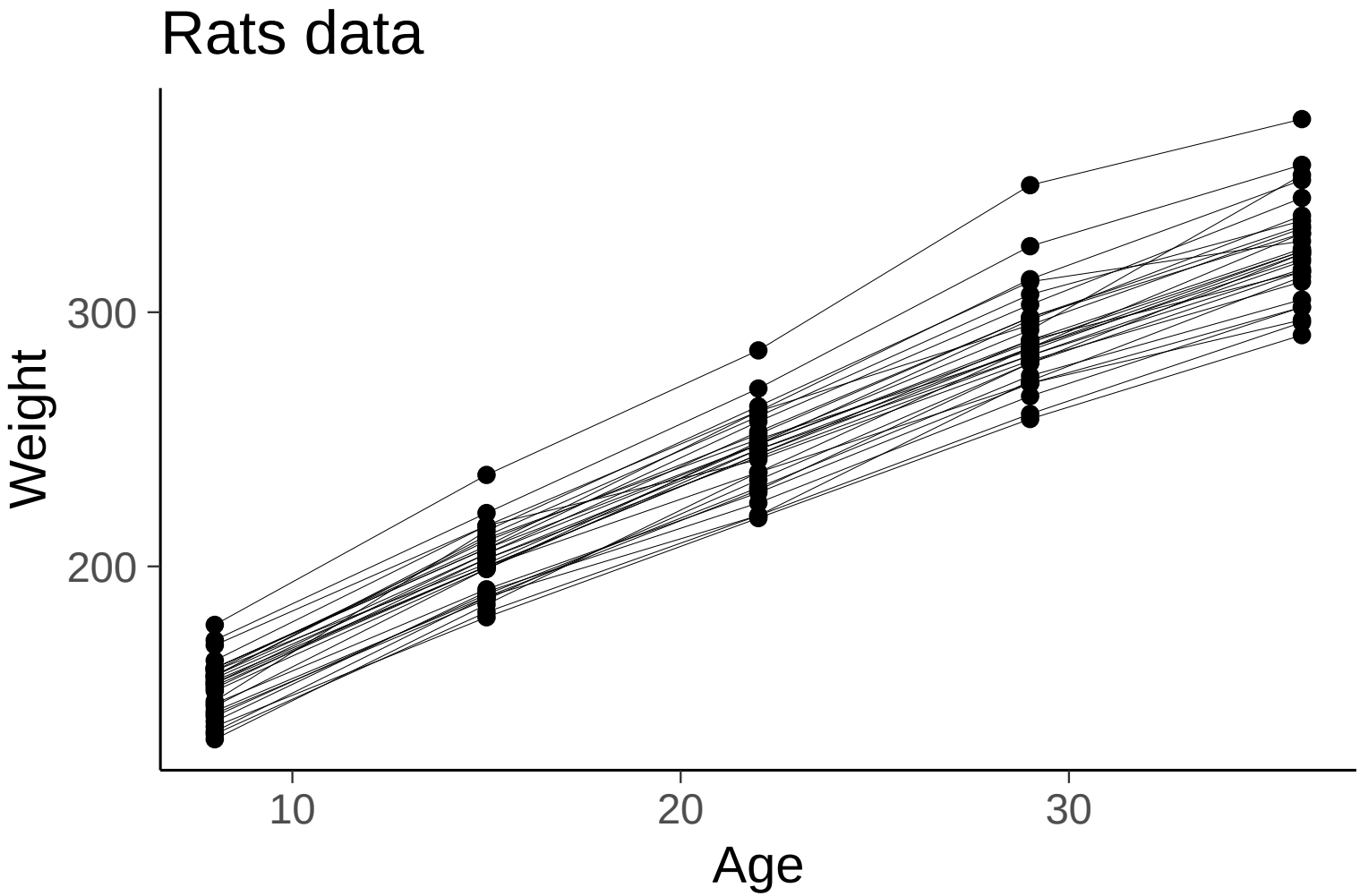
1-step-ahead cross-validation is better for predicting future



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

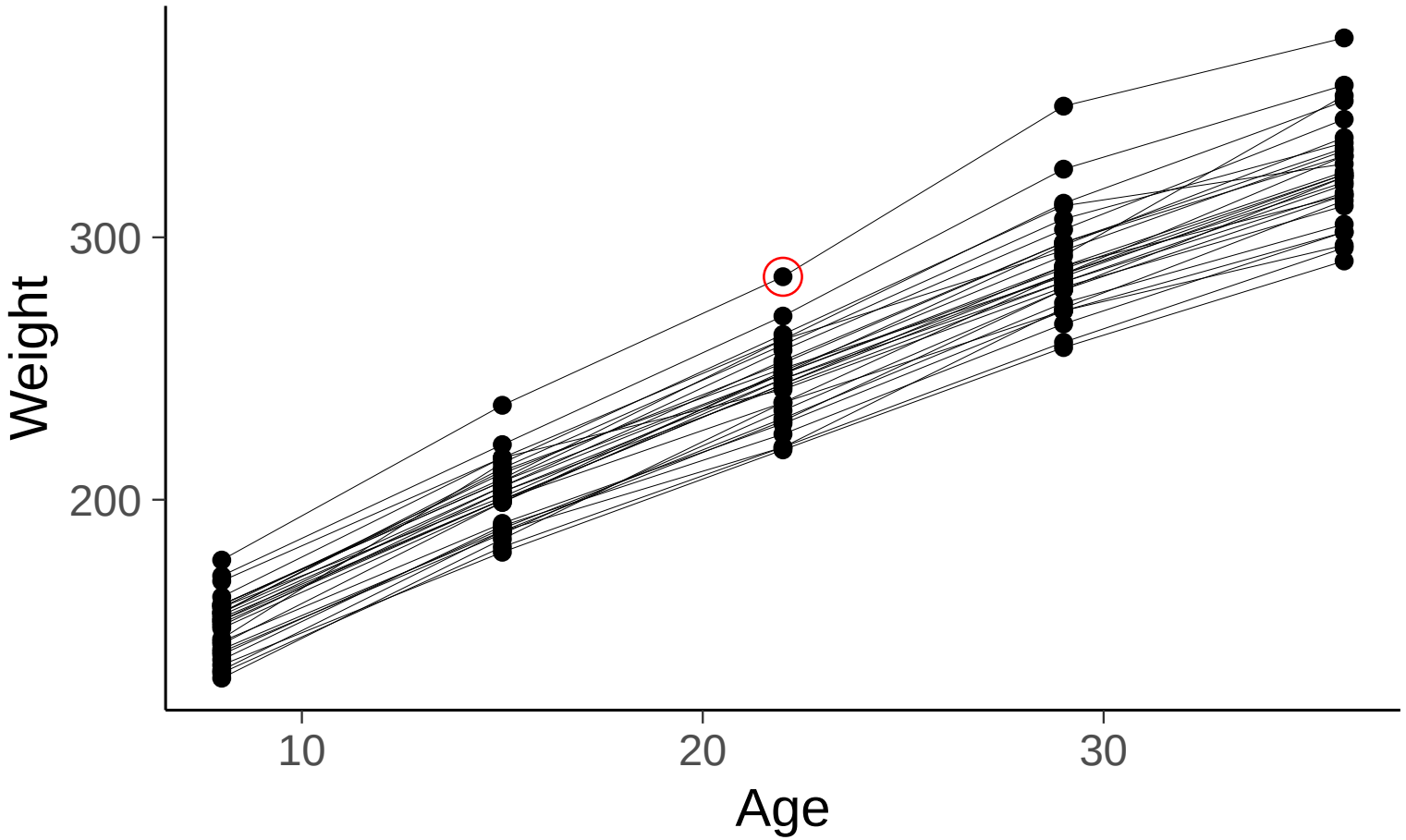


m-step-ahead leave-a-block-out cross-validation



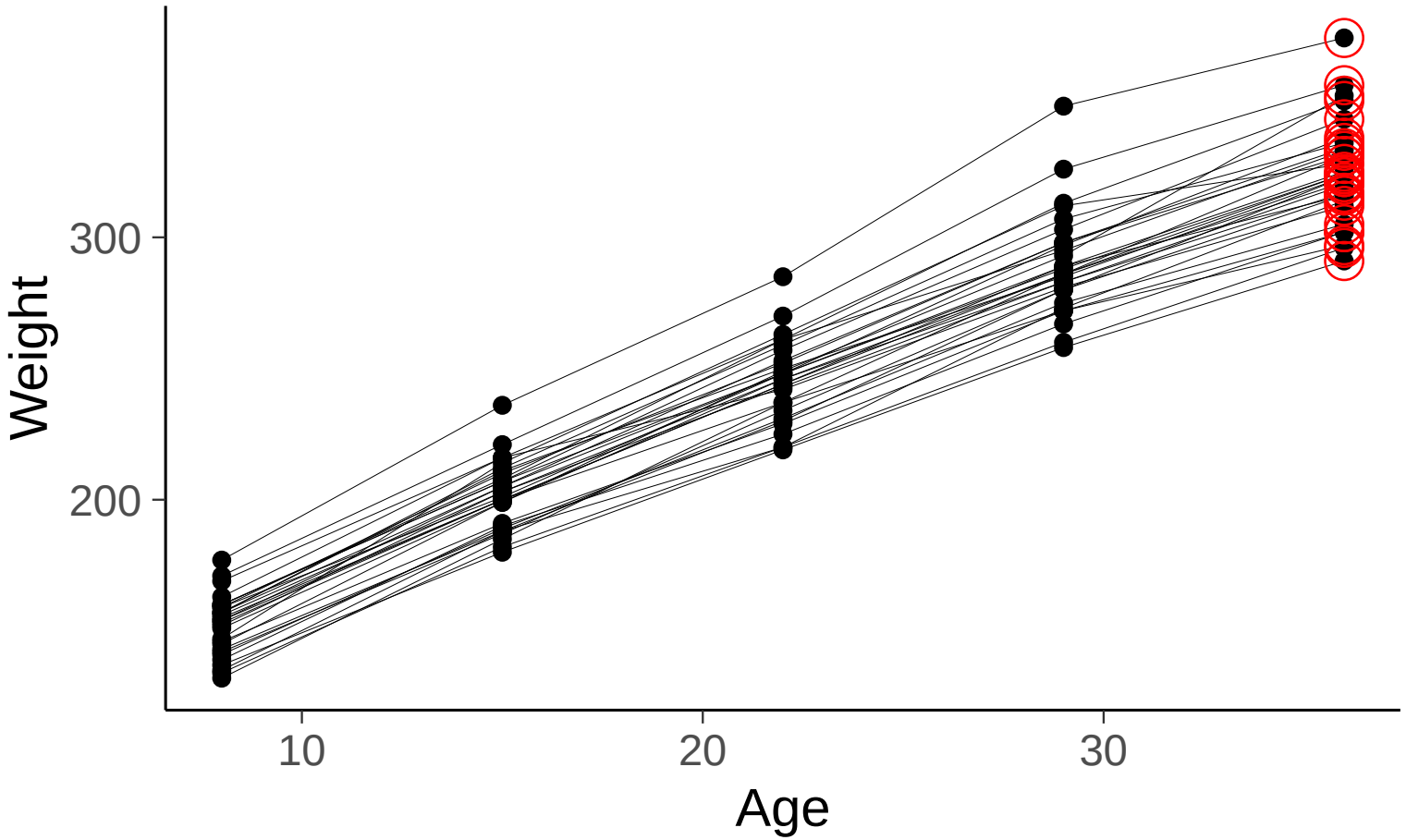
Can LOO or other cross-validation be used with hierarchical data?

Leave−one−out?



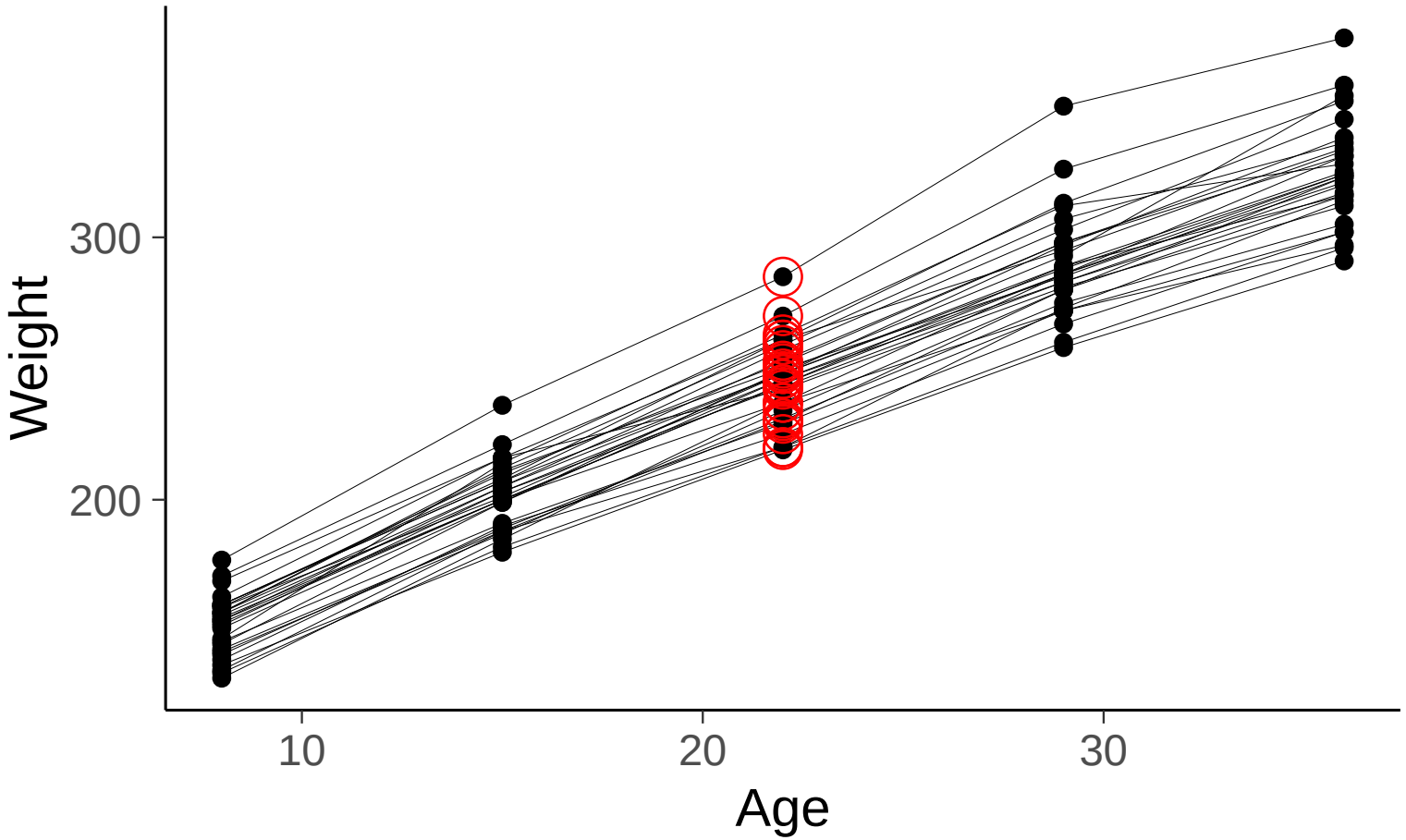
Yes!

1−step−ahead?



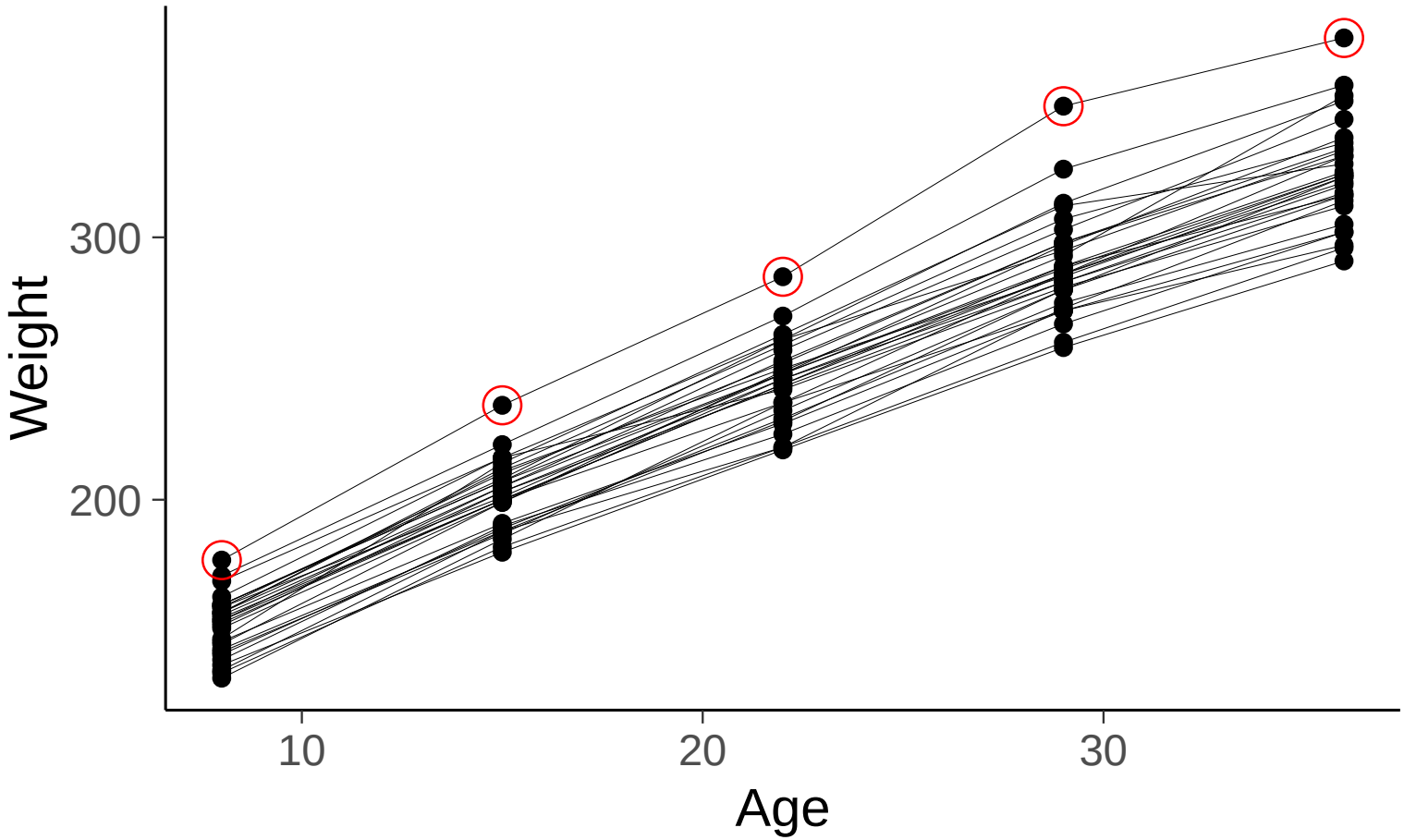
Yes!

Leave−one−time−point−out?



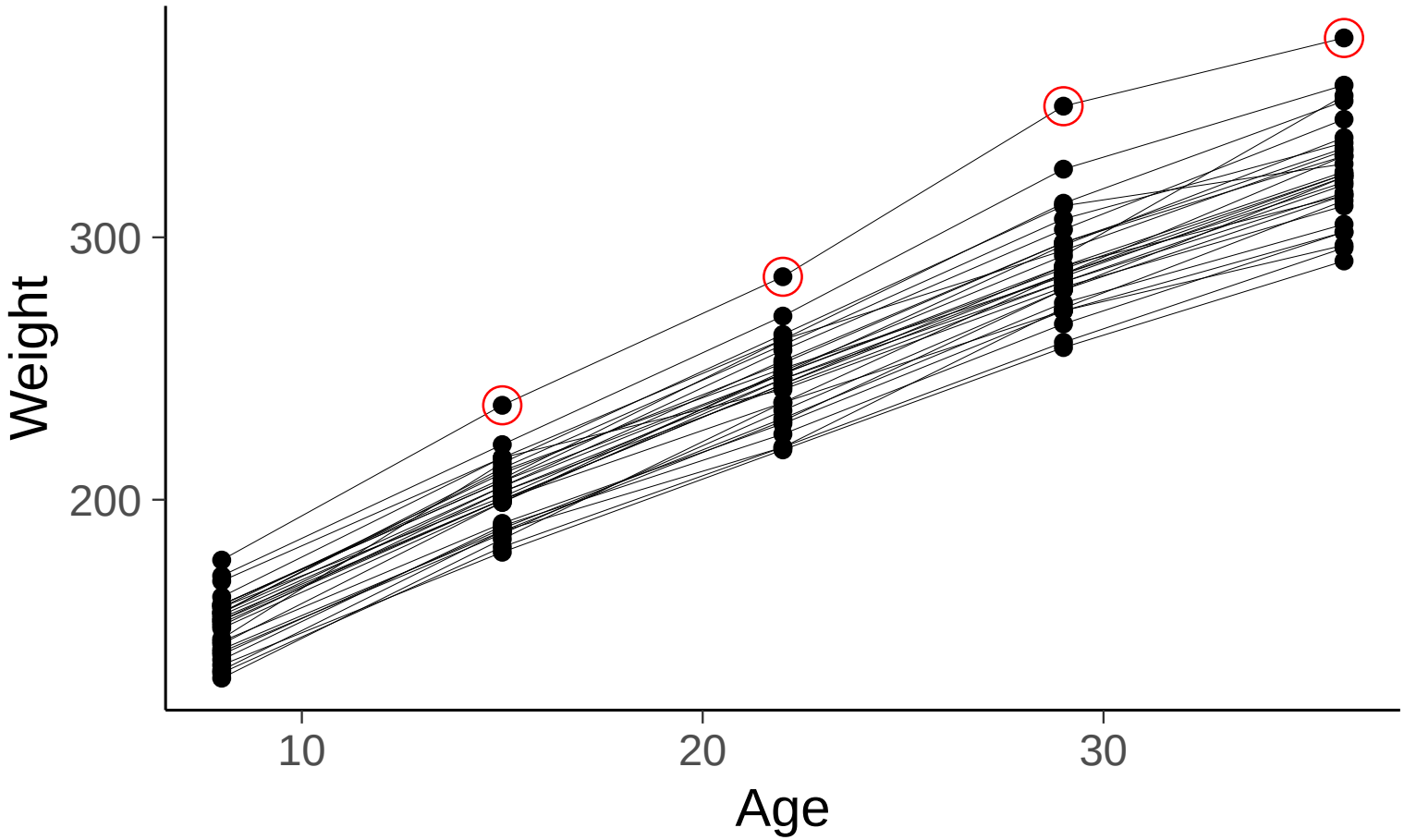
Yes!

Leave−one−rat−out?



Yes!

Predict given initial weight?



Yes!

Summary of data generating mechanisms and prediction tasks

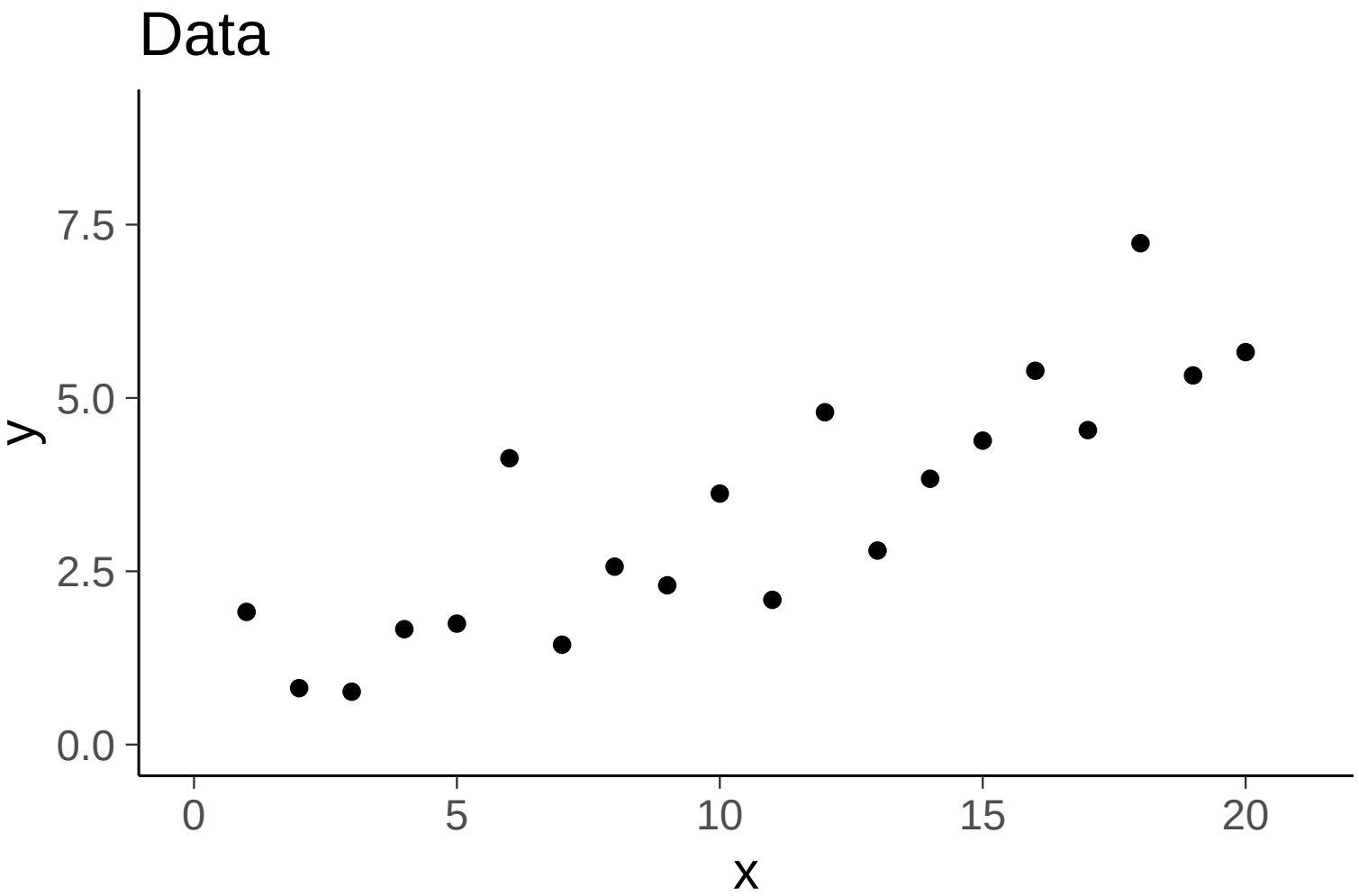
- You have to make some assumptions on data generating mechanism
- Use the knowledge prediction task if available
- Cross-validation can be used to analyse different parts, even if there is no clear prediction task

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

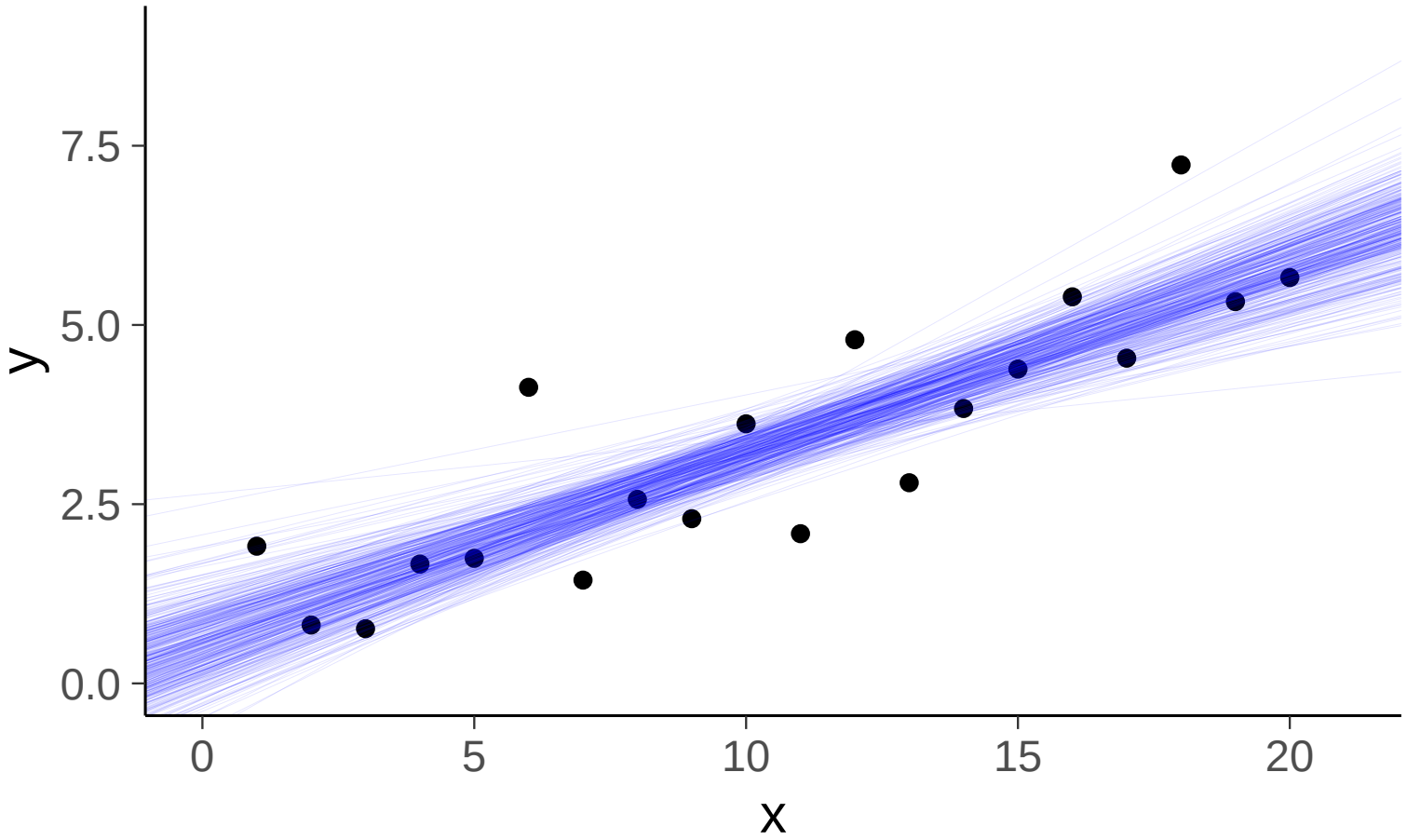
Fast cross-validation

- Pareto smoothed importance sampling LOO
- K-fold cross-validation

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

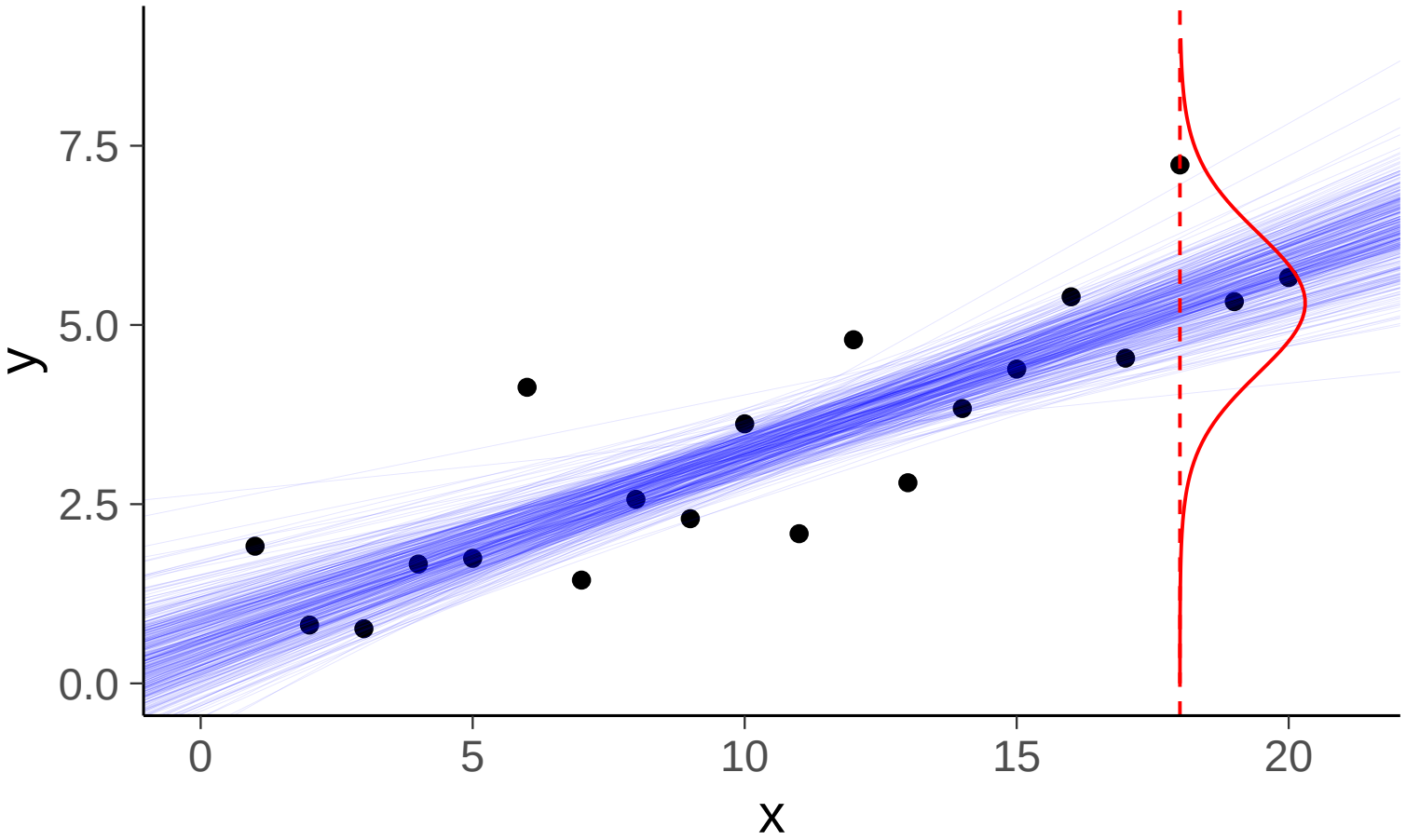

Data
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

Posterior draws



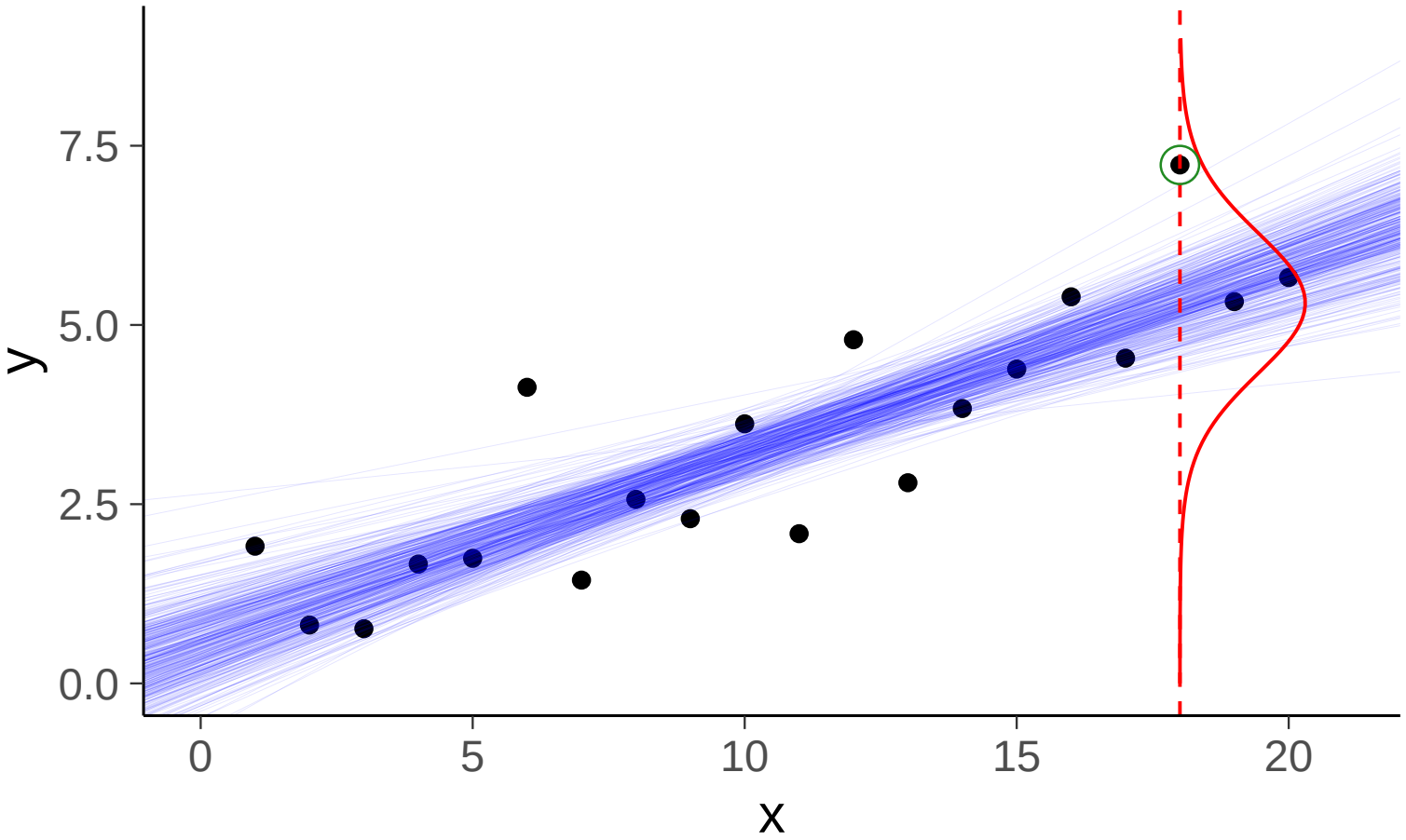
$$\theta^{(s)} \sim p(\theta|x, y)$$

Posterior predictive distribution



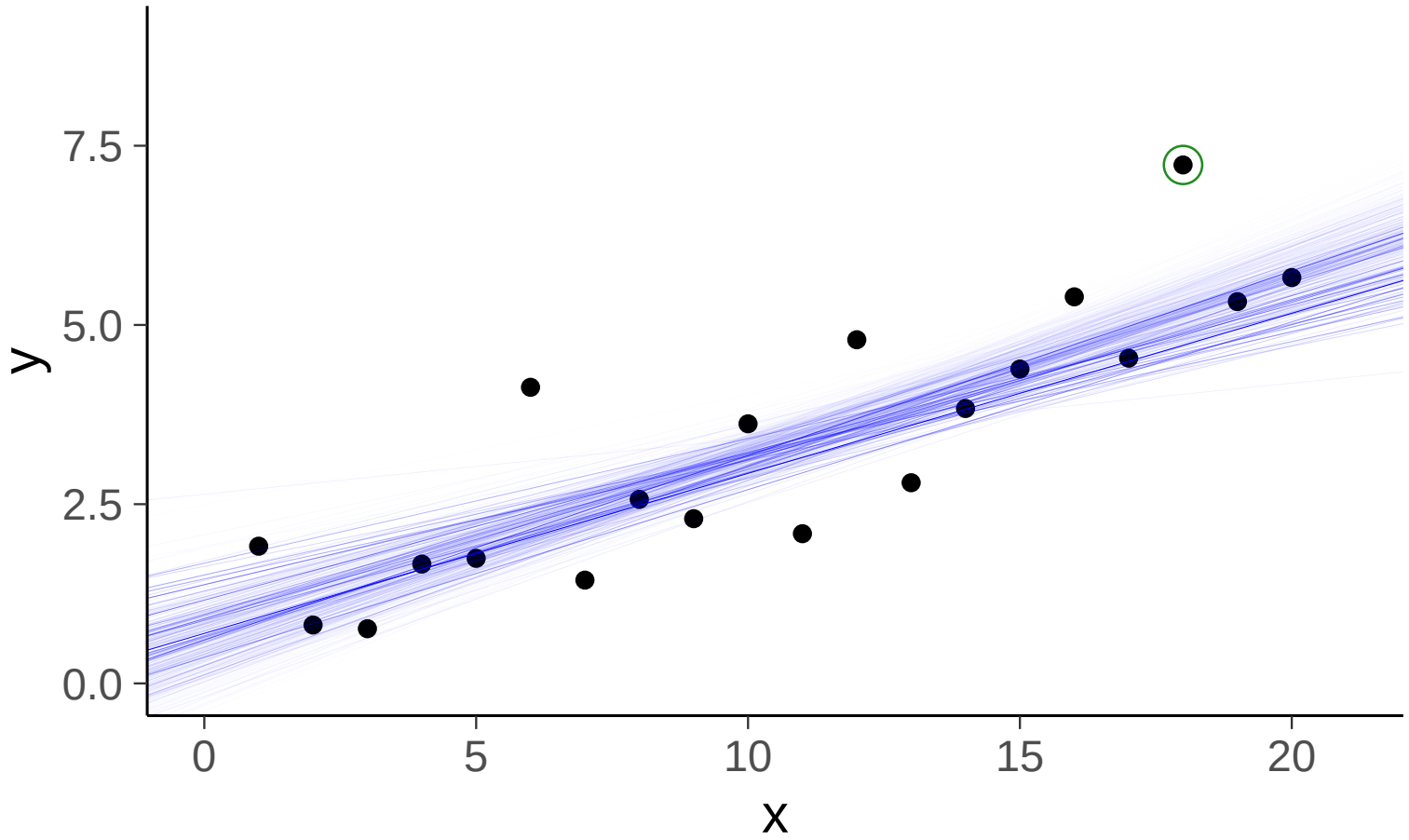
$$\theta^{(s)} \sim p(\theta|x,y), \quad p(\tilde{y}|\tilde{x},x,y) \approx \sum_{s=1}^{S} p(\tilde{y}|\tilde{x},\theta^{(s)})$$

Posterior predictive distribution



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

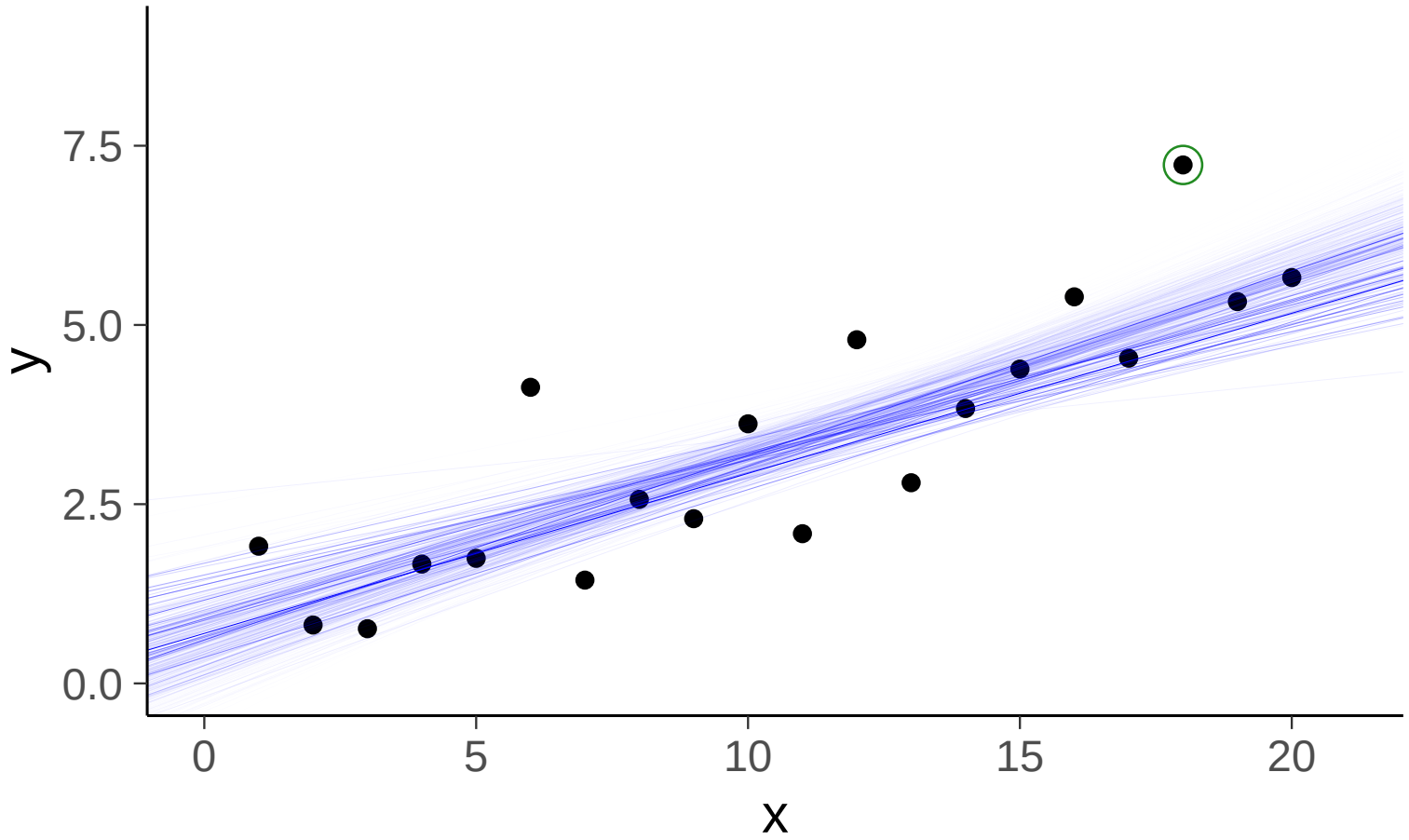
PSIS−LOO weighted draws



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

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

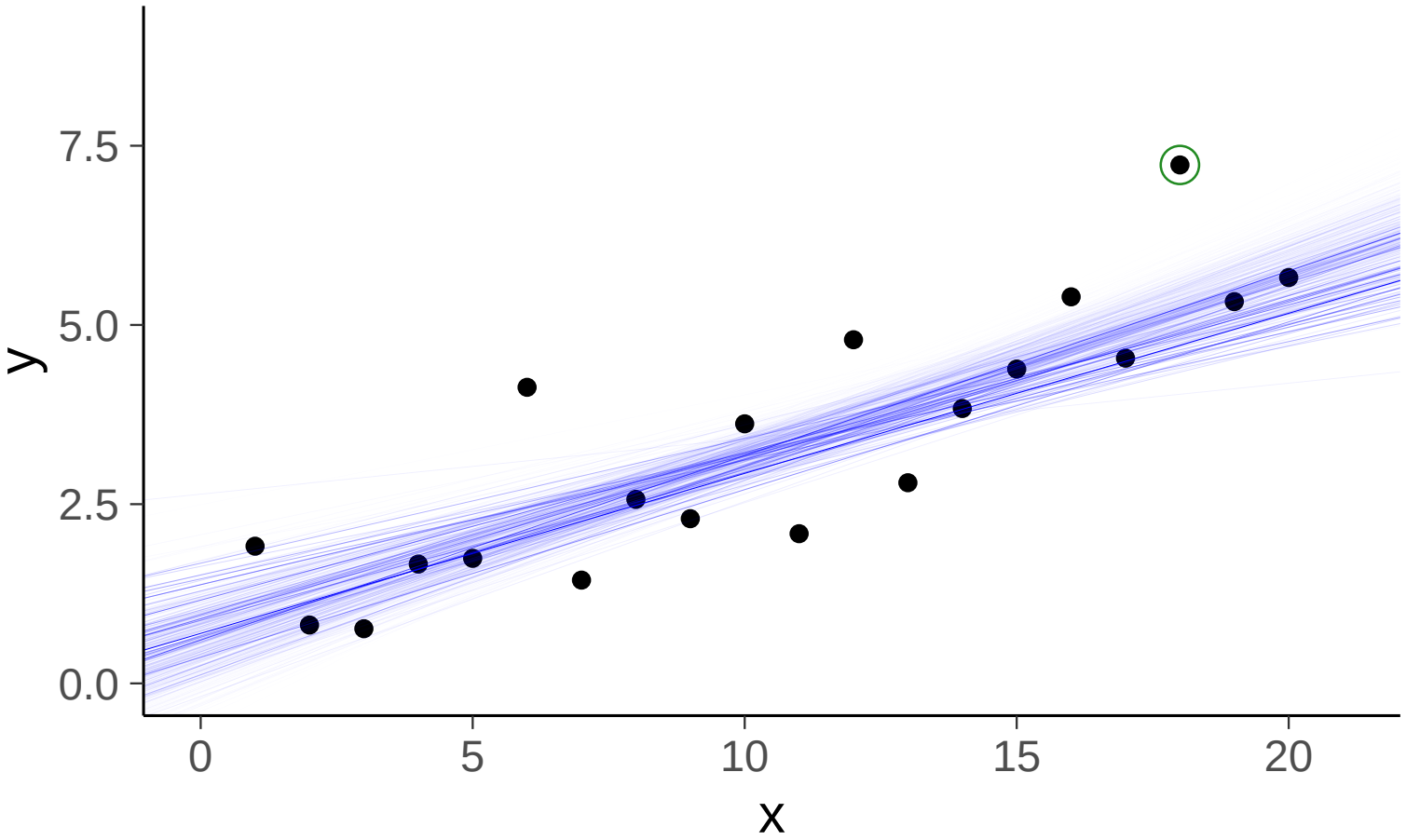
PSIS−LOO weighted draws



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

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

PSIS−LOO weighted draws

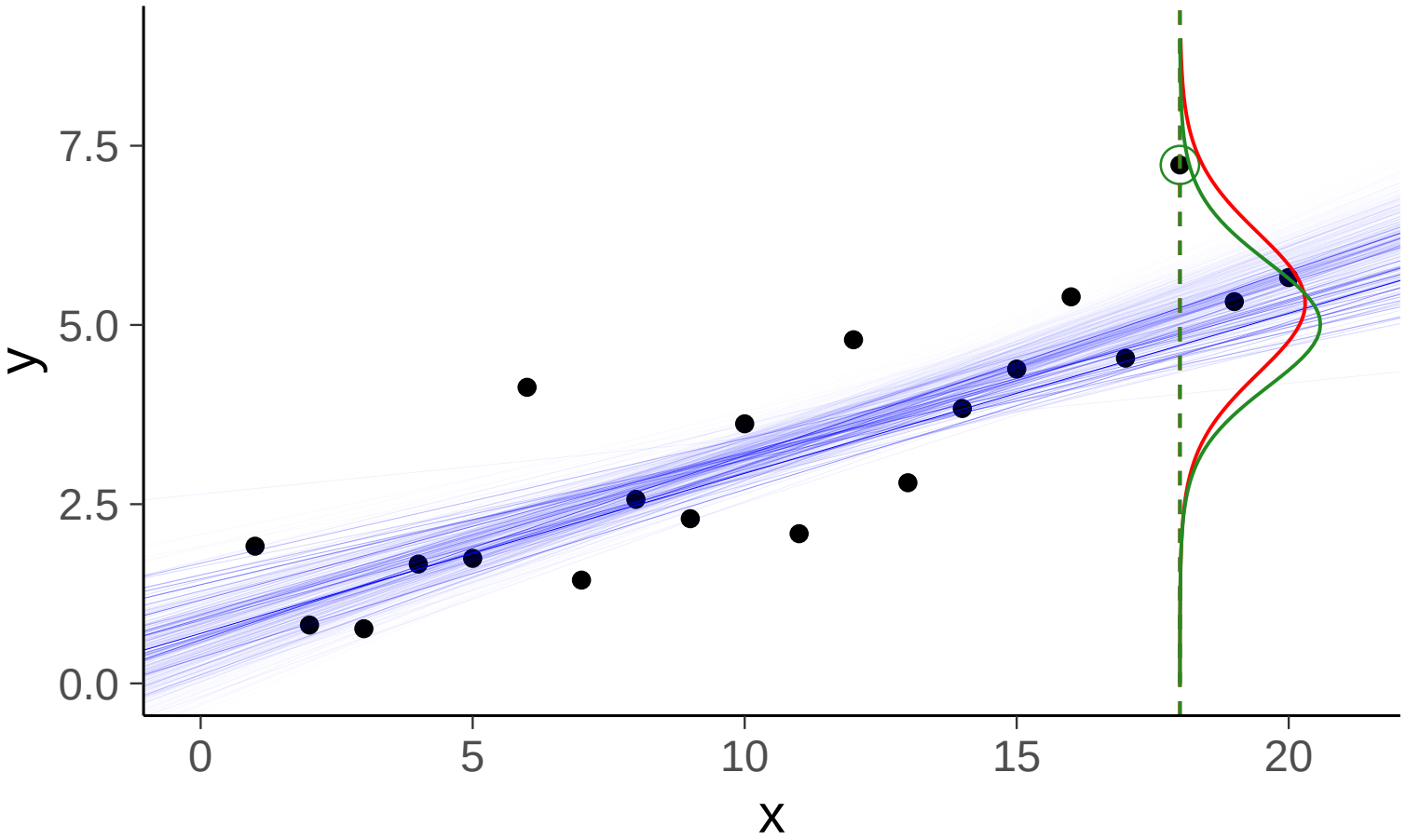


$\theta^{(s)} \sim p(\theta|x,y)$

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

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

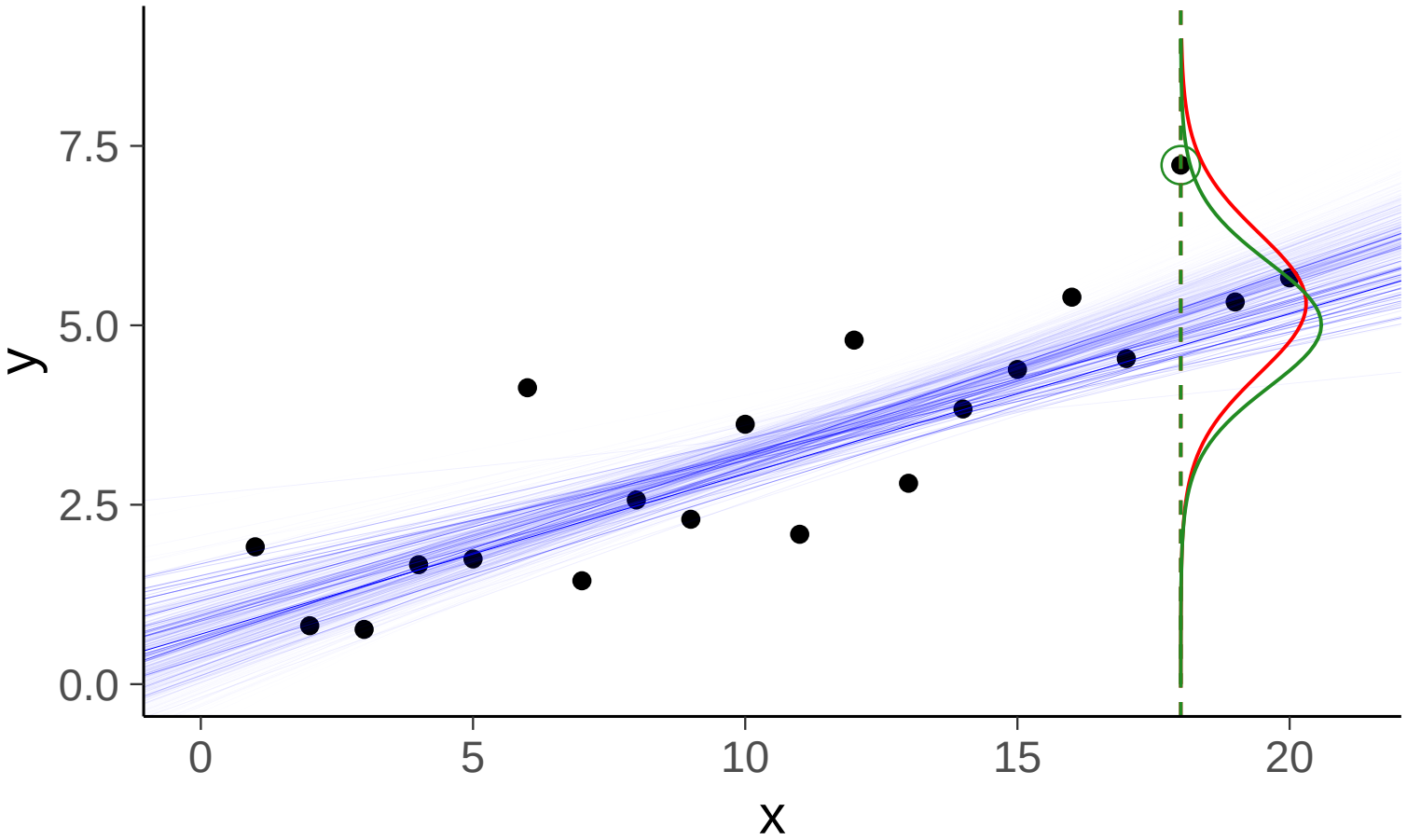
PSIS−LOO weighted predictive distribution



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

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

PSIS−LOO weighted predictive distribution

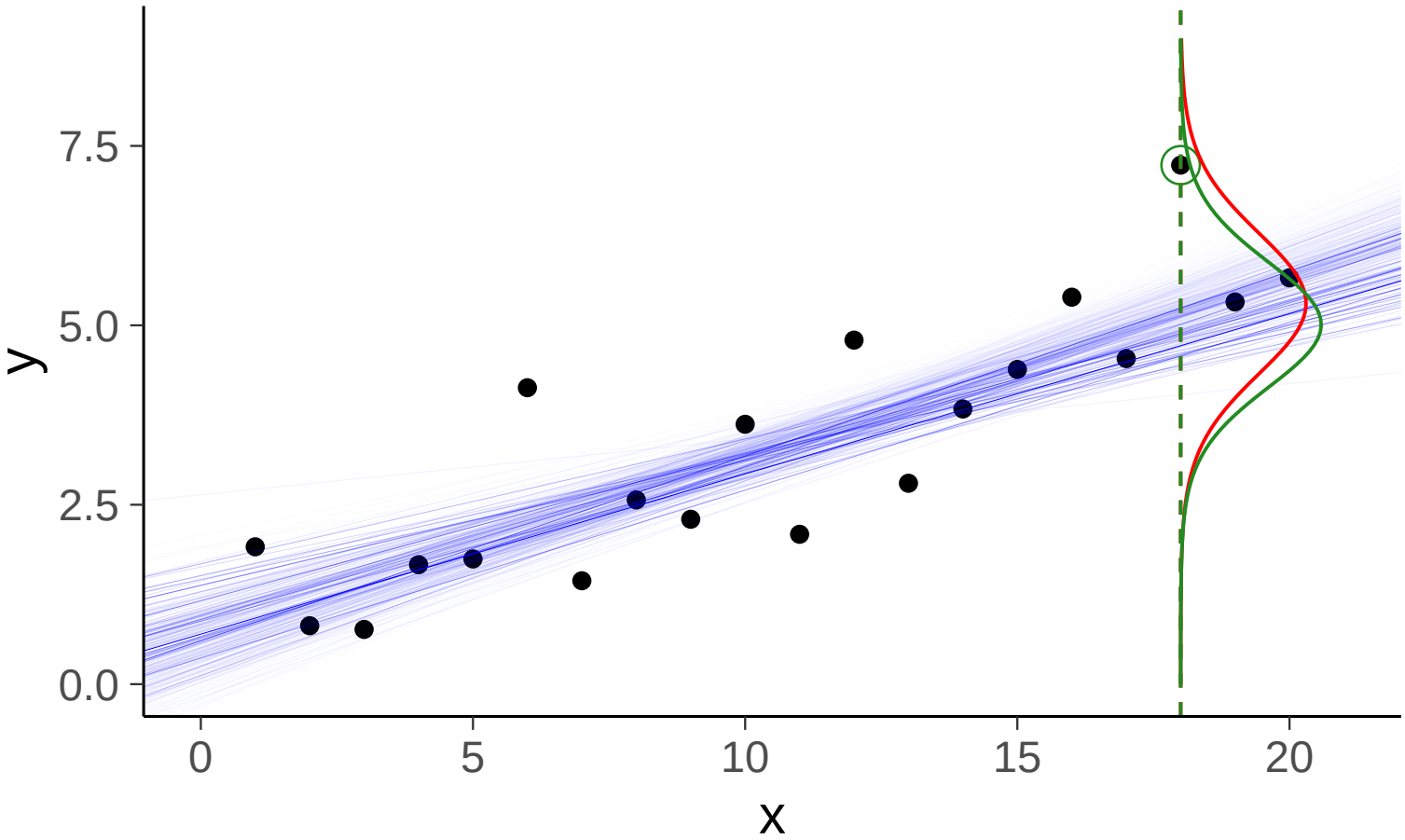


$$\theta^{(s)} \sim p(\theta|x,y)$$

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

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

PSIS−LOO weighted predictive distribution

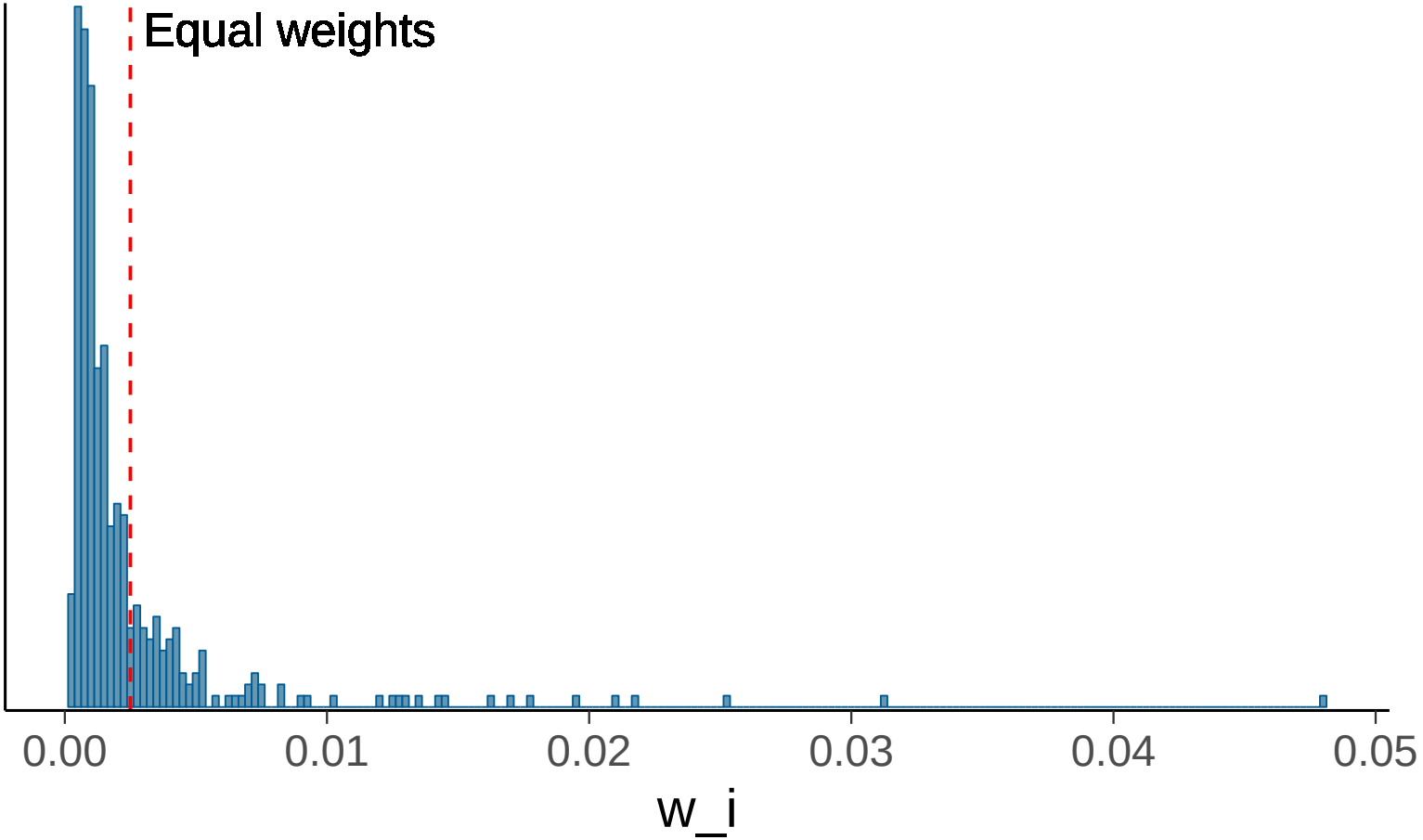


$\theta^{(s)} \sim p(\theta|x,y)$

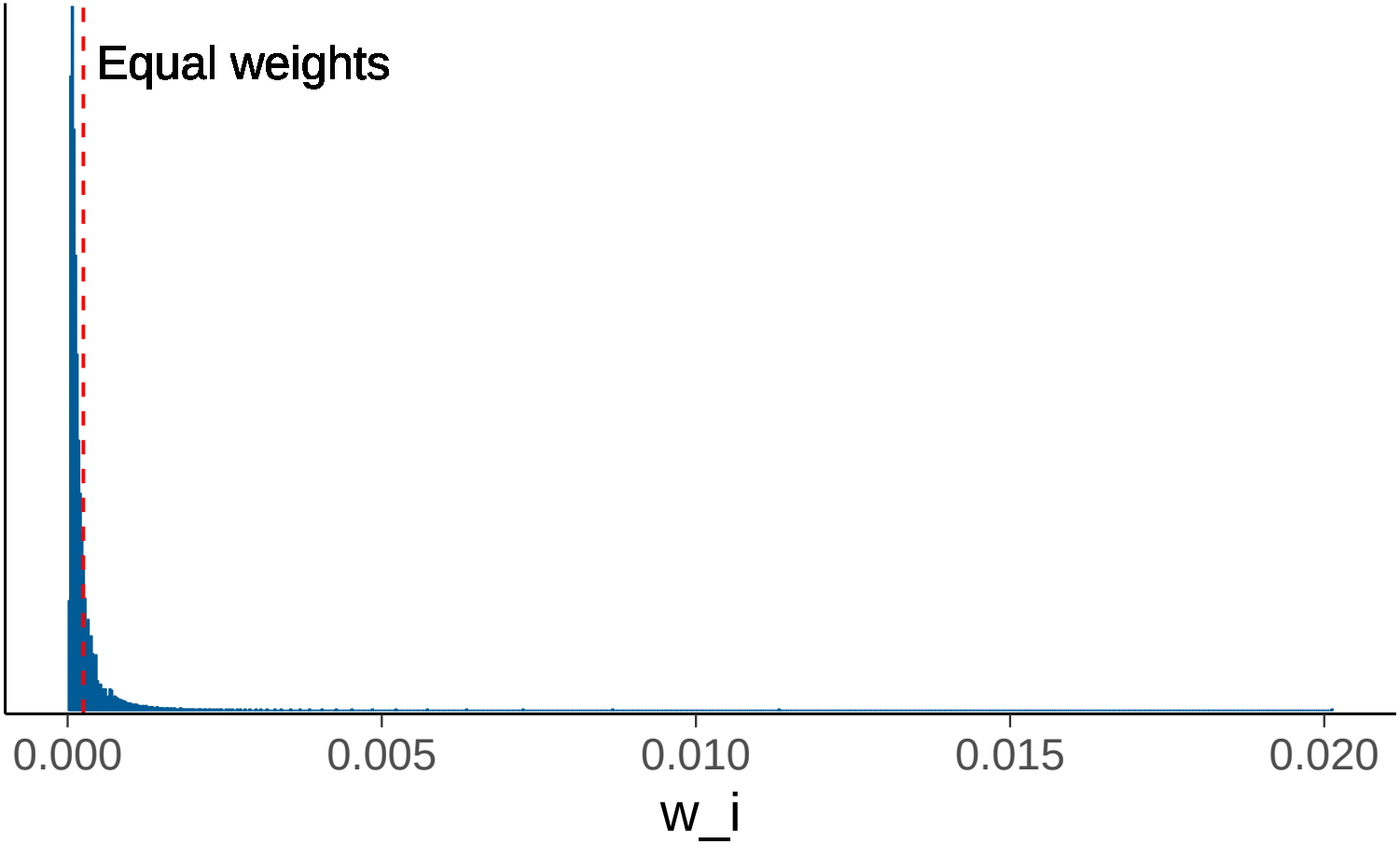
$r_i^{(s)} = p(\theta^{(s)}|x_{-i},y_{-i})/p(\theta^{(s)}|x,y) \propto 1/p(y_i|x_i,\theta^{(s)})$

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

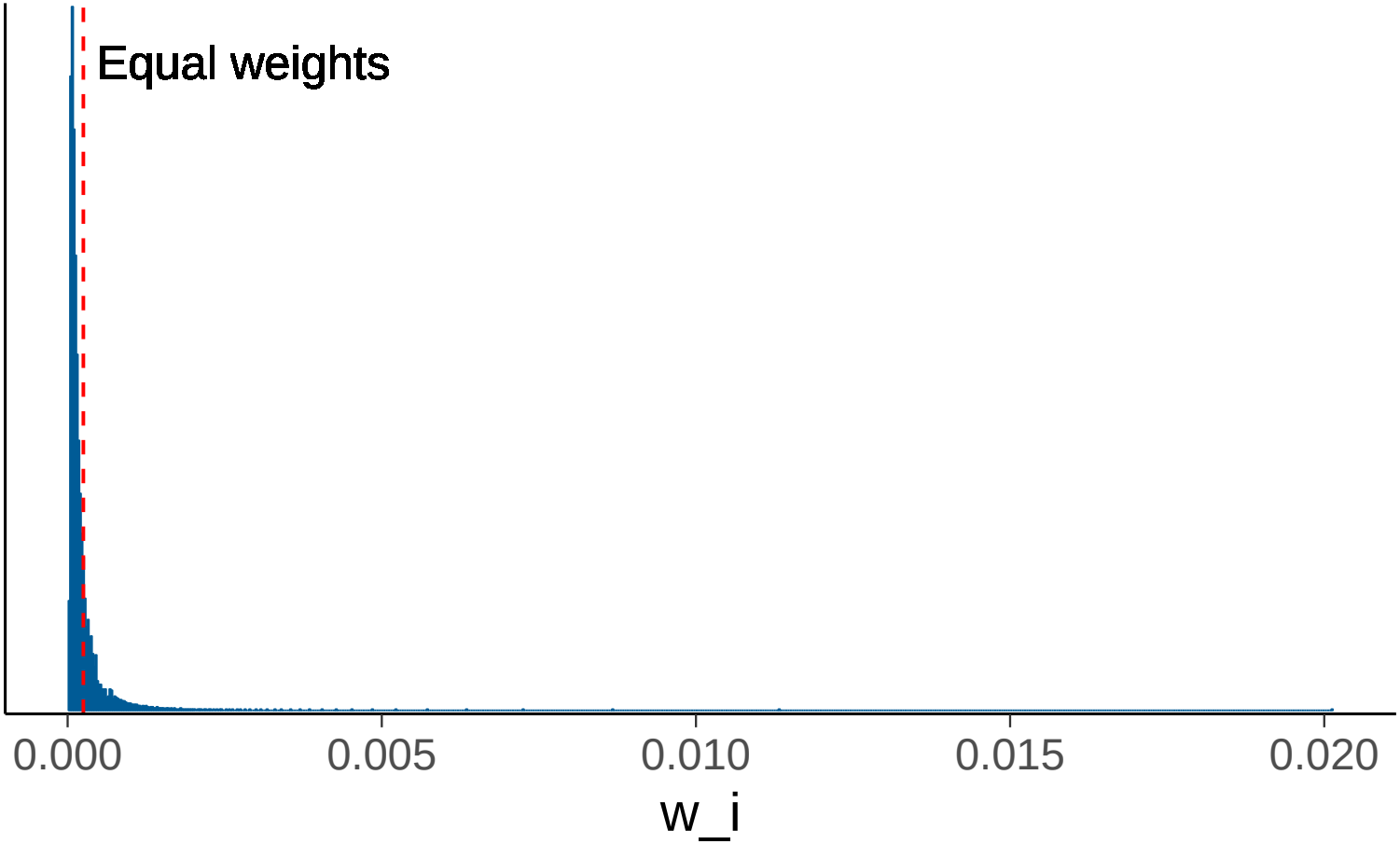
400 importance weights for leave−18th−out



4000 importance weights for leave−18th−out



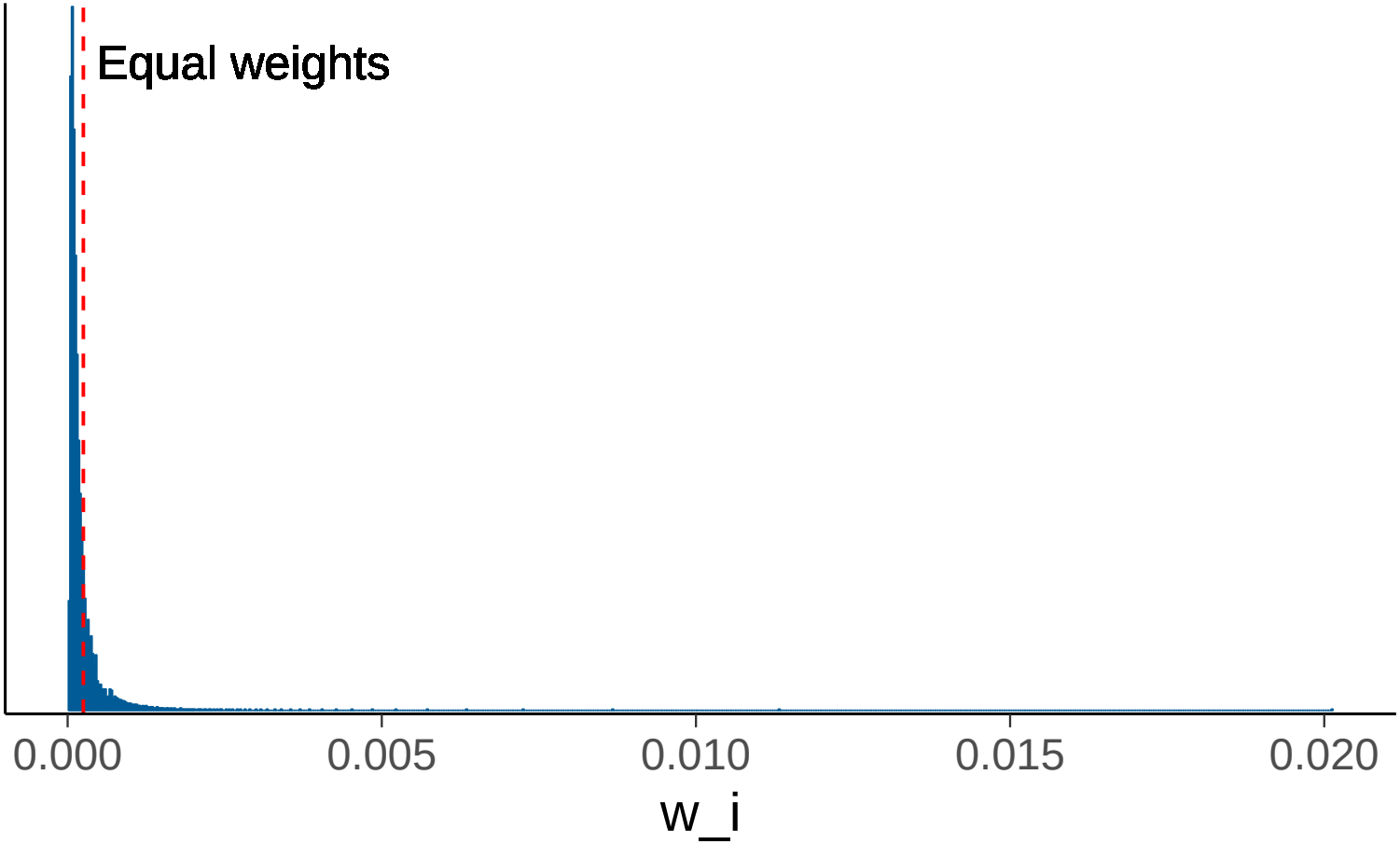
4000 importance weights for leave−18th−out



n_eff $\approx$ 459

see Vehtari, Gelman & Gabry (2017b)

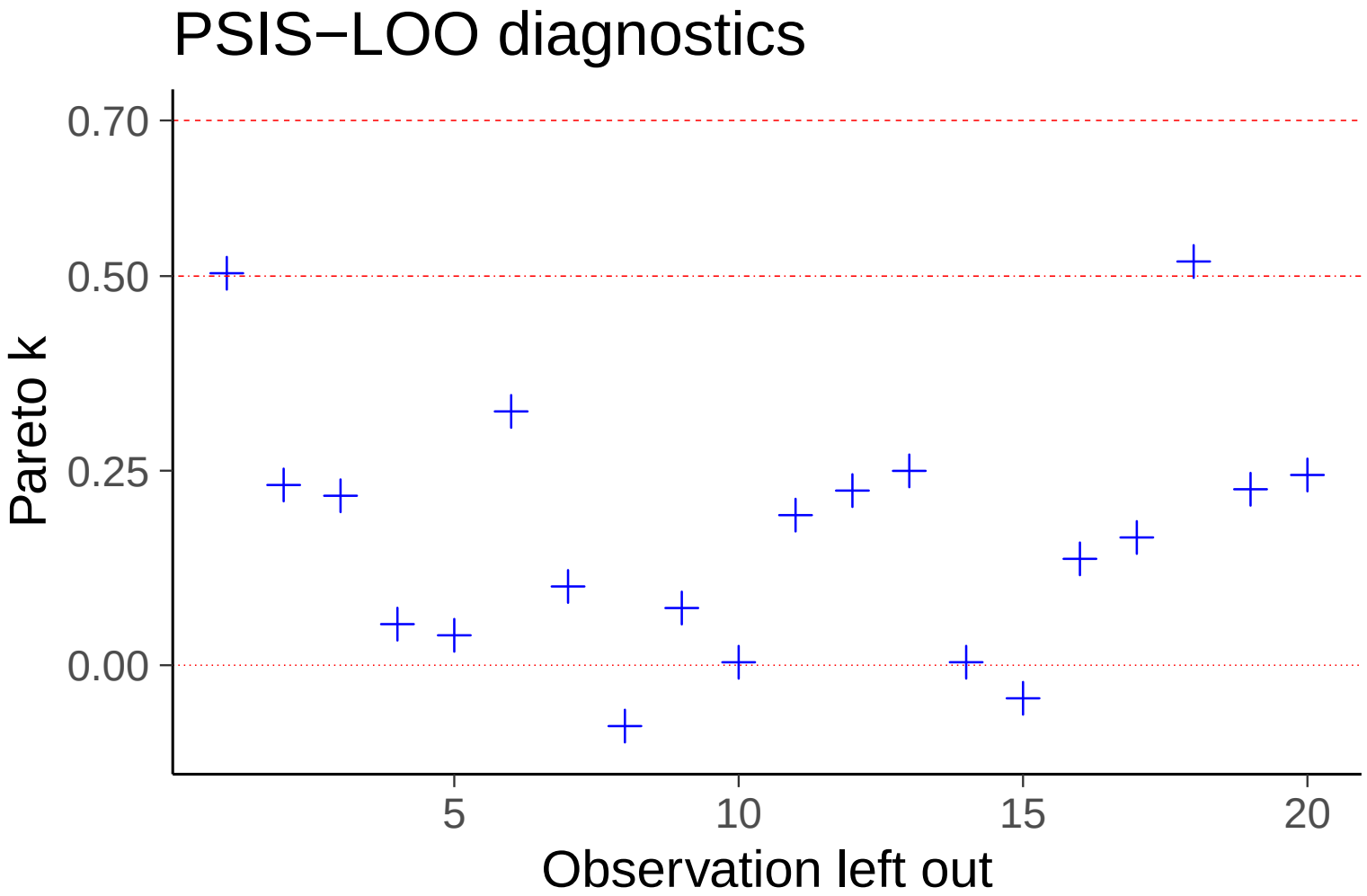
4000 importance weights for leave−18th−out

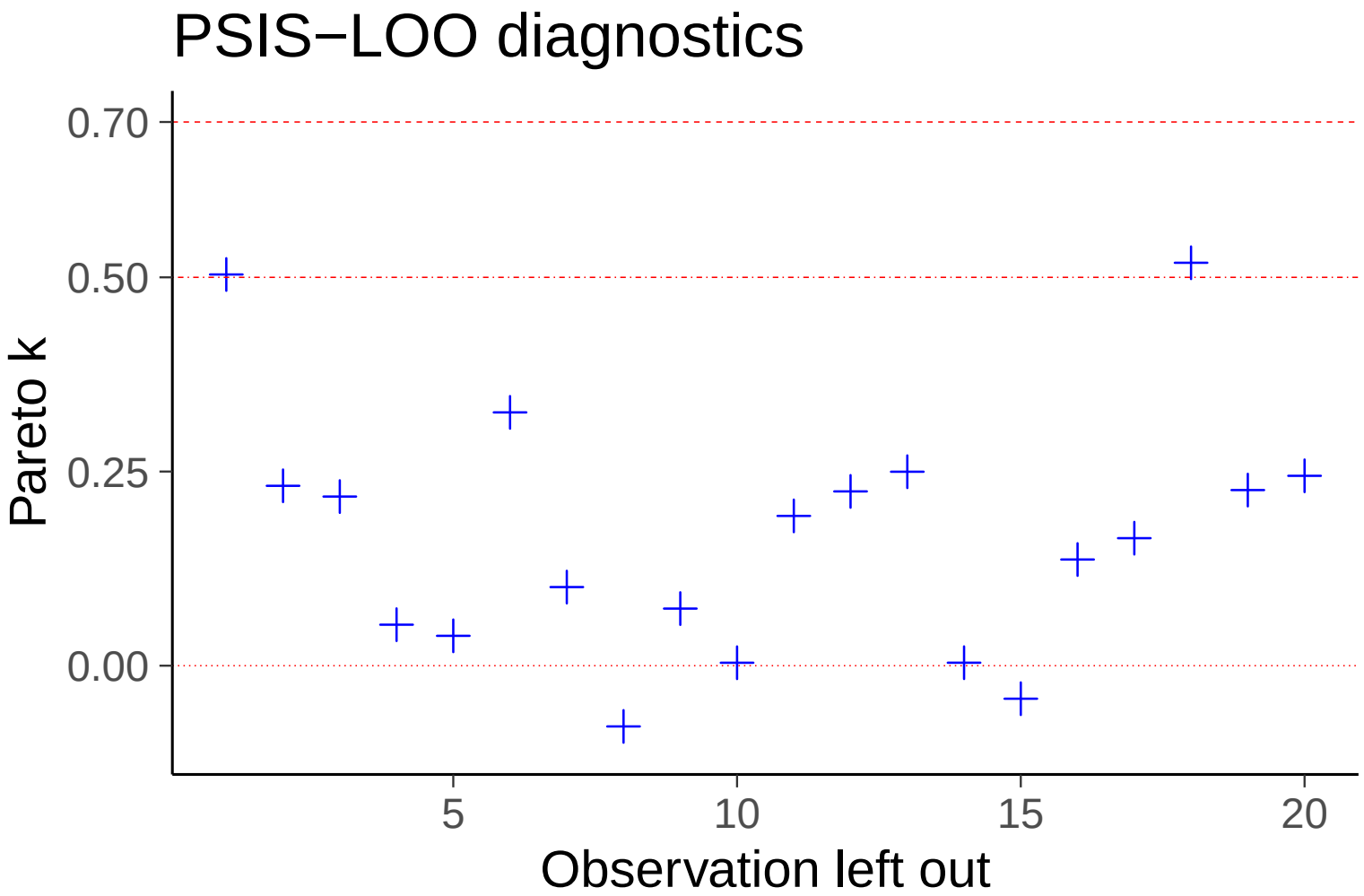


n_eff $\approx$ 459

Pareto k $\approx$ 0.52 (less than 0.7 is ok)

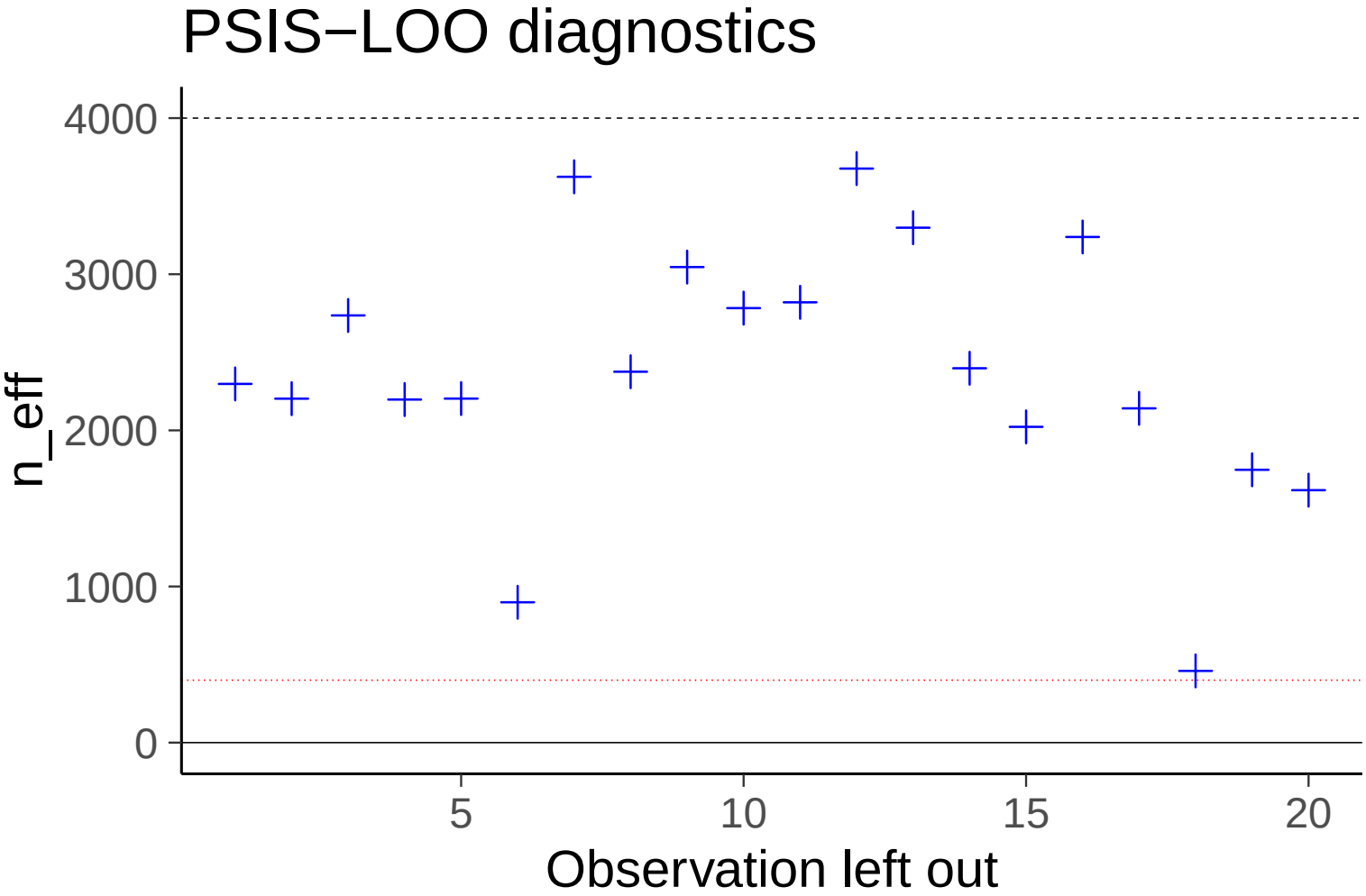
see Vehtari, Gelman & Gabry (2017b)


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>



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
looic        58.9 6.7

------

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.
```

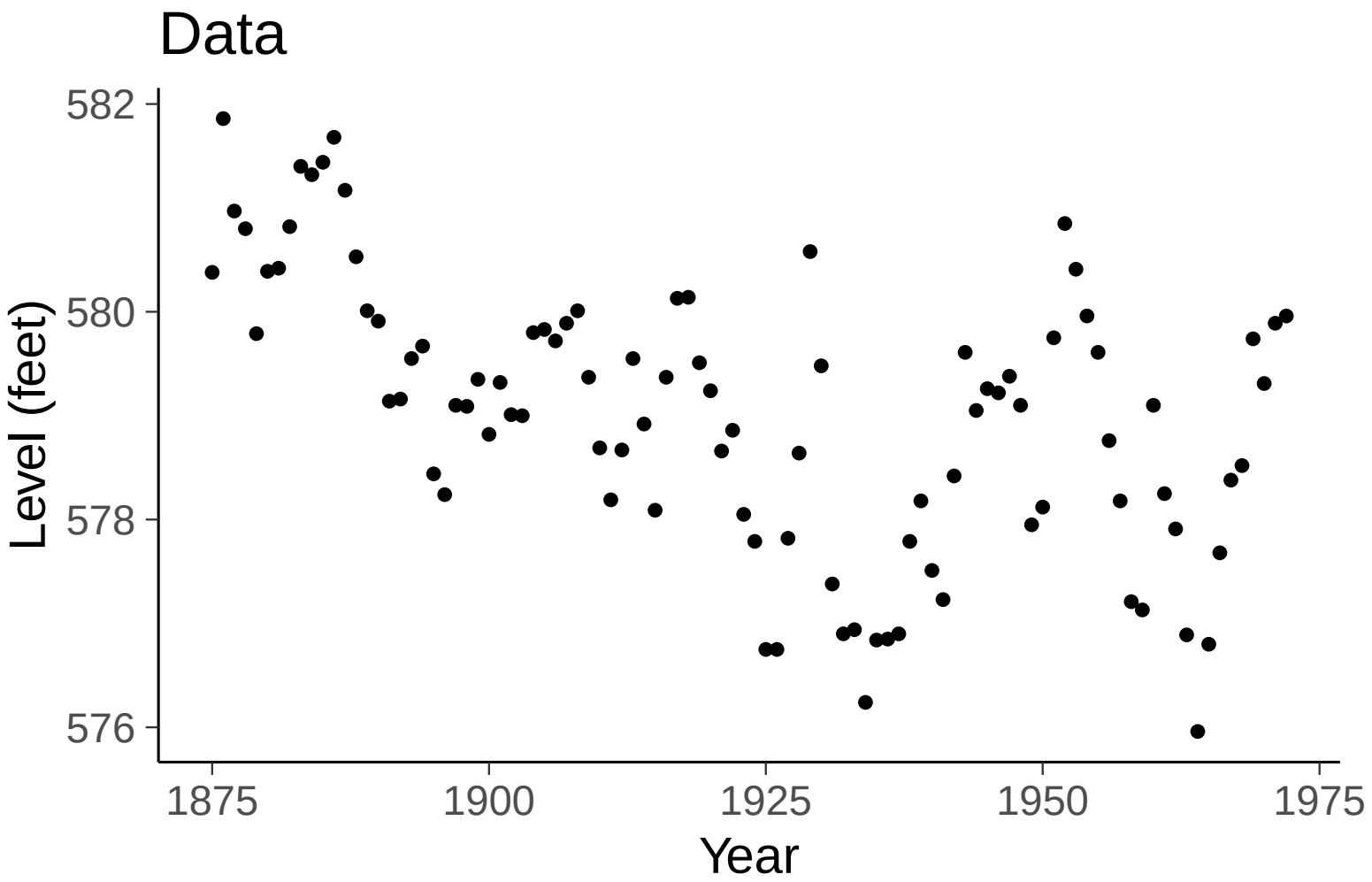
see more in Vehtari, Gelman & Gabry (2017b)

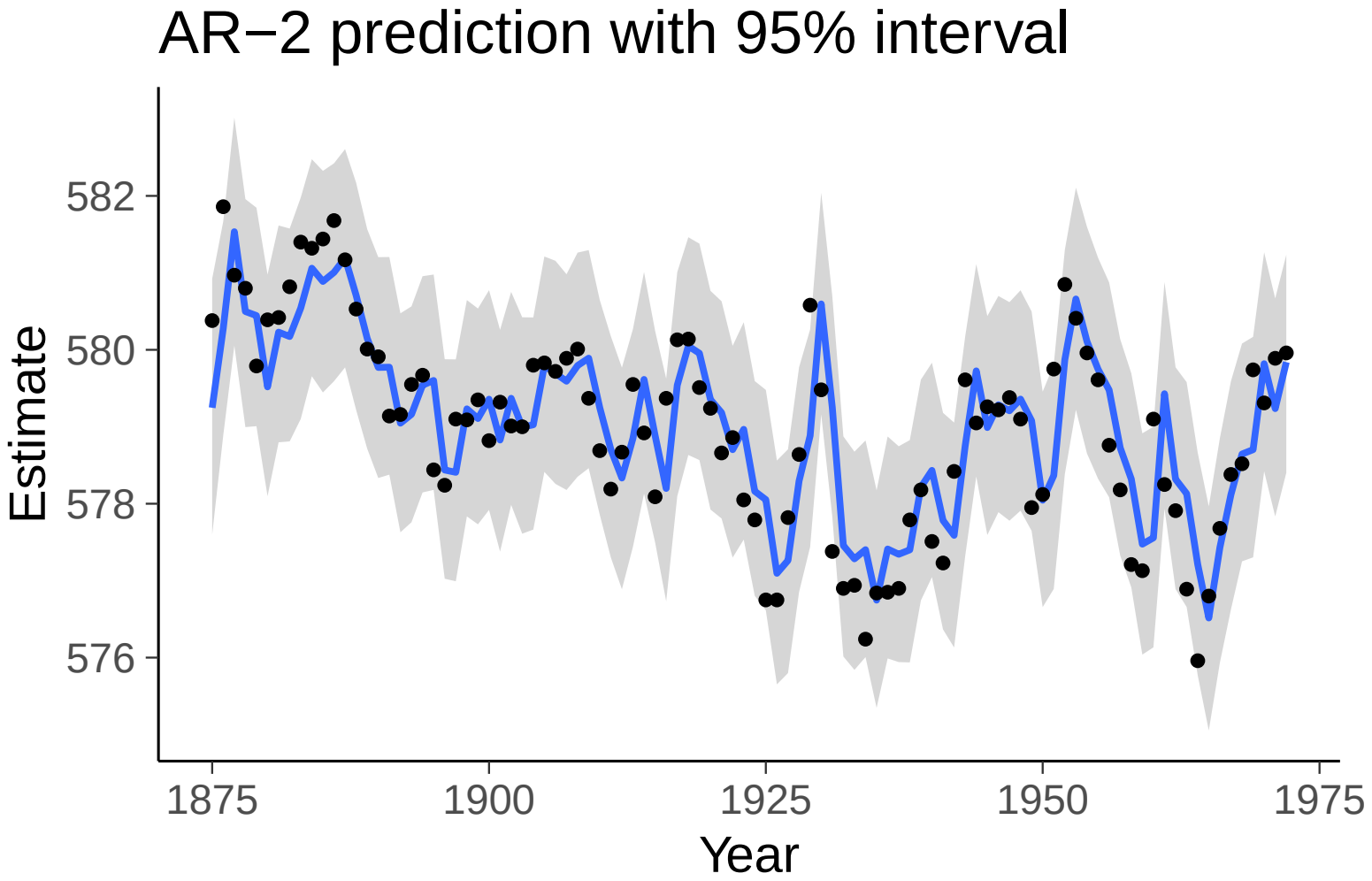
Pareto smoothed importance sampling LOO

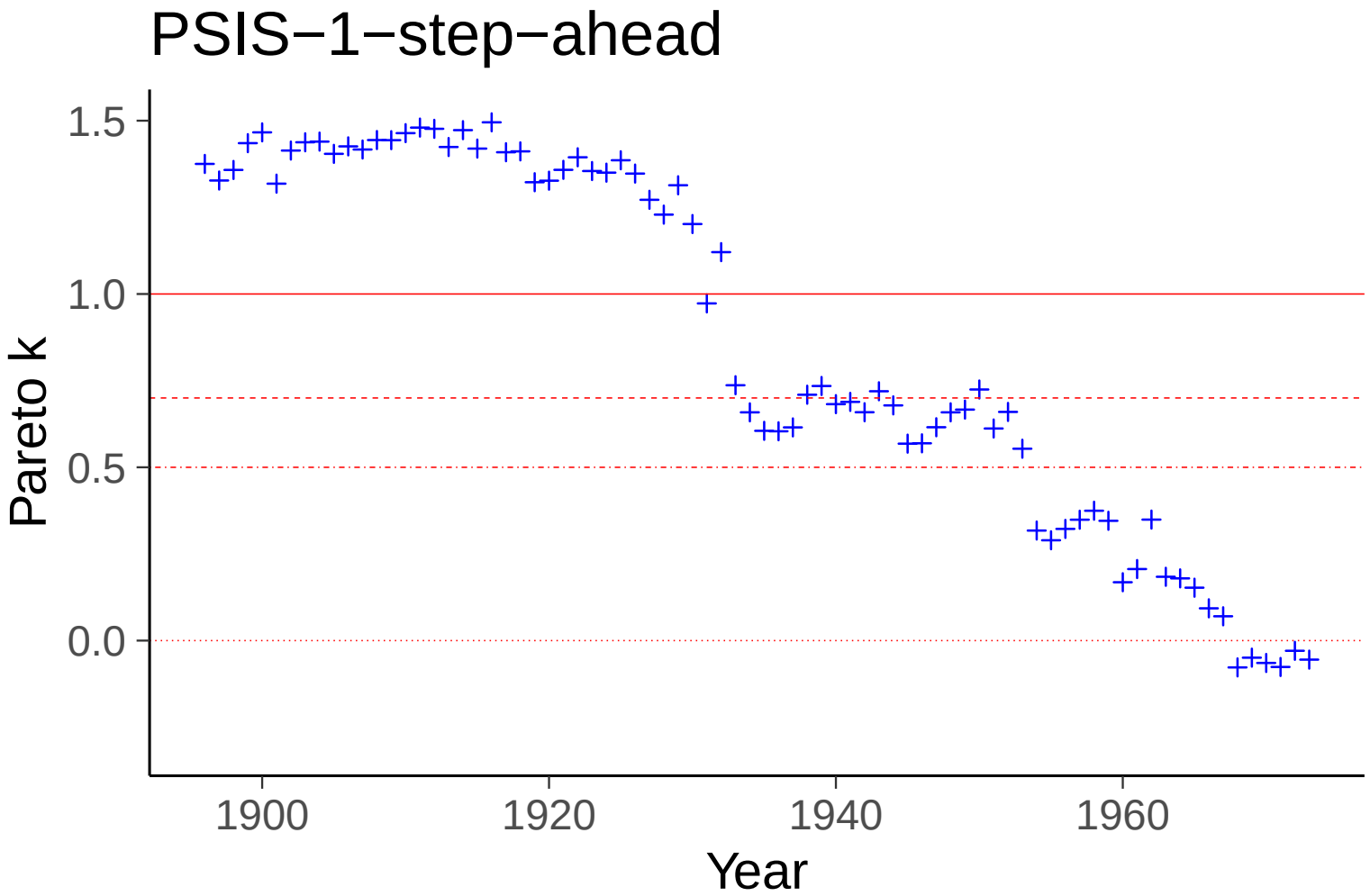
- PSIS-LOO for hierarchical models
  - leave-one-group out is challenging for PSIS-LOO

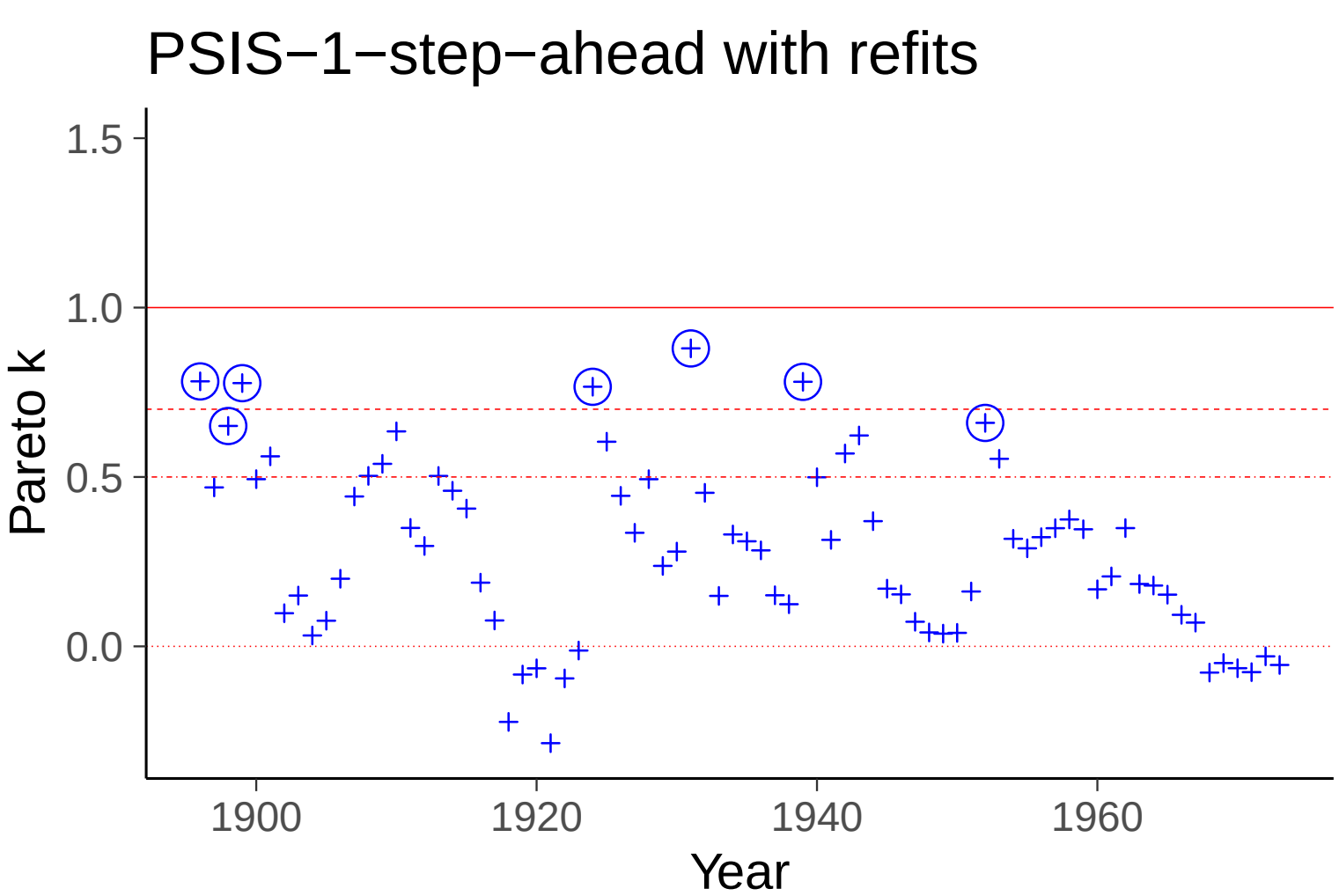
    see Merkel, Furr and Rabe-Hesketh (2018) for an approach using quadrature integration
- PSIS-LOO for time series
  - m-step-ahead works

    mc-stan.org/loo/articles/m-step-ahead-predictions.html


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


AR−2 prediction with 95% interval
Estimate
582
580
578
576
1875
1900
1925
1950
1975
Year


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

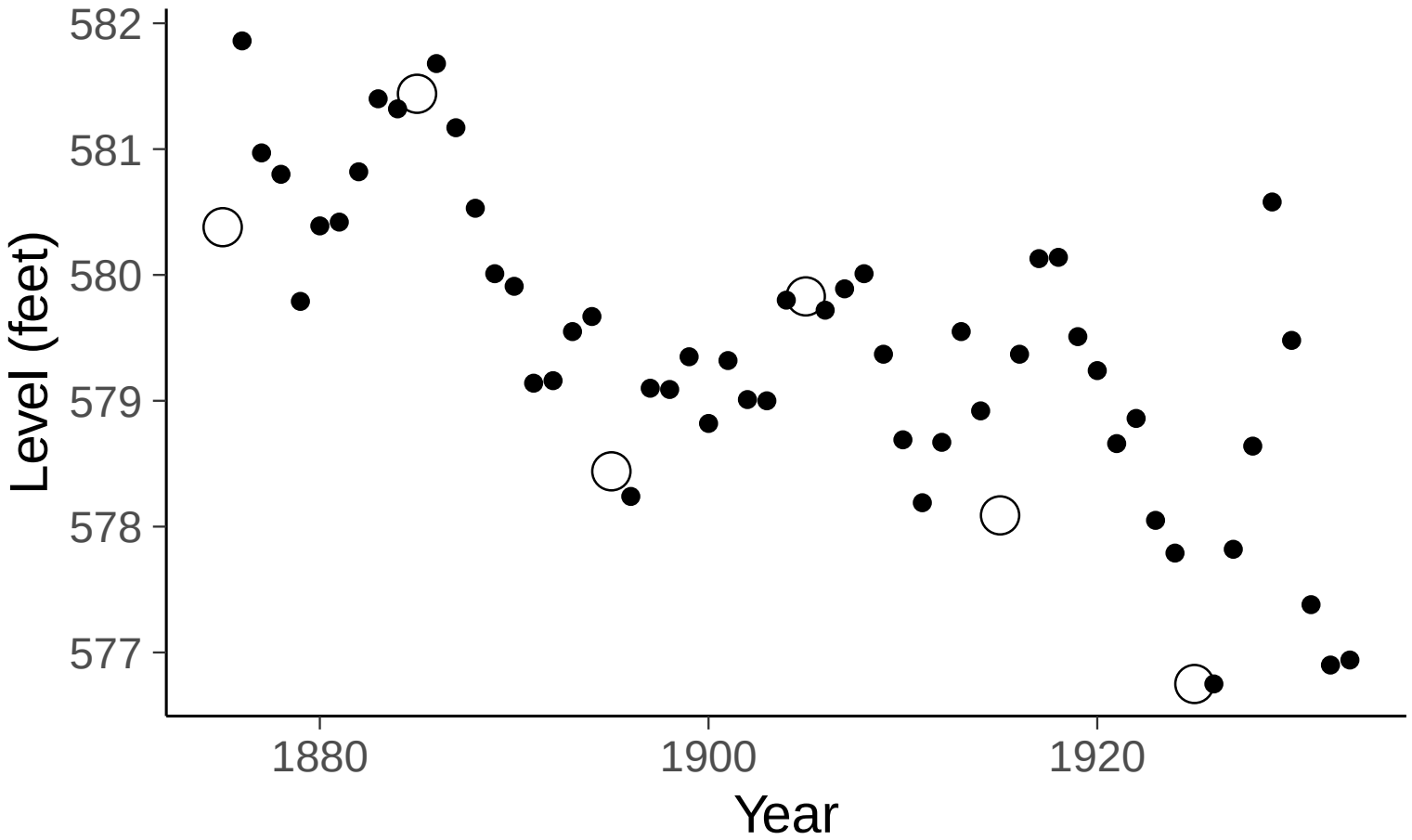


mc-stan.org/loo/articles/m-step-ahead-predictions.html

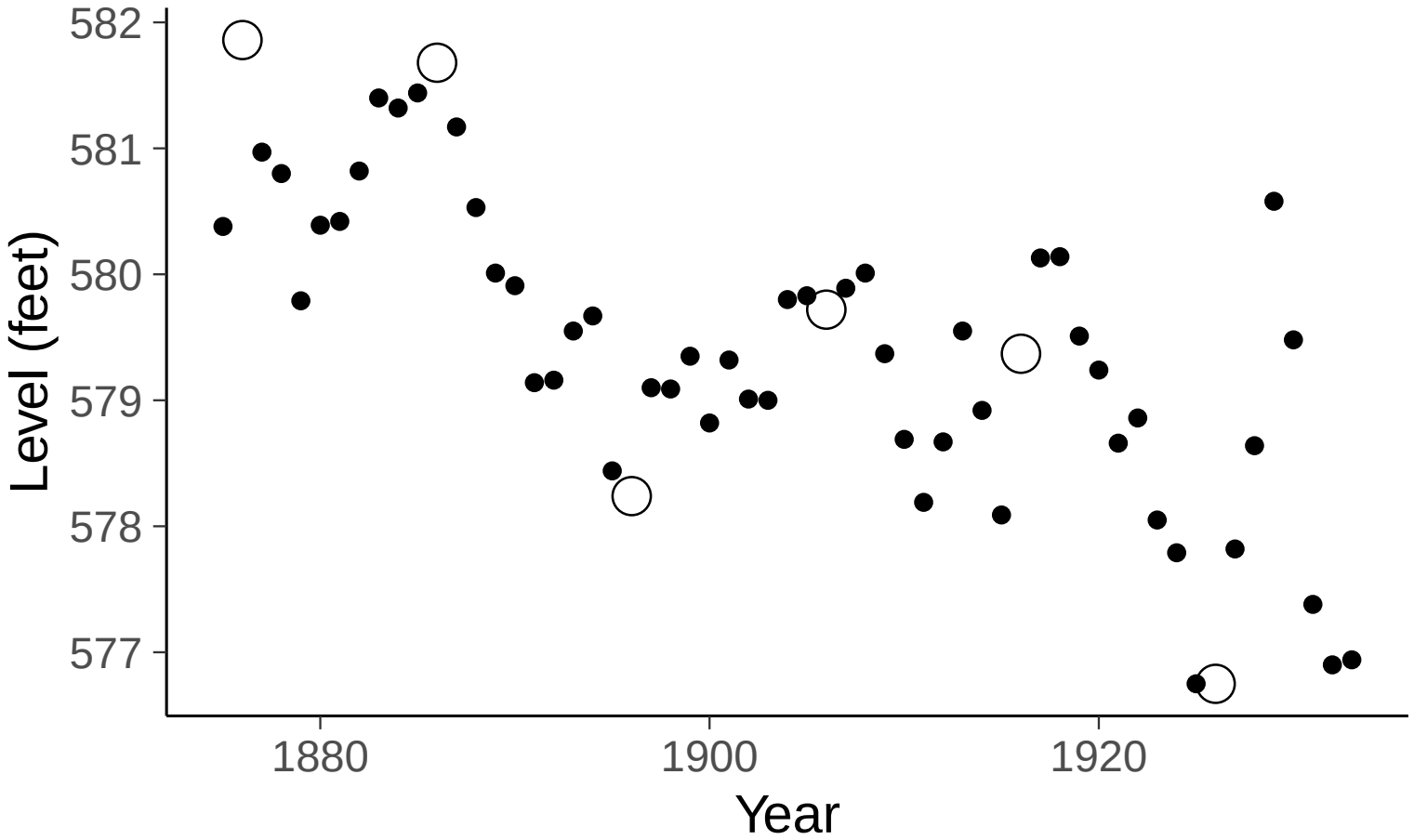
K-fold cross-validation

- 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

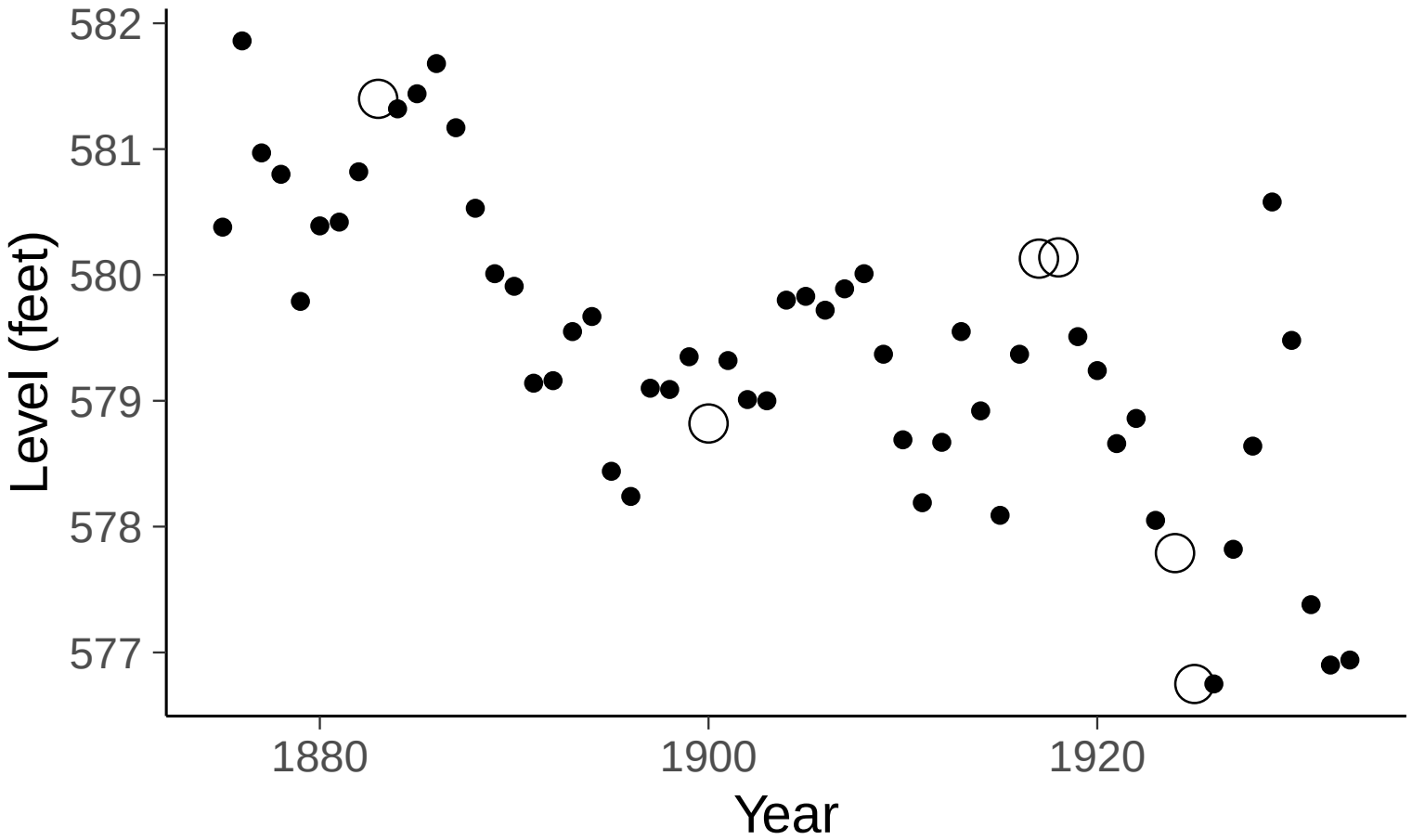
Balance k−fold approximation of LOO



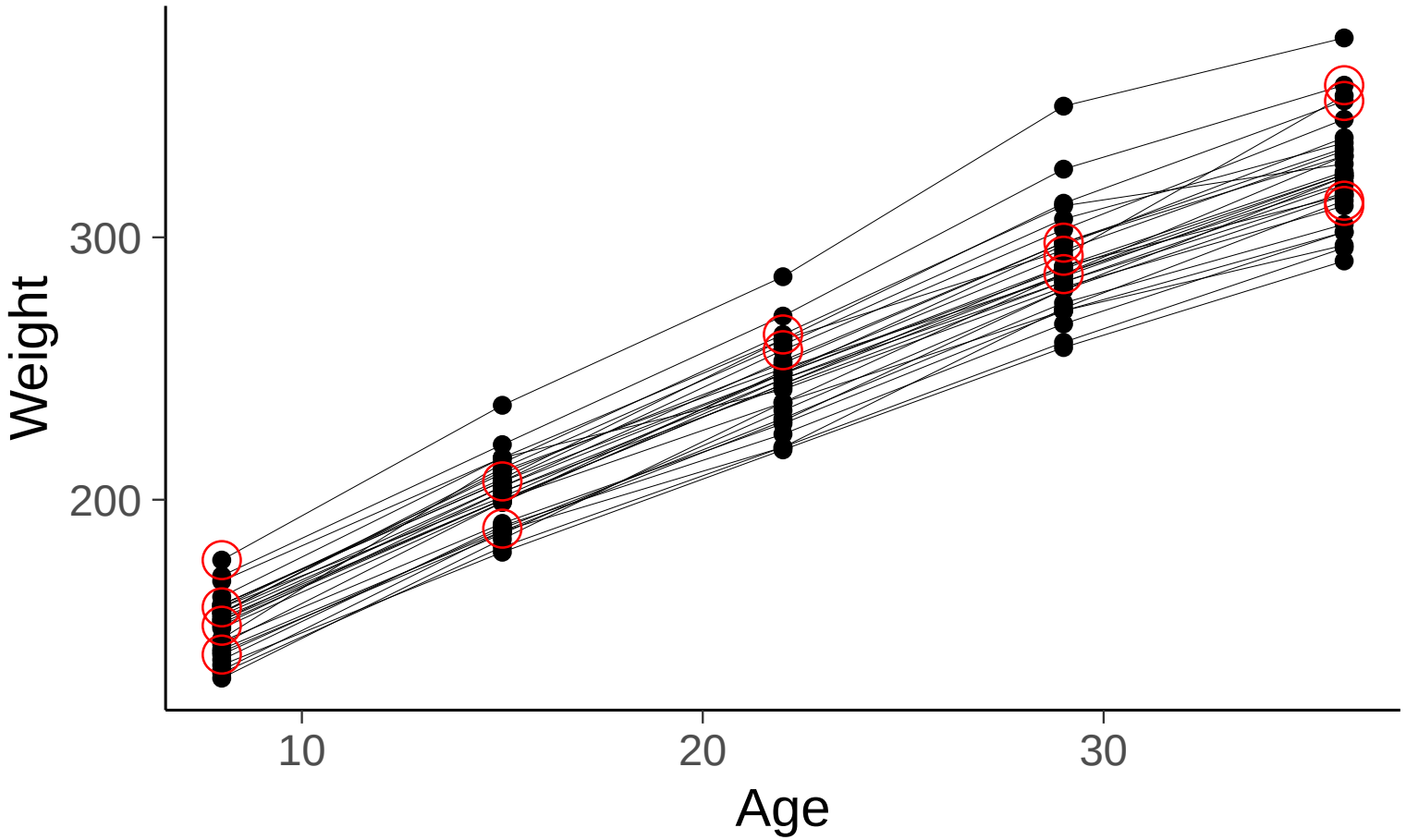
Balance k−fold approximation of LOO



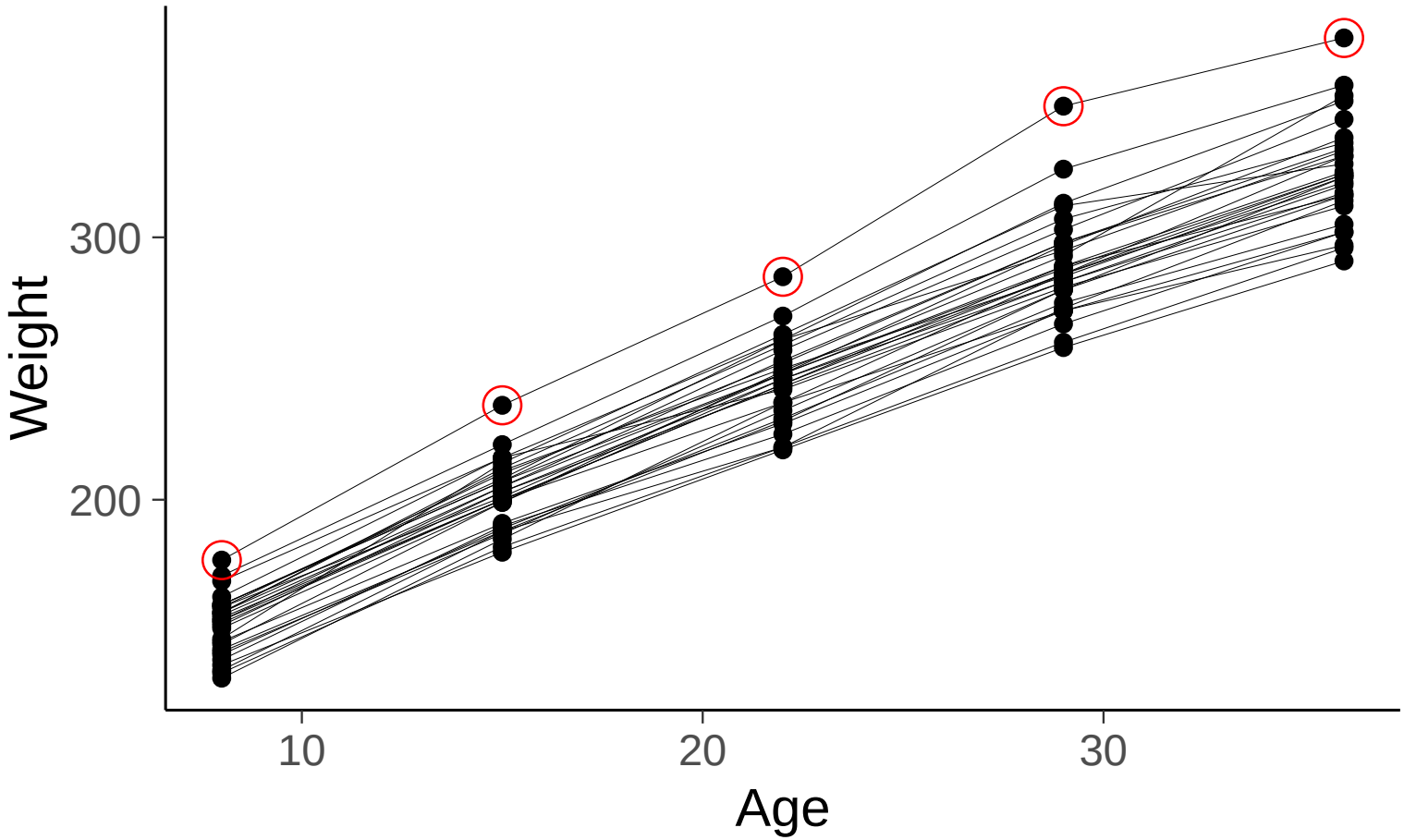
Random k−fold approximation of LOO

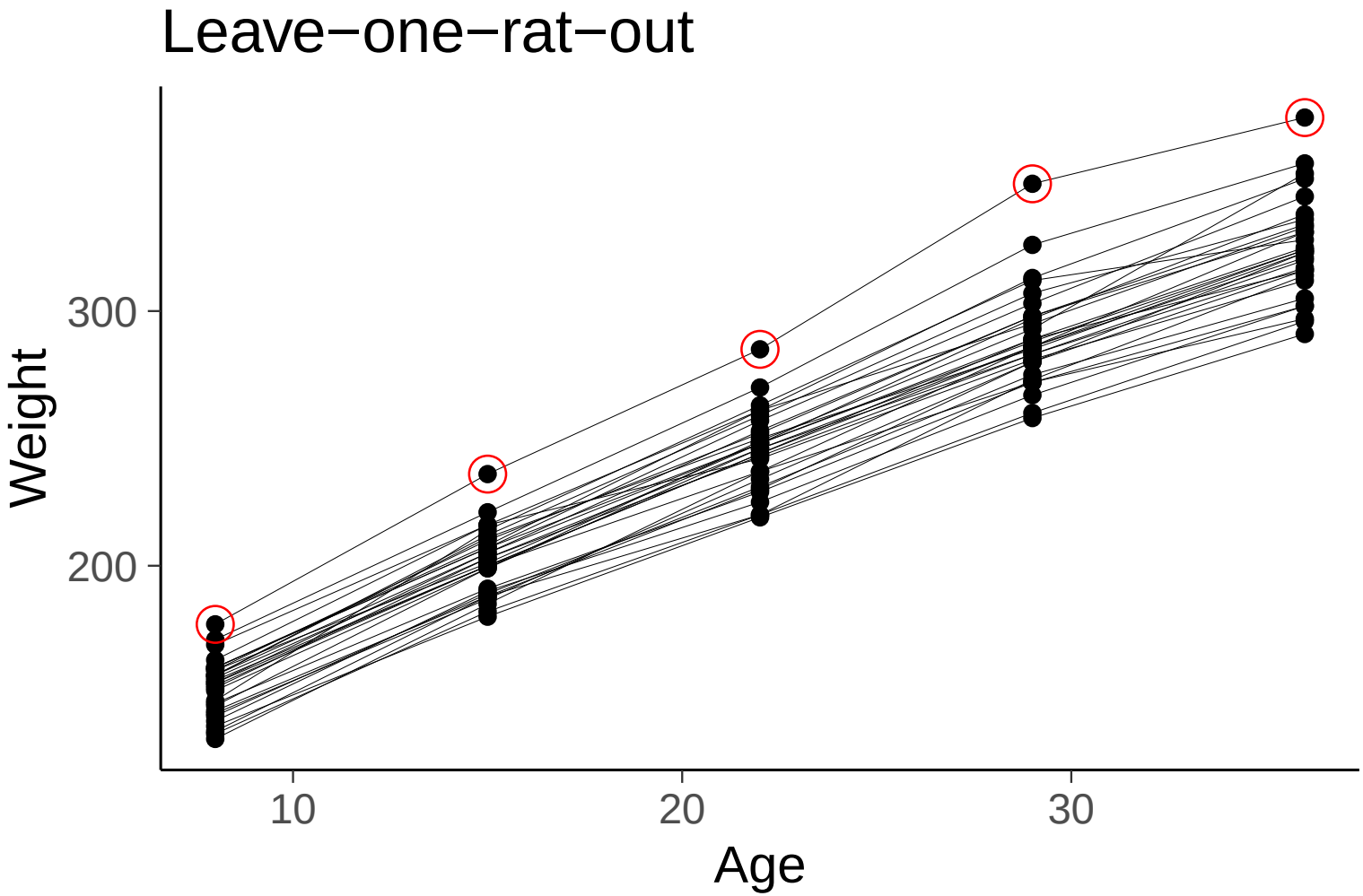


Random kfold approximation of LOO



Leave−one−rat−out





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

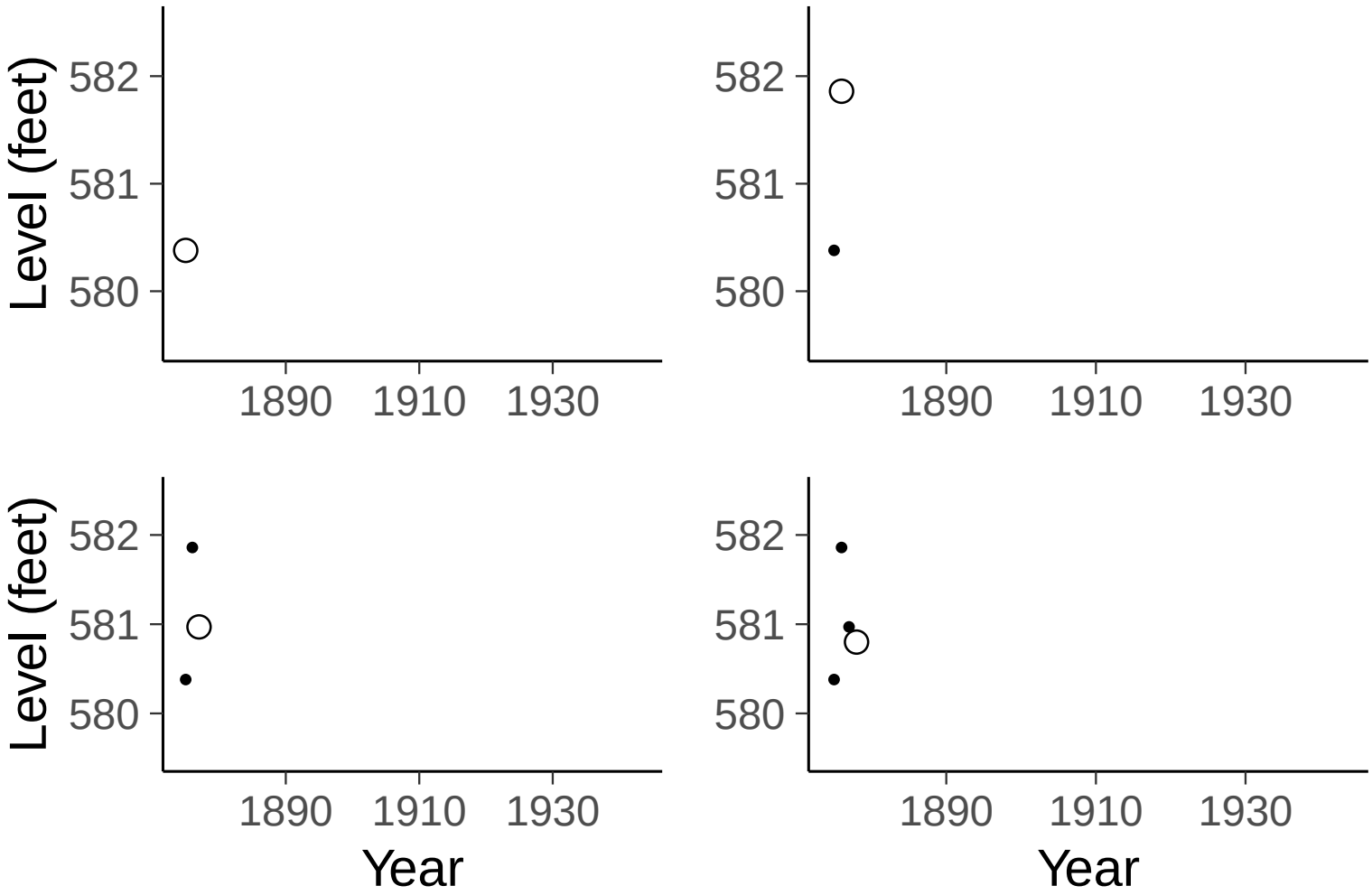
- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction
- BIC is an approximation for marginal likelihood
- TIC, NIC, RIC, PIC, BPIC, QIC, AICc, ...

Marginal likelihood / Bayes factor

- Like 1-step-ahead but starting with 0 observations

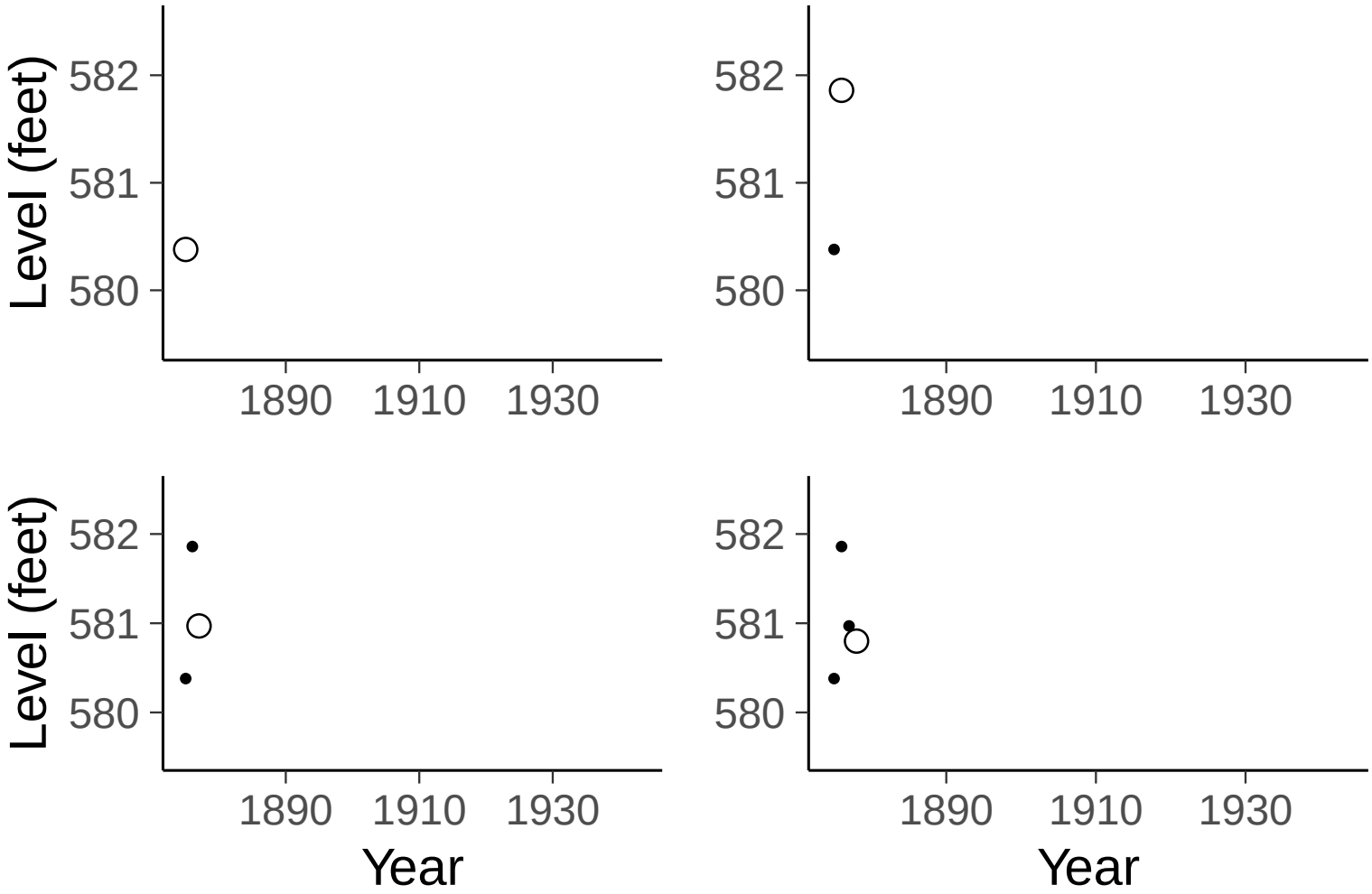
Marginal likelihood / Bayes factor

- Like 1-step-ahead but starting with 0 observations



Marginal likelihood / Bayes factor

- Like 1-step-ahead but starting with 0 observations which makes it very sensitive to prior



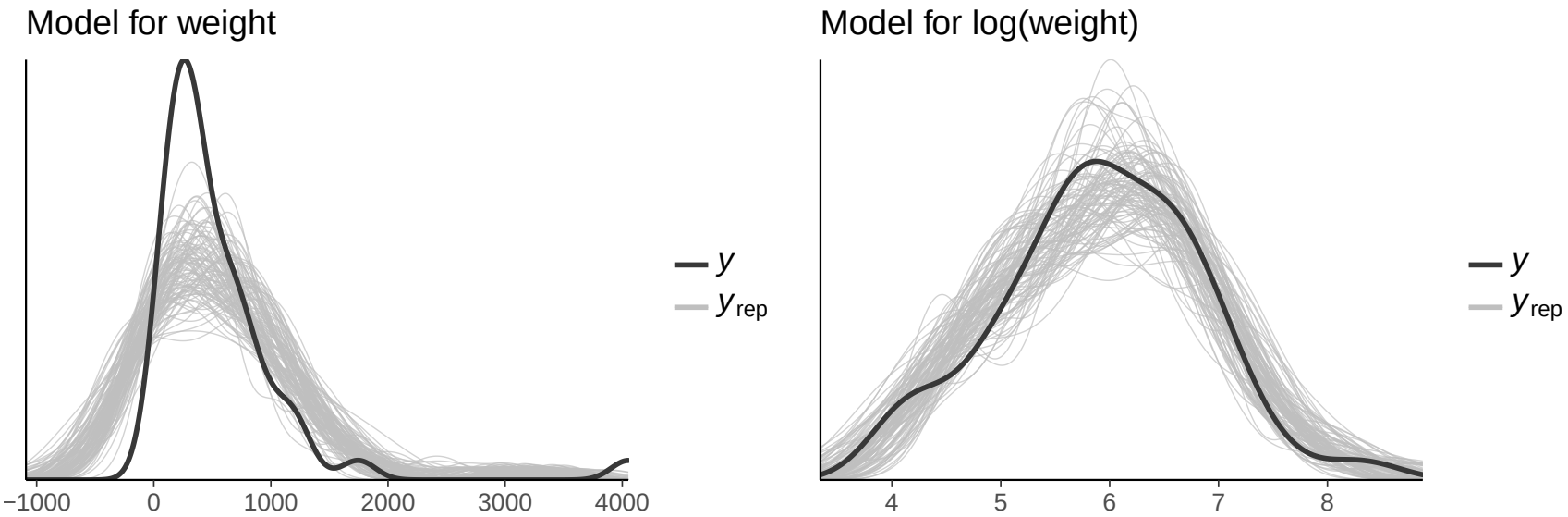
Cross-validation for model assessment

- 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)

  see demos avehtari.github.io/modelselection/

Sometimes cross-validation is not needed

Sometimes cross-validation is not needed



Predicting the yields of mesquite bushes.

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

Model comparison

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

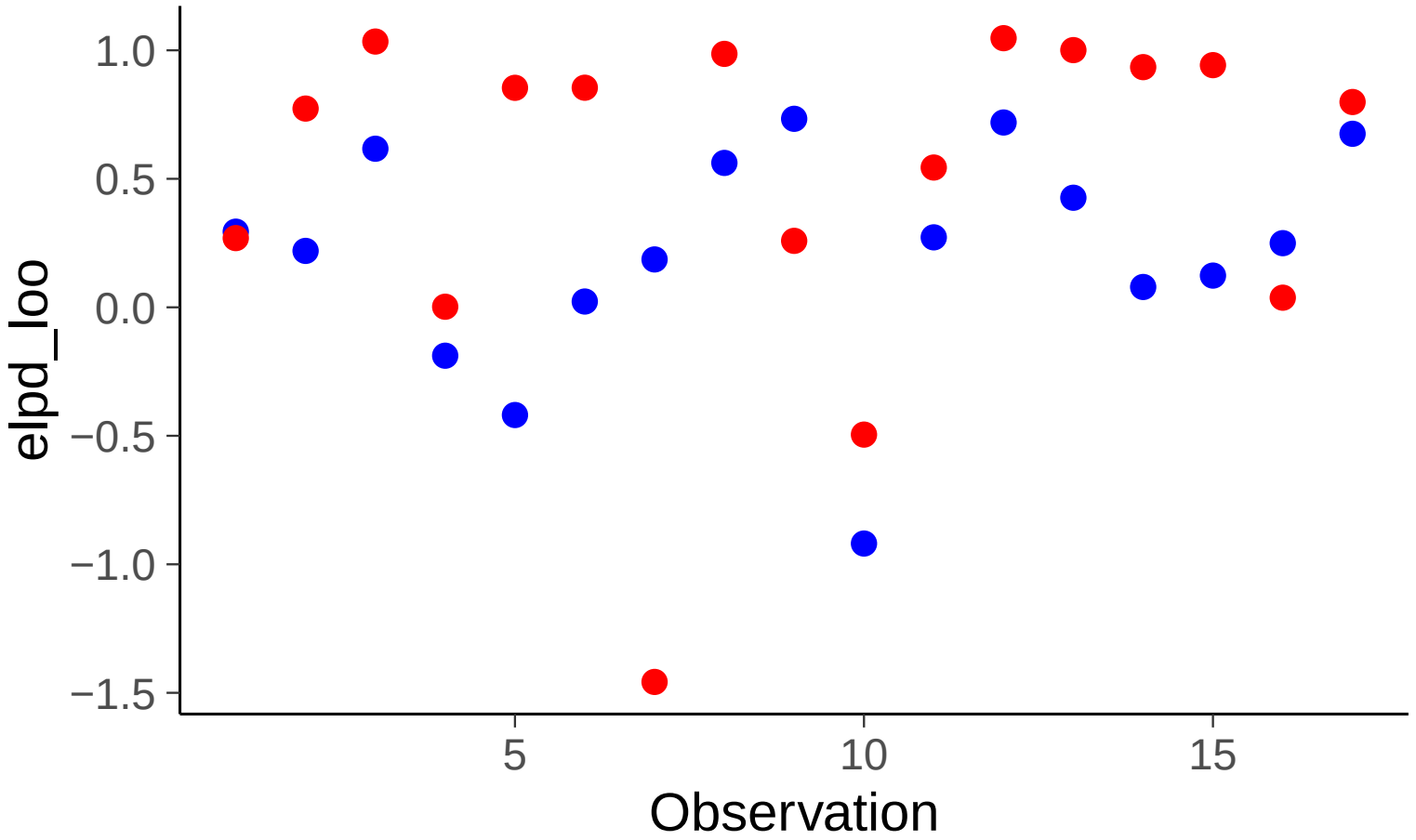
  avehtari.github.io/modelselection/betablockers.html

Model comparison

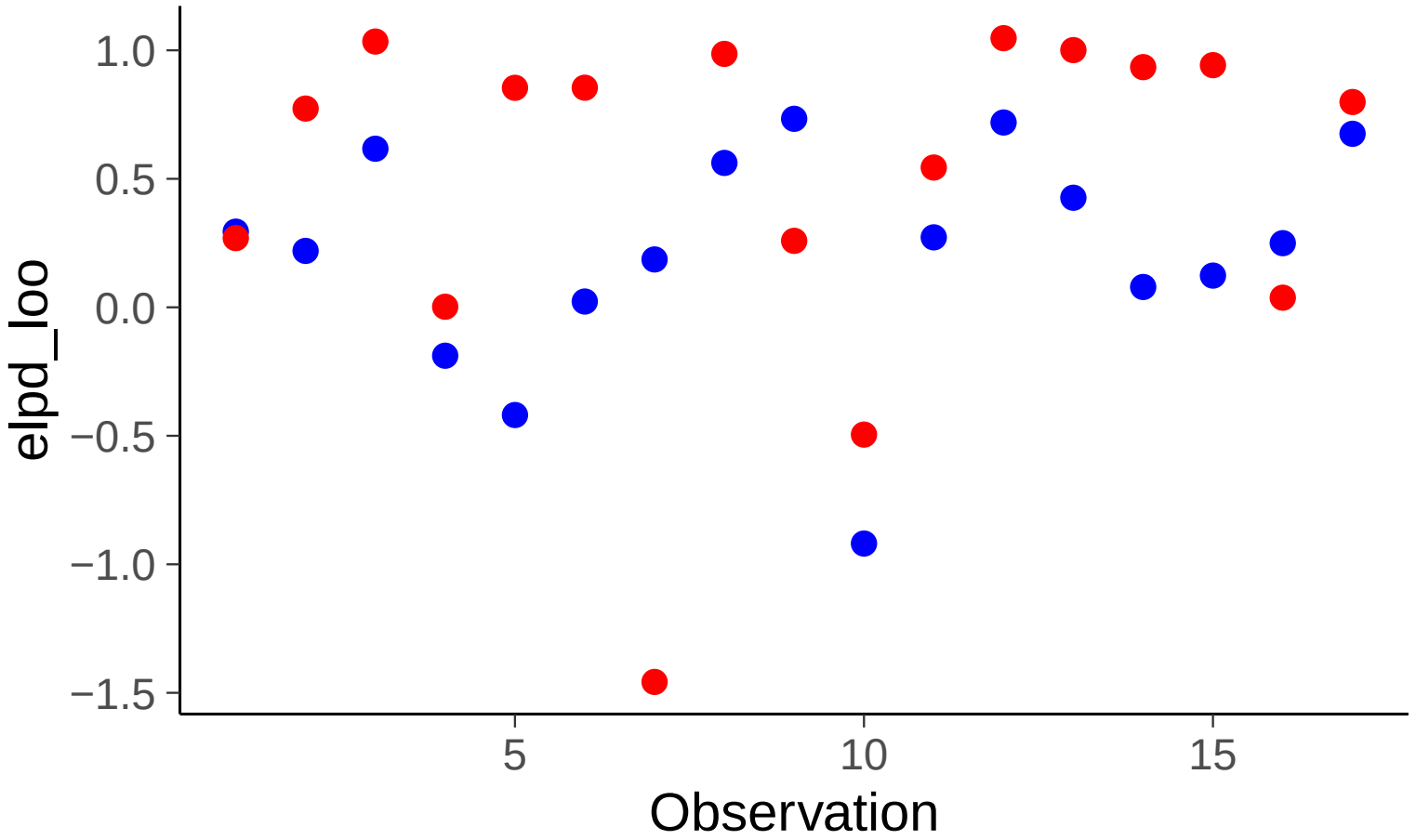
- “A popular hypothesis has it that primates with larger brains produce more energetic milk, so that brains can grow quickly” (from Statistical Rethinking)
  - Model 1: formula = kcal.per.g $\sim$ neocortex
  - Model 2: formula = kcal.per.g $\sim$ neocortex + log(mass)

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

Pointwise comparison LOO models: Model 1

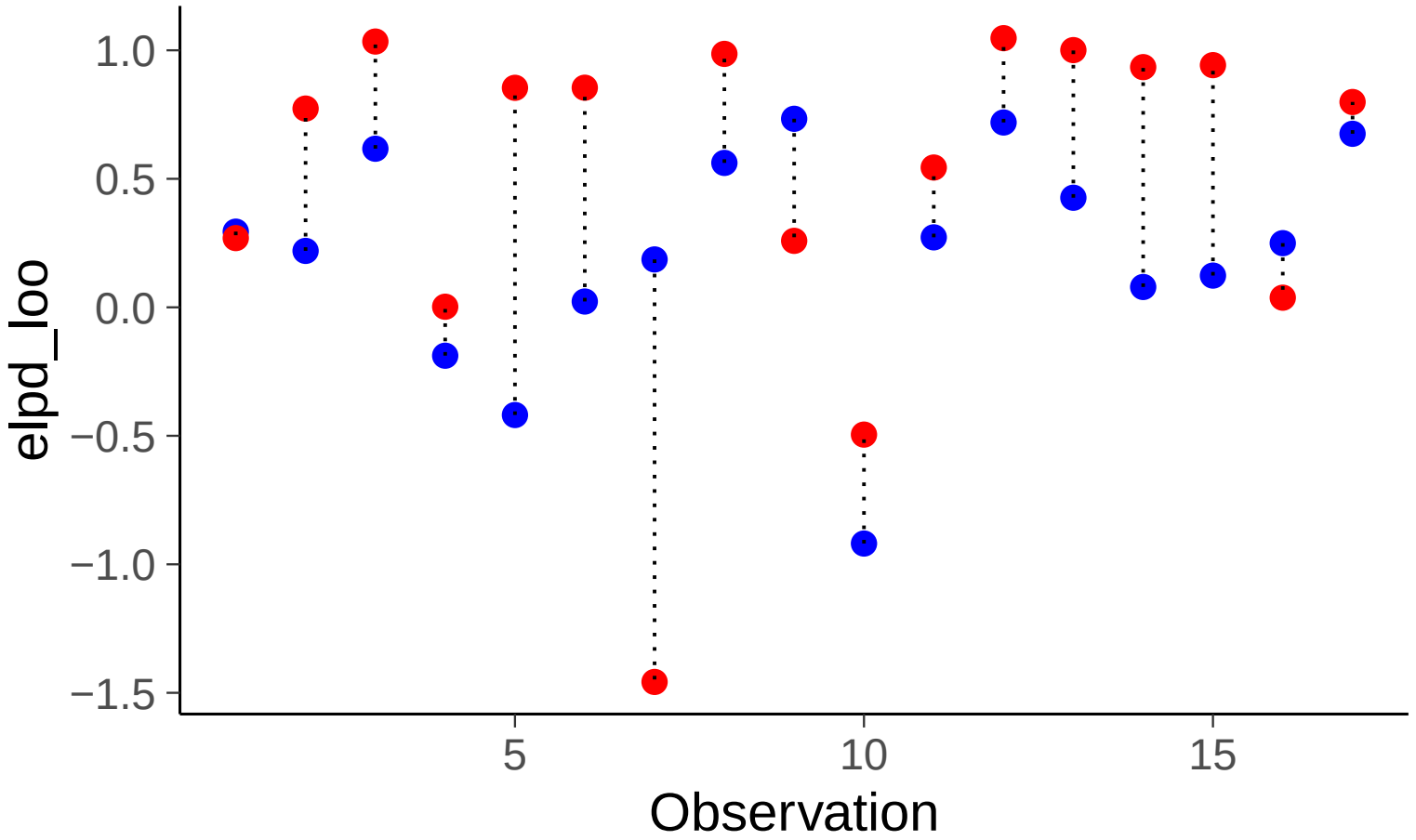


Pointwise comparison LOO models: Model 1

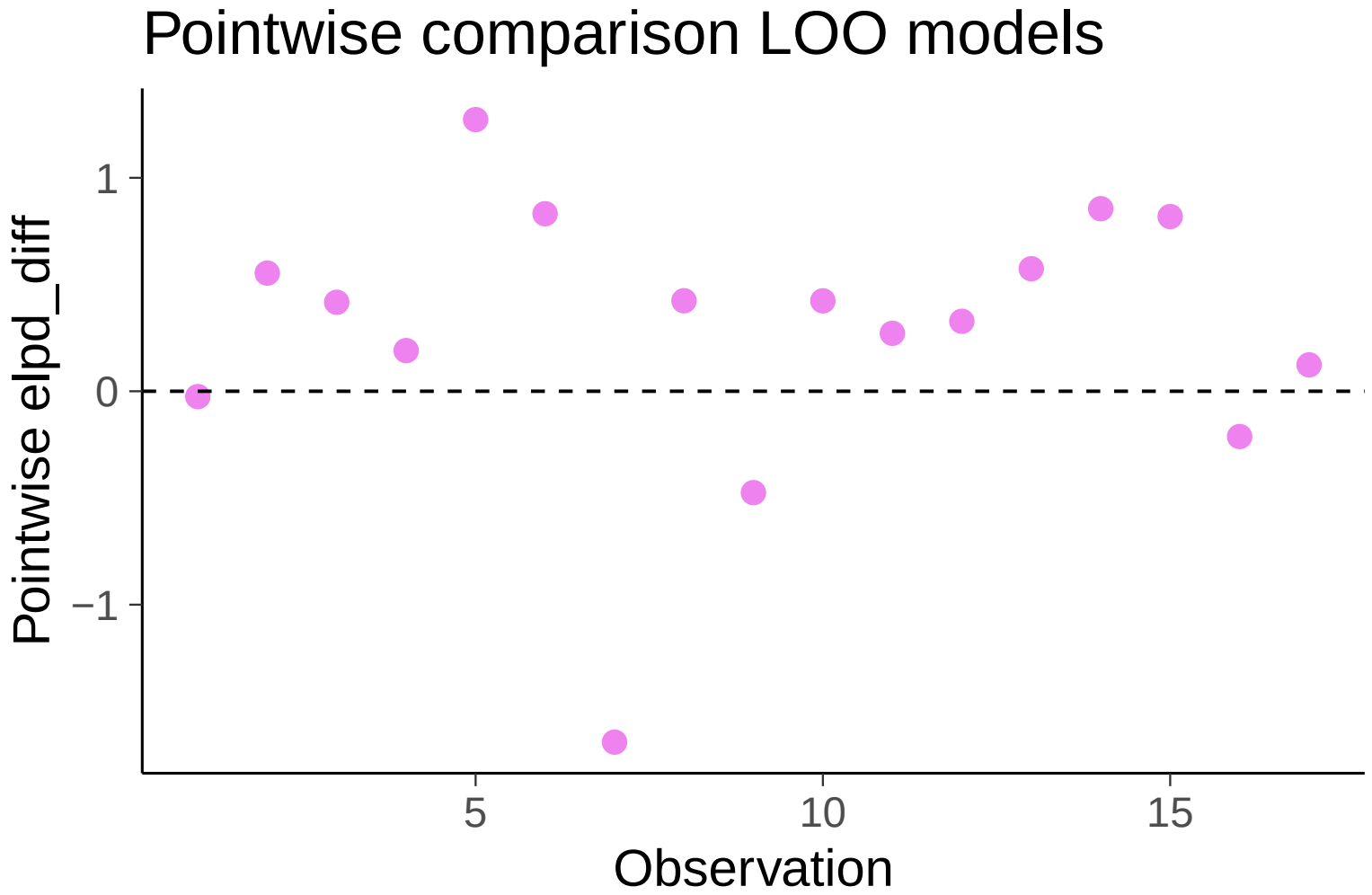


Model 1 elpd_loo $\approx$ 3.7, SE=1.8
Model 2 elpd_loo $\approx$ 8.4, SE=2.8

Pointwise comparison LOO models: Model 1



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




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

elpd_diff        se
      4.7       2.7
```

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
  - see 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
  - see 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
  - see 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
  - see 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/

When not to use cross-validation

- Do not use cross-validation to choose from a large set of models!
  - selection process leads to overfitting!
  - you may use projection predictive approach
  - useful when correlating variables make the posterior distribution analysis difficult

    video, refs and demos at avehtari.github.io/modelselection/

    and Piironen & Vehtari (2017)

Bayesian stacking LOO weights

- Bayesian stacking and Pseudo-BMA+ should be used only for model averaging
  - you may drop models with 0 weights
  - you shouldn't choose the model with largest weight unless it's 1

see Yao, Vehtari, Simpson, & Gelman (2018)

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

References

All references and more at avehtari.github.io/modelselection/

- Model selection tutorial at StanCon 2018 Asilomar
  - more about projection predictive variable selection
- Regularized horseshoe talk at StanCon 2018 Asilomar
- Several case studies
- References with links to open access pdfs