Chapter 7

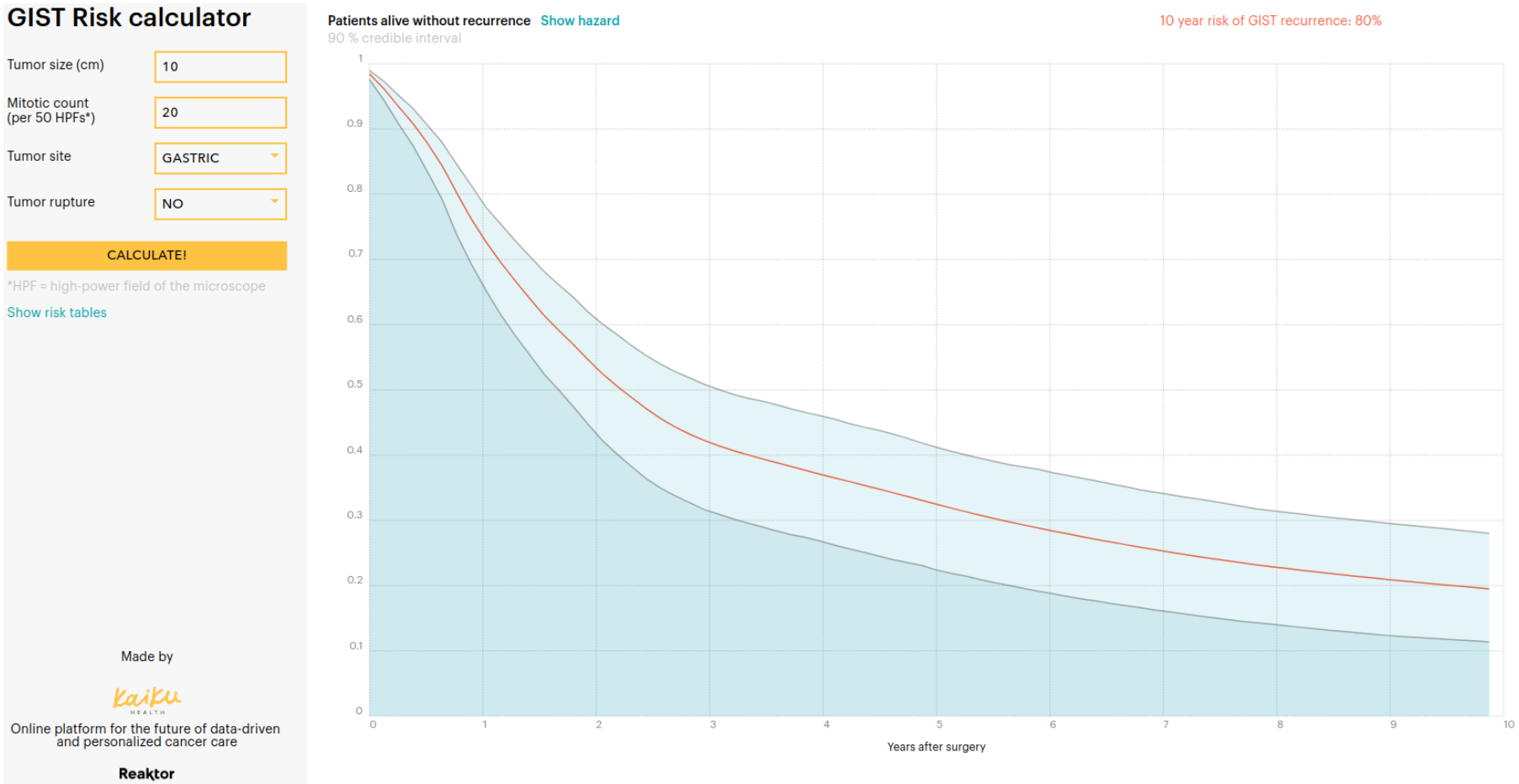
- 7.1 Measures of predictive accuracy
- 7.2 Information criteria and cross-validation
  - Instead of 7.2, read:
    Vehtari, A., Gelman, A., Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing. 27(5):1413–1432. arXiv preprint.
- 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 assessment, 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 goodness of the predictive distribution, or we don't know (yet) the application specific utility, then good information theoretically justified choice is log-score

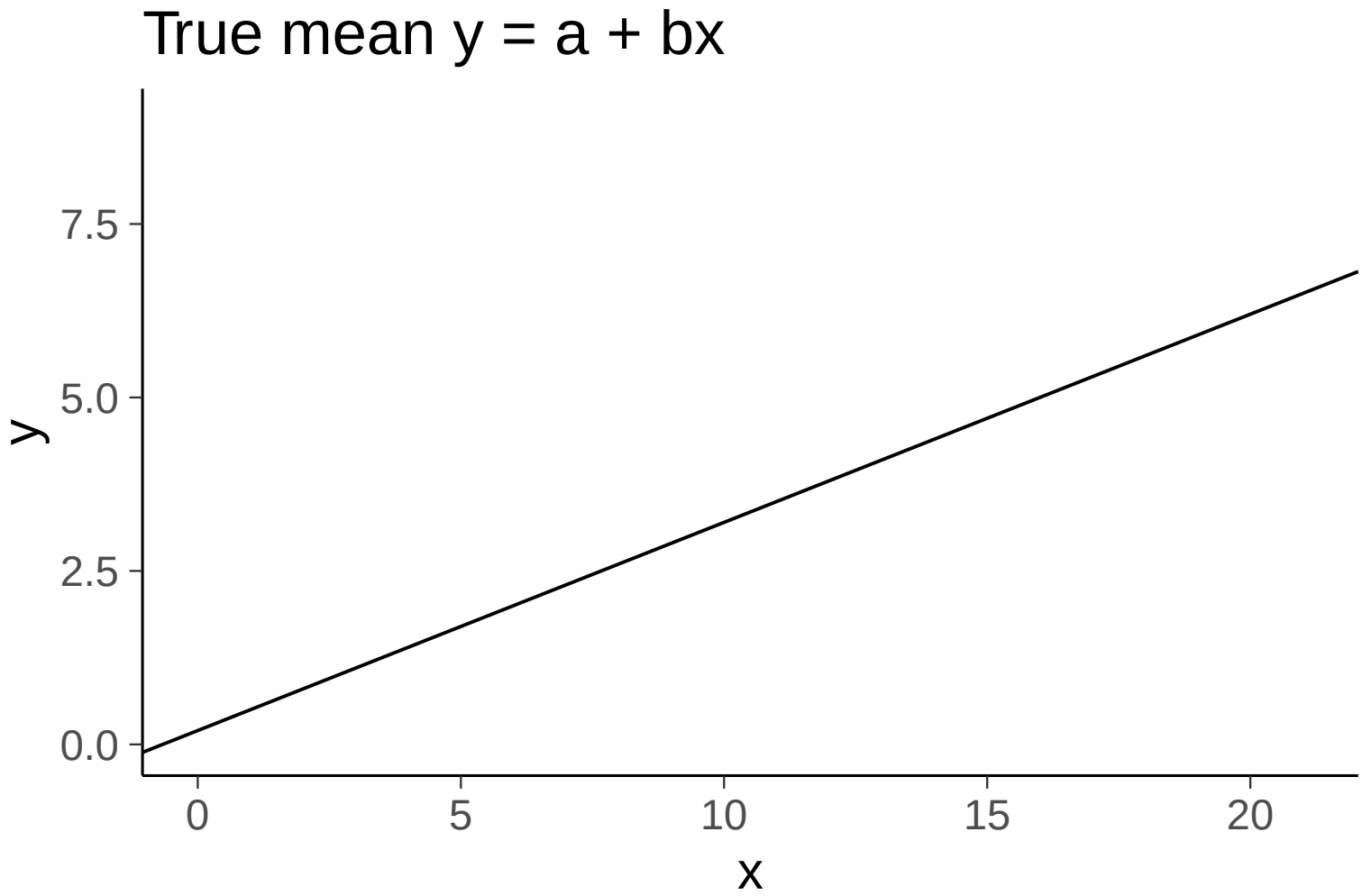
$$\log p(y^{\text{rep}}|y, M),$$

Outline

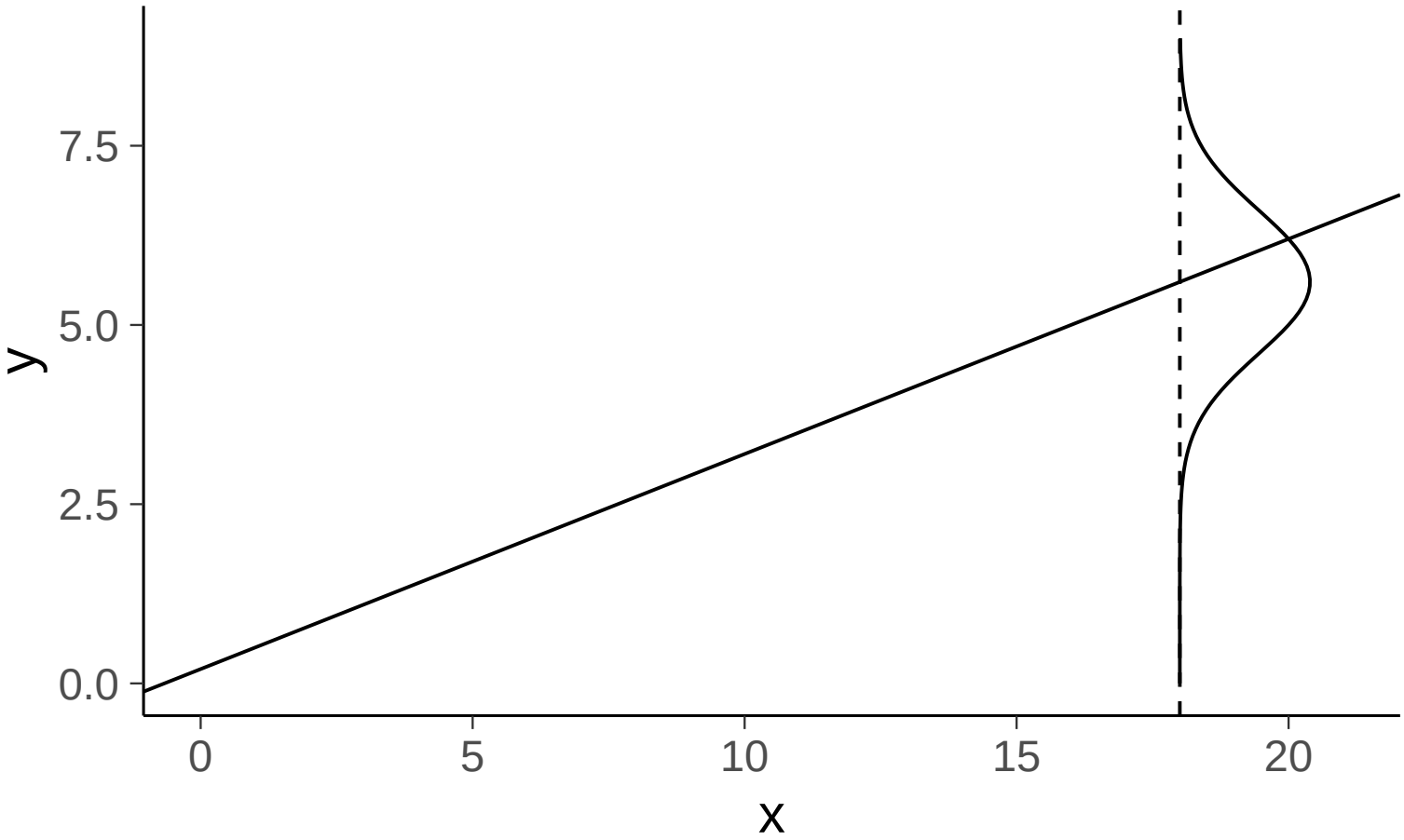
- What is cross-validation
  - Leave-one-out cross-validation (elpd_loo, p_loo)
  - Uncertainty in LOO (SE)
- When is cross-validation applicable?
  - data generating mechanisms and prediction tasks
  - leave-many-out cross-validation
- Fast cross-validation
  - PSIS and diagnostics in loo package (Pareto 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

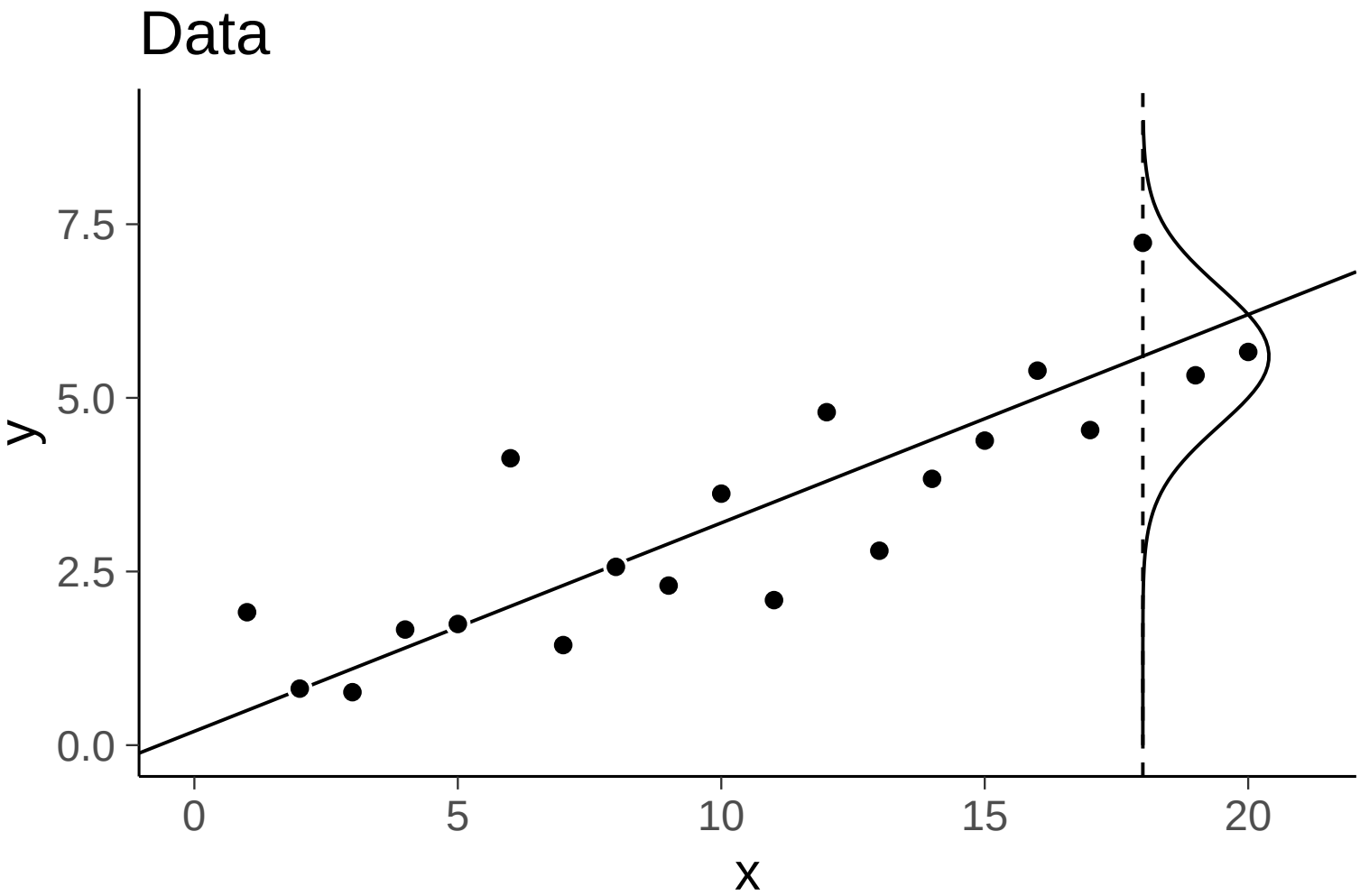
Stan and `loo` package

```
 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
```

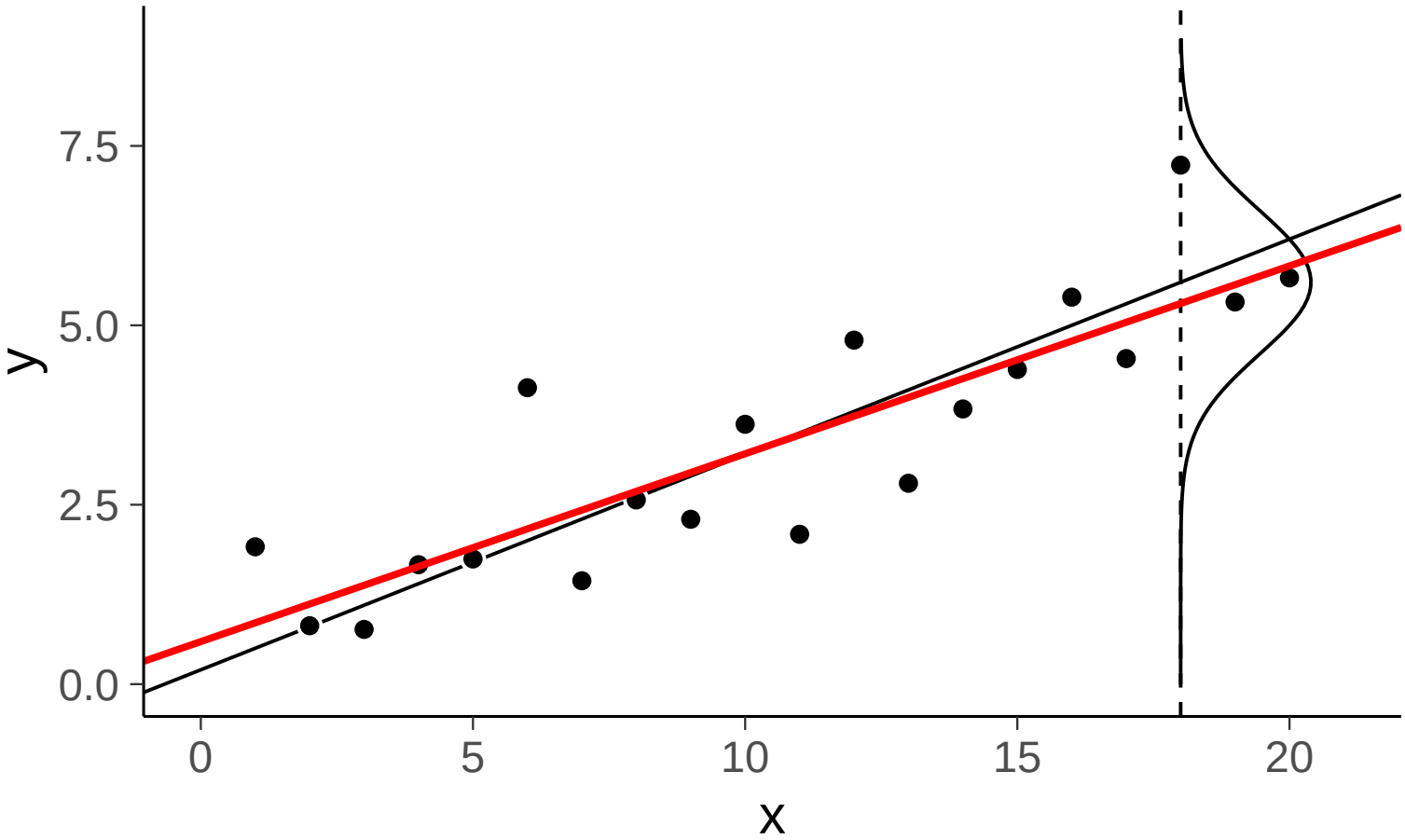

True mean y = a + bx
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

True mean and sigma

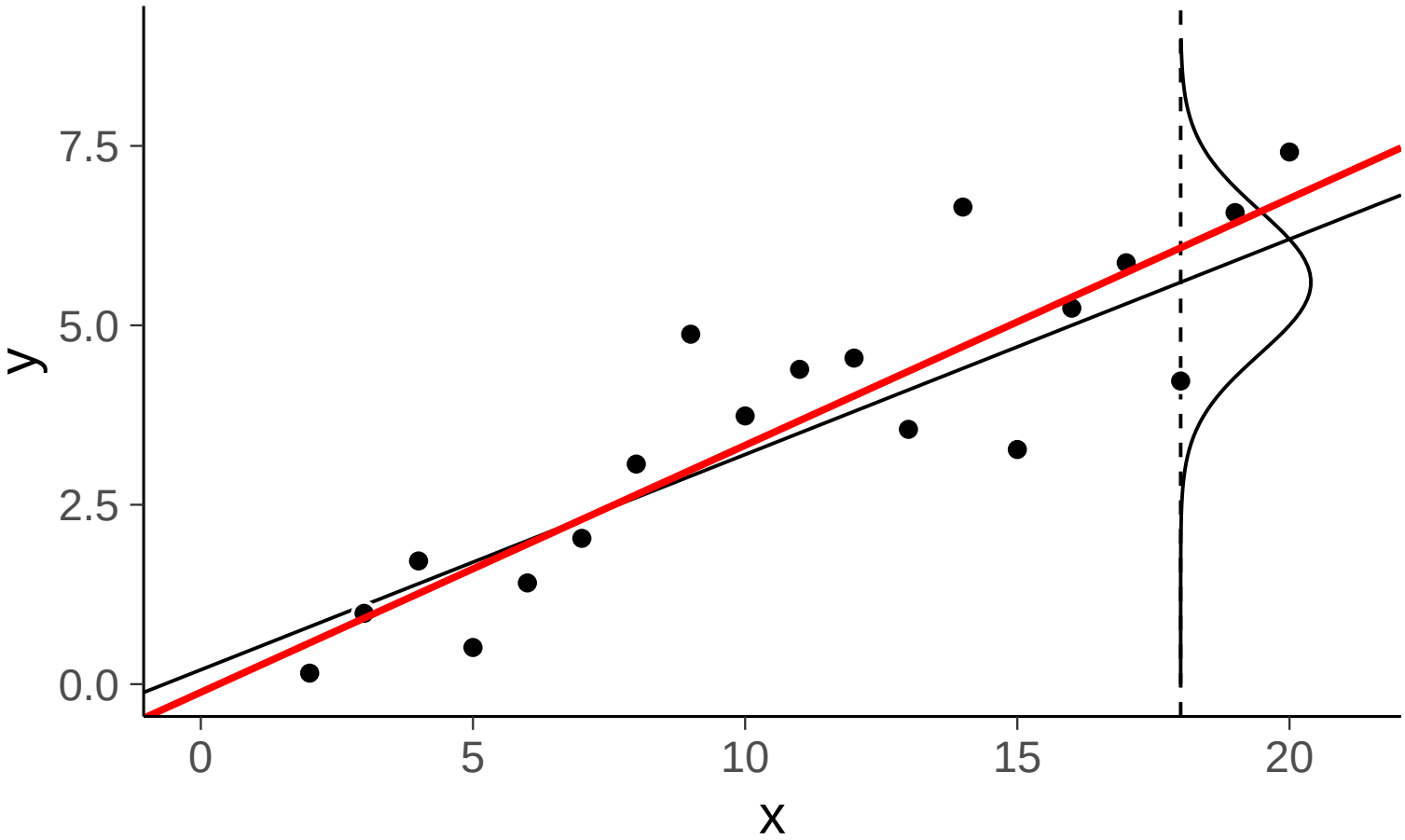



Data
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

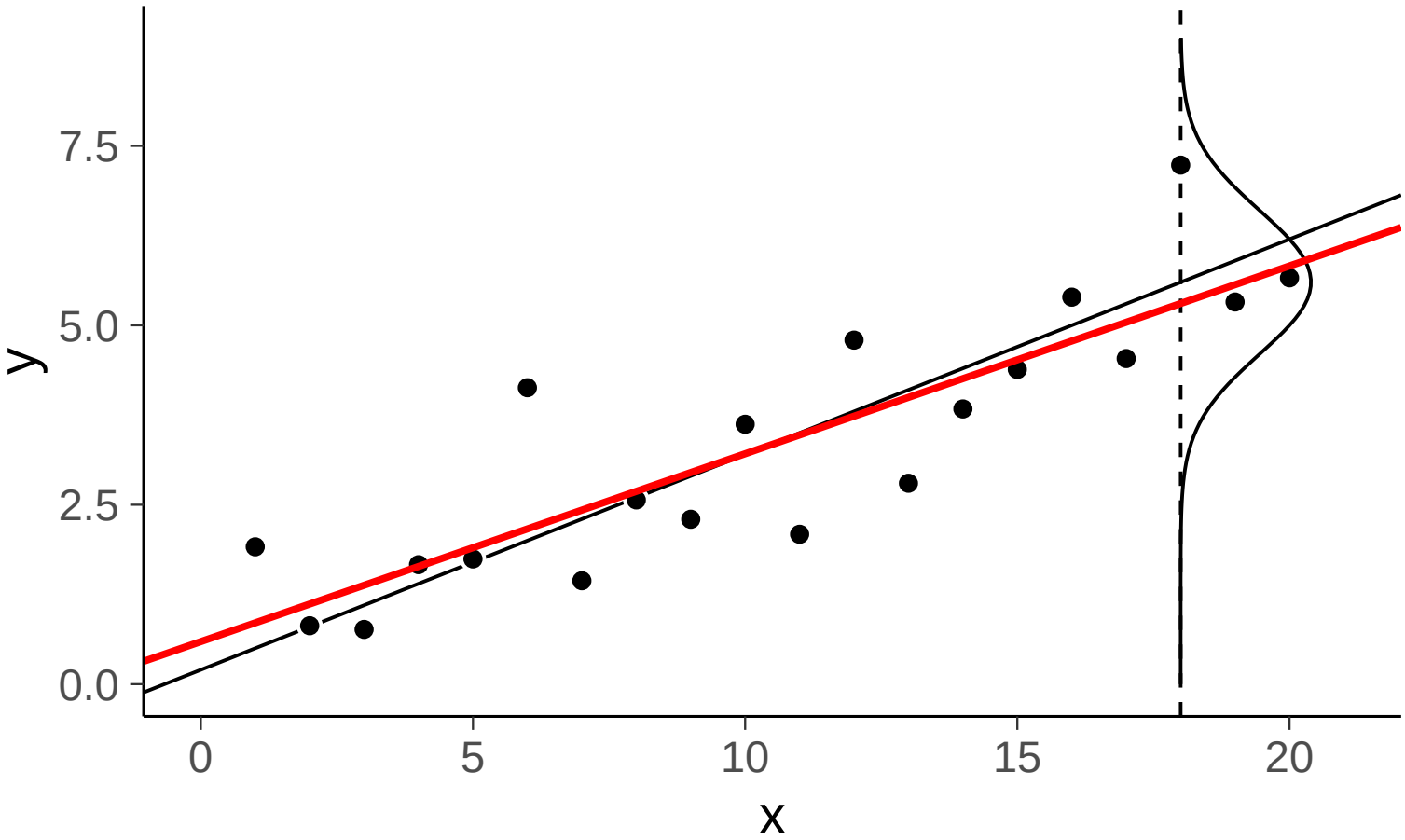
Posterior mean

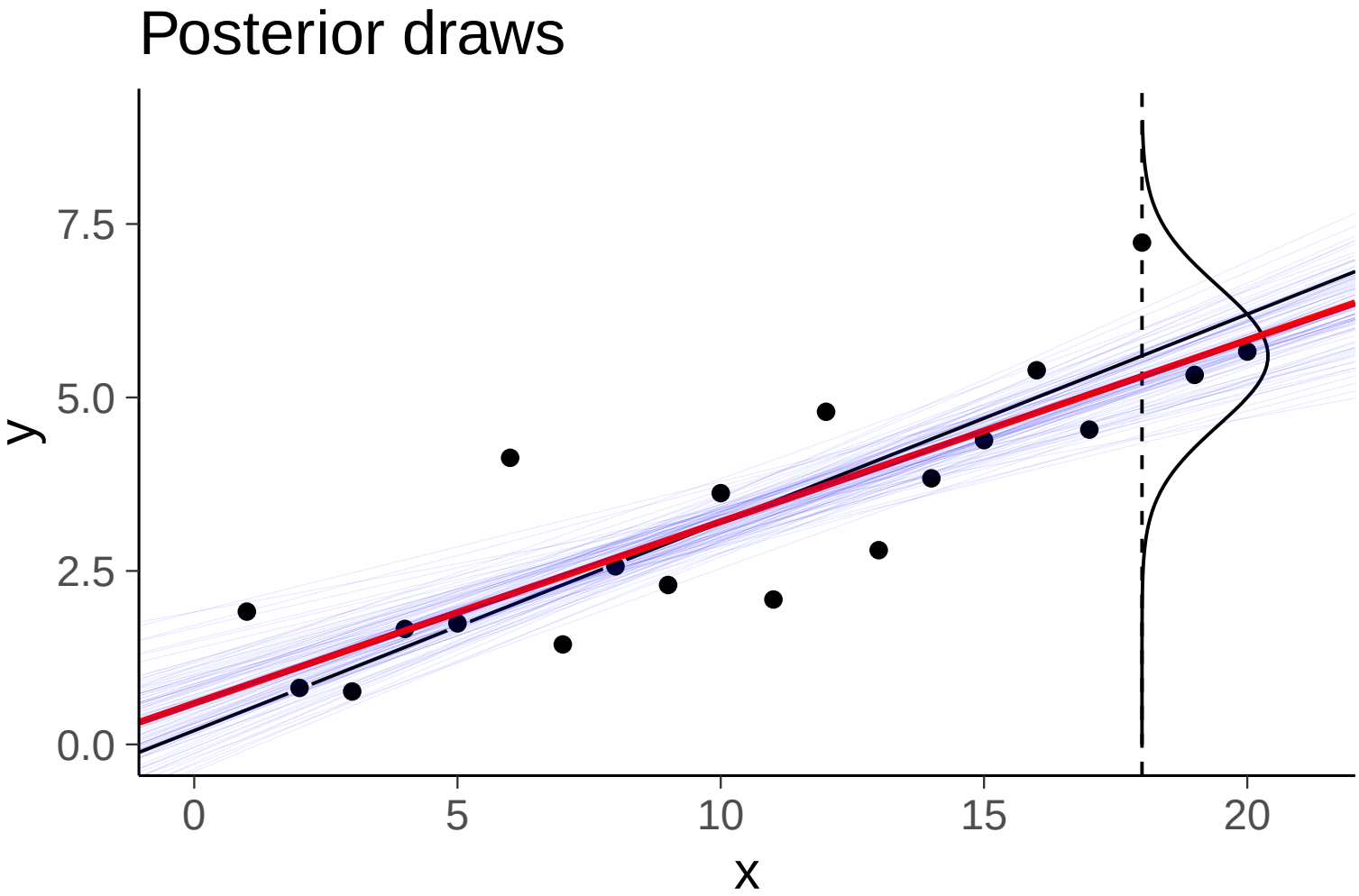


Posterior mean, alternative data realisation

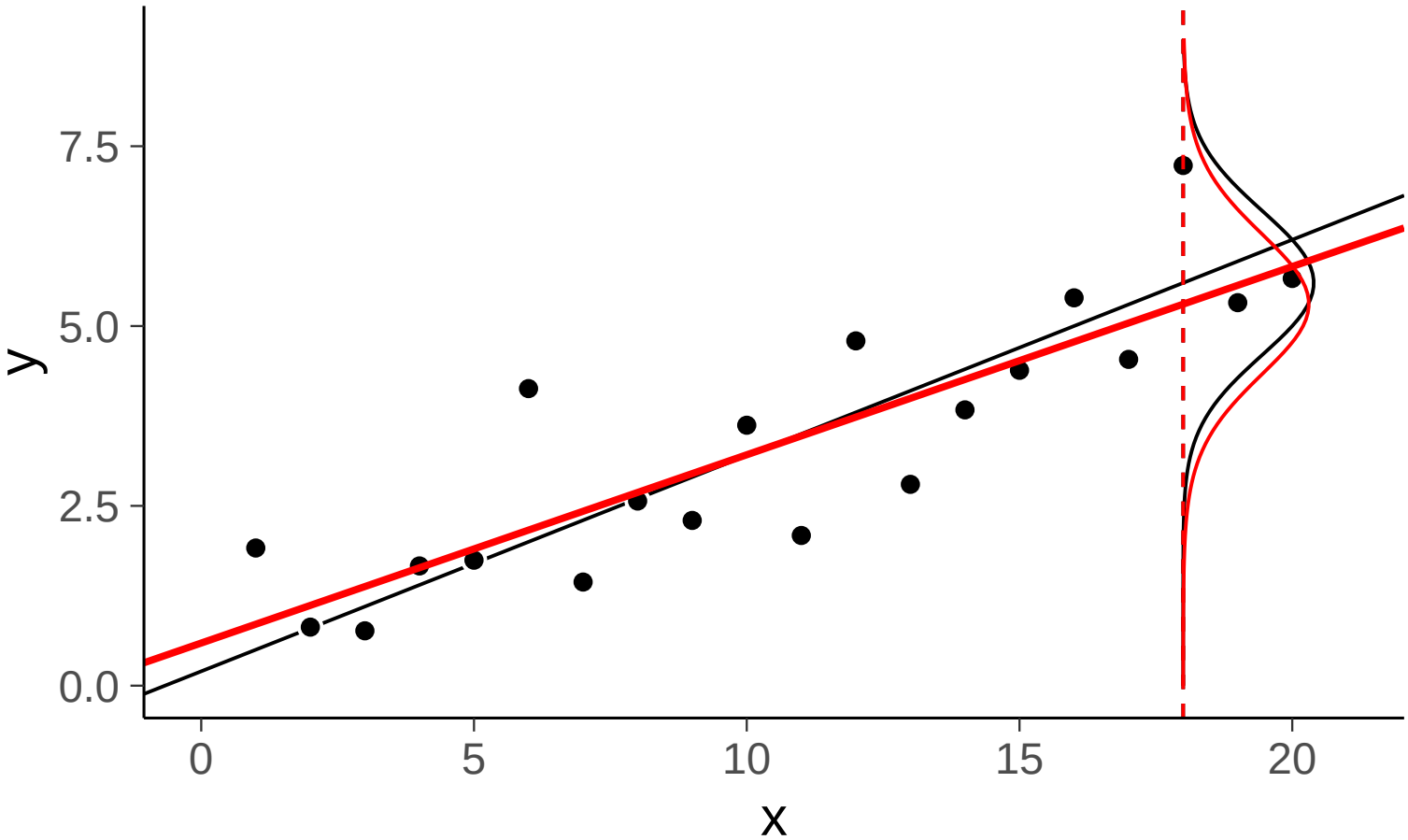


Posterior mean

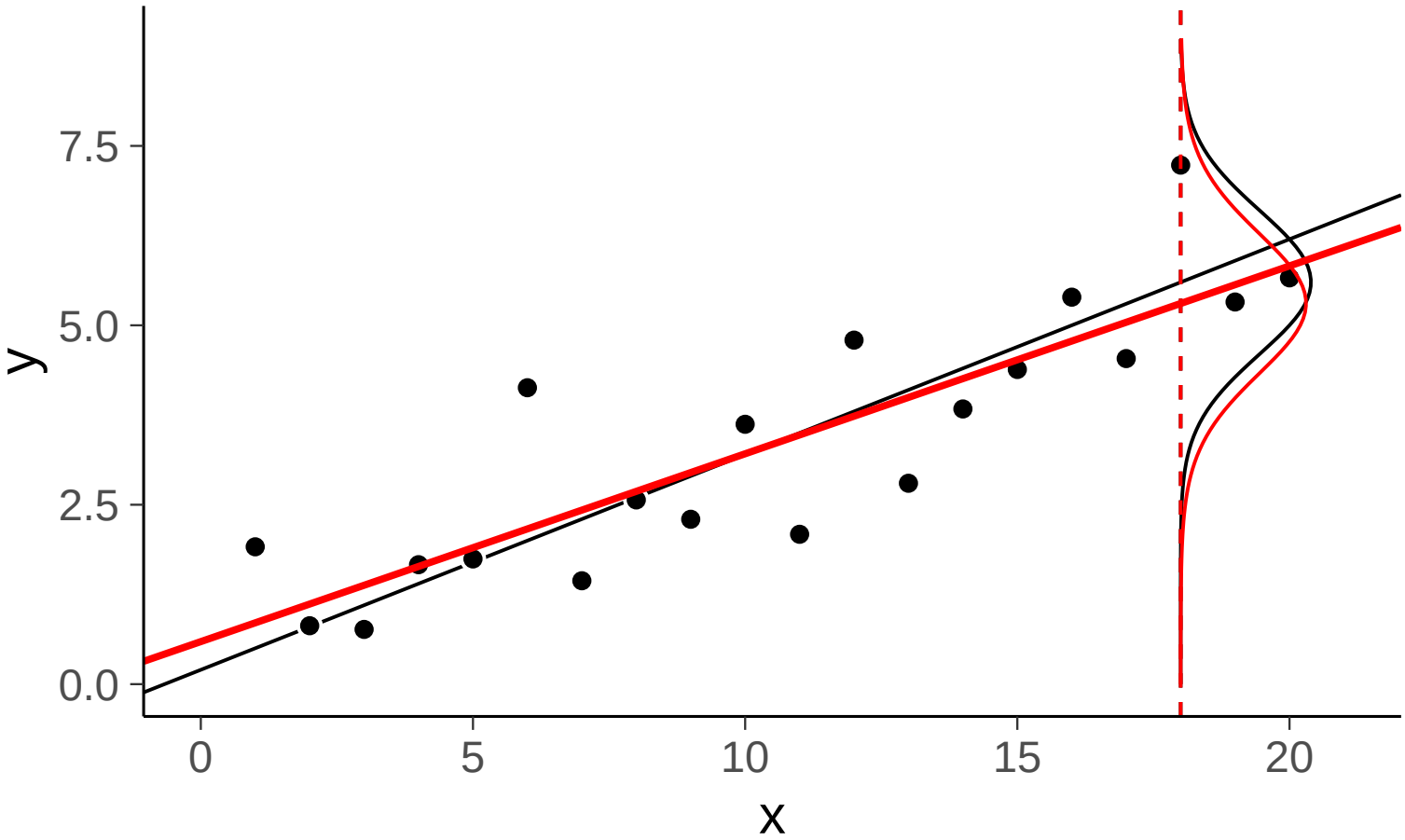



Posterior draws
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x

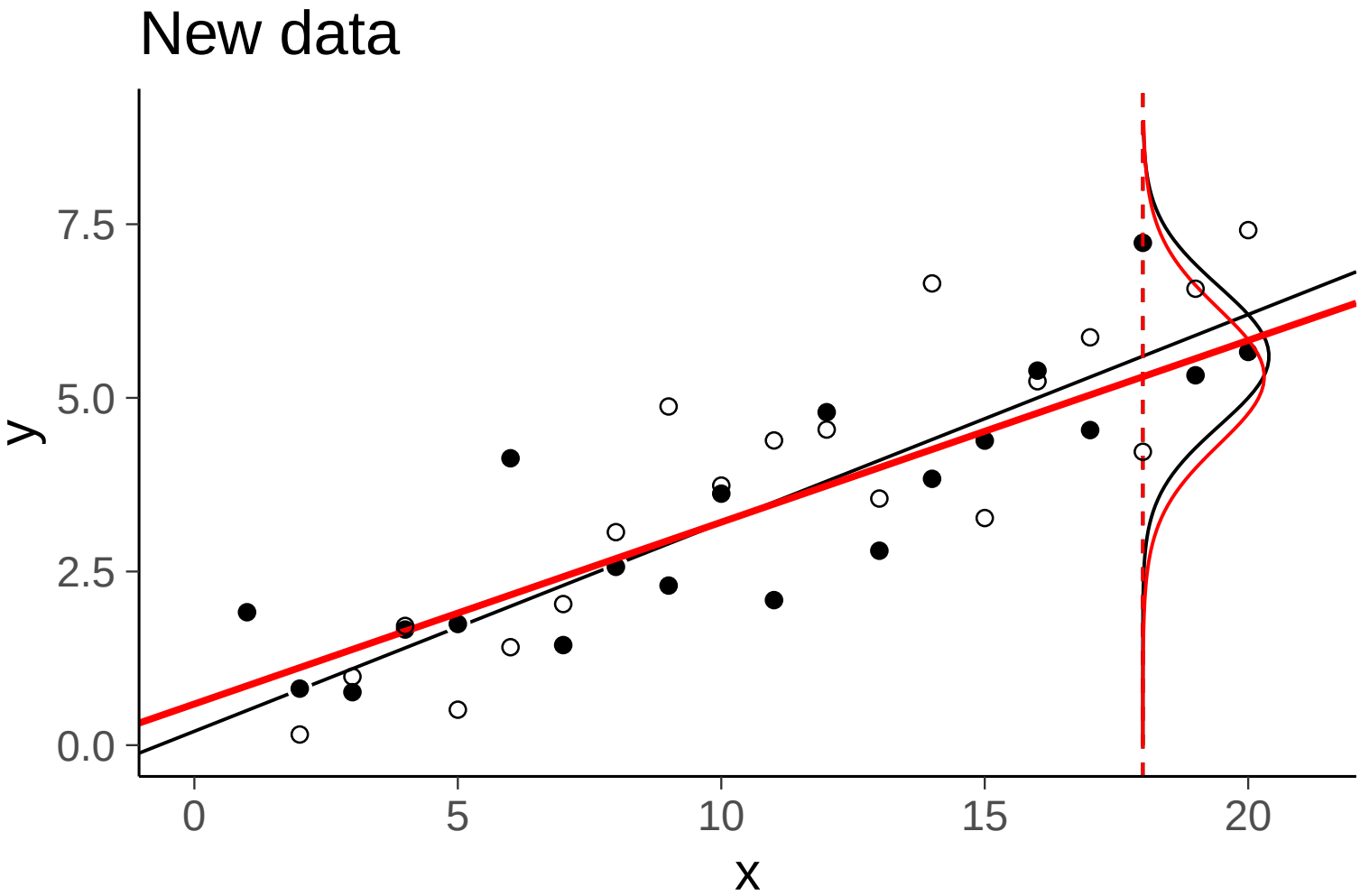
Posterior predictive distribution



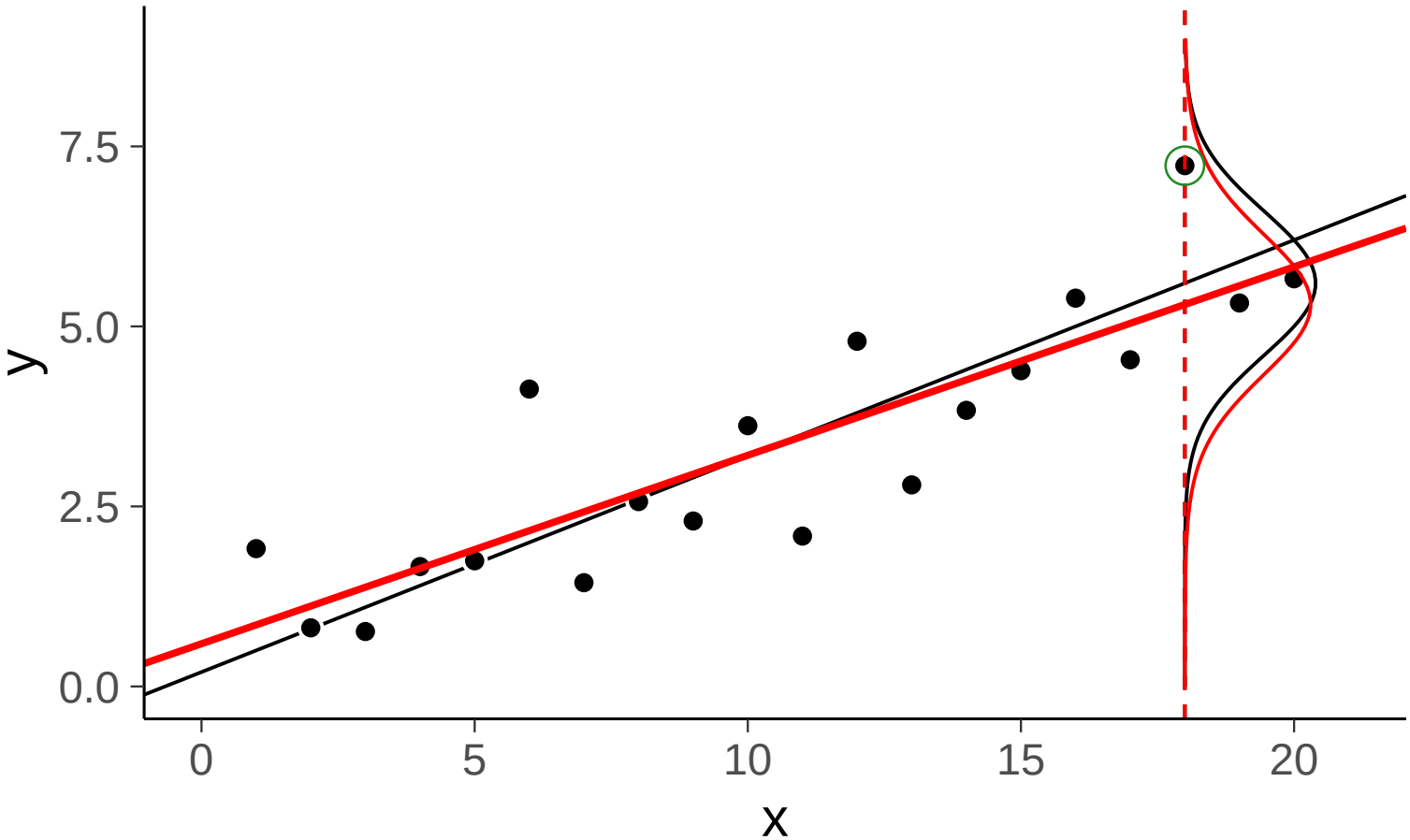
Posterior predictive distribution



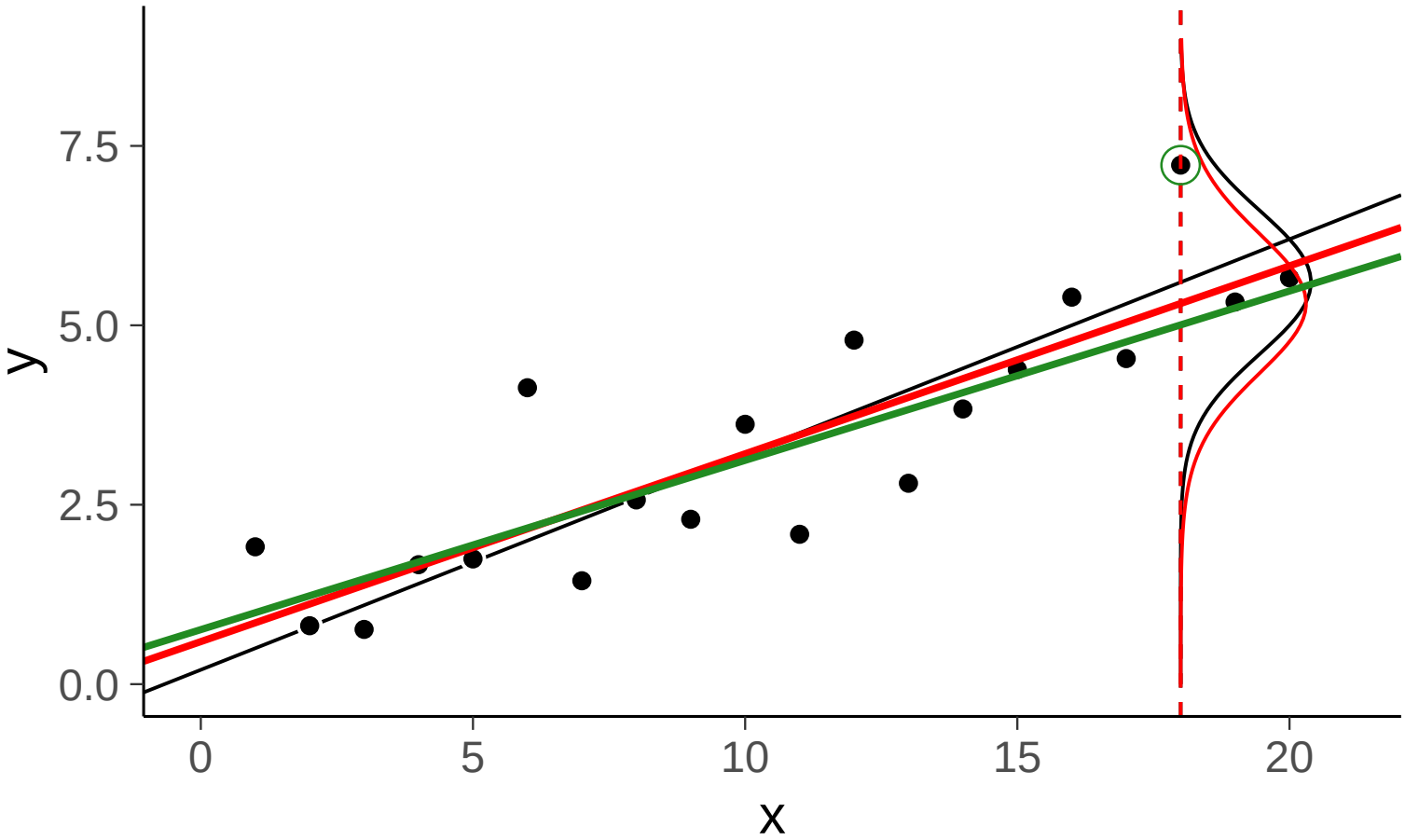
$$p(\tilde{y}|\tilde{x} = 18, x, y) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x, y)d\theta$$


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

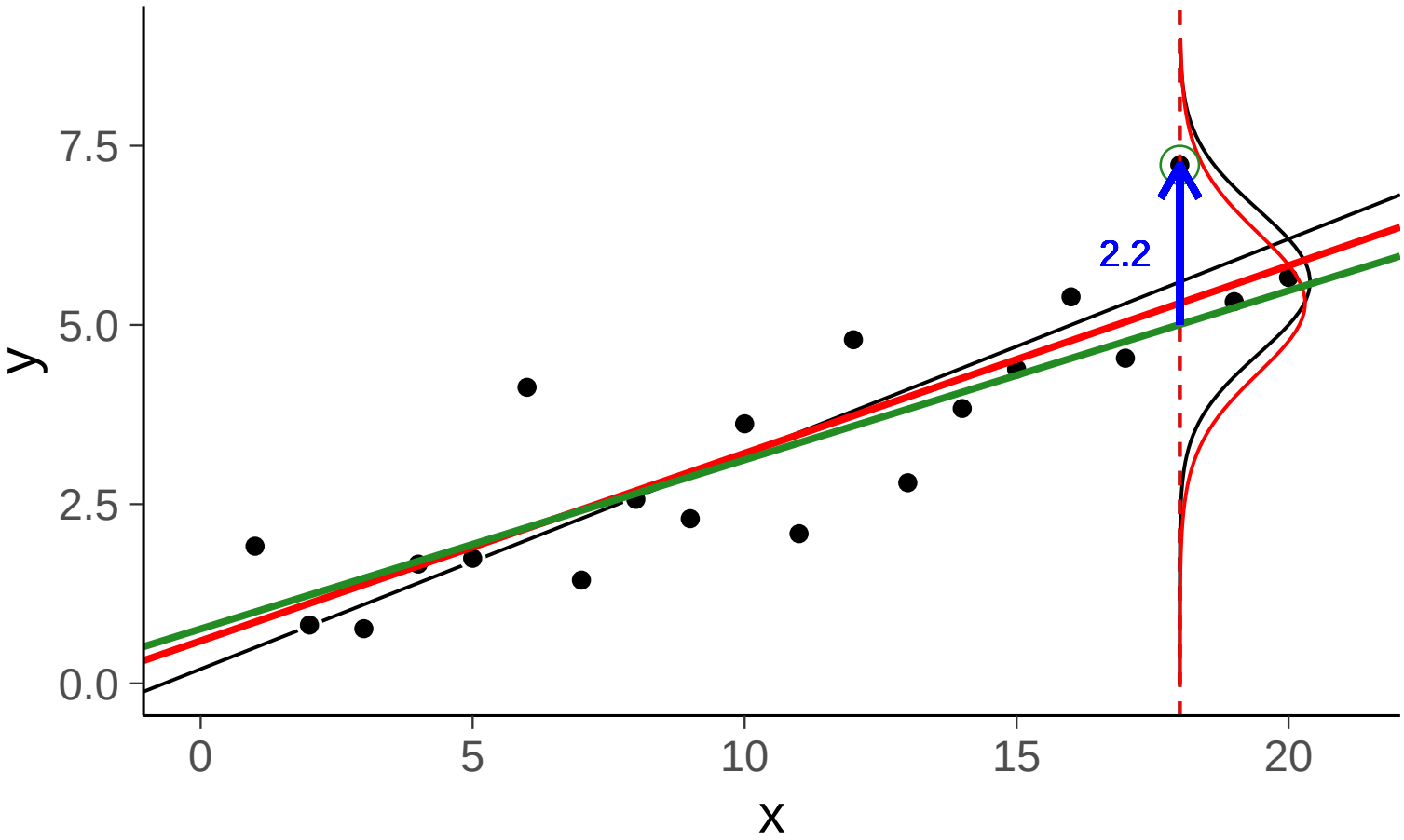
Posterior predictive distribution



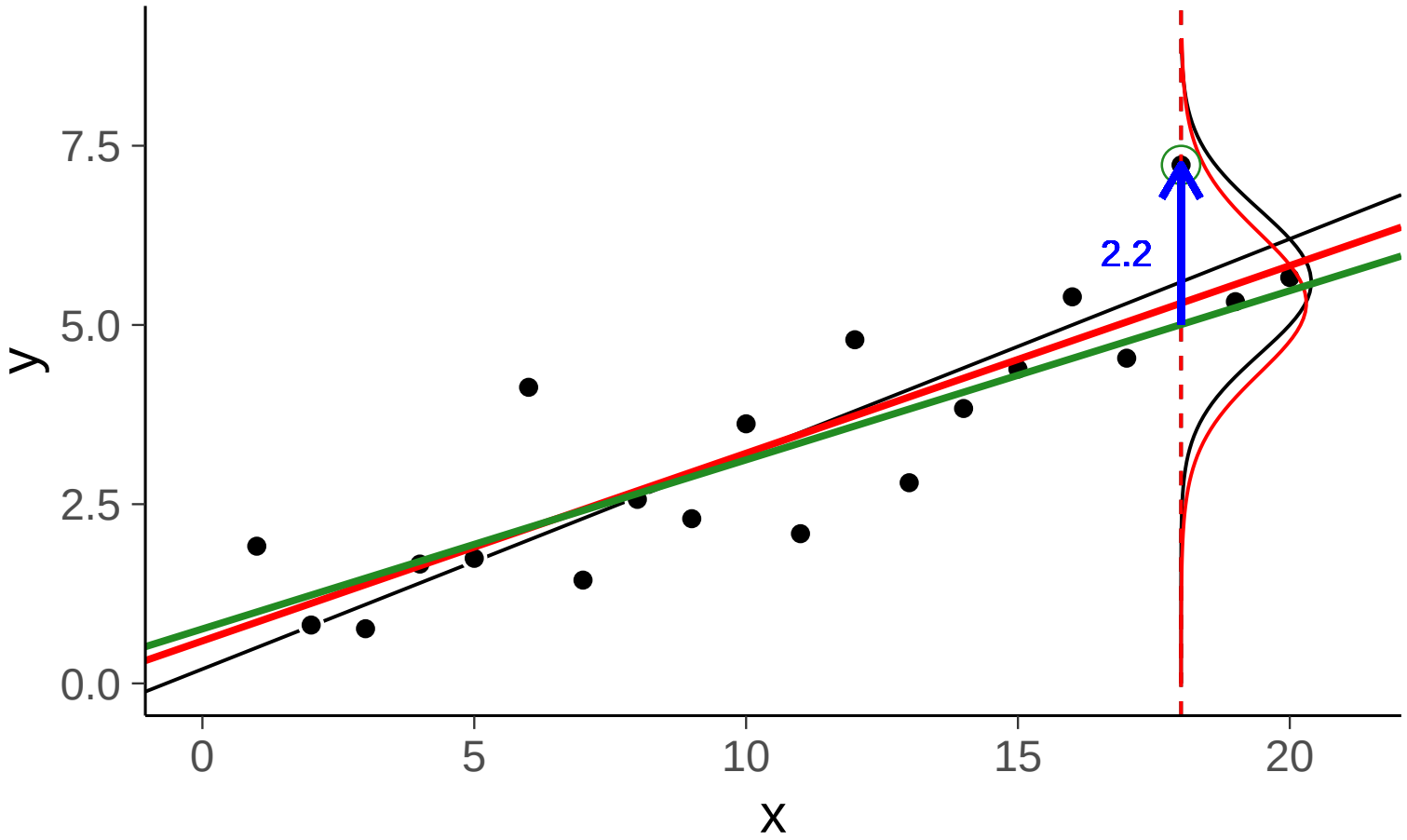
Leave−one−out mean



Leave−one−out residual

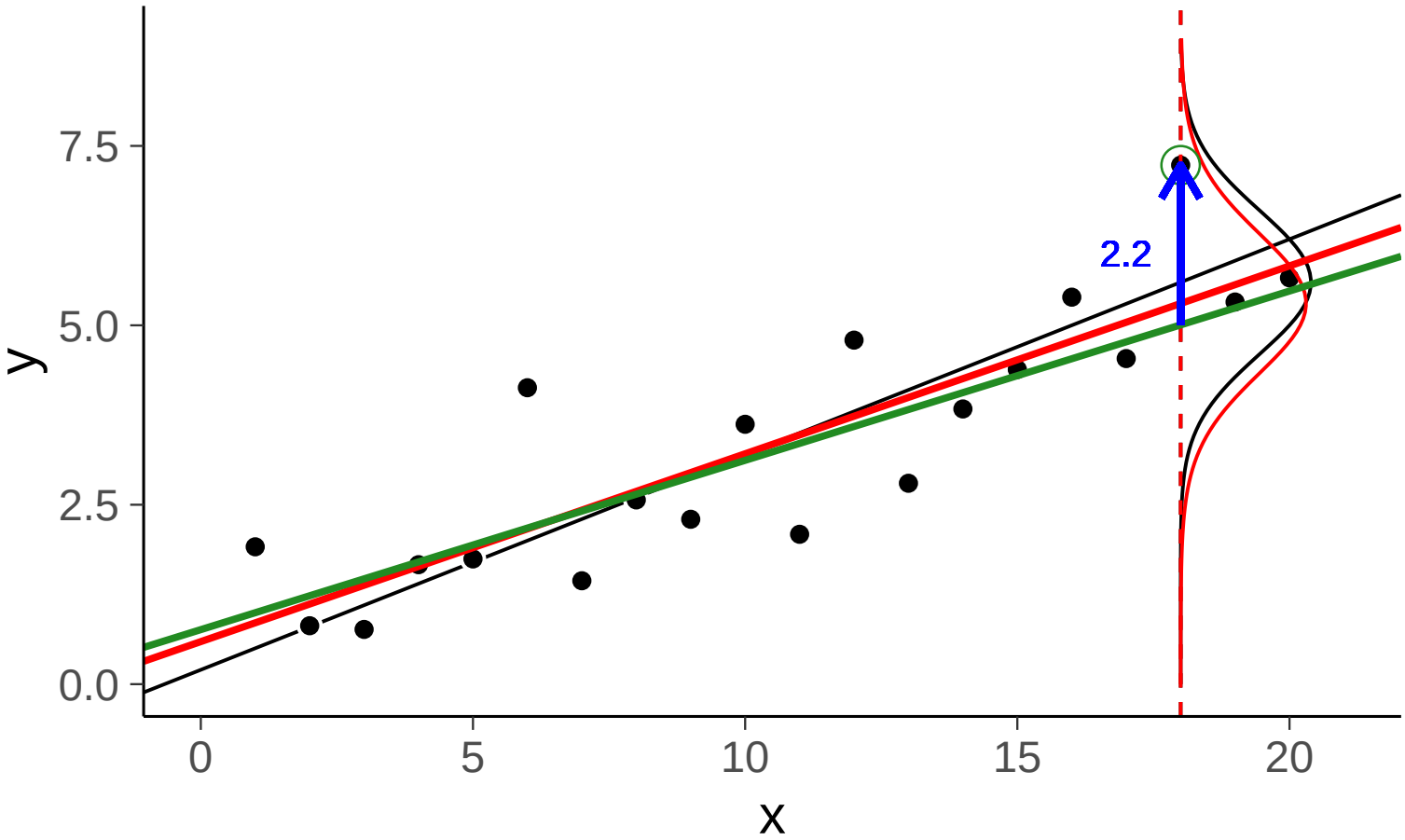


Leave−one−out residual



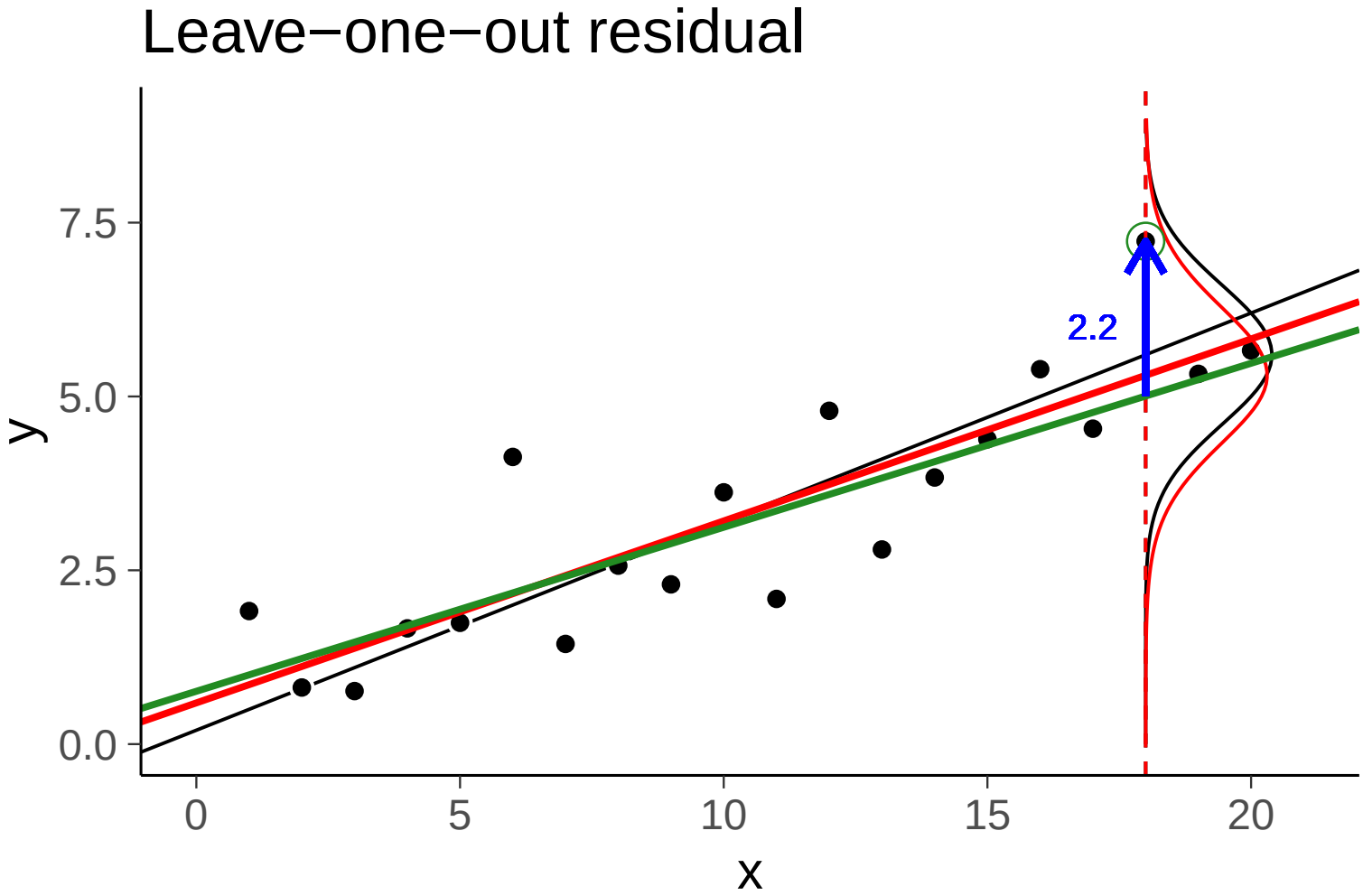
$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

Leave−one−out residual



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

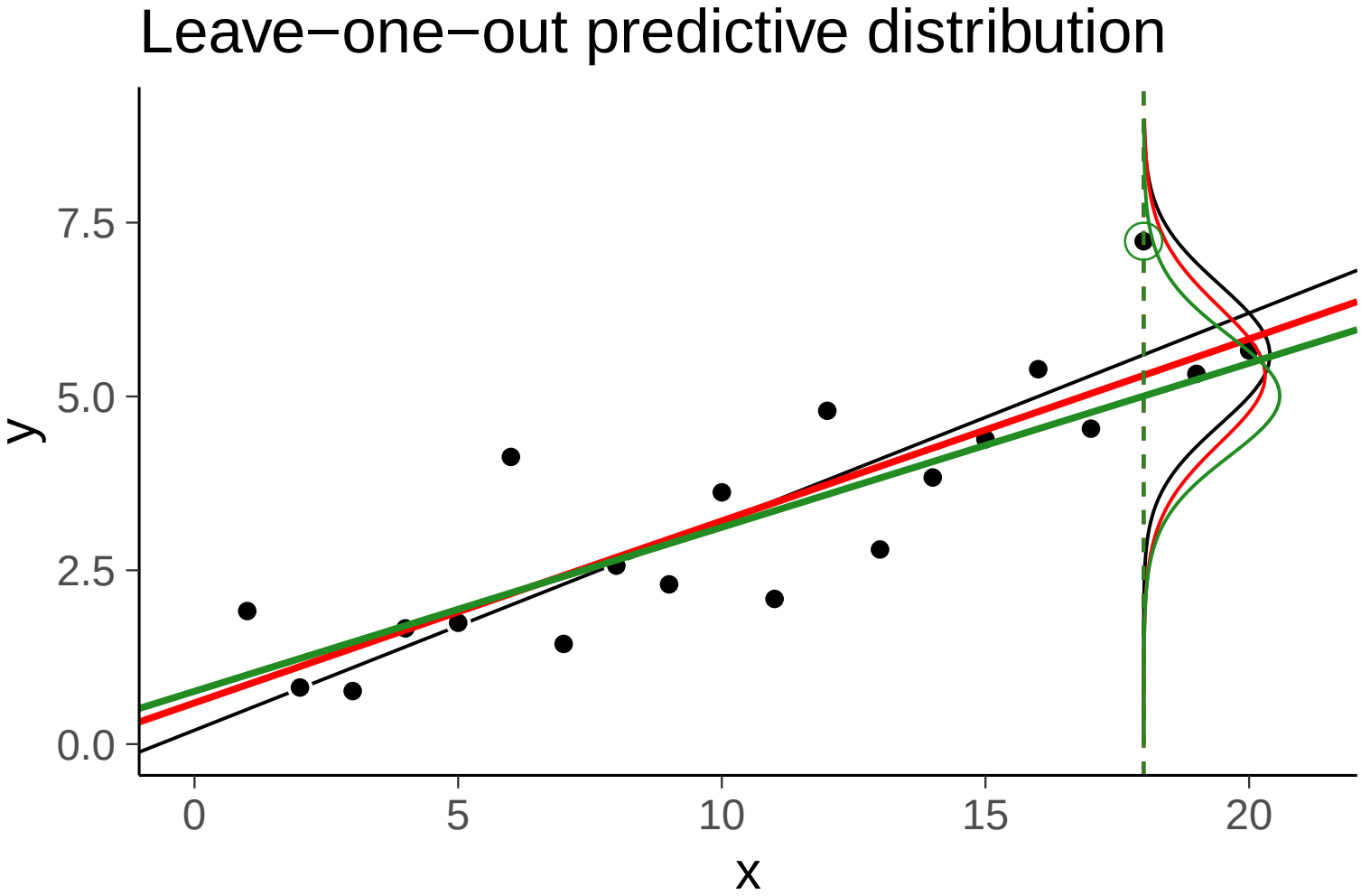
Can be use to compute, e.g., RMSE, R^2, 90% error



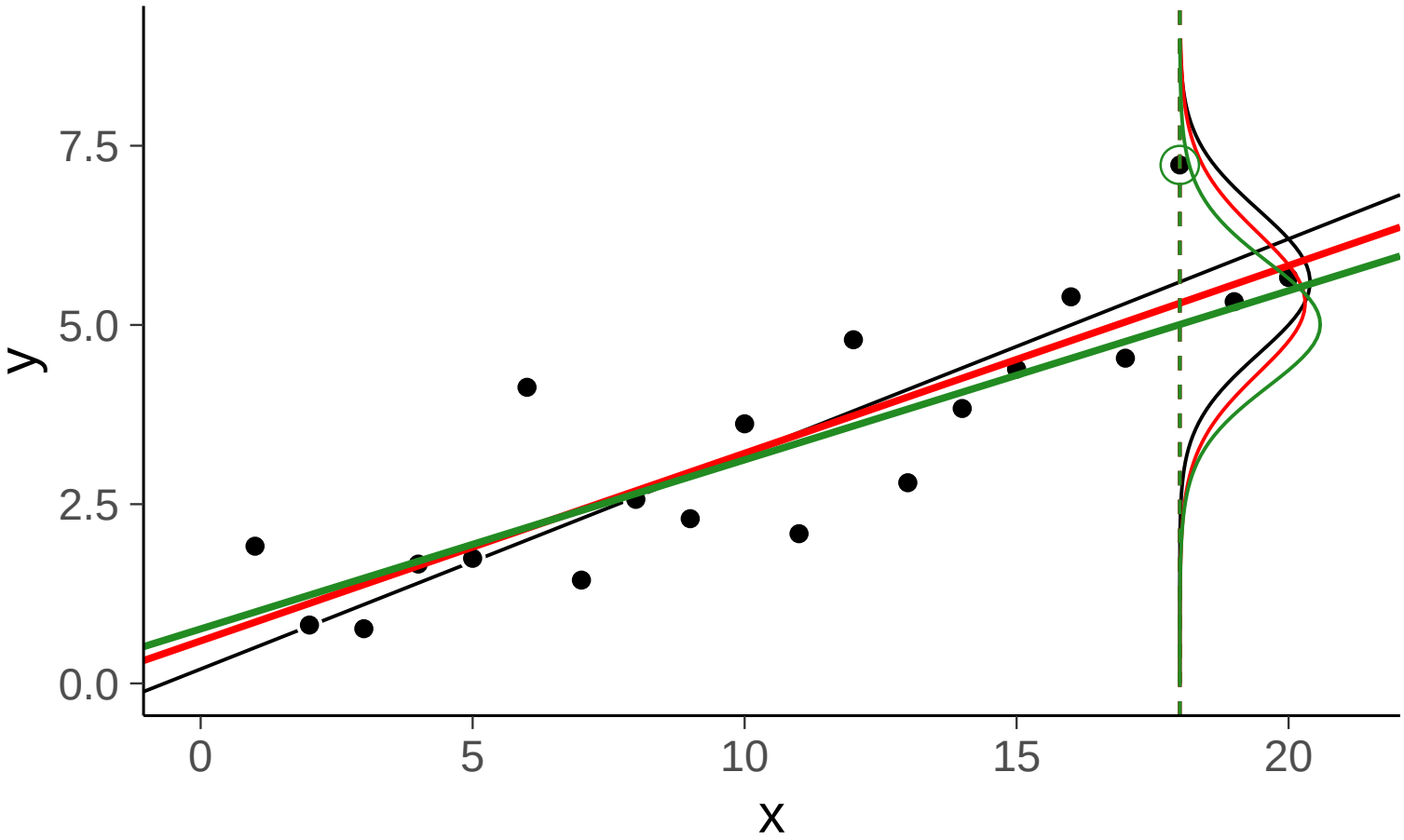
$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

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

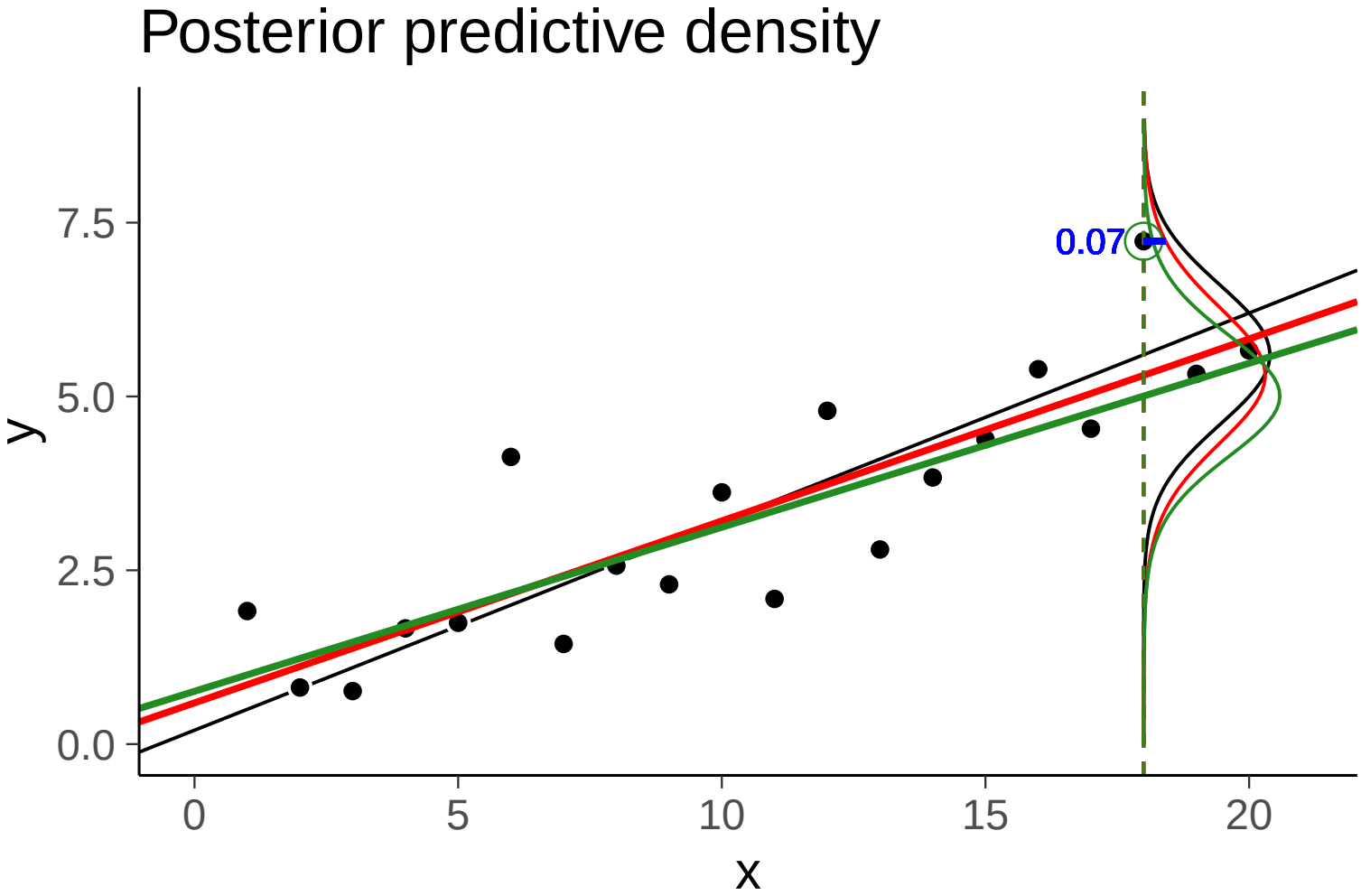
See LOO-R^2 at avehtari.github.io/bayes_R2/bayes_R2.html


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

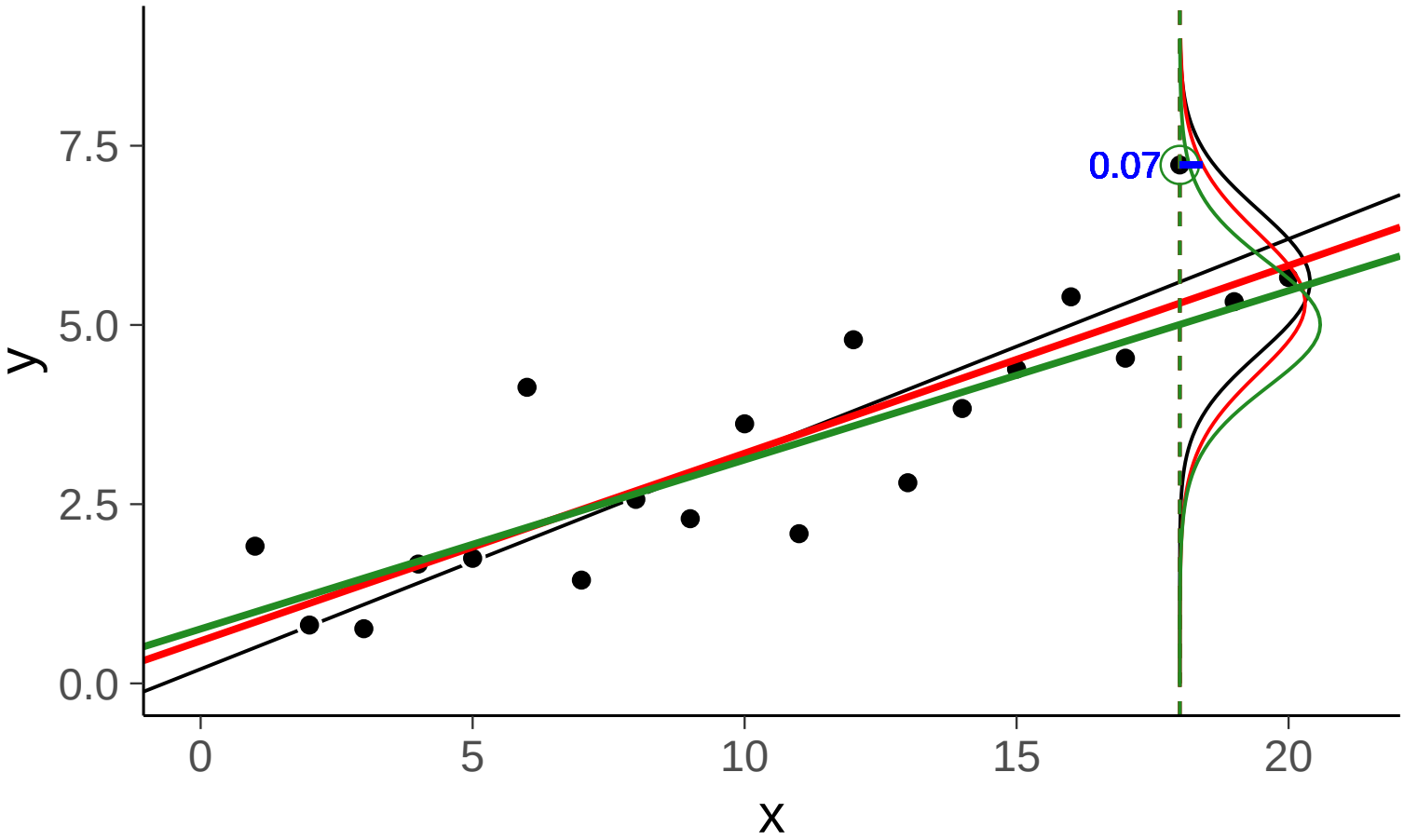
Leave−one−out predictive distribution



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

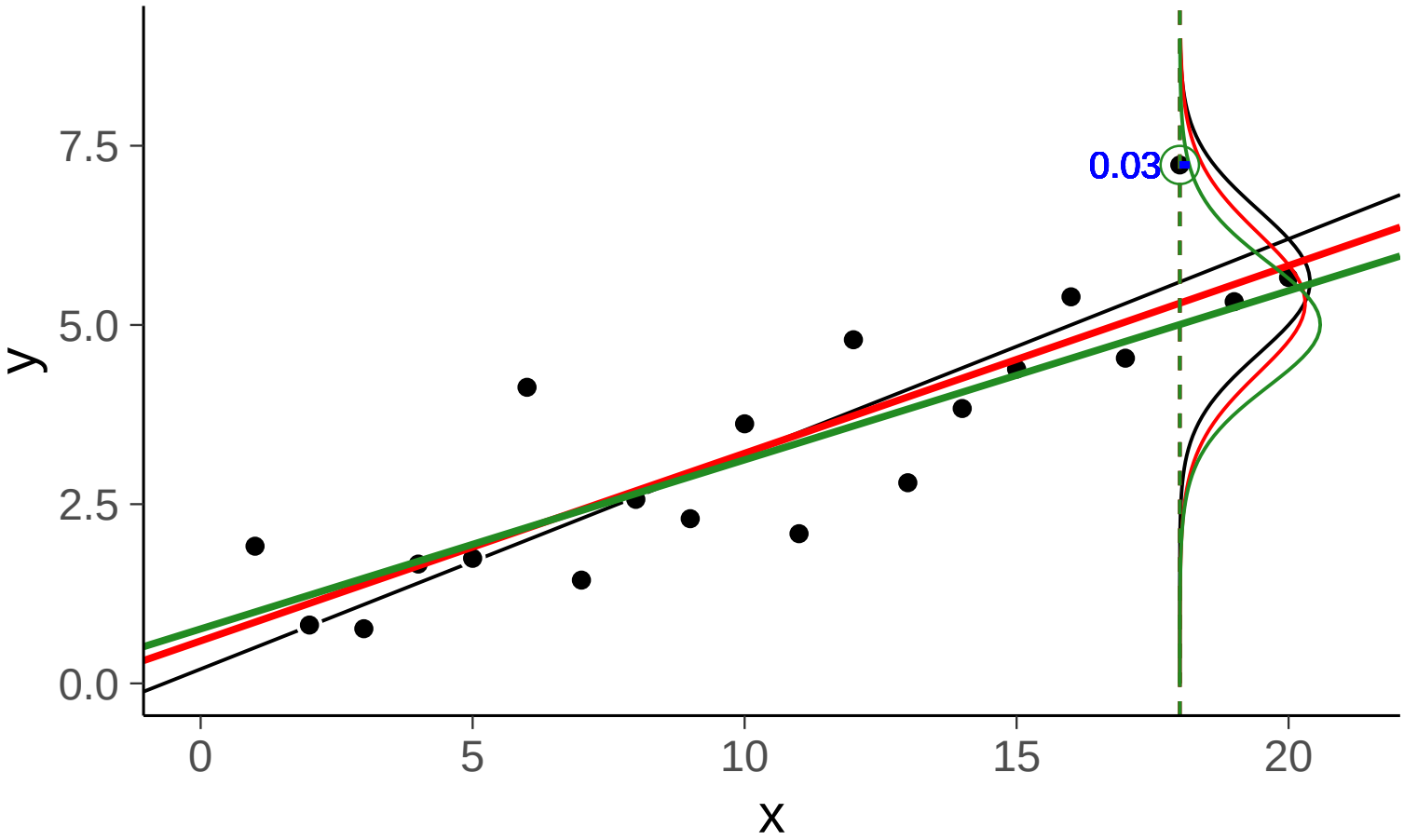

Posterior predictive density
7.5
5.0
2.5
0.0
y
0
5
10
15
20
x
0.07

Posterior predictive density



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

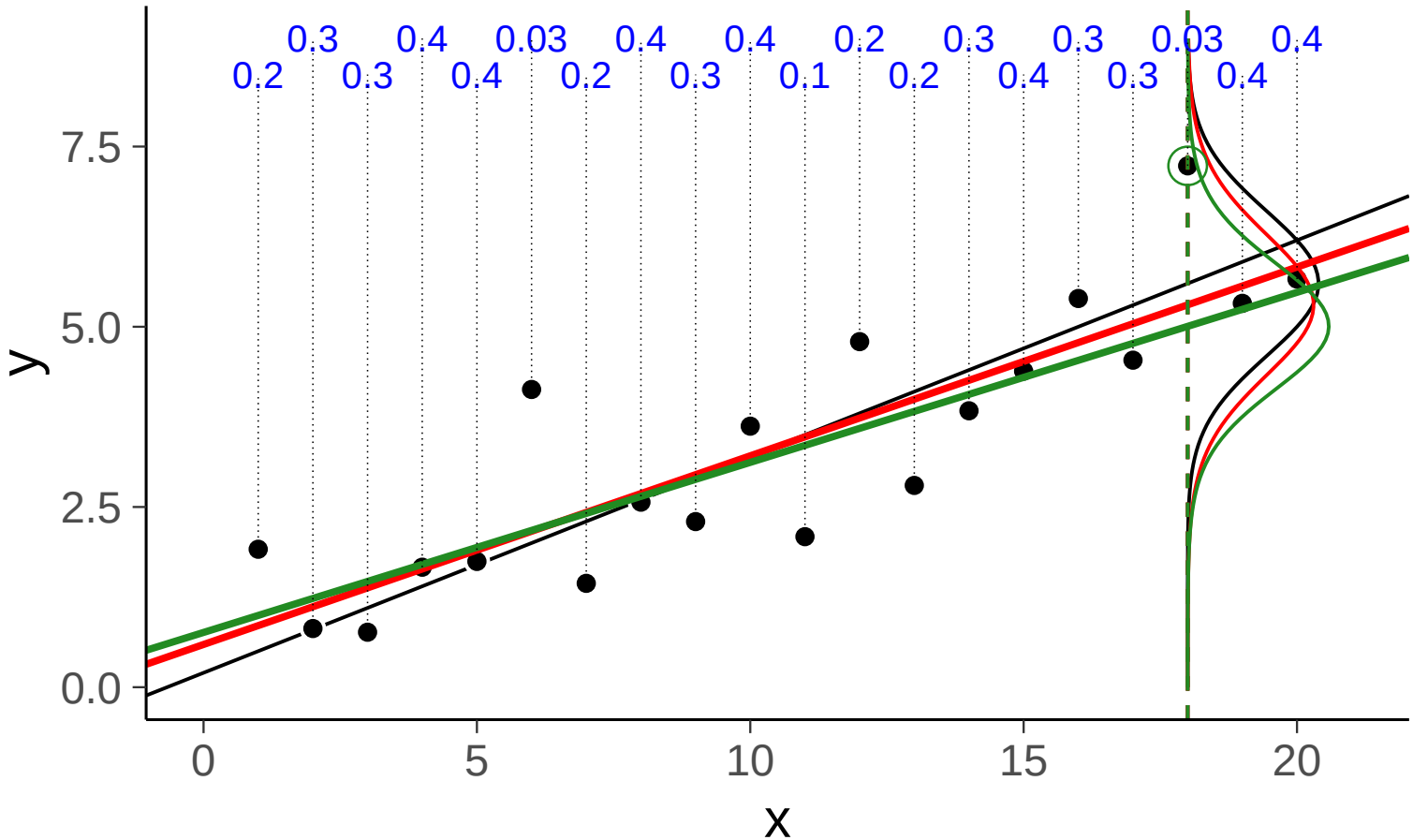
Leave−one−out predictive density



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

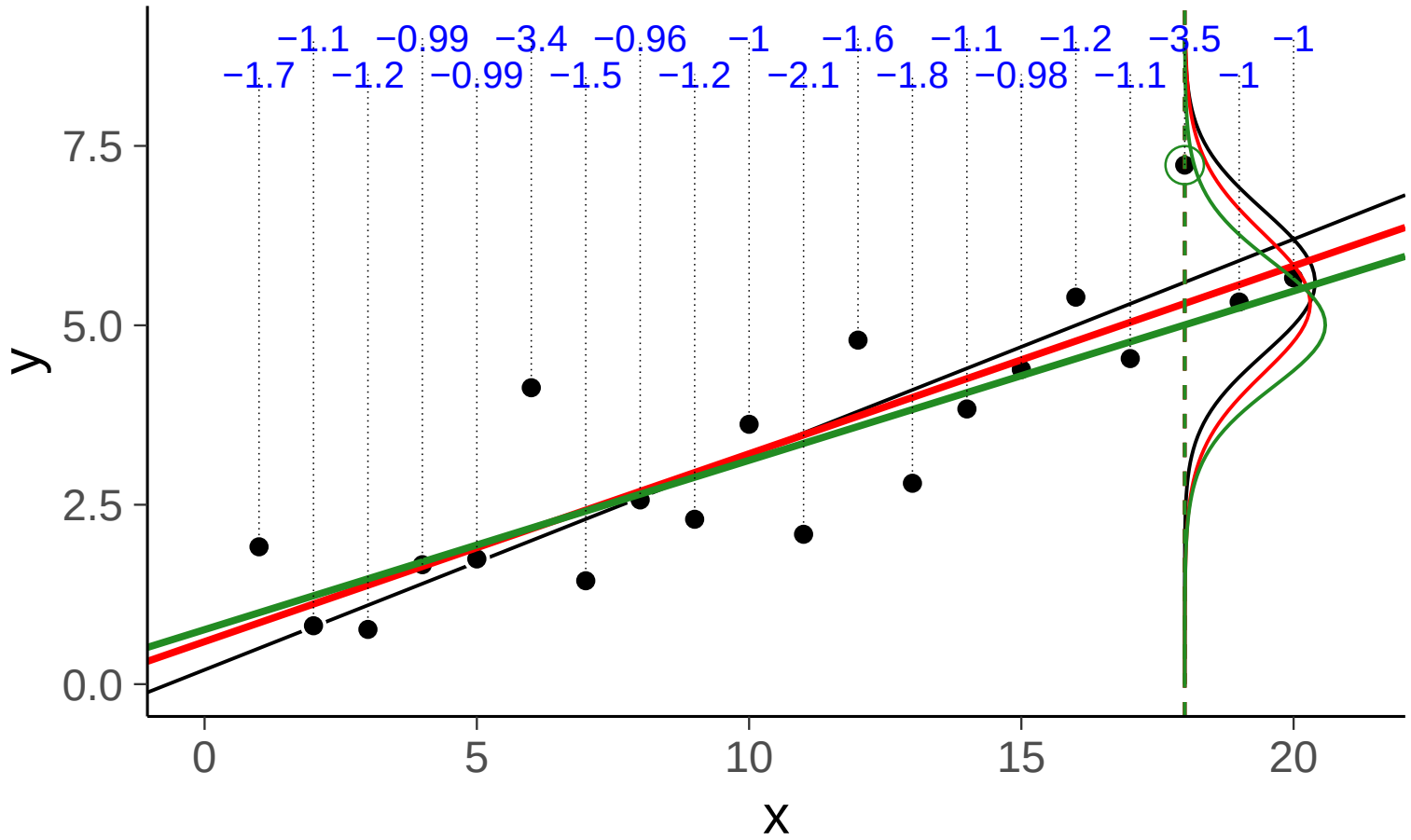
$p(\tilde{y} = y_{18} | \tilde{x} = 18, x_{-18}, y_{-18}) \approx 0.03$

Leave−one−out predictive densities



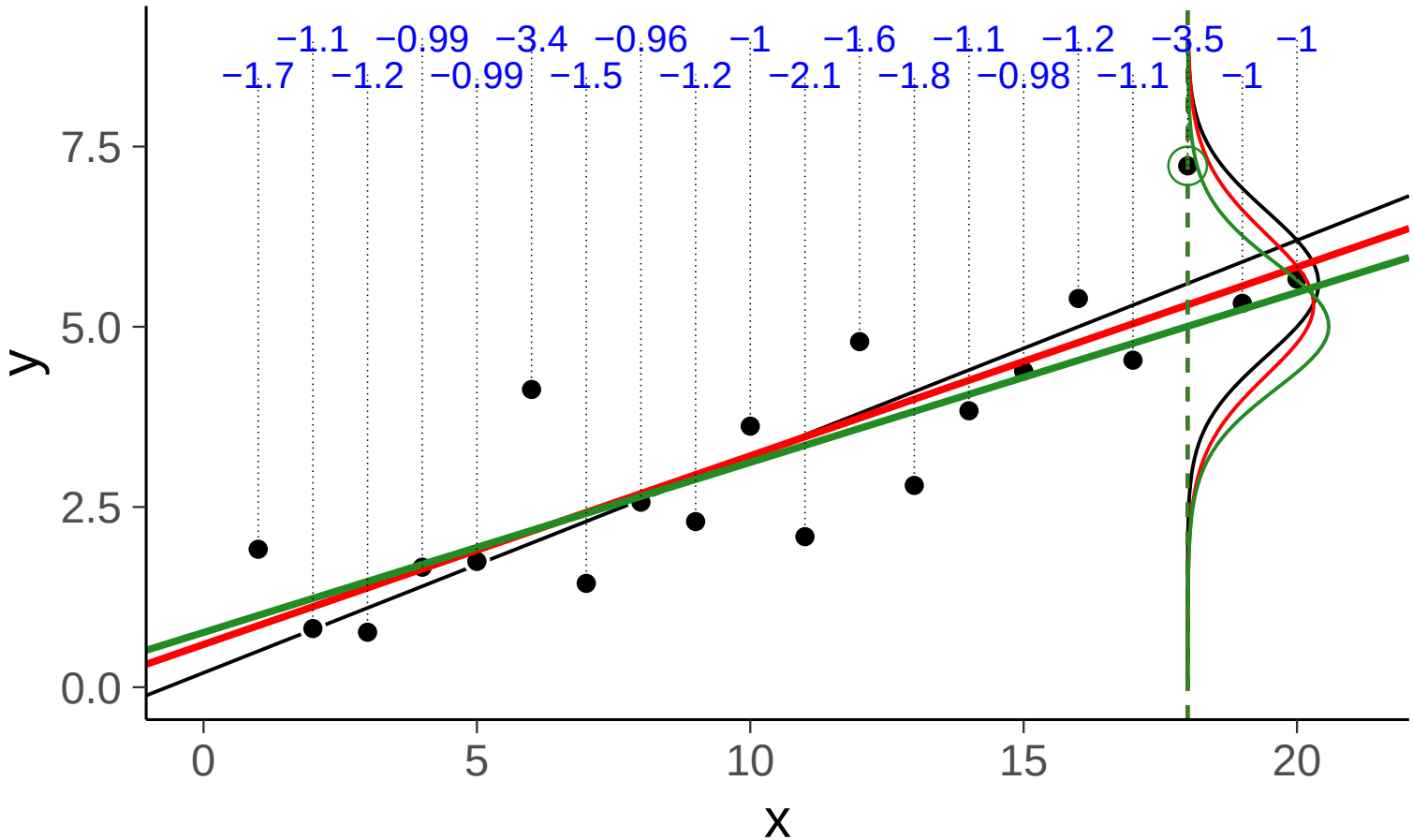
$p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

Leave−one−out log predictive densities



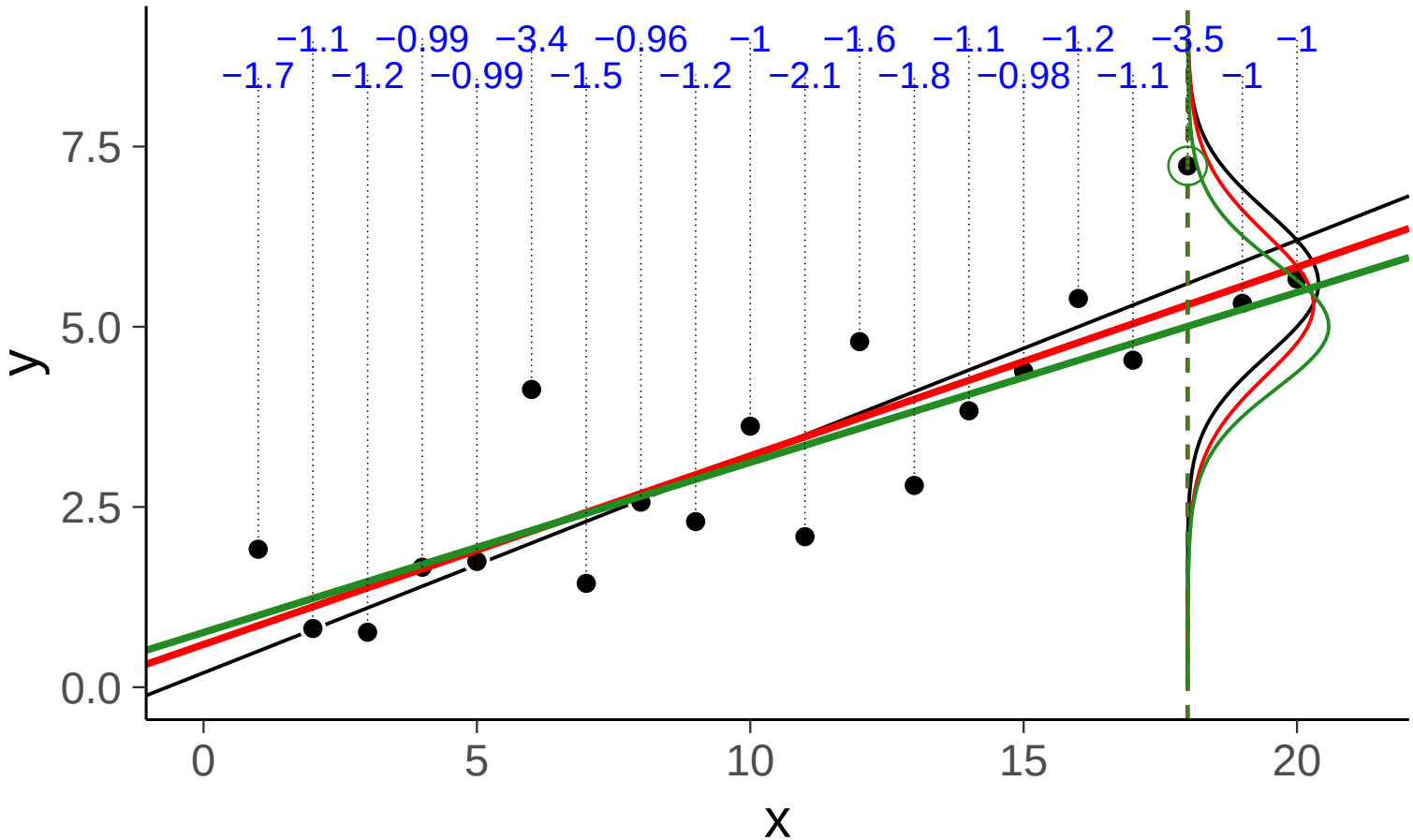
$\log p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

Leave−one−out log predictive densities



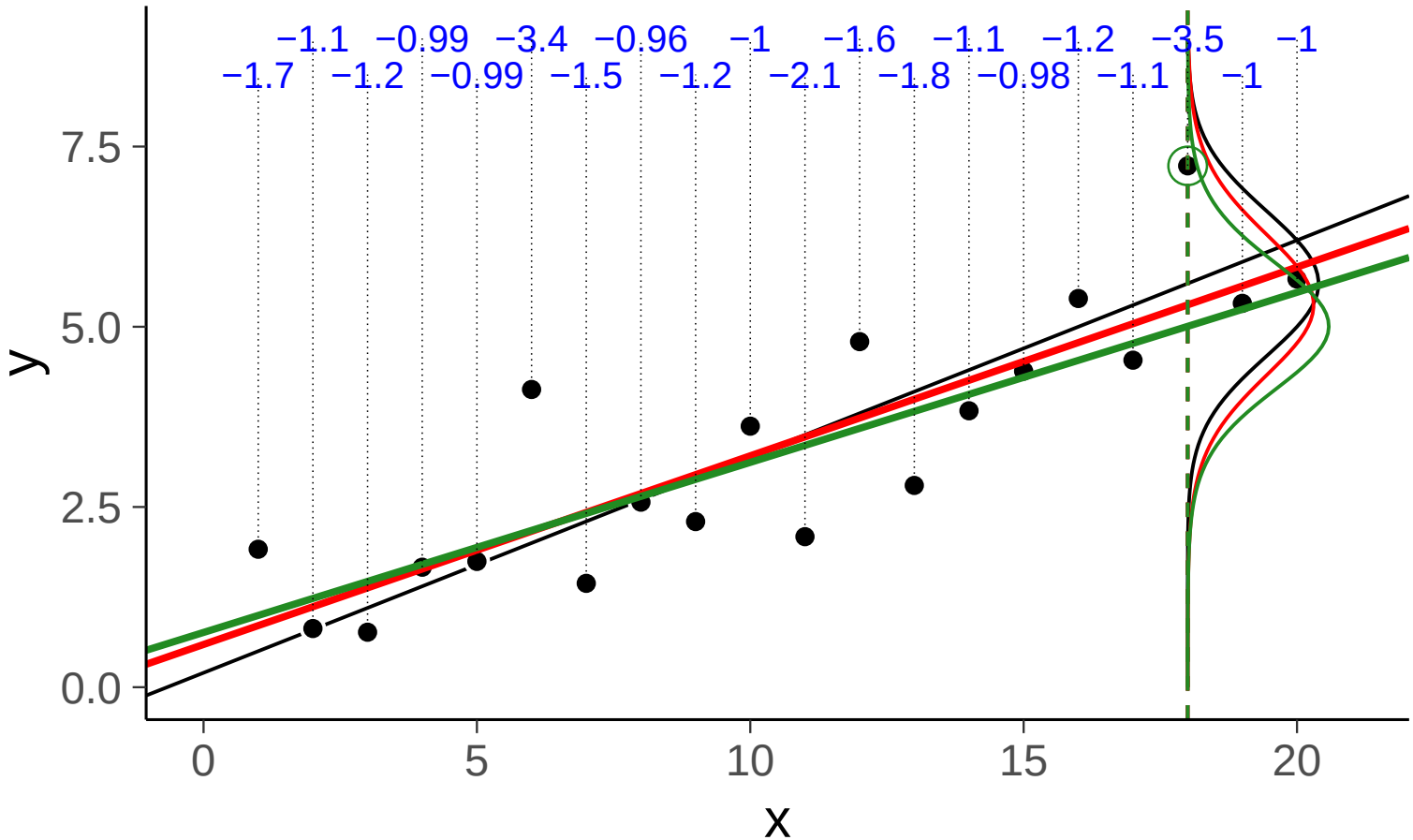
$\sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

Leave−one−out log predictive densities



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

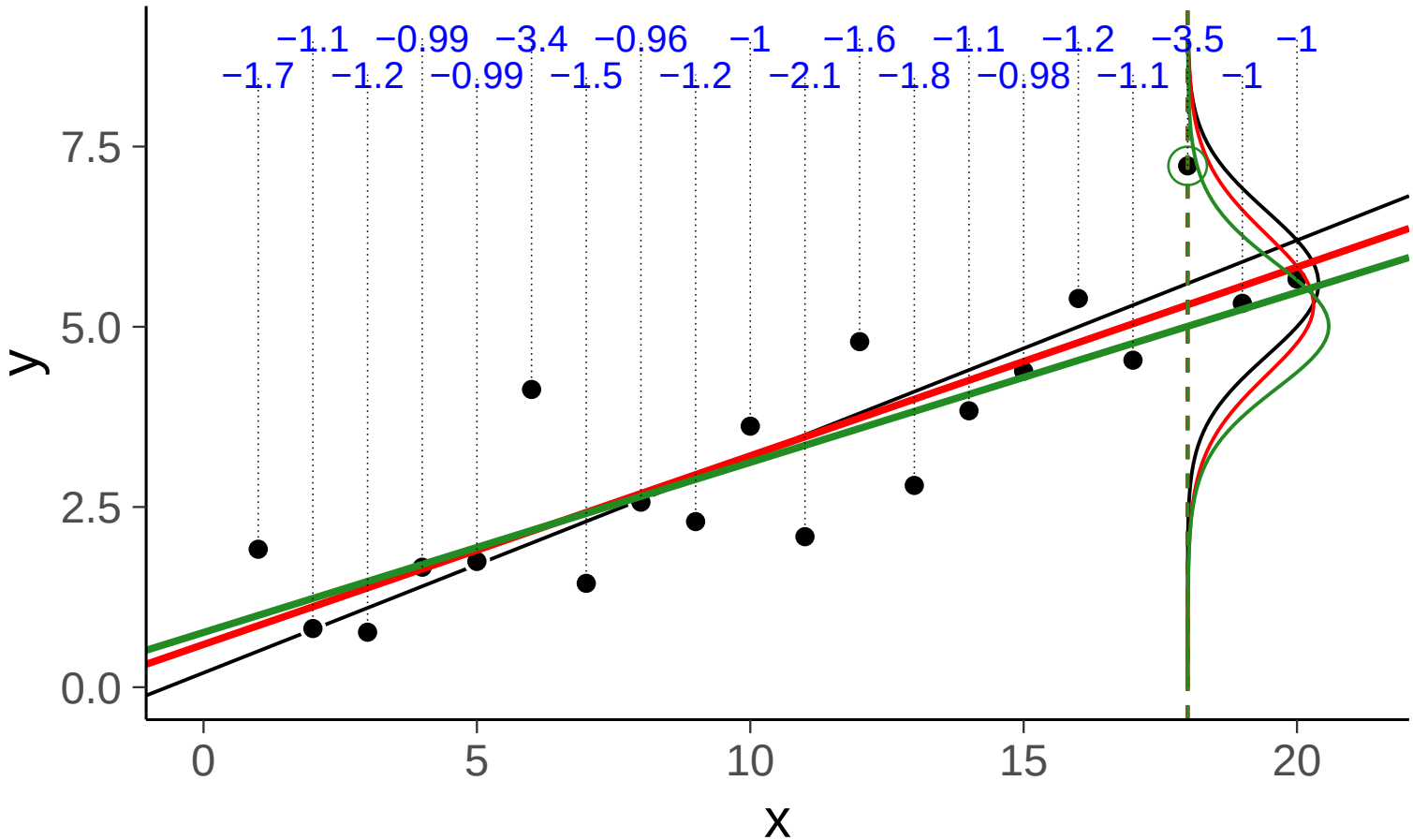
Leave−one−out log predictive densities



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

unbiased estimate of log posterior pred. density for new data

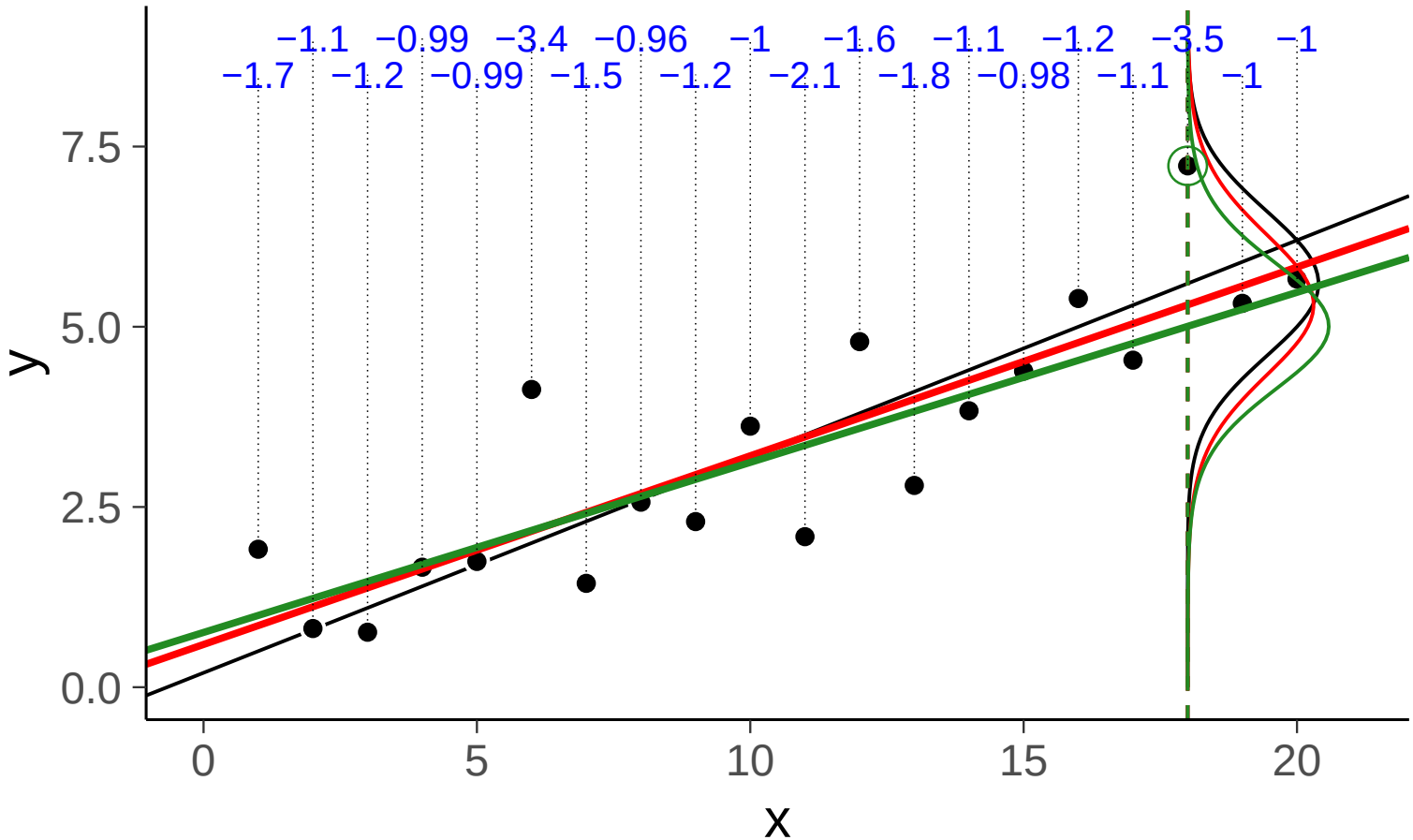
Leave−one−out log predictive densities



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

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

Leave−one−out log predictive densities

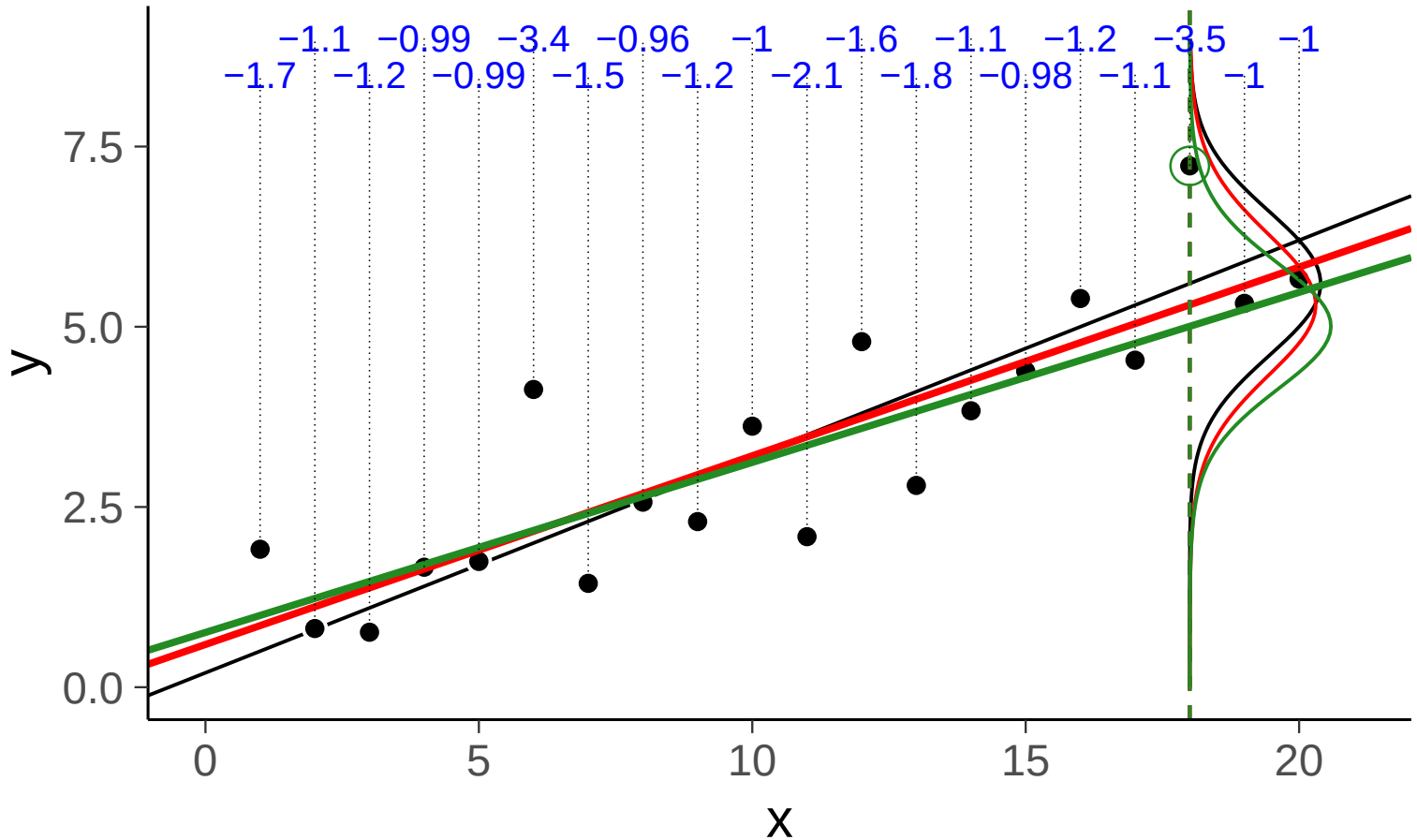


$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

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

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

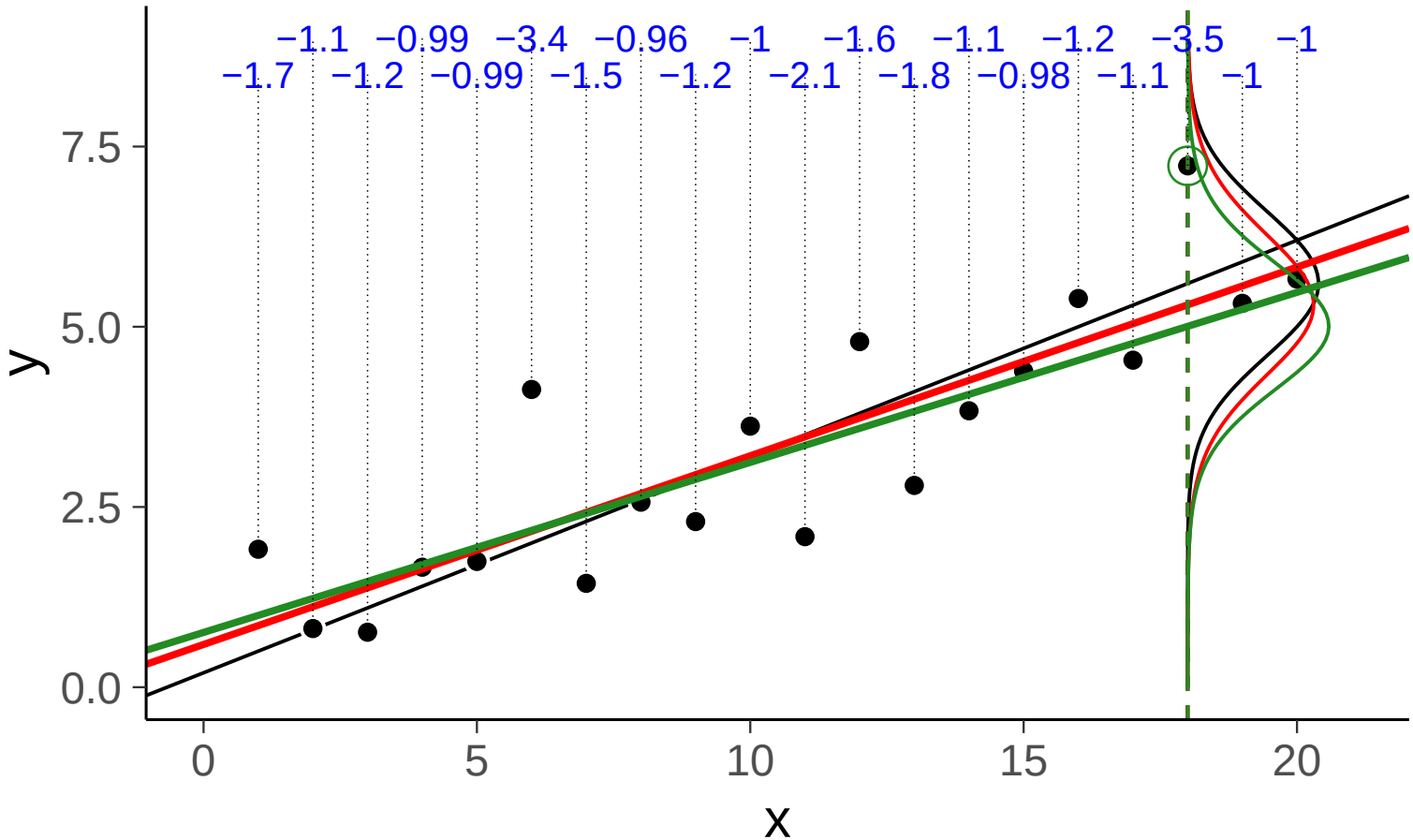
Leave−one−out log predictive densities



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

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

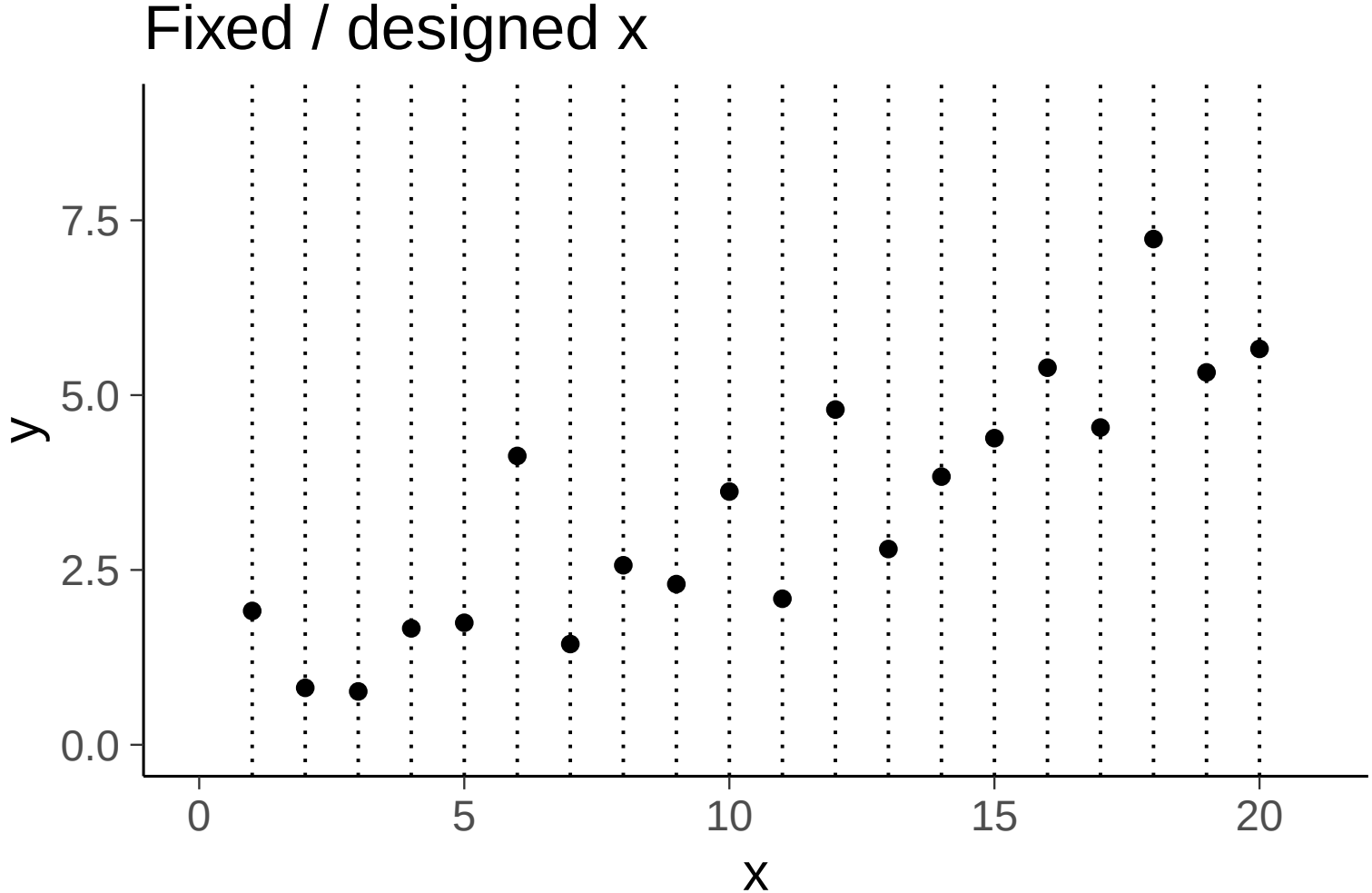
Leave−one−out log predictive densities



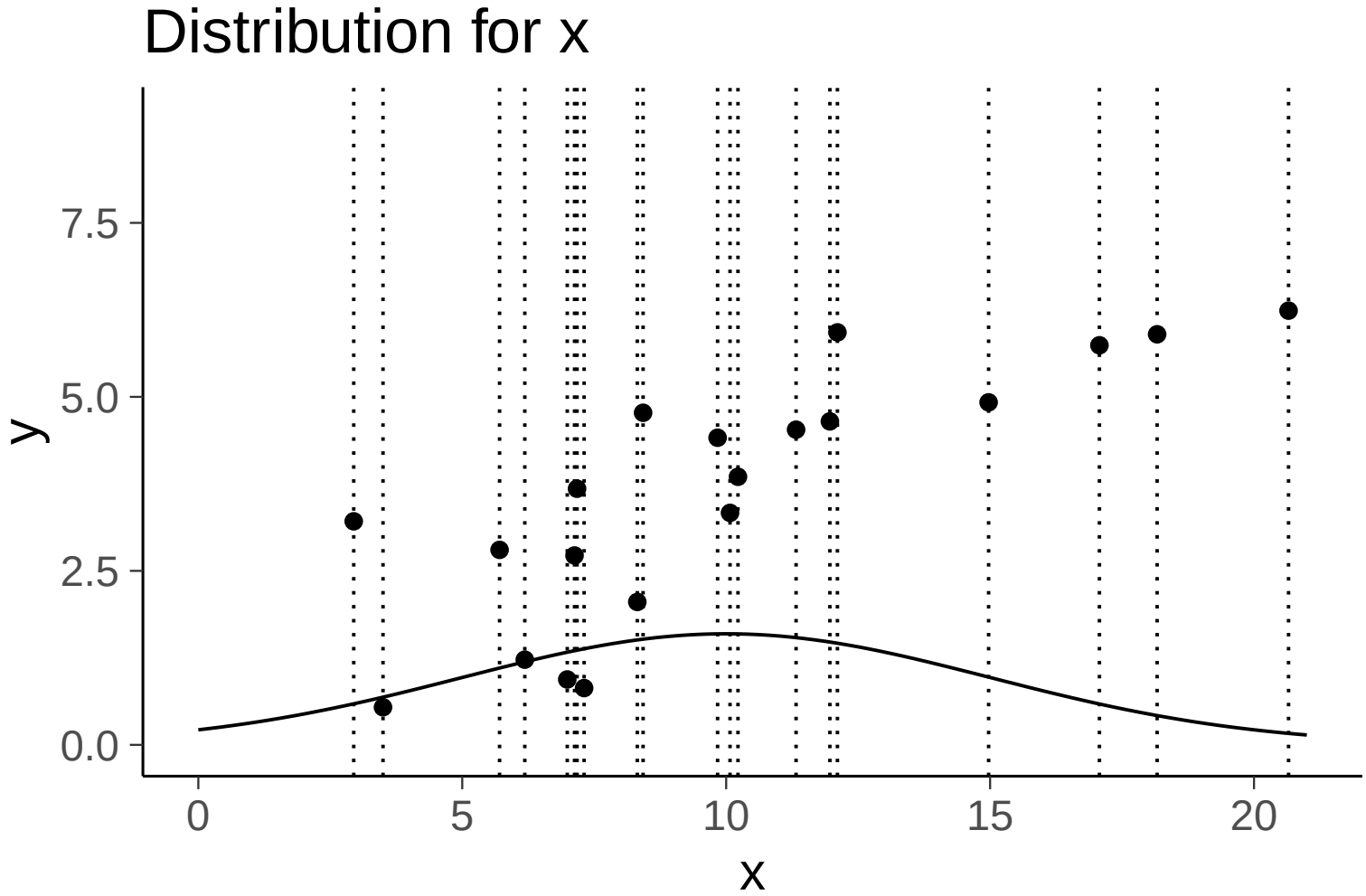
$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$$

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

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)



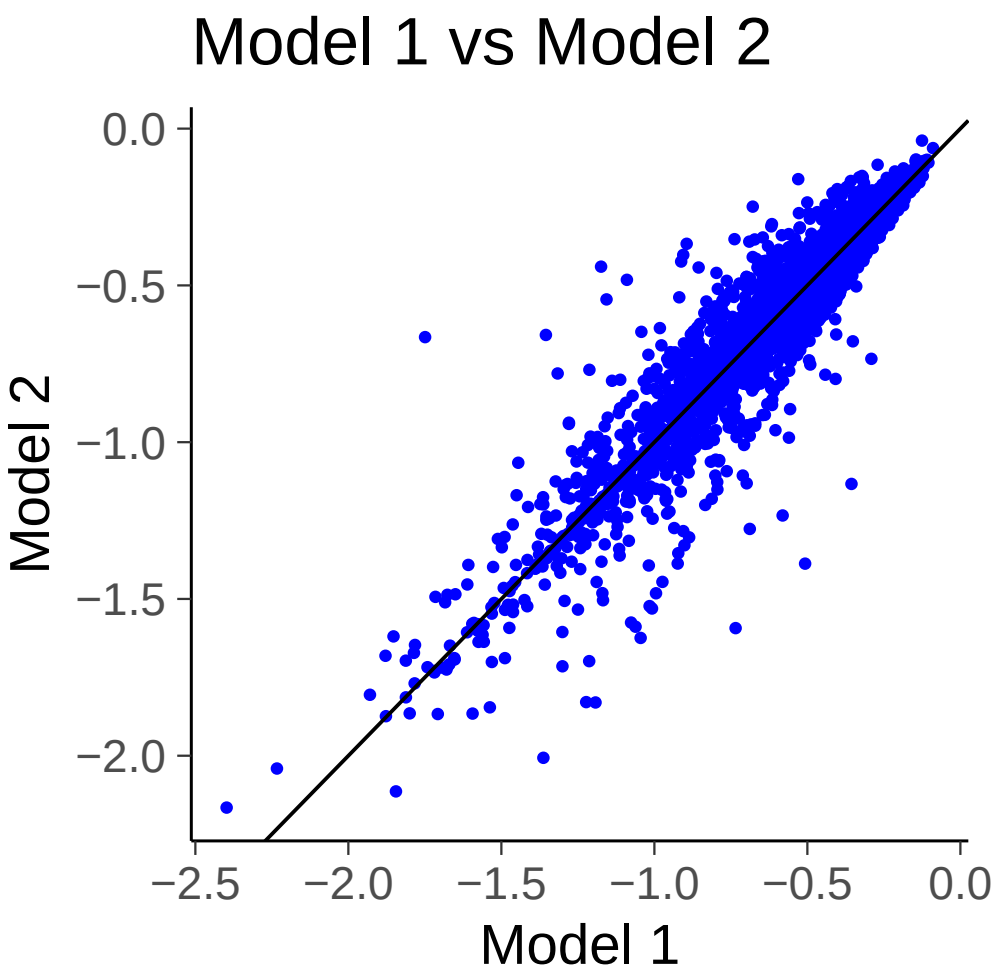
LOO is ok for random x. SE is uncertainty about $y|x$ and x.
see Vehtari & Ojanen (2012)

Arsenic well example – Model comparison

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

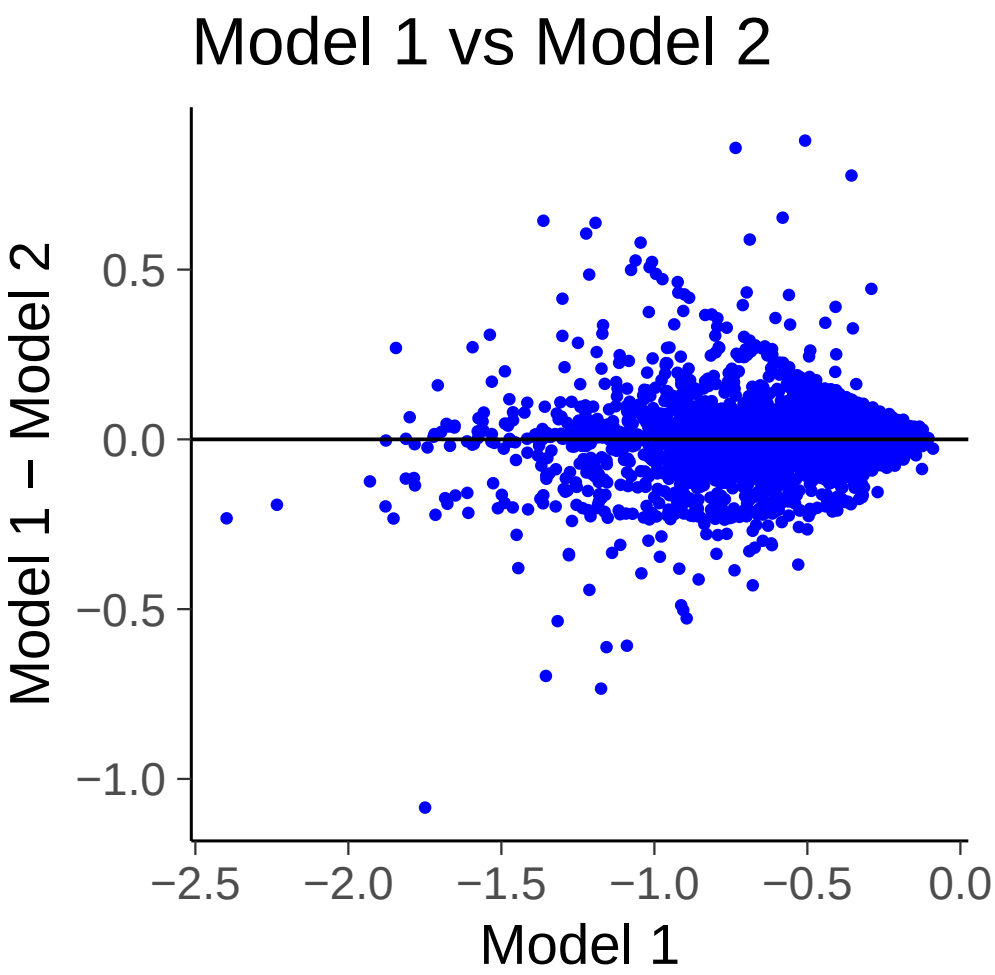
Gelman, Hill & Vehtari (2020): Regression and Other Stories, Chapter 13.

Arsenic well example – Model comparison



Model 1 elpd_loo $\approx$ -1952, SE=16
Model 2 elpd_loo $\approx$ -1938, SE=17

Arsenic well example – Model comparison




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

see Vehtari, Gelman & Gabry (2017a)

Arsenic well example – Model comparison

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

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

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)

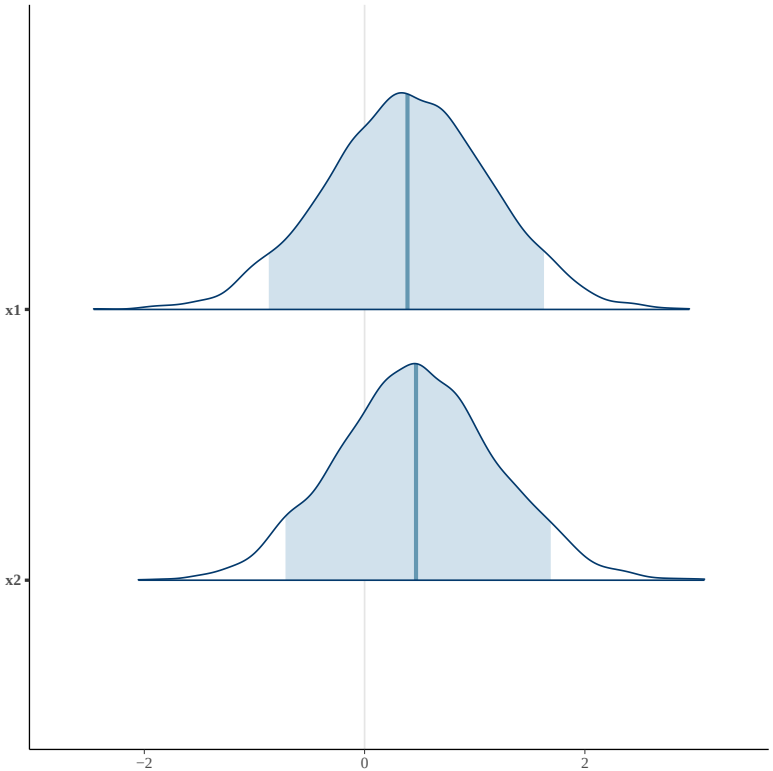
Sometimes cross-validation is not needed

- 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

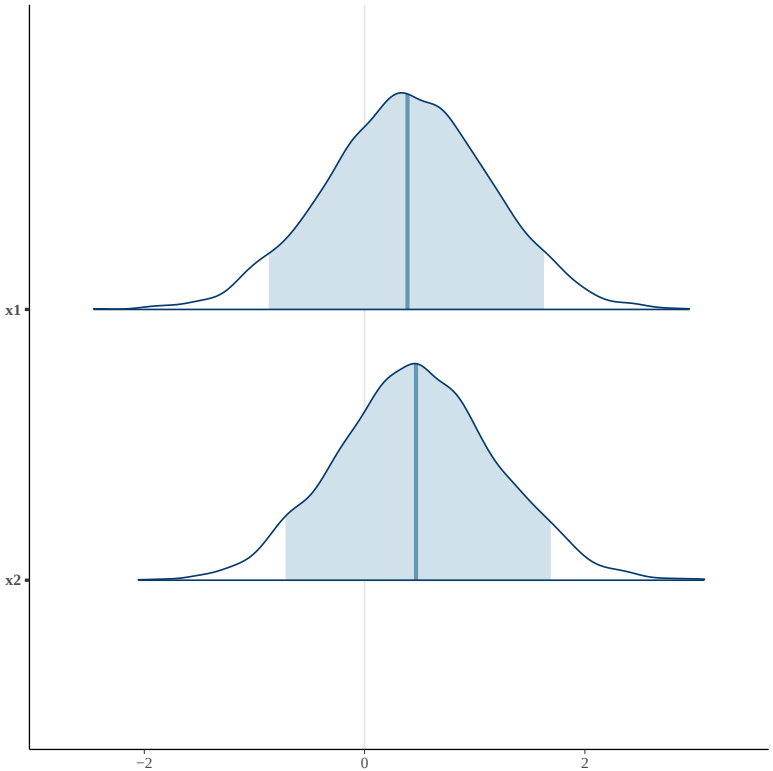
- For some very simple cases you may assume that true model is included in the list of models considered (*M*-closed)
  - see predictive model selection in *M*-closed case by San Martini and Spezzaferri (1984)
  - but you should not force your design of experiment or analysis to stay in the simplified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

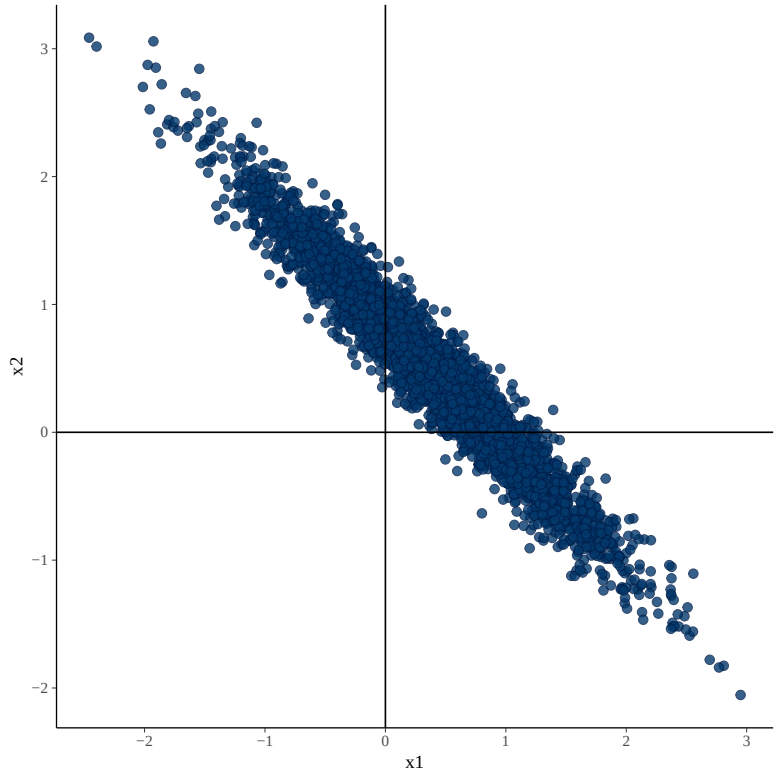


Marginal posterior intervals

Sometimes predictive model comparison can be useful



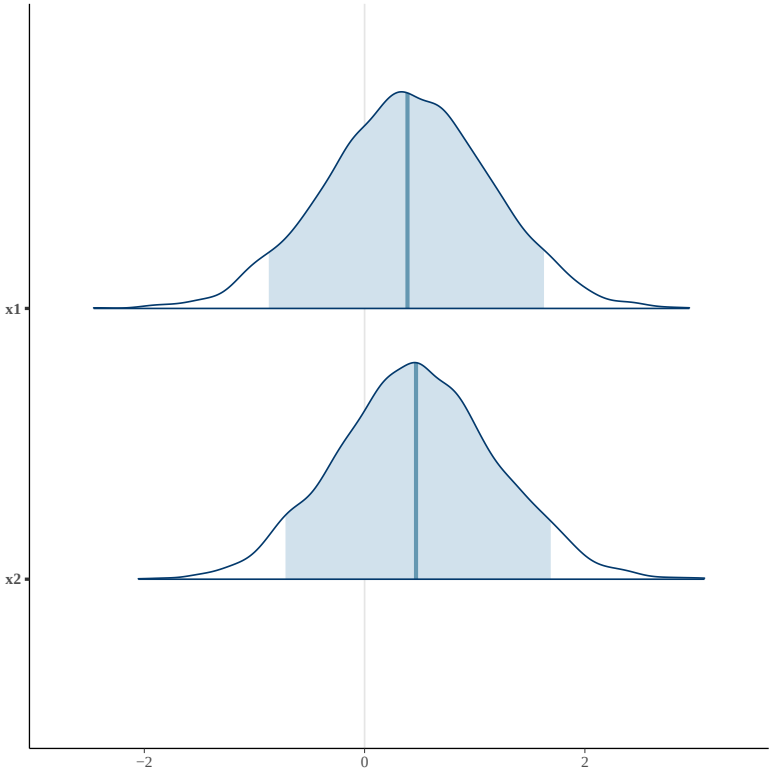
Marginal posterior intervals



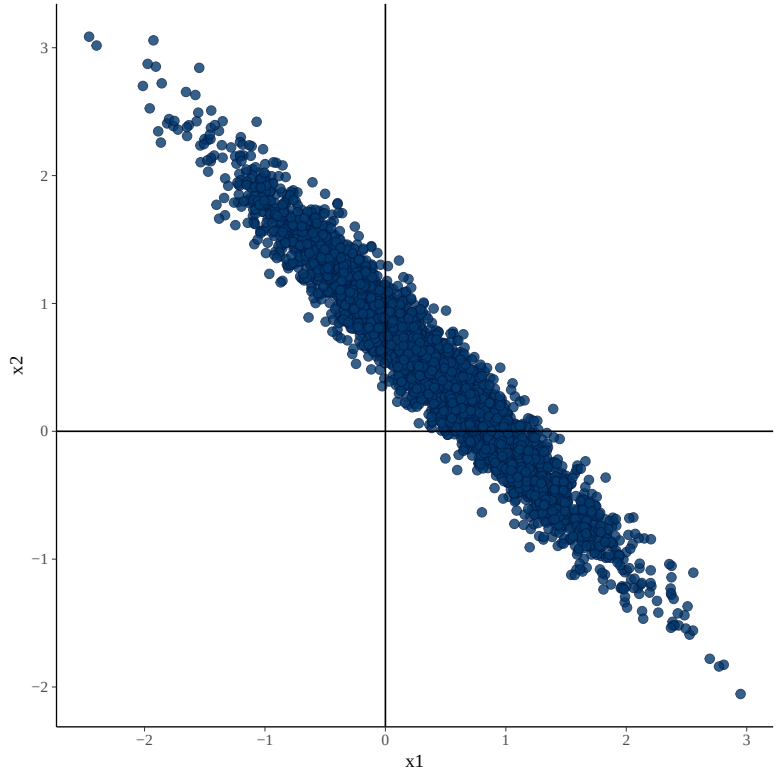
Joint posterior density

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
  - recognized already, e.g., by Stone (1974)

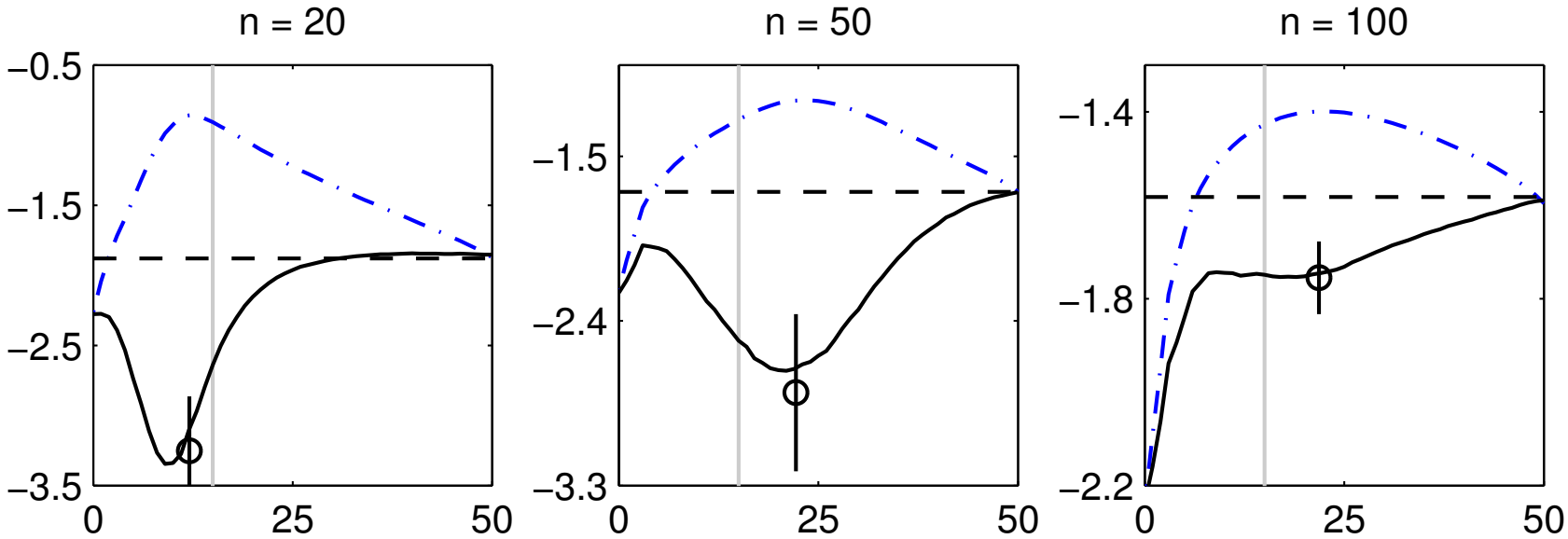
Selection induced bias and overfitting

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

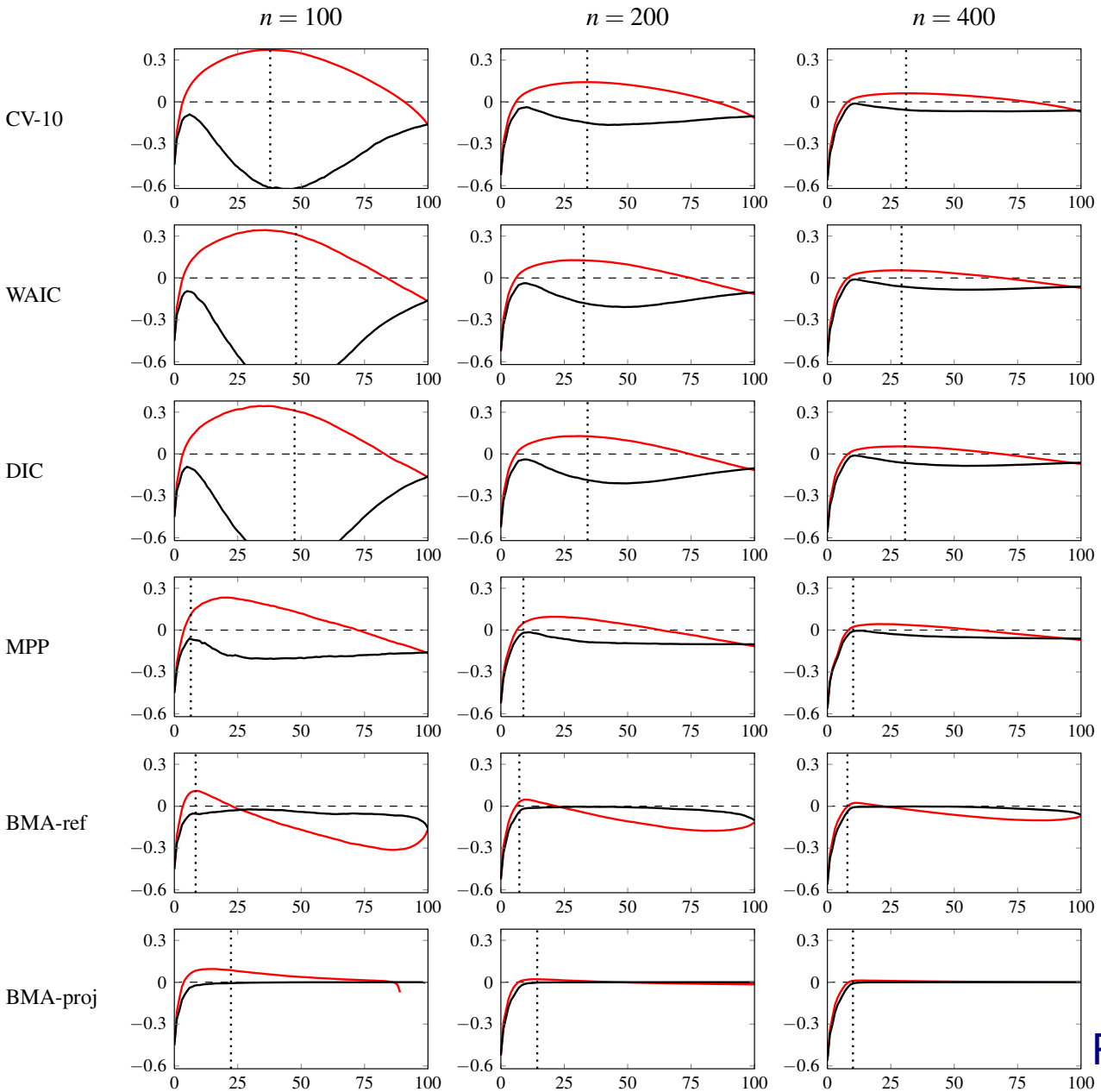
Selection induced bias and overfitting

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

Selection induced bias in variable selection



Selection induced bias in variable selection



Piironen & 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