Announcement

- HW 3 will be posted by 11/28(Thur). You will have 10 days to work on it
- Project Instruction will be posted by 11/1 (Mon)
- Midterm: 11/9
- Review: Class 11/2 (Some sample questions, homework feedback)
- Scope: Chap 1-7, Chap 10, 11 (Slides + Notes will be enough)

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

- 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
- If interested in the overall goodness of the predictive distribution, then a general choice is log-score

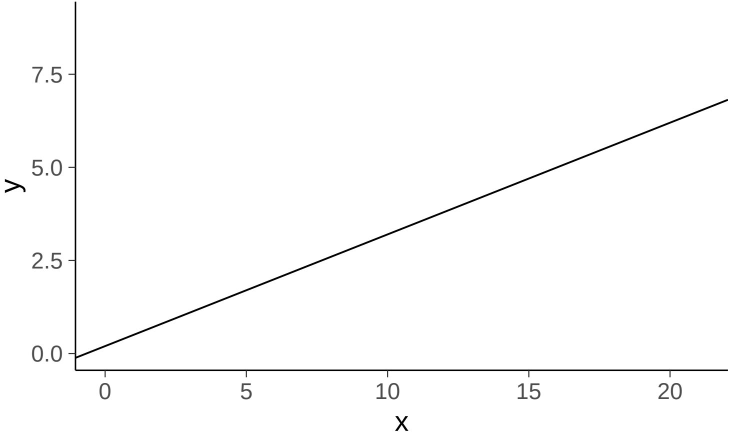
$$\log p(y^{\mathrm{rep}}|y, M),$$

- For specific applications, we can choose utility/cost function

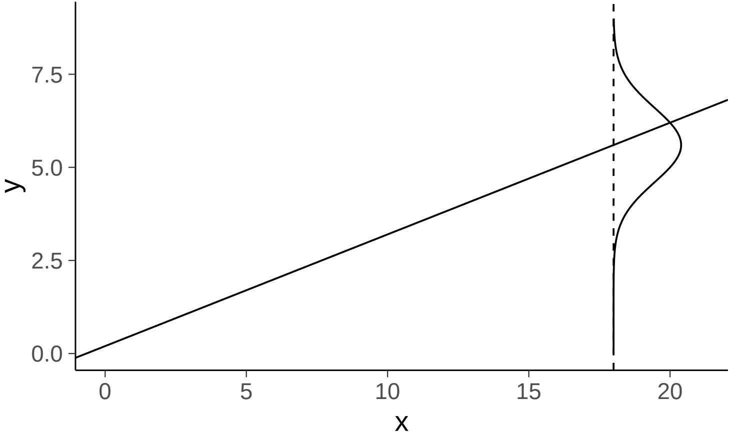
Outline

- 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
  - K-fold cross-validation
- Related methods (WAIC, *IC, BF)
- Model comparison and selection (elpd_diff, se)
- Model averaging with Bayesian stacking

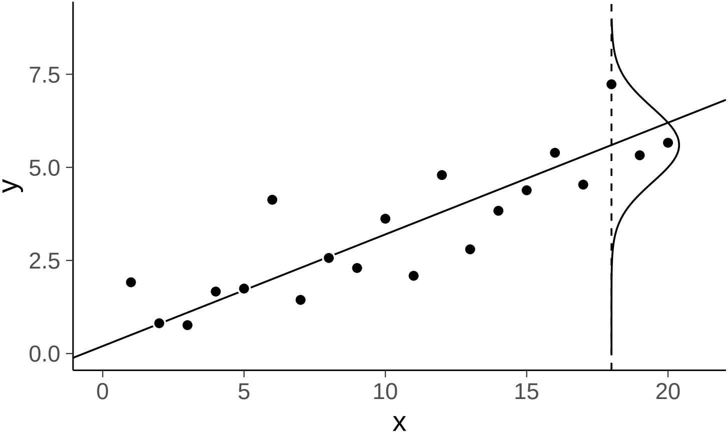
True mean y = a + bx



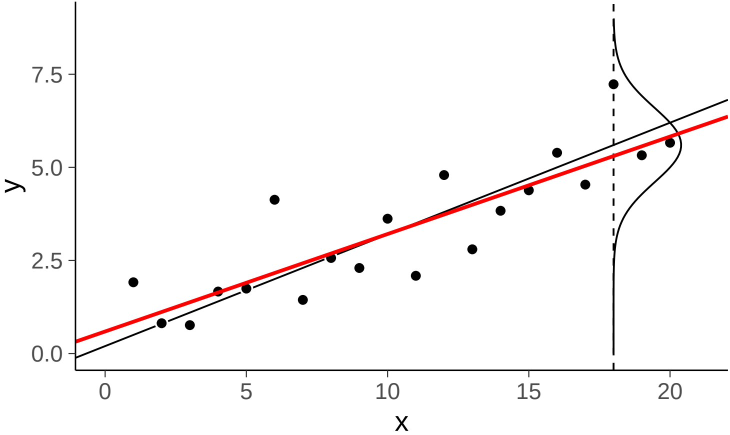
True mean and sigma



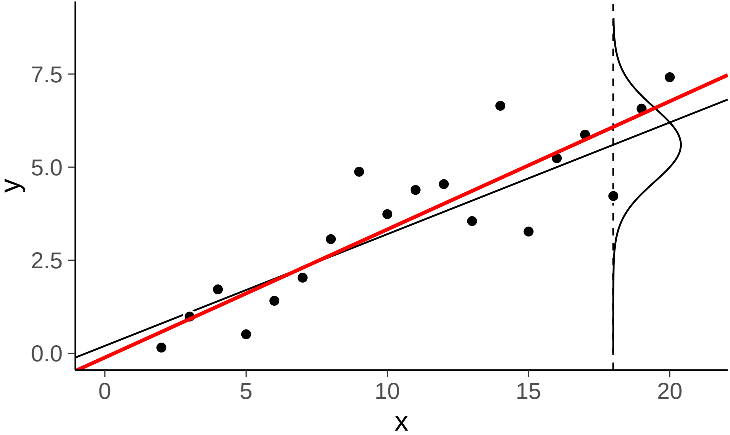
Data



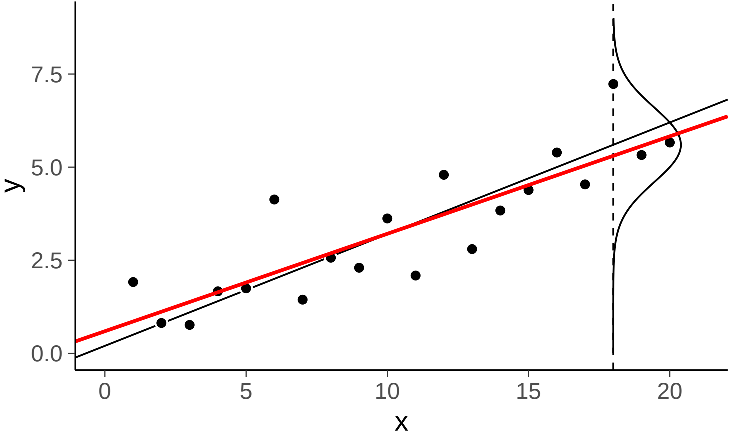
Posterior mean



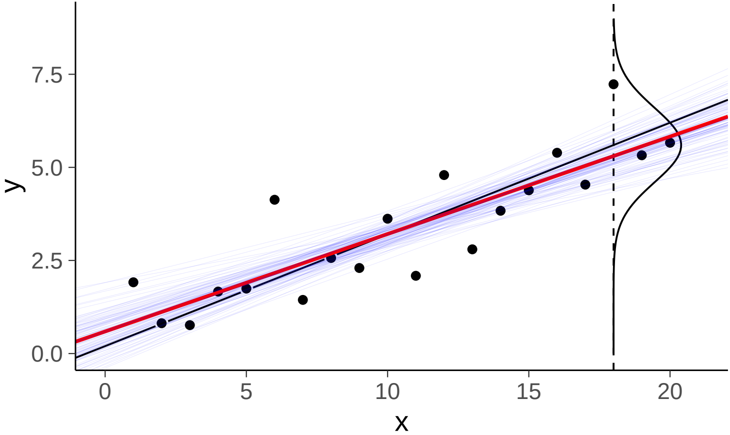
Posterior mean, alternative data realisation



Posterior mean

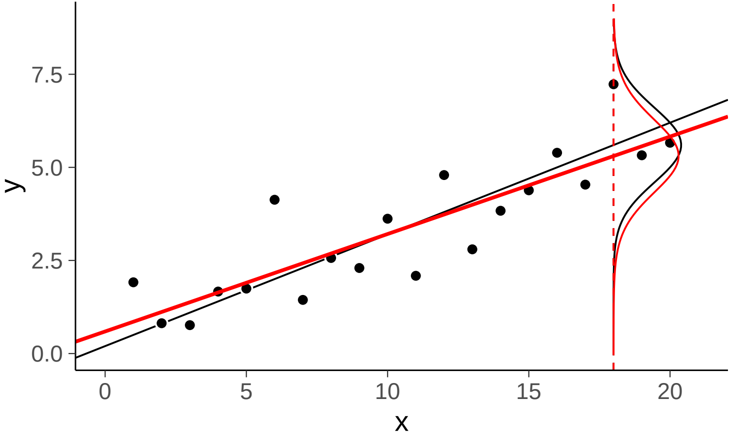


Posterior draws

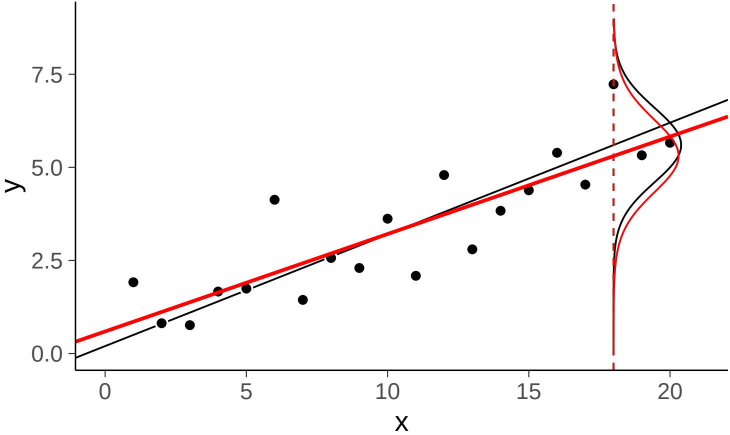


1) Draw σ^2 from $\sigma^2|y \sim Inv-\chi^2(n-2, s^2)$
2) Draw β from $\beta|y, \sigma^2 \sim N(\hat{\beta}, V_\beta \sigma^2)$
3) Repeat step 1) and step 2)

Posterior predictive distribution

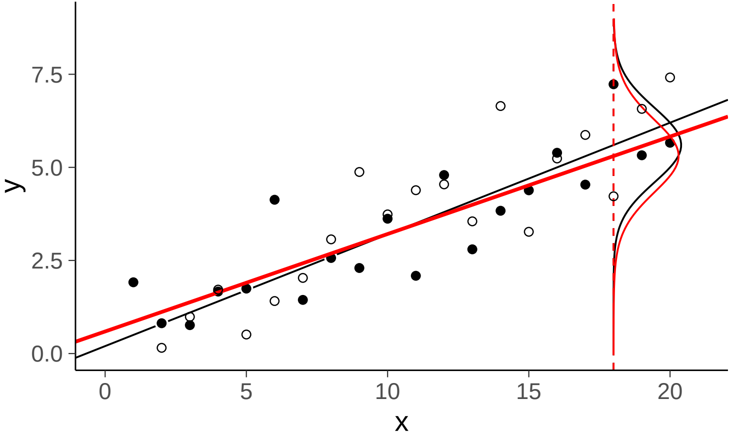


Posterior predictive distribution

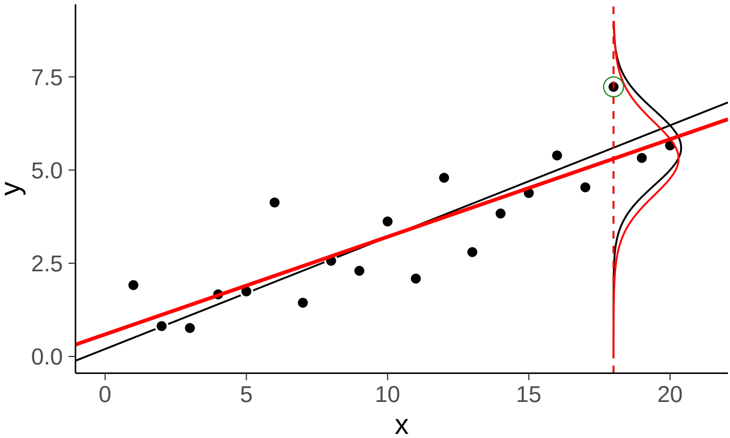


$$p(\tilde{y}|\tilde{x} = 18, x, y) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x, y)d\theta$$

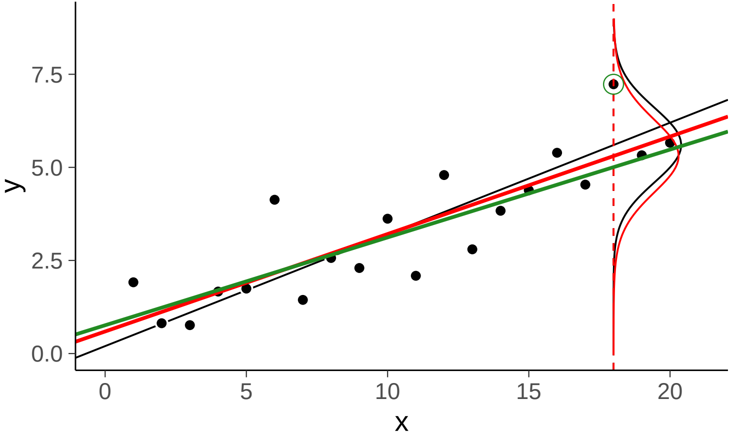
New data



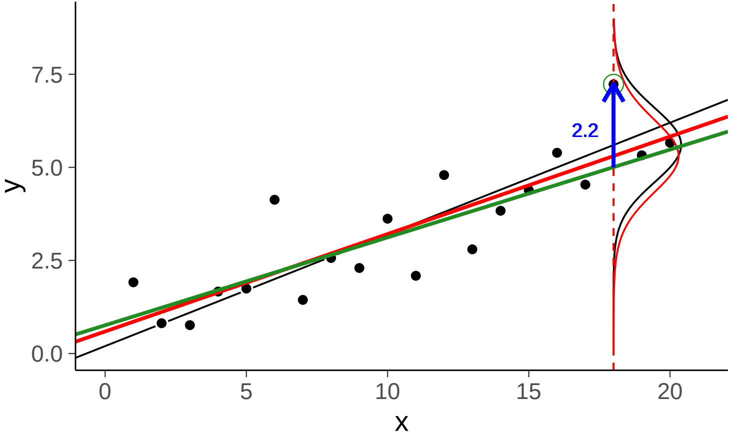
Posterior predictive distribution

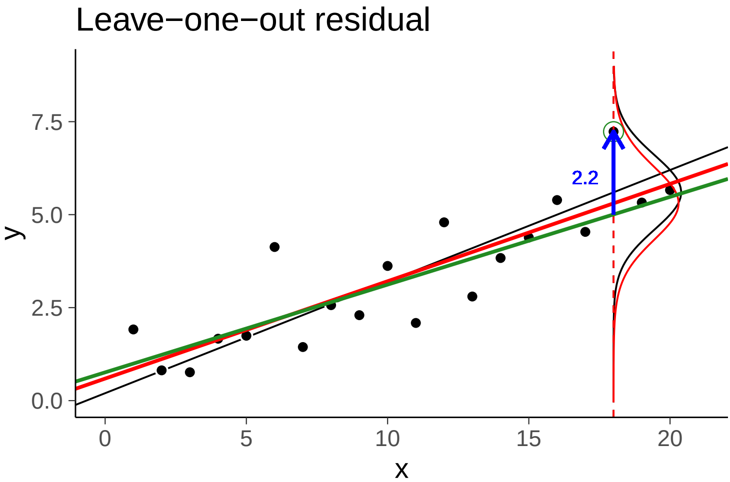


Leave−one−out mean

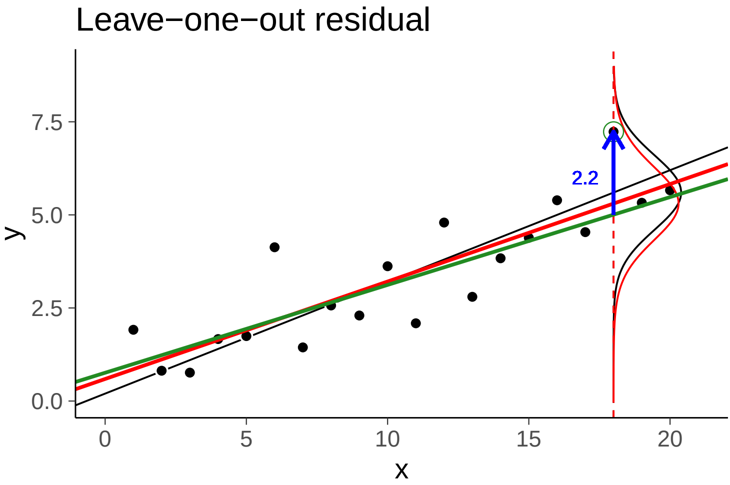


Leave−one−out residual





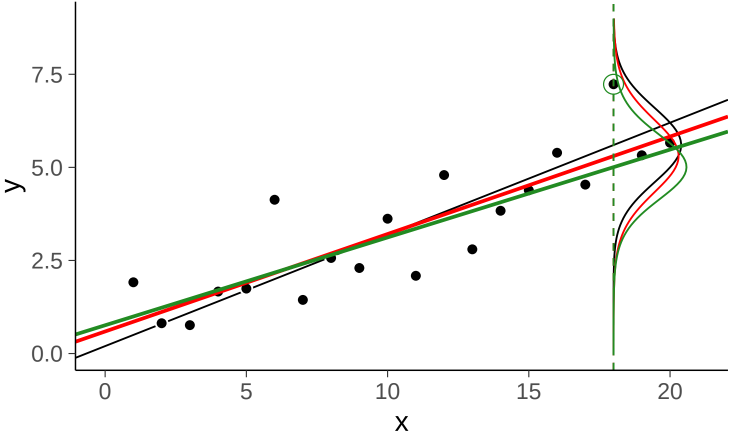
$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$



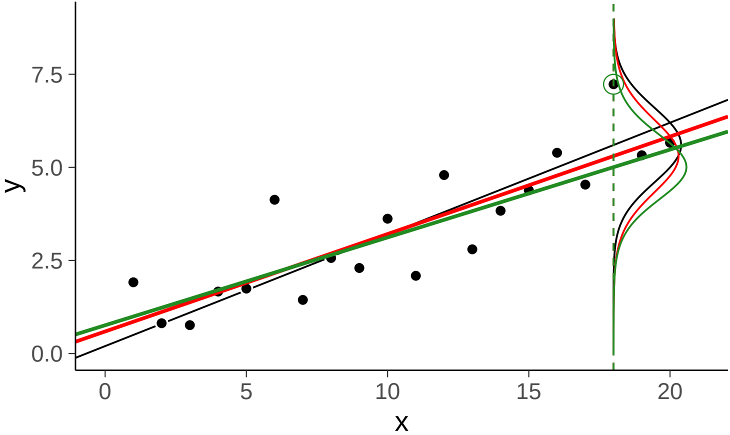
$y_{18} - E[p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18})]$

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

Leave−one−out predictive distribution

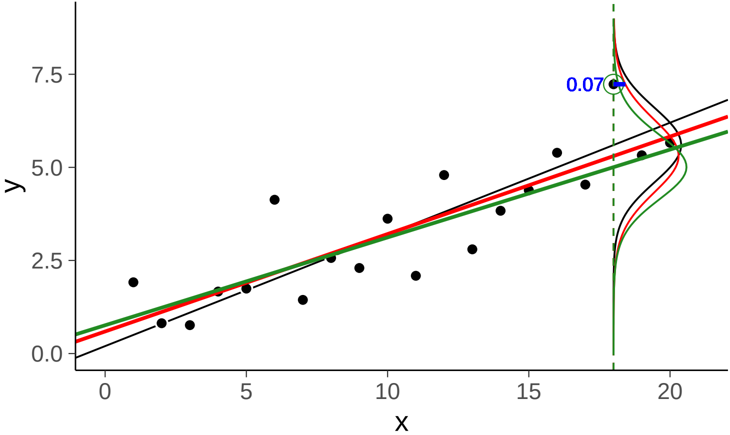


Leave−one−out predictive distribution

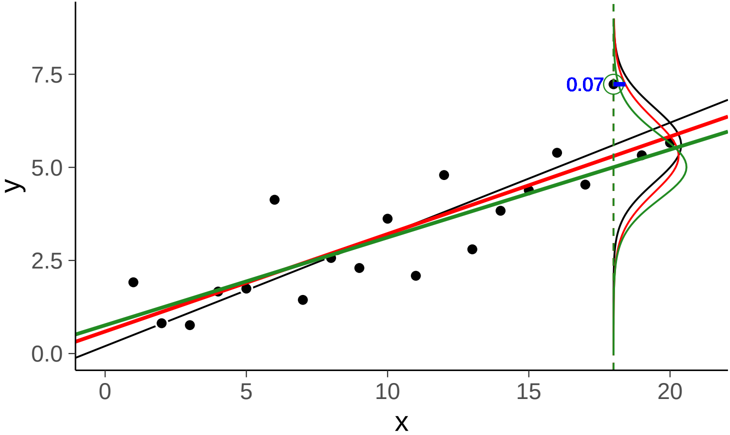


$$p(\tilde{y}|\tilde{x} = 18, x_{-18}, y_{-18}) = \int p(\tilde{y}|\tilde{x} = 18, \theta)p(\theta|x_{-18}, y_{-18})d\theta$$

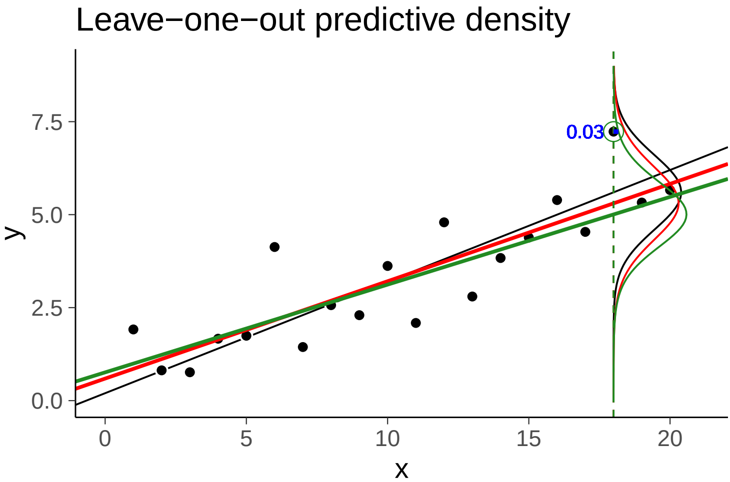
Posterior predictive density



Posterior predictive density

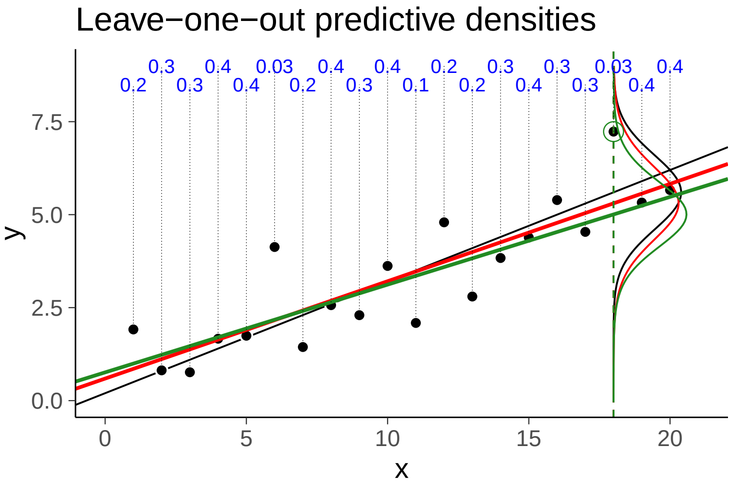


$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$

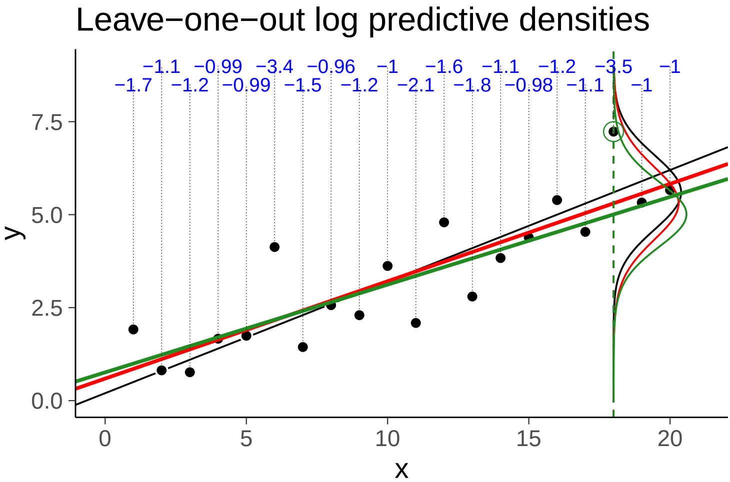


$p(\tilde{y} = y_{18} | \tilde{x} = 18, x, y) \approx 0.07$

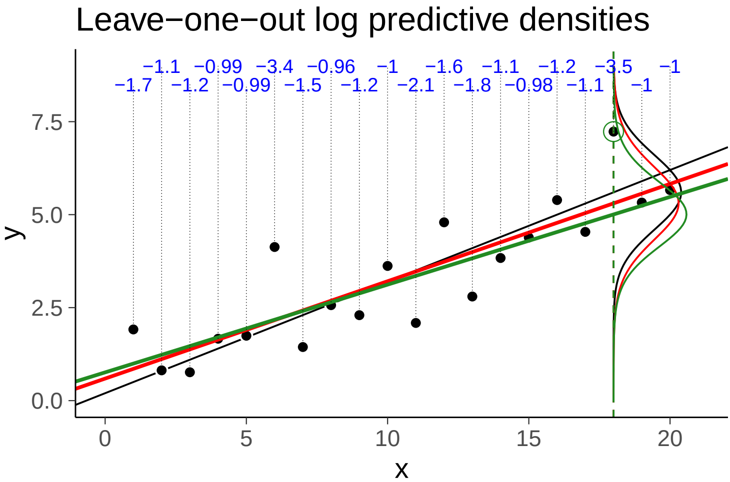
$p(\tilde{y} = y_{18} | \tilde{x} = 18, x_{-18}, y_{-18}) \approx 0.03$



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

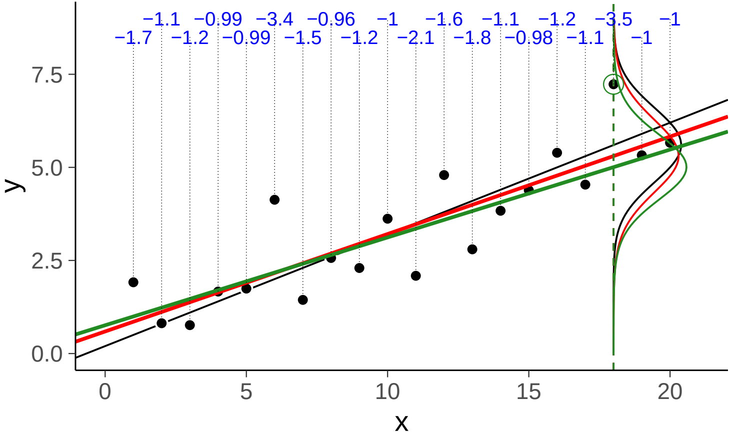


$\log p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$



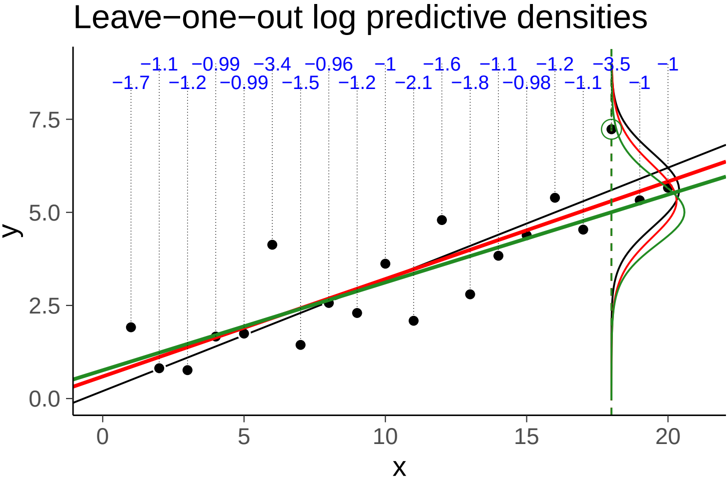
$\sum_{i=1}^{20} \log p(y_i | x_i, x_{-i}, y_{-i}) \approx -29.5$

Leave−one−out log predictive densities



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

Expected log posterior predictive density

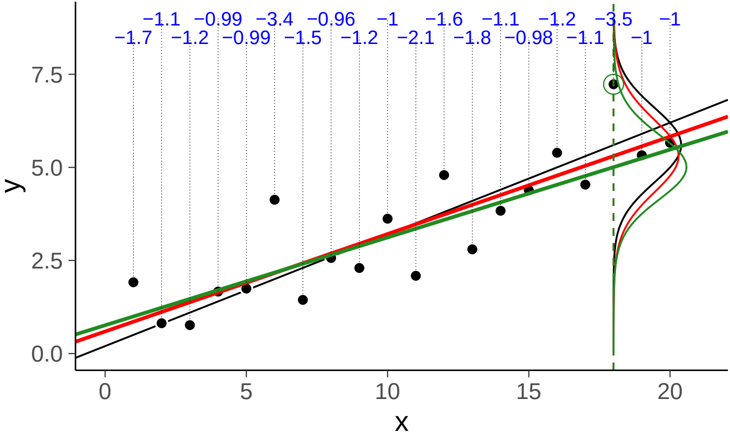


$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

Expected log posterior predictive density

Estimate of log posterior predictive density for new data

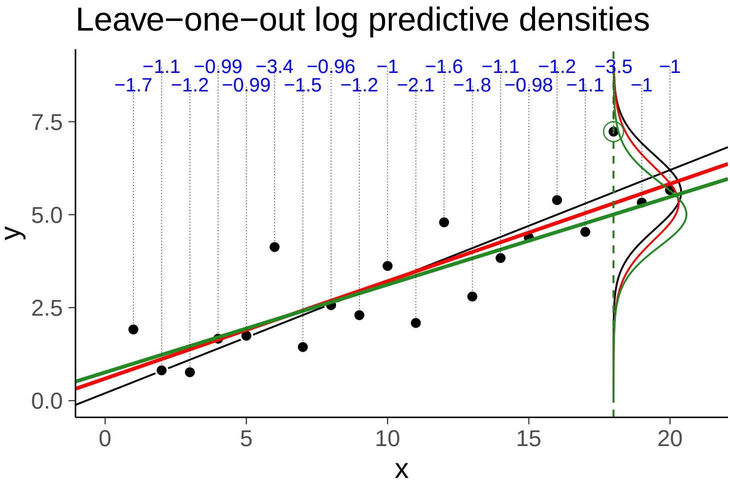
Leave−one−out log predictive densities



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

Expected log posterior predictive density

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

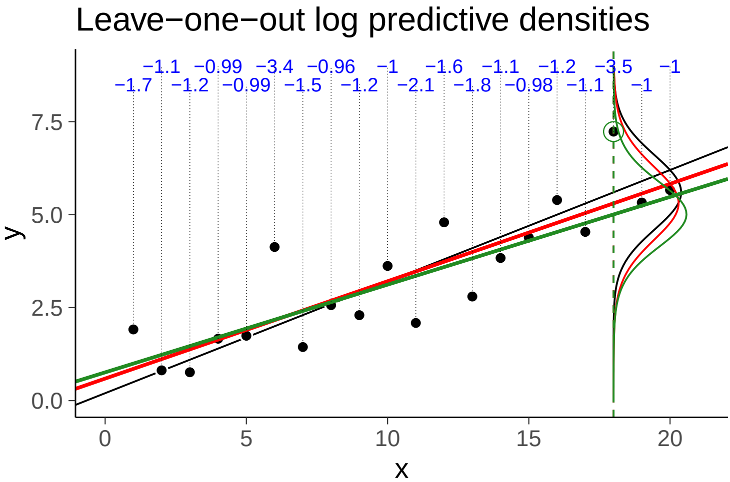


$$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$$

Expected log posterior predictive density

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

$\text{p\_loo} = \text{lpd} - \text{elpd\_loo} \approx 2.7$ Effective number of parameters

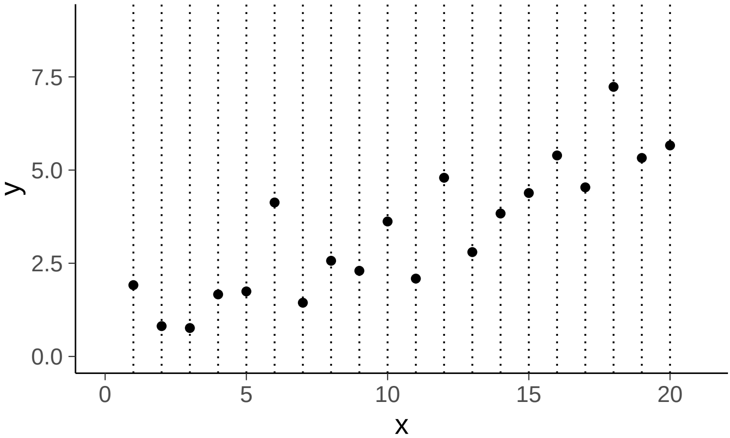


$\text{elpd\_loo} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

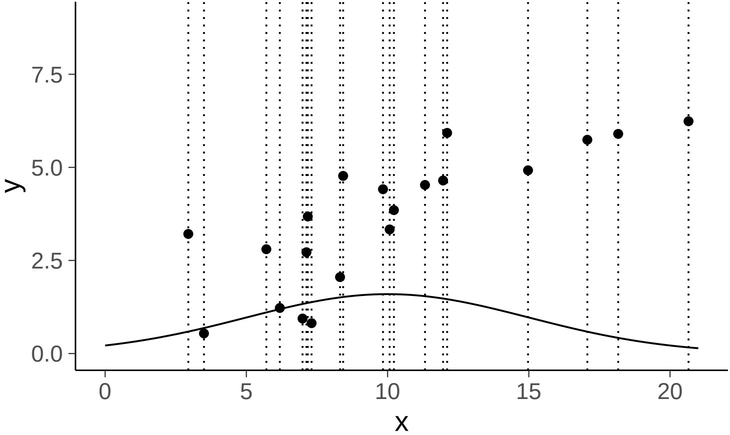
Expected log posterior predictive density

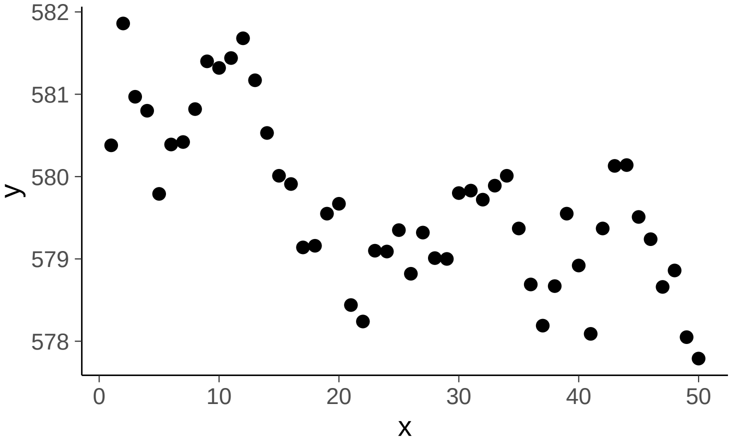
$\text{SE} = \text{sd}(\log p(y_i|x_i, x_{-i}, y_{-i})) \cdot \sqrt{20} \approx 3.3$

Fixed / designed x

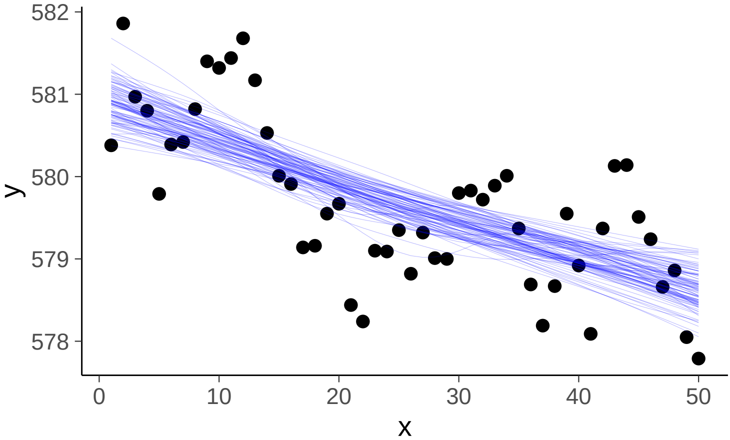


Distribution for x

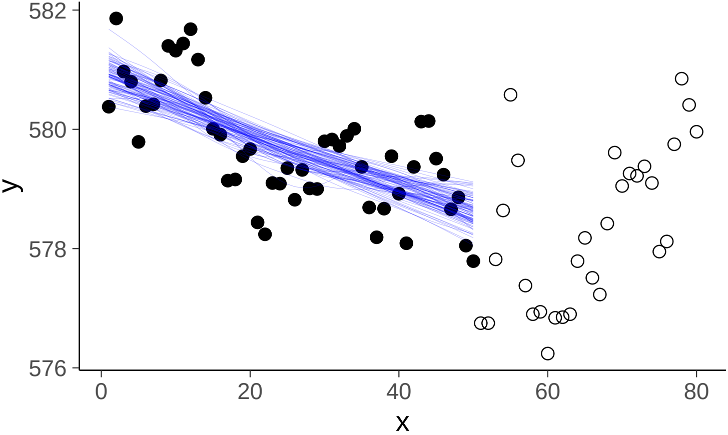


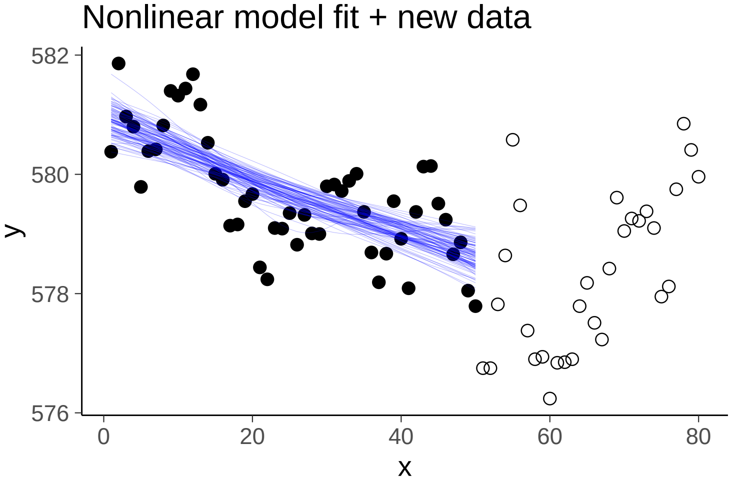

582
581
580
579
578
y
0
10
20
30
40
50
x

Nonlinear model fit

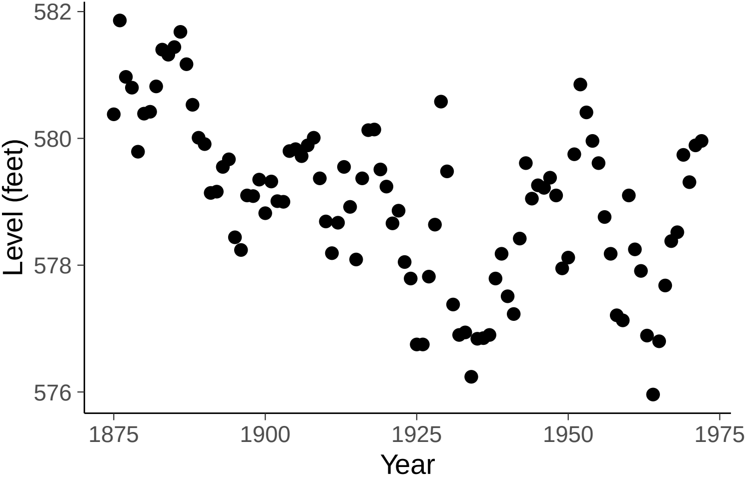


Nonlinear model fit + new data

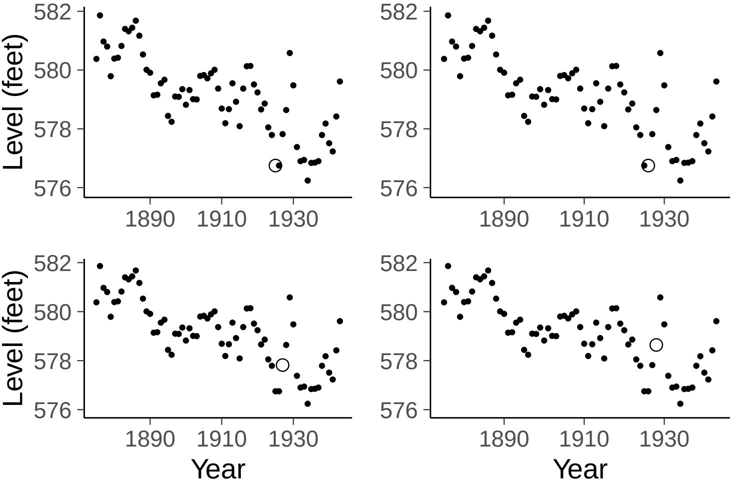




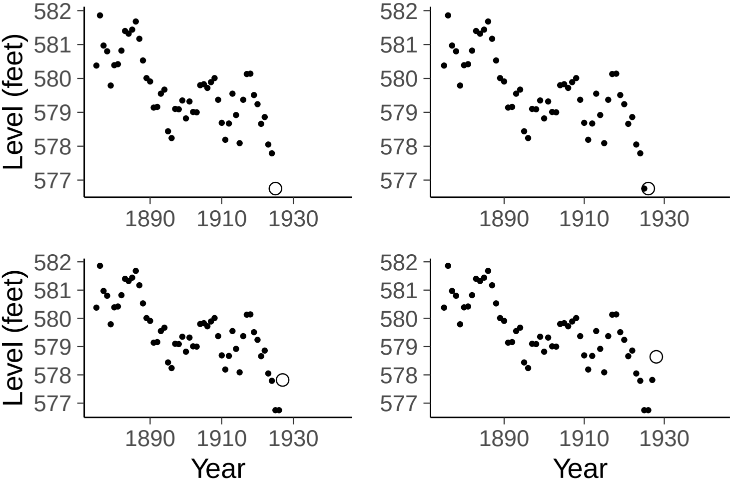
Extrapolation is more difficult



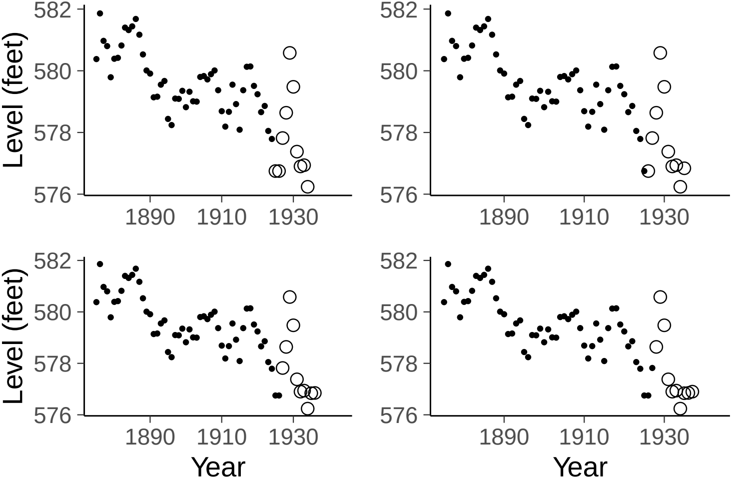
Can LOO or other cross-validation be used with time series?



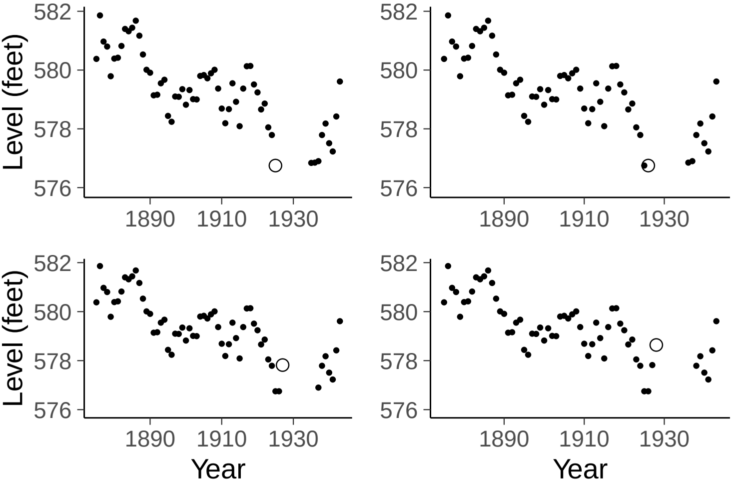
Leave-one-out cross-validation is ok for assessing conditional model



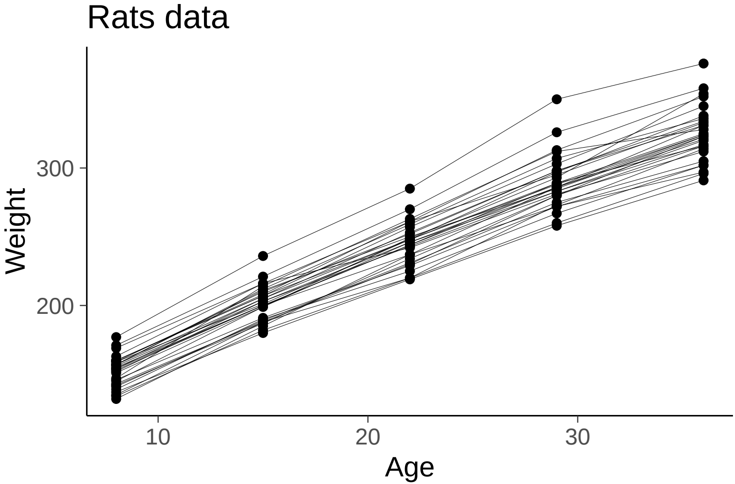
Leave-future-out cross-validation is better for predicting future



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

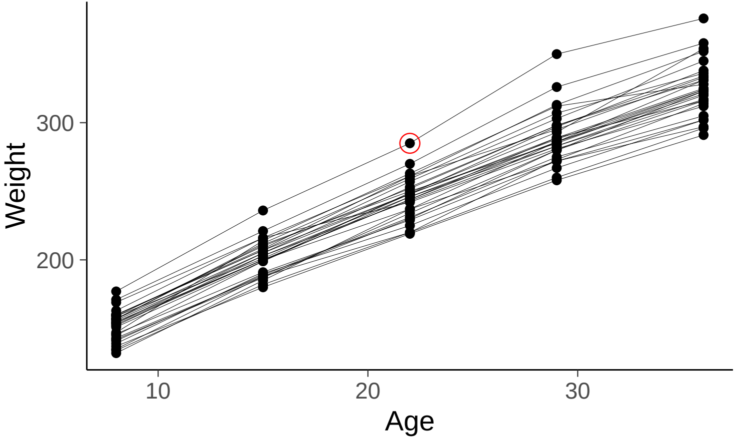


m-step-ahead leave-a-block-out cross-validation



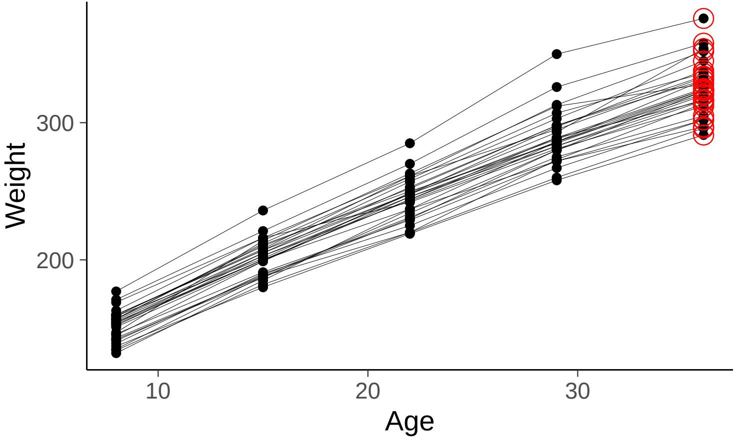
Can LOO or other cross-validation be used with hierarchical data?

Leave−one−out?



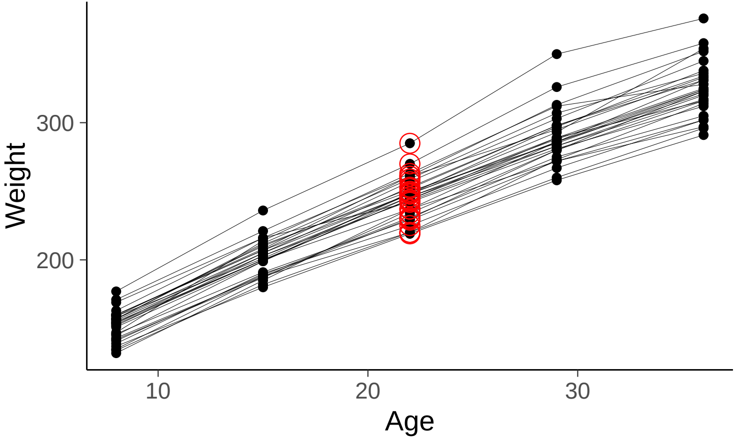
Yes!

1−step−ahead?

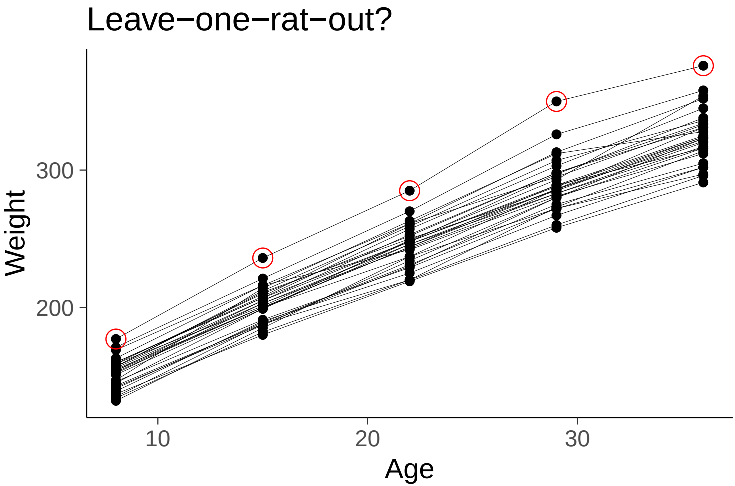


Yes!

Leave−one−time−point−out?

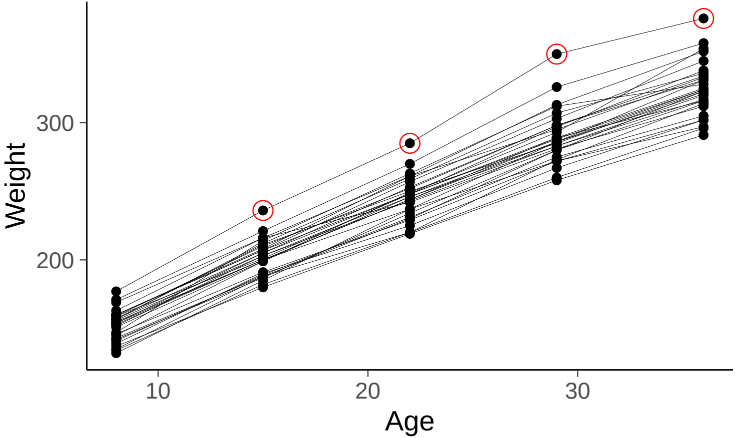


Yes!



Yes!

Predict given initial weight?



Yes!

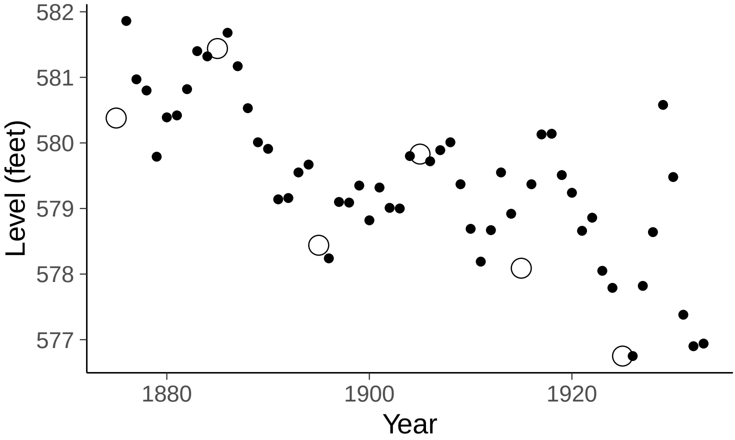
Fast cross-validation

- K-fold cross-validation

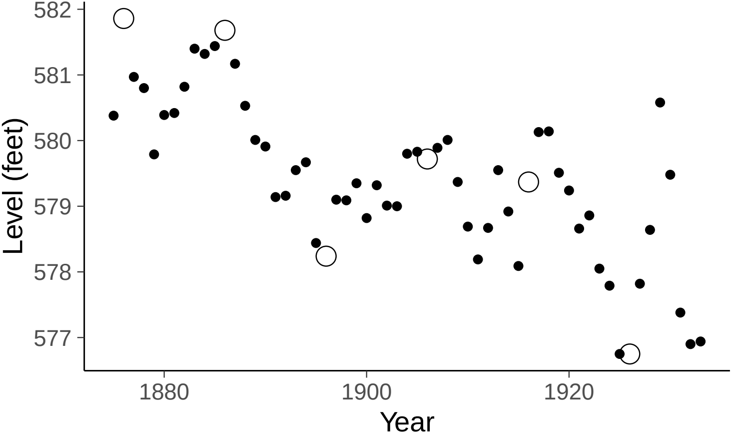
K-fold cross-validation

- 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

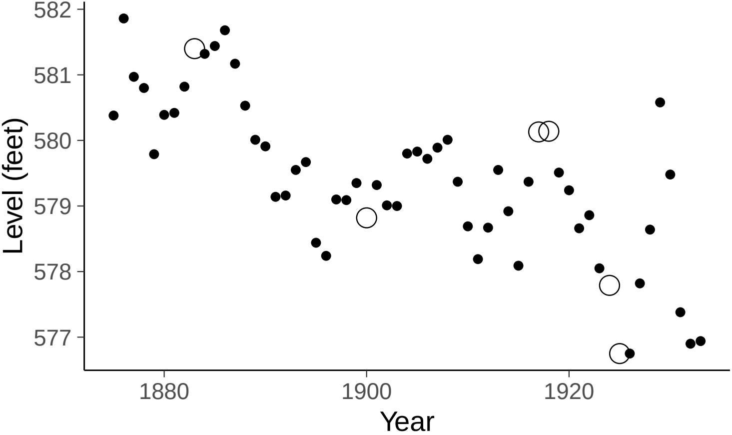
Balance k−fold approximation of LOO



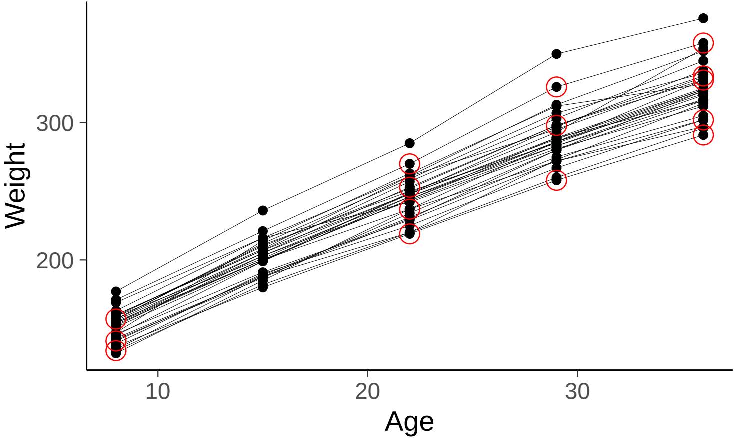
Balance k−fold approximation of LOO



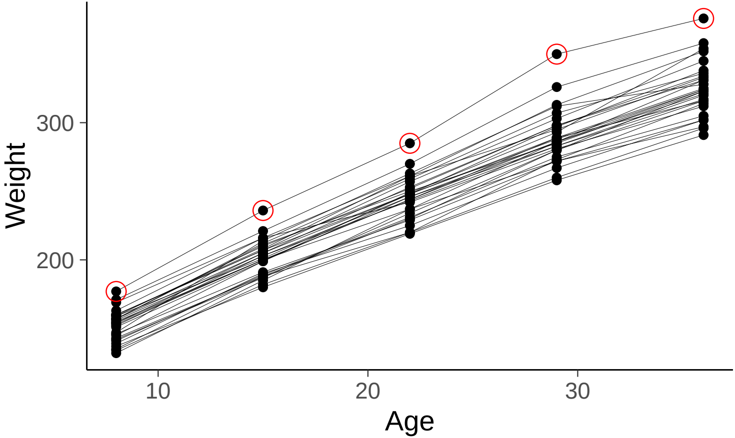
Random k−fold approximation of LOO



Random kfold approximation of LOO



Leave−one−rat−out



Information Criteria

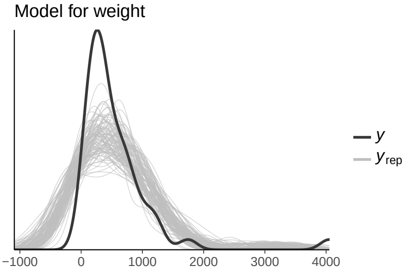
Measures of predictive accuracy are referred to as information criteria

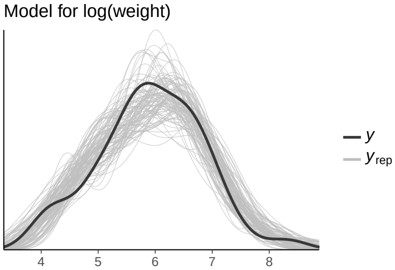
- AIC uses maximum likelihood estimate for prediction
- DIC uses posterior mean for prediction
- BIC is an approximation for marginal likelihood
- TIC, NIC, RIC, PIC, BPIC, QIC, AICc, ...

Note: ICs may be improper when you have a very complex model (e.g. NNs with thousands of parameters)

How to compare different models?

- Posterior predictive checking is often sufficient





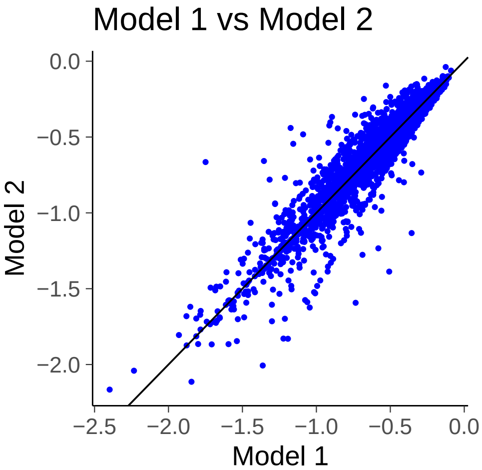
E.g. Predicting the yields of mesquite bushes.

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 et al., (2020): Regression and Other Stories, Chapter 13.

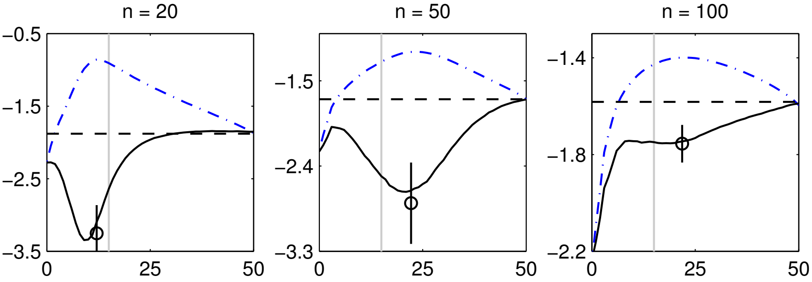
Arsenic well example – Model comparison



Model 1 elpd_loo $\approx$ -1952
Model 2 elpd_loo $\approx$ -1938

Model 2 has a bigger *elpd_loo*, thus it is better than Model 1

Selection induced bias in variable selection



Take-home messages

- It's good to think about predictions of current data, especially those are the ones we can observe
- Cross-validation can simulate predicting and observing new data
- 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