Outline

Gaussian processes – integration and model selection

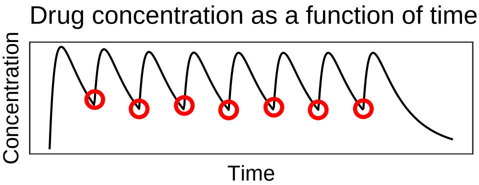
- Background
- Rasmussen & Williams Chapter 5
- Point estimate vs. integration
  - motorcycle crash g-forces
- Using GPs as components
  - motorcycle crash g-forces
  - birthdays
- Model selection

How I started working on GPs

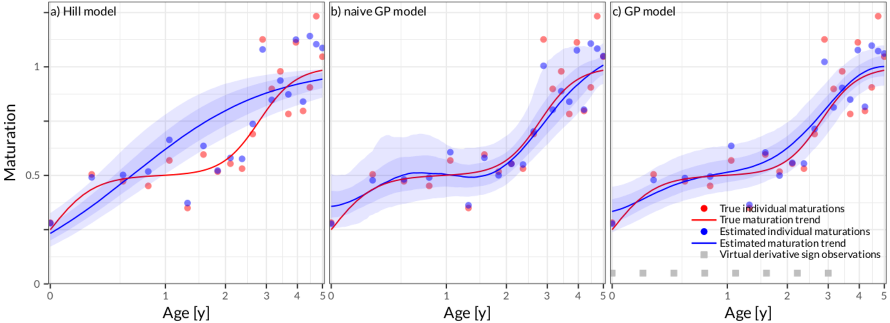
→

GPs as priors for model components

GPs as priors for model components

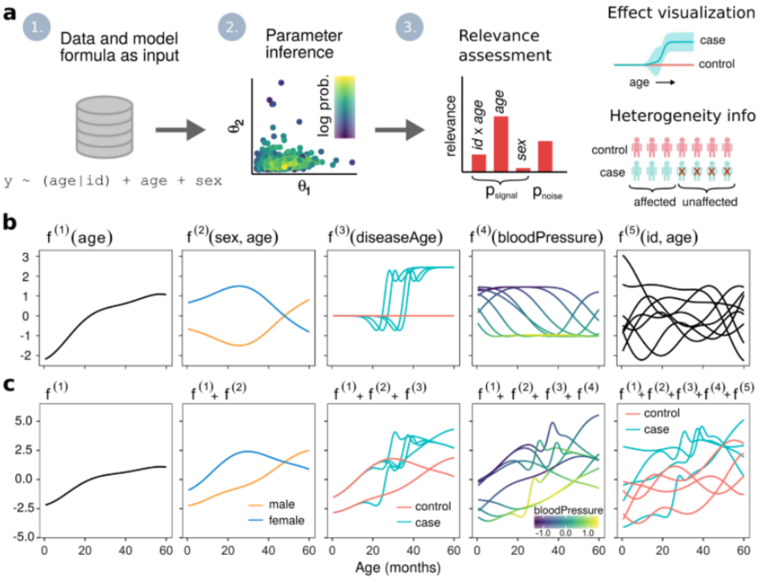


Monotonic maturation effect



lgpr – longitudinal Gaussian process regression

R package for **L**ongitudinal **G**aussian **P**rocess **R**egression.



“Model selection”

- Lecture 3
- Rasmussen & Williams Chapter 5

Hyperparameters & model selection (I)

- Almost all covariance functions have hyperparameters
- How do we choose values for them?
- Ideally, we would like to put prior distributions on the hyperparameters and compute the posterior
- Let $\boldsymbol{\theta}$ be the hyperparameters of interest, then

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{38}$$

but in this case the marginal likelihood is almost always intractable

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{39}$$

Hyperparameters & model selection (II)

- Approximation: We will use the MAP (Maximum a posterior estimate)
- $p(\boldsymbol{y})$ is constant wrt. $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{40}$$

- The MAP estimate is defined as

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \tag{41}$$

- If the prior $p(\boldsymbol{\theta}) \propto 1$ is uniform

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}|\boldsymbol{\theta}) + \ln k = \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}|\boldsymbol{\theta}) = \hat{\theta}_{\text{ML}} \tag{42}$$

- This is also sometimes called the maximum likelihood type II estimate

The marginal likelihood computation (I)

- Marginal likelihood for Gaussian likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{f} \tag{43}$$

$$= \int \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{f}, \sigma_{obs}^2 \boldsymbol{I}\right) \mathcal{N}\left(\boldsymbol{f}|0, \boldsymbol{K}\right) \mathrm{d}\boldsymbol{f} \tag{44}$$

$$= \mathcal{N}\left(\boldsymbol{y}|0, \sigma_{obs}^2 \boldsymbol{I} + \boldsymbol{K}\right) \tag{45}$$

- Then

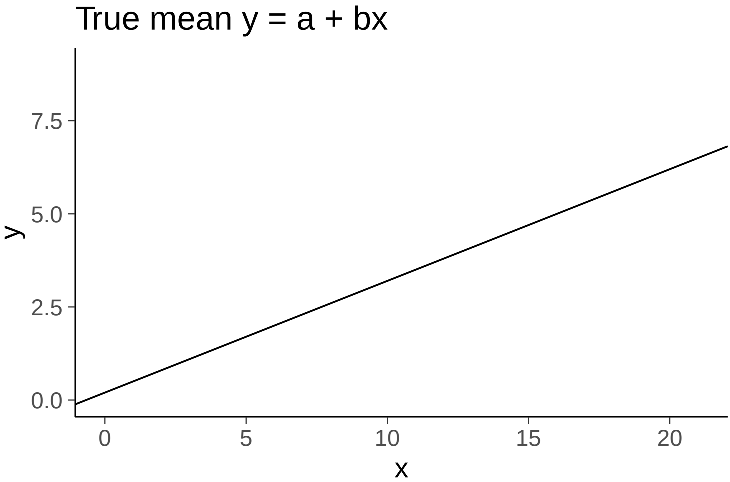
$$\ln p(\boldsymbol{y}|\boldsymbol{\theta}) = \ln \mathcal{N}\left(\boldsymbol{y}|0, \sigma_{obs}^2 \boldsymbol{I} + \boldsymbol{K}\right) \tag{46}$$

$$= \ln \left[(2\pi)^{-\frac{N}{2}} \left|\sigma_{obs}^2 \boldsymbol{I} + \boldsymbol{K}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \left(\sigma_{obs}^2 \boldsymbol{I} + \boldsymbol{K}\right)^{-1} \boldsymbol{y}\right)\right] \tag{47}$$

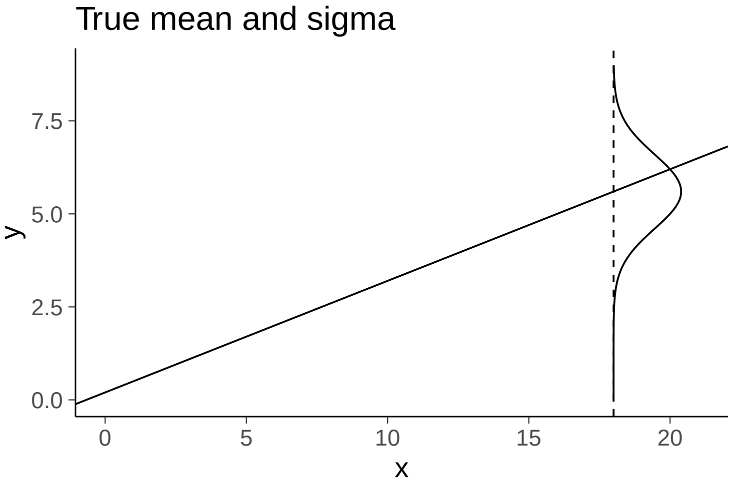
$$= -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln\left|\sigma_{obs}^2 \boldsymbol{I} + \boldsymbol{K}\right| - \frac{1}{2}\boldsymbol{y}^T \left(\sigma_{obs}^2 \boldsymbol{I} + \boldsymbol{K}\right)^{-1} \boldsymbol{y} \tag{48}$$

- Motorcycle crash g-forces
- Birthdays
- Traffic deaths

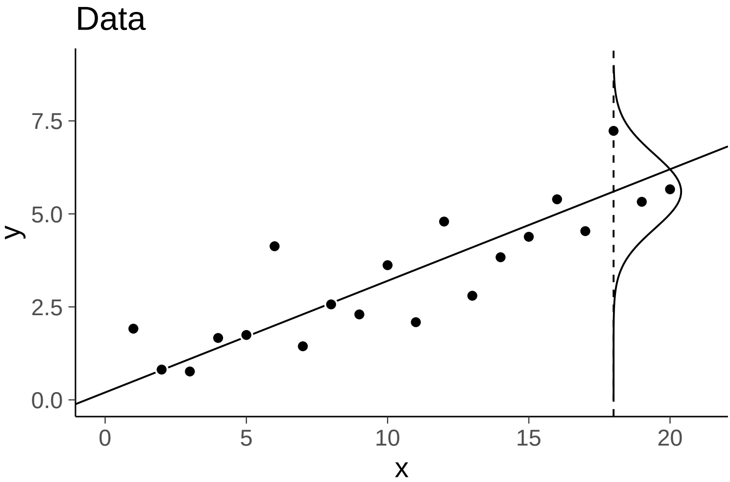
Leave-one-out cross-validation



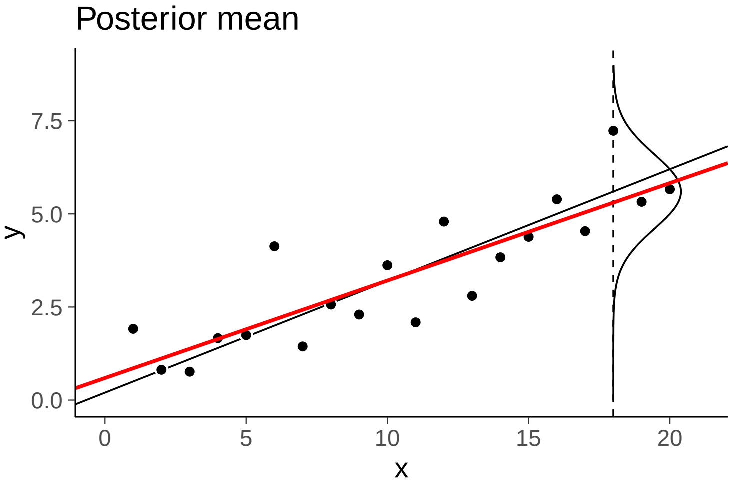
Leave-one-out cross-validation



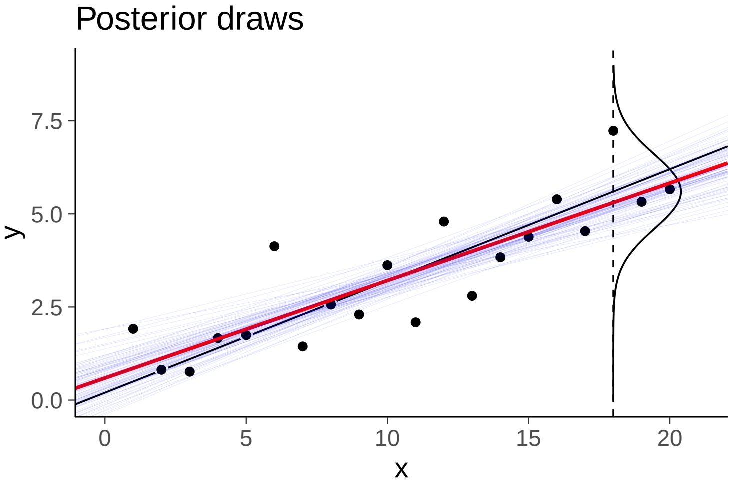
Leave-one-out cross-validation



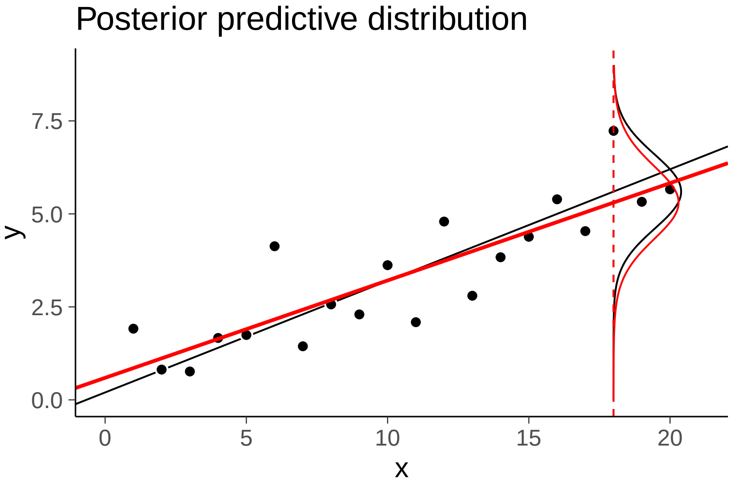
Leave-one-out cross-validation



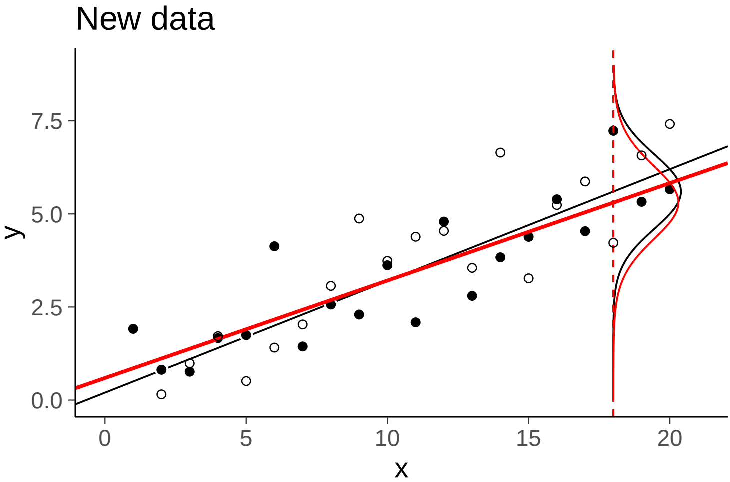
Leave-one-out cross-validation



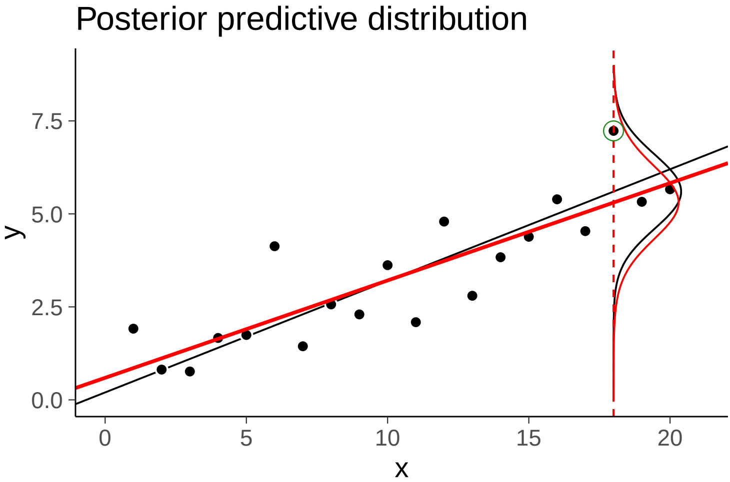
Leave-one-out cross-validation



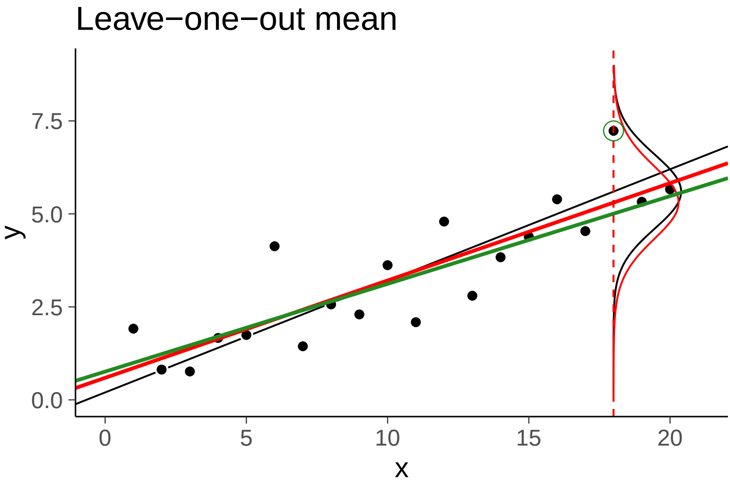
Leave-one-out cross-validation



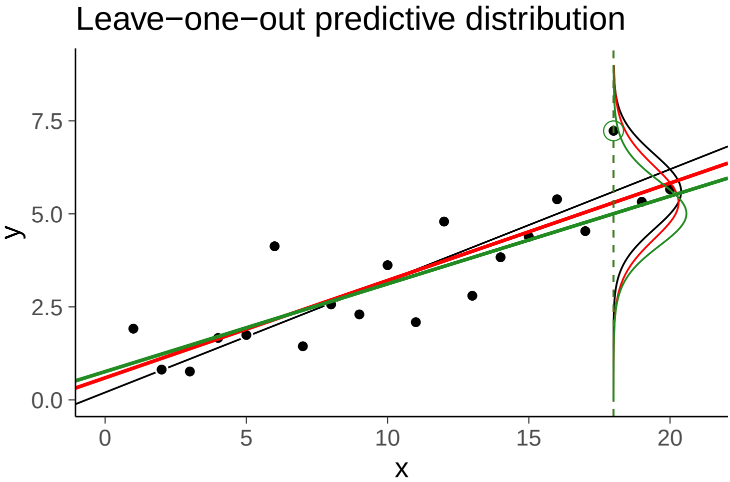
Leave-one-out cross-validation



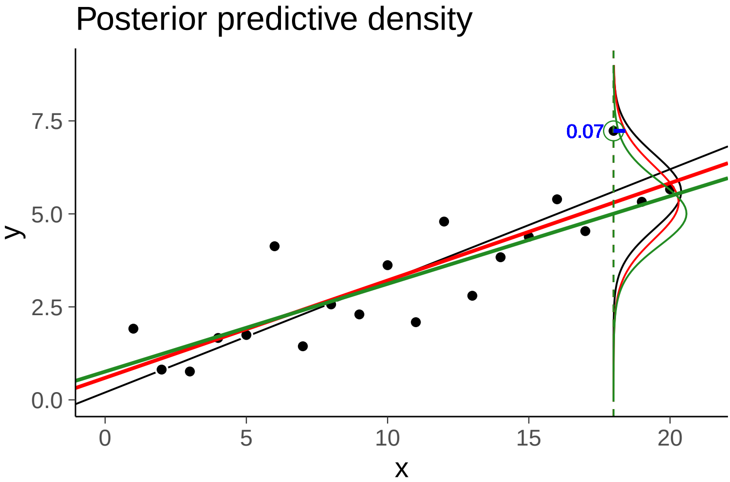
Leave-one-out cross-validation



Leave-one-out cross-validation – log score

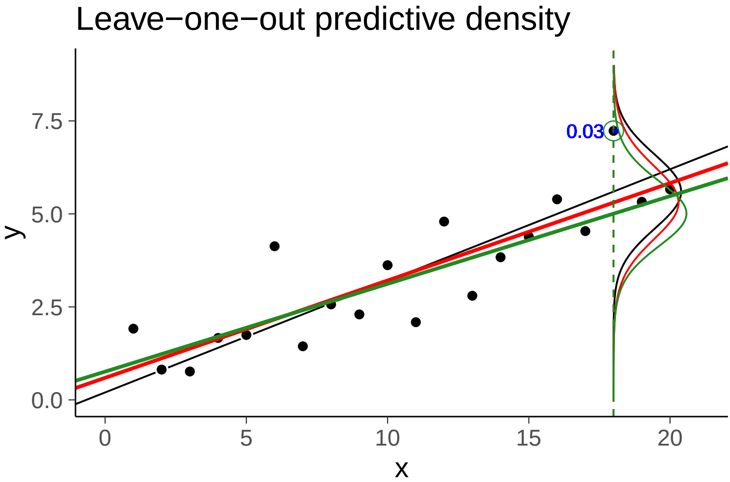


Leave-one-out cross-validation – log score



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

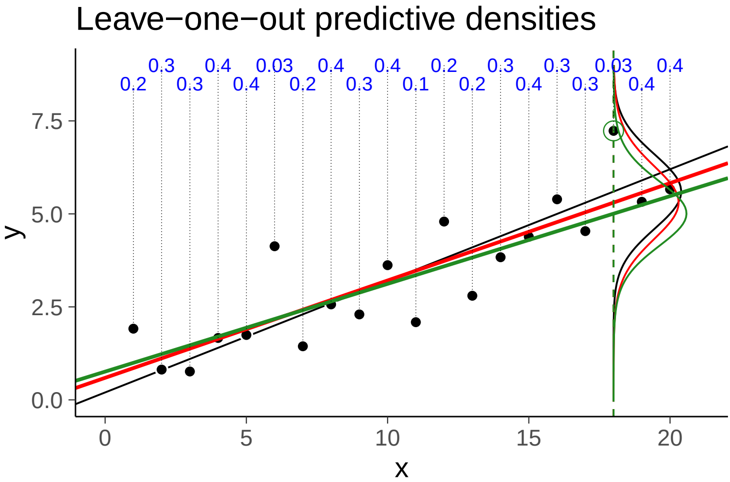
Leave-one-out cross-validation – log score



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

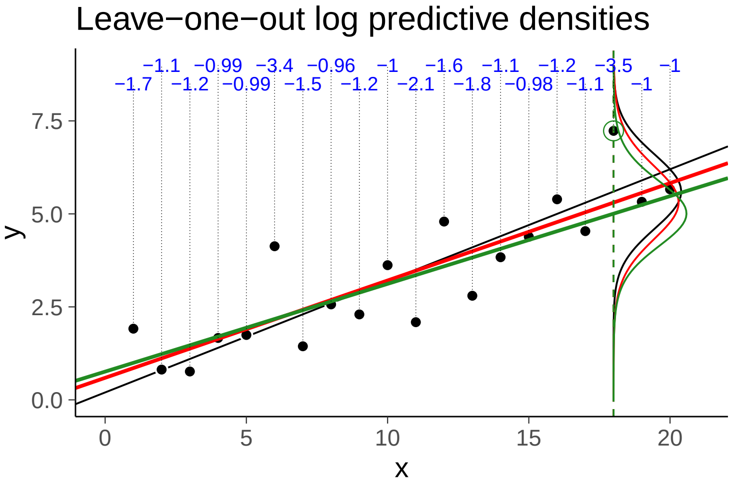
$p(\tilde{y} = y_{18} | \tilde{x} = 18, x_{-18}, y_{-18}) \approx 0.03$

Leave-one-out cross-validation – log score



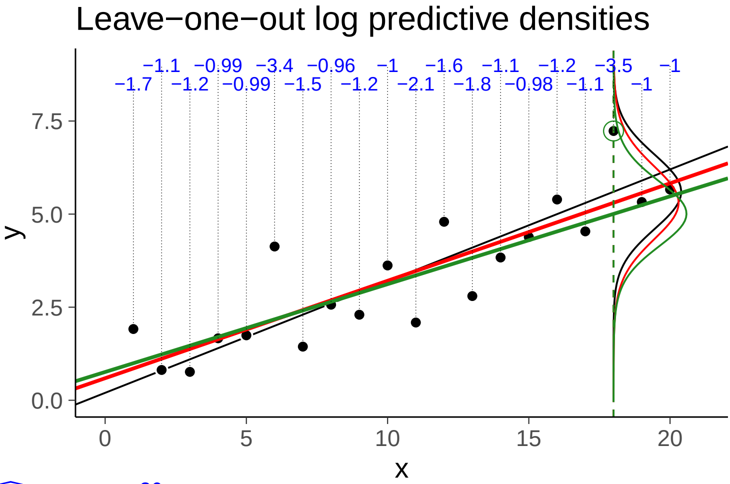
$p(y_i|x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

Leave-one-out cross-validation – log score



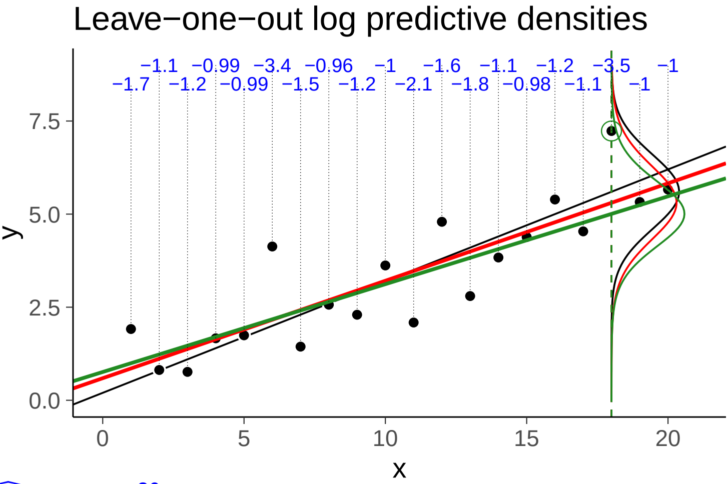
$\log p(y_i | x_i, x_{-i}, y_{-i}), \quad i = 1, \ldots, 20$

Leave-one-out cross-validation – log score



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

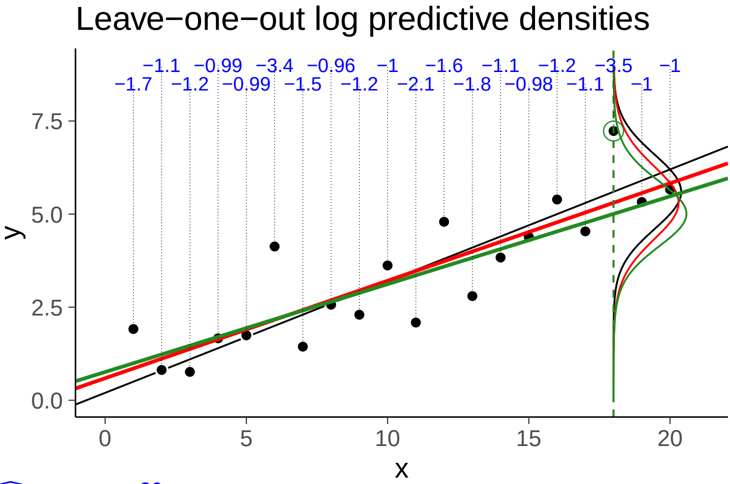
Leave-one-out cross-validation – log score



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

almost unbiased estimate of elpd for new data

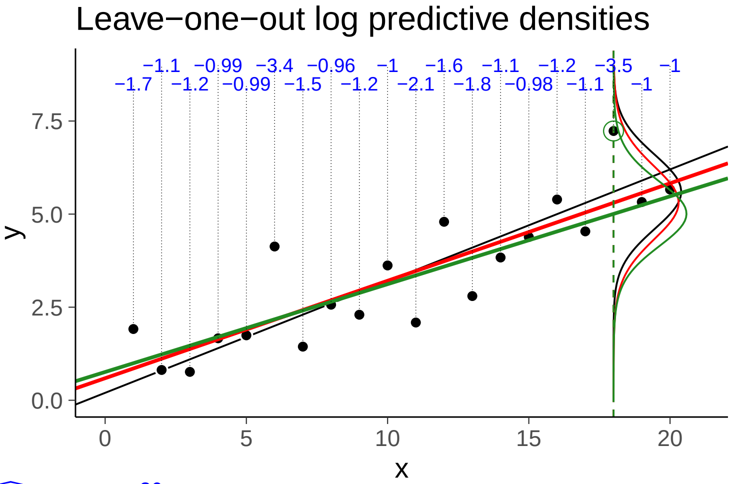
Leave-one-out cross-validation – log score



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

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

Leave-one-out cross-validation – log score



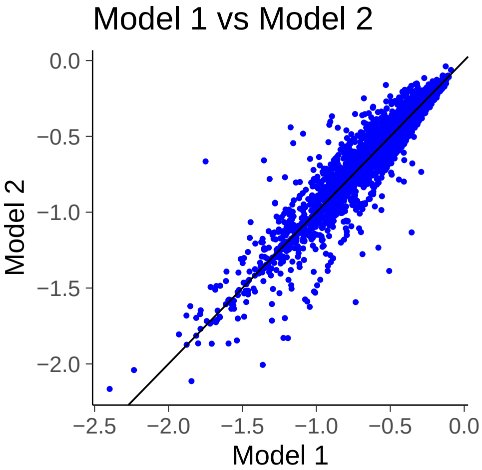
$\widehat{\mathrm{elpd}}_{\mathrm{LOO}} = \sum_{i=1}^{20} \log p(y_i|x_i, x_{-i}, y_{-i}) \approx -29.5$

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

Arsenic well example – Model comparison

- Logistic regression for predicting probability of switching well with high arsenic level in rural Bangladesh
  - Model 1:
    log(arsenic) + distance
  - Model 2:
    log(arsenic) + distance + education level

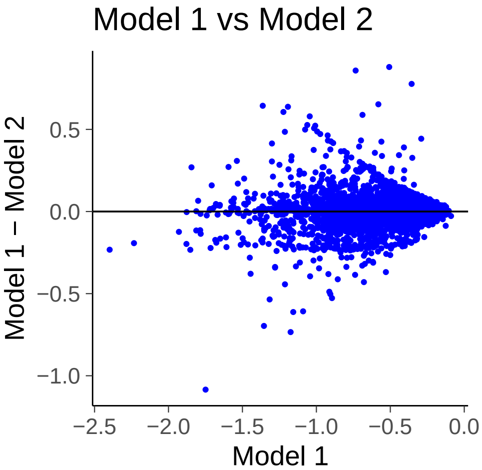
Arsenic well example – Model comparison



Model 1: $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a \mid y^{\mathrm{obs}}\right) \approx$ -1952, SE=16

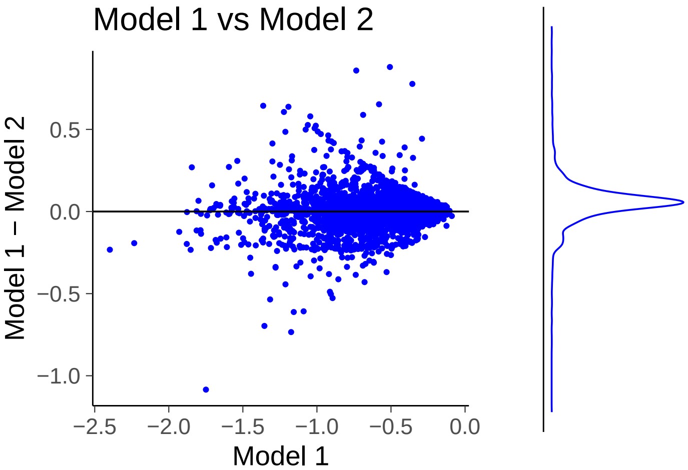
Model 2: $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_b \mid y^{\mathrm{obs}}\right) \approx$ -1938, SE=17

Arsenic well example – Model comparison



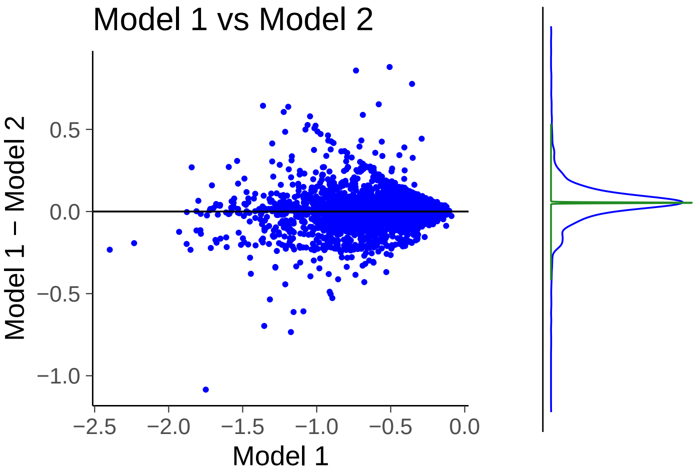
Difference: $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right) \approx$ -14.4, SE = 6.1

Arsenic well example – Model comparison



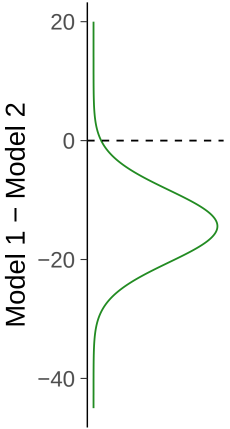
Difference: $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right) \approx$ -14.4, SE = 6.1

Arsenic well example – Model comparison



Difference: $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right) \approx$ -14.4, SE = 6.1

Arsenic well example – Model comparison



Difference: $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right) \approx$ -14.4, SE = 6.1

Cross-validation variants

- leave-group-out
- leave-future-out
- K-fold

References

- Vehtari, Mononen, Tolvanen, Sivula and Winther (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. Journal of Machine Learning Research, 17(103):1–38.
- Riutort-Mayol, Bürkner, Andersen, Solin, and Vehtari (2020). Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. arXiv:2004.11408.
- Siivola, Weber, and Vehtari (2021). Qualifying drug dosing regimens in pediatrics using Gaussian processes. Statistics in Medicine, 10.1002/sim.8907, in press.
- Timonen, Mannerström, Vehtari, and Lähdesmäki (2021). lgpr: An interpretable nonparametric method for inferring covariate effects from longitudinal data. Bioinformatics, accepted for publication. arXiv:1912.03549.
- Järvenpää, Gutmann, Vehtari, and Marttinen (2021). Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. Bayesian Analysis, 16(1):147-148.
- Koistinen, Ásgeirsson, Vehtari, and Jónsson (2019). Nudged elastic band calculations accelerated with Gaussian process regression based on inverse inter-atomic distances. Journal of Chemical Theory and Computation, 15:6738-6751,