ECON5120 Bayesian Data Analysis

Week 6: Forecasting and Model Uncertainty

Santiago Montoya-Blandón

University of Glasgow
Adam Smith Business School

February 25, 2026

Outline

1. Predictive accuracy (external validation)
2. Estimating predictive accuracy
   - 2.1 Naive in-sample estimator
   - 2.2 Information criteria (bias correction)
   - 2.3 Cross-validation
3. Model comparison and hypothesis testing
4. Bayes factors
   - 4.1 Definitions
   - 4.2 Pitfalls
   - 4.3 Alternatives

Predictive accuracy

- Before, we focused on *model checking*
- Model checking involves evaluating in-sample accuracy and so deals with *interval validation*
- We went through a few ways to gauge in-sample accuracy

Predictive accuracy

- Before, we focused on *model checking*
- Model checking involves evaluating in-sample accuracy and so deals with *interval validation*
- We went through a few ways to gauge in-sample accuracy
- We will now focus on *predictive accuracy* instead
- Predictive accuracy evaluates out-of-sample accuracy and so deals with *external validation*
- We want to know how good is our fitted model for predicting new data
- True predictive performance is found out by using model to make predictions and comparing predictions to new observations

UofG

Fitting vs Prediction

- Main issue is that prediction is inherently harder than fitting

Fitting vs Prediction

- Main issue is that prediction is inherently harder than fitting
- To see this, let's look at a simple example
- True model: $y = f(x; \theta_0) + u$ with fixed x
- Fitted model: $\hat{y} = f(x; \hat{\theta})$
- Define $f_0 := f(x; \theta_0)$ and $\hat{f} := f(x; \hat{\theta})$
- We use mean squared error as our measure of quality of prediction:

$$\mathrm{MSE}(\hat{y}) = \mathbb{E}[(\hat{y} - y)^2]$$

Fitting vs Prediction

- We can easily show

$$\text{MSE}(\hat{y}) = \text{MSE}(\hat{f}) + \text{Var}(u)$$

- Prediction accuracy depends on two factors:
  1. Fitting accuracy as measured by $\text{MSE}(\hat{f}) = \mathbb{E}[(\hat{f} - f_0)^2]$
  2. Variance of the error term $\text{Var}(u)$

Fitting vs Prediction

- We can easily show

$$\text{MSE}(\hat{y}) = \text{MSE}(\hat{f}) + \text{Var}(u)$$

- Prediction accuracy depends on two factors:
  1. Fitting accuracy as measured by $\text{MSE}(\hat{f}) = \mathbb{E}[(\hat{f} - f_0)^2]$
  2. Variance of the error term $\text{Var}(u)$
- We also have $\text{MSE}(\hat{f}) = \text{Var}(\hat{f}) + \text{Bias}(\hat{f}, f_0)^2$
- We can try to minimize $\text{MSE}(\hat{f})$ by improving our fit of $\hat{\theta}$ (variance and bias trade-off)
- However, the error term variance $\text{Var}(u)$ *cannot* be reduced via modelling
- This shows $\text{MSE}(\hat{y}) > \text{MSE}(\hat{f})$

Digression: Measures of predictive accuracy

- How to select your cost or utility function in order to evaluate predictions?
- *Loss functions* $L(\hat{\theta}, \theta_0)$ capture the cost of using $\hat{\theta}$ to make decisions when the truth is θ_0

Digression: Measures of predictive accuracy

- How to select your cost or utility function in order to evaluate predictions?
- *Loss functions* $L(\hat{\theta}, \theta_0)$ capture the cost of using $\hat{\theta}$ to make decisions when the truth is θ_0
- In class and theory we focus on measures of *statistical* predictive accuracy
  - Squared loss $\Rightarrow L(\hat{\theta}, \theta_0) = (\hat{\theta} - \theta_0)^2$
  - Absolute loss $\Rightarrow L(\hat{\theta}, \theta_0) = |\hat{\theta} - \theta_0|$
  - Zero-one loss $\Rightarrow L(\hat{\theta}, \theta_0) = \mathbb{I}(\hat{\theta} \neq \theta_0) = \begin{cases} 0 & \text{if } \hat{\theta} = \theta_0 \\ 1 & \text{if } \hat{\theta} \neq \theta_0 \end{cases}$
  - Asymmetric losses $\Rightarrow L(\hat{\theta}, \theta_0) = \begin{cases} a|\hat{\theta} - \theta| & \text{if } \hat{\theta} > \theta_0 \\ b|\hat{\theta} - \theta| & \text{if } \hat{\theta} < \theta_0 \end{cases}$

Objective Functions: Example

- The loss function can (and should when possible) be application-specific:
  - Money
  - Consumption
  - Quality-adjusted life expentancy
- Example: how to choose between more money or more time?

Objective Functions: Example

- The loss function can (and should when possible) be application-specific:
  - Money
  - Consumption
  - Quality-adjusted life expentancy
- Example: how to choose between more money or more time?
- Consider the following simple exercise highlighting objective functions
- **Task:** Summarize the distribution of random variable θ (e.g., returns to education), distributed as $f(\theta)$, according to some scalar measure
- That is, provide a single number that best describes the distribution of θ

Question for you

What would you choose?

Objective Functions: Example

- As we saw: answer depends on what you mean by *best*
- Different summaries are optimal according to different objectives

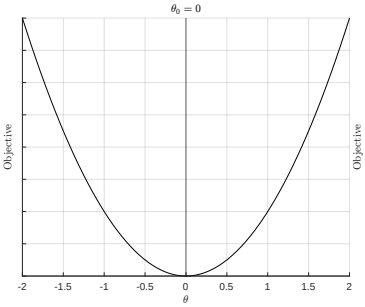
Objective Functions: Example

- As we saw: answer depends on what you mean by *best*
- Different summaries are optimal according to different objectives
- To formalize the example:
  - $\theta_0 \in \mathbb{R}$ is a single realization from θ; i.e., $\theta_0 \sim f(\theta)$ (imagine $\theta_0 = 0$ for now)
  - $\hat{\theta} \in \mathbb{R}$ represents any scalar summary measure
  - $L(\hat{\theta}, \theta_0)$ is the objective (loss) function of using $\hat{\theta}$ to summarize data θ_0
- Use expected loss $\mathbb{E}[L(\hat{\theta}, \theta_0)]$ to remove dependence on θ_0, where the expectation is with respect to $f(\theta)$
- The best summary $\hat{\theta}$, denoted as θ^*, should be the one that minimizes this expected loss
- The form and properties of θ^* depend on the loss function itself

Objective Functions: Example

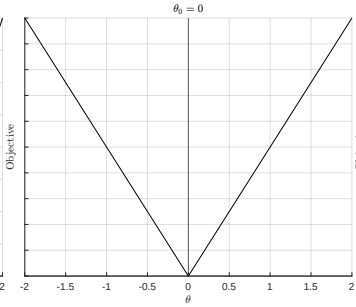
Solutions to $\theta^* = \arg\min_{\hat{\theta} \in \mathbb{R}} \mathbb{E}[L(\hat{\theta}, \theta_0)]$ for different objectives (losses) L:

Figure: Squared loss



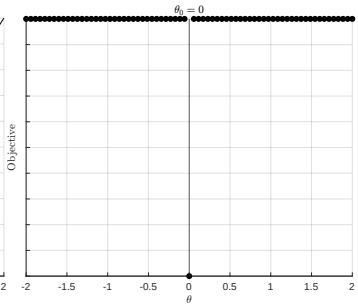
$$L(\hat{\theta}, \theta_0) = (\hat{\theta} - \theta_0)^2$$
$$\theta^* = \mathbb{E}[\theta]$$

Figure: Absolute loss



$$L(\hat{\theta}, \theta_0) = |\hat{\theta} - \theta_0|$$
$$\theta^* = \text{median}(\theta)$$

Figure: Zero-one loss



$$L(\hat{\theta}, \theta_0) = \mathbb{I}(\hat{\theta} \neq \theta_0)$$
$$\theta^* = \text{mode}(\theta)$$

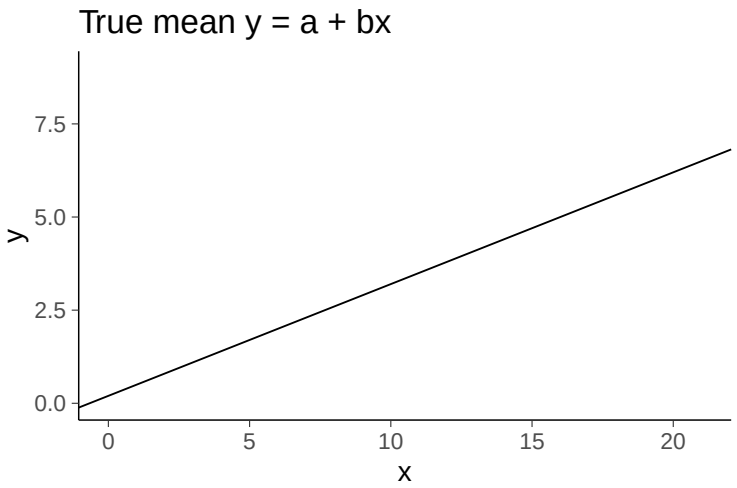
Bayesian predictive accuracy

- Our initial example was frequentist in nature and focused on providing a *point prediction* $\hat{y}$ (i.e., a number)
- We will focus on a Bayesian measure of predictive accuracy and so we will provide a *probabilistic prediction* (a distribution over new data $\tilde{y}$)
- Our Bayesian probabilistic prediction is the posterior $p(\tilde{y}|y)$

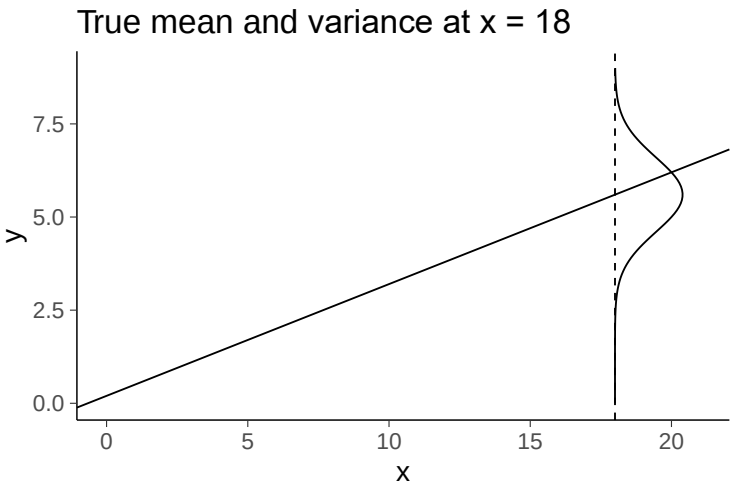
Bayesian predictive accuracy

- Our initial example was frequentist in nature and focused on providing a *point prediction* $\hat{y}$ (i.e., a number)
- We will focus on a Bayesian measure of predictive accuracy and so we will provide a *probabilistic prediction* (a distribution over new data $\tilde{y}$)
- Our Bayesian probabilistic prediction is the posterior $p(\tilde{y}|y)$
- When interested in overall predictive accuracy for a distribution or when there is no application-specific cost function, use the *logarithmic score*
- For model M based on data y that yields a posterior distribution $p(\tilde{y}|y, M)$, the logarithmic score is simply $\log p(\tilde{y}|y, M)$
- This measure has some decision-theoretic optimality properties and is connected to MSE

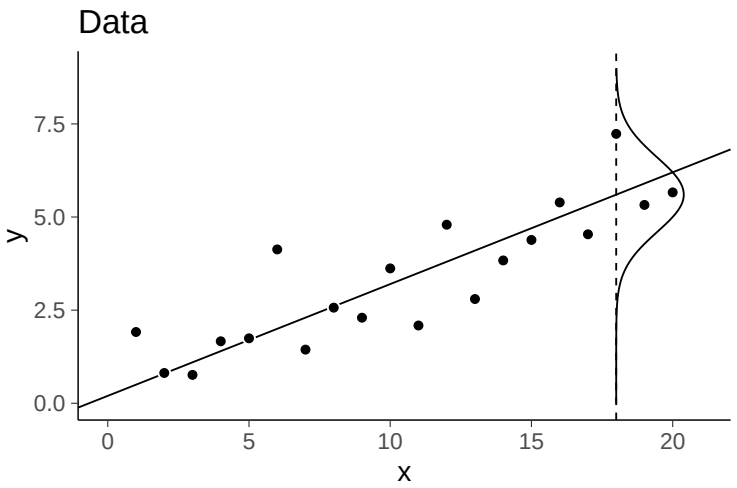
Example: Bayesian predictive accuracy



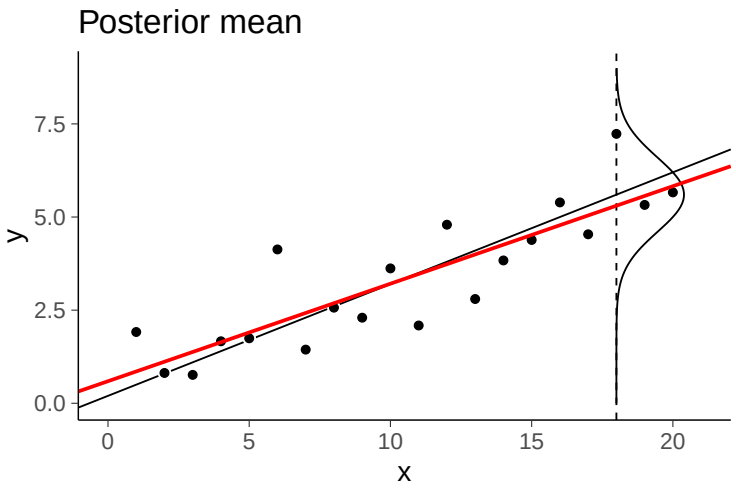
Example: Bayesian predictive accuracy



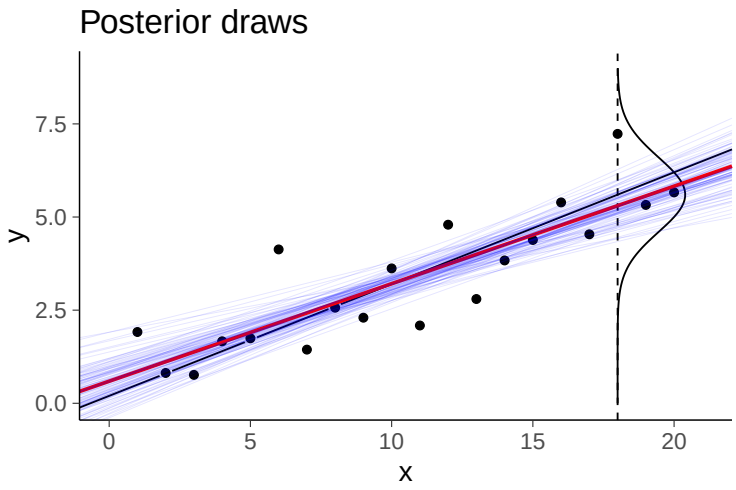
Example: Bayesian predictive accuracy



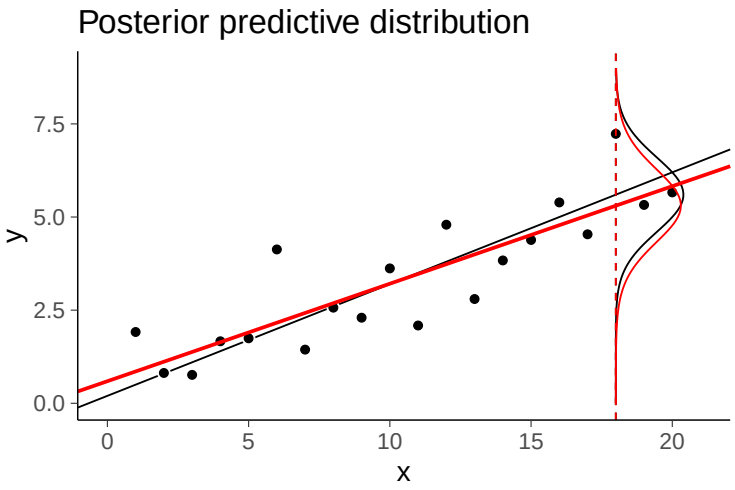
Example: Bayesian predictive accuracy



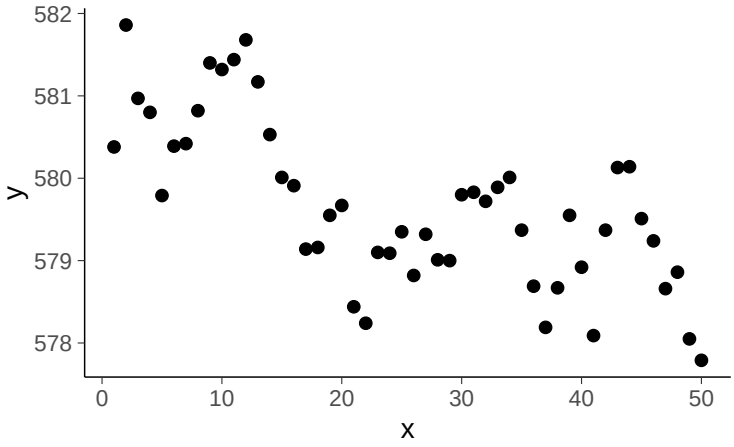
Example: Bayesian predictive accuracy



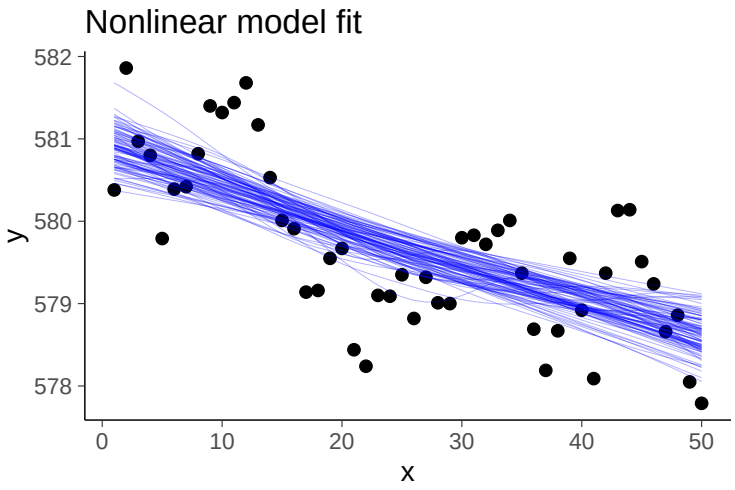
Example: Bayesian predictive accuracy



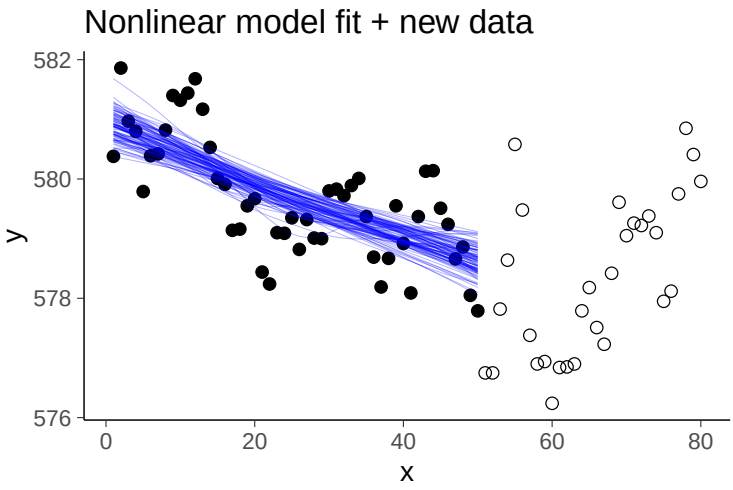
Example: Bayesian predictive accuracy in nonlinear models



Example: Bayesian predictive accuracy in nonlinear models



Example: Bayesian predictive accuracy in nonlinear models



Bayesian predictive accuracy

- For a given $\tilde{y}$, our measure of quality would be

$$\log p(\tilde{y}|y) = \log \int p(\tilde{y}|\theta)p(\theta|y)d\theta = \log \mathbb{E}_{p(\theta|y)}[p(\tilde{y}|\theta)]$$

Bayesian predictive accuracy

- For a given $\tilde{y}$, our measure of quality would be

$$\log p(\tilde{y}|y) = \log \int p(\tilde{y}|\theta)p(\theta|y)d\theta = \log \mathbb{E}_{p(\theta|y)}[p(\tilde{y}|\theta)]$$

- We want a measure that works across the distribution of $\tilde{y}$ given by $f(\tilde{y})$
- We define the expected out-of-sample log predictive density (elpd)

$$\text{elpd} = \mathbb{E}_f[\log p(\tilde{y}|y)] = \int \log p(\tilde{y}|y) f(\tilde{y}) d\tilde{y}$$

Bayesian predictive accuracy

- For a given $\tilde{y}$, our measure of quality would be

$$\log p(\tilde{y}|y) = \log \int p(\tilde{y}|\theta)p(\theta|y)d\theta = \log \mathbb{E}_{p(\theta|y)}[p(\tilde{y}|\theta)]$$

- We want a measure that works across the distribution of $\tilde{y}$ given by $f(\tilde{y})$
- We define the expected out-of-sample log predictive density (elpd)

$$\text{elpd} = \mathbb{E}_f[\log p(\tilde{y}|y)] = \int \log p(\tilde{y}|y) f(\tilde{y}) d\tilde{y}$$

- This is our measure of *predictive accuracy* that we want to maximize across different models
- For convenience and comparability, we usually transform elpd to a minimization and so we work with $-2\,\text{elpd}$

Estimating predictive accuracy

- As elpd depends on the unknown $f(\tilde{y})$, we will need to estimate it
- Ideally, we would collect new data $\tilde{y}$ and fit our model to the new data to measure predictive accuracy directly

Estimating predictive accuracy

- As elpd depends on the unknown $f(\tilde{y})$, we will need to estimate it
- Ideally, we would collect new data $\tilde{y}$ and fit our model to the new data to measure predictive accuracy directly
- Question becomes: how to best estimate elpd without new data?
  1. Naive estimate: Estimate elpd using data y. Problem is this tends to overfit (optimistic estimate)
  2. Information criteria: use the naive estimate but correct the bias that is created by overfitting
  3. Cross-validation: artificially split data into training and evaluation sets. Fit model on the training data and then test on the evaluation data. Repeat on different ways to split data

UofG

Estimating predictive accuracy: Naive estimator

- A simple yet naive way of estimating elpd is simply to compute it over the actual data y
- That is, compute the expectation in elpd using draws $\theta^{(1)}, \ldots, \theta^{(S)}$:

$$\widehat{\text{elpd}}_{naive} = \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y|\theta^{(s)})\right)$$

- If data are independent then $p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$ and so the previous computation reduces to

$$\widehat{\text{elpd}}_{naive} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^{(s)})\right)$$

Estimating predictive accuracy: Information criteria

- As noted before, main issue with the naive estimator is that it is too optimistic (gives an accuracy that is too large)
- As model was fit on data y, then overfitting is likely and so we are overestimating out-of-sample predictive accuracy

Estimating predictive accuracy: Information criteria

- As noted before, main issue with the naive estimator is that it is too optimistic (gives an accuracy that is too large)
- As model was fit on data y, then overfitting is likely and so we are overestimating out-of-sample predictive accuracy
- Idea: take the naive estimate and correct for the bias caused by overfitting
- This is precisely what most information criteria are designed to accomplish
  1. Akaike (AIC)
  2. Deviance (DIC)
  3. Watanabe-Akaike or widely available IC (WAIC)

Estimating predictive accuracy: AIC

- We will need a way to measure predictive accuracy for a given point estimate $\hat{\theta}$
- As this estimate does not capture the uncertainty in θ, then only uncertainty present in $\tilde{y}$ is relevant:

$$\text{elpd}_{\hat{\theta}} = \mathbb{E}_f[\log p(\tilde{y}|\hat{\theta})]$$

Estimating predictive accuracy: AIC

- We will need a way to measure predictive accuracy for a given point estimate $\hat{\theta}$
- As this estimate does not capture the uncertainty in θ, then only uncertainty present in $\tilde{y}$ is relevant:

$$\text{elpd}_{\hat{\theta}} = \mathbb{E}_f[\log p(\tilde{y}|\hat{\theta})]$$

- AIC estimates elpd at data y by using the maximum likelihood estimate $\hat{\theta}_{mle}$ and then corrects for the number of estimated parameters k

$$\widehat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k$$

- AIC is then defined as $\text{AIC} = -2\widehat{\text{elpd}}_{\text{AIC}} = -2\log p(y|\hat{\theta}_{mle}) + 2k$

Estimating predictive accuracy: AIC

- For AIC, bias is only in terms of the number of estimated parameters
- Intuitively, the more parameters you fit in a model the better that fit should be, since you can capture more features of the distribution of y
- AIC corrects for how much the fitting of k parameters will increase predictive accuracy by chance alone

Estimating predictive accuracy: AIC

- For AIC, bias is only in terms of the number of estimated parameters
- Intuitively, the more parameters you fit in a model the better that fit should be, since you can capture more features of the distribution of y
- AIC corrects for how much the fitting of k parameters will increase predictive accuracy by chance alone
- However, the number of parameters alone is generally not enough to capture the improvement in fit
- Focus instead on *effective* number of parameters
- This depends both on the model and data (not just model as for AIC)

Estimating predictive accuracy: DIC

- DIC has two changes compared to DIC:
  1. Use posterior mean $\hat{\theta}_{\text{Bayes}} = \mathbb{E}_{p(\theta|y)}[\theta]$ as point estimate
  2. Compute effective number of parameters p_{DIC} as bias correction
- Two possibilities for p_{DIC}:

$$p_{\text{DIC},1} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - \mathbb{E}_{p(\theta|y)}[\log p(y|\theta)]\right)$$
$$p_{\text{DIC},2} = \text{Var}_{p(\theta|y)}[\log p(y|\theta)]$$

- Compute predictive accuracy according to DIC using either form of p_{DIC} as

$$\widehat{\text{elpd}}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - \hat{p}_{\text{DIC}}$$

- DIC is then defined as $\text{DIC} = -2\widehat{\text{elpd}}_{\text{DIC}}$

Estimating predictive accuracy: DIC

- As always, compute all expectations over the posterior by using draws $\theta^{(1)}, \ldots, \theta^{(S)}$
- The effective number of parameters p_{DIC} simplifies to k when using a normal likelihood with known variance and uniform priors over the coefficients
- In practice, it is recommended to use $p_{\text{DIC},1}$ as it is more numerically stable

Estimating predictive accuracy: WAIC

- WAIC is a more fully Bayesian alternative to previous criteria
- It is based on an independence (or hierarchical) assumption over the data
- Instead of computing elpd at a given $\hat{\theta}$, it computes the naive Bayesian estimate
- The corrections p_{WAIC} are similar to DIC:

$$p_{\text{WAIC},1} = 2\sum_{i=1}^{n}\left(\log(\mathbb{E}_{p(\theta|y)}[p(y_i|\hat{\theta})]) - \mathbb{E}_{p(\theta|y)}[\log p(y_i|\theta)]\right)$$

$$p_{\text{WAIC},2} = \sum_{i=1}^{n}\text{Var}_{p(\theta|y)}[\log p(y_i|\theta)]$$

Estimating predictive accuracy: WAIC

- Compute predictive accuracy according to WAIC using either form of p_{WAIC} as

$$\widehat{\text{elpd}}_{\text{WAIC}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{(s)}) \right) - \hat{p}_{\text{WAIC}}$$

- WAIC is then defined as $\text{WAIC} = -2\widehat{\text{elpd}}_{\text{WAIC}}$

Estimating predictive accuracy: WAIC

- Compute predictive accuracy according to WAIC using either form of p_{WAIC} as

$$\widehat{\text{elpd}}_{\text{WAIC}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{(s)}) \right) - \hat{p}_{\text{WAIC}}$$

- WAIC is then defined as $\text{WAIC} = -2\widehat{\text{elpd}}_{\text{WAIC}}$
- In practice, $p_{\text{WAIC},2}$ is recommended for theoretical and practical reasons
- BDA recommend the use of WAIC as a simple predictive measure when cross-validation is too expensive to compute
- A problem with WAIC is that it assumes some sort of independence in the data, which is not done by AIC or DIC

Estimating predictive accuracy: Cross-validation

- *Cross-validation* is a widely used technique in data analysis for evaluating prediction accuracy
- Remember that the main issue with computing predictive accuracy is that we do not observe out-of-sample data
- Cross-validation solves this issue by using the following steps:
  1. Split y into y_{train} and y_{test}: y_{train} represents actual data, y_{test} represents out-of-sample data
  2. Fit model to training data by getting the posterior $p(\theta|y_{train})$
  3. Compute predictive accuracy of the fitted model on test data y_{test} by using $\log p(y_{test}|y_{train}) = \log \mathbb{E}_{p(\theta|y_{train})}[p(y_{test}|\theta)]$
  4. Repeat previous steps by varying y_{train} and y_{test} and then aggregate the results across partitions

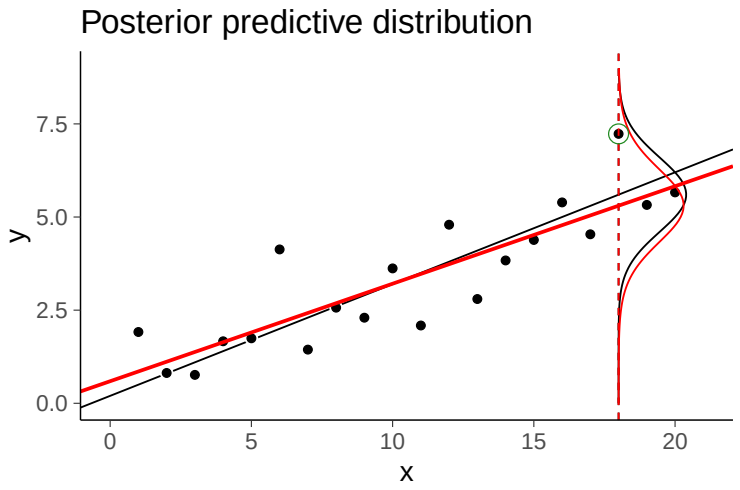
Estimating predictive accuracy: Cross-validation

- Let's focus on leave-one-out cross-validation (LOO-CV)
- This variation sets $y_{test} = y_i$ and $y_{train} = y_{-i}$ for $i = 1, \ldots, n$
- That is, take the full data y and fit the model to all but one observation. Compute test accuracy for this observation and repeat for all observations
- After the process, you will have n estimates of $\log p(\tilde{y}|y)$ given by $\log p(y_i|y_{-i})$ for $i = 1, \ldots, n$
- The final estimate of out-of-sample log predicted fit is given by

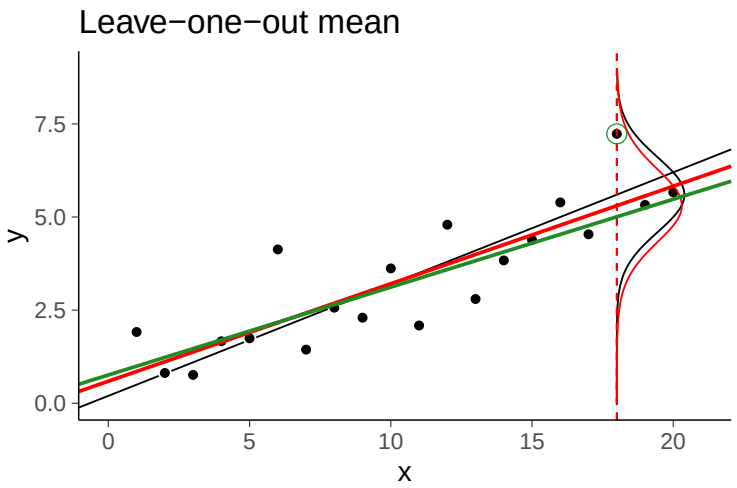
$$\sum_{i=1}^{n} \log \hat{\mathbb{E}}_{p(\theta|y_{-i})}[p(y_i|\theta)]$$

- As usual $\hat{\mathbb{E}}_{p(\theta|y_{-i})}[p(y_i|\theta)]$ is estimated using posterior draws

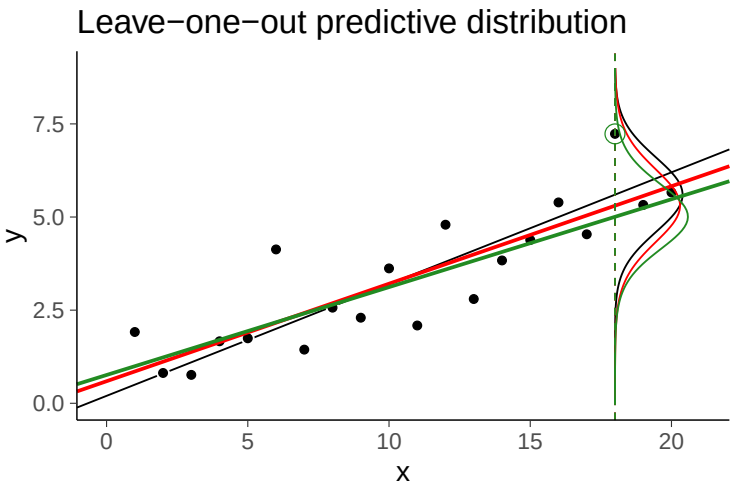
Example: Bayesian predictive accuracy



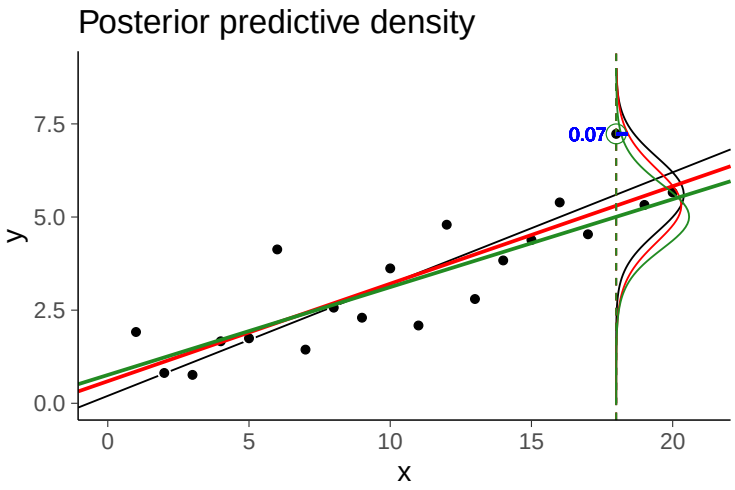
Example: Bayesian predictive accuracy



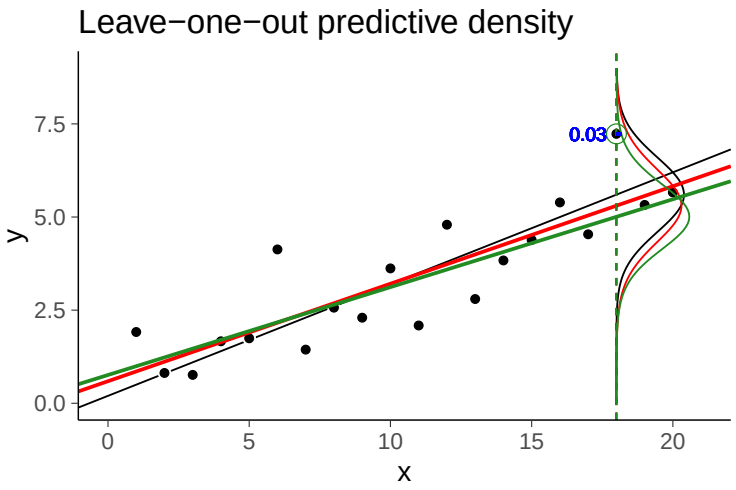
Example: Bayesian predictive accuracy



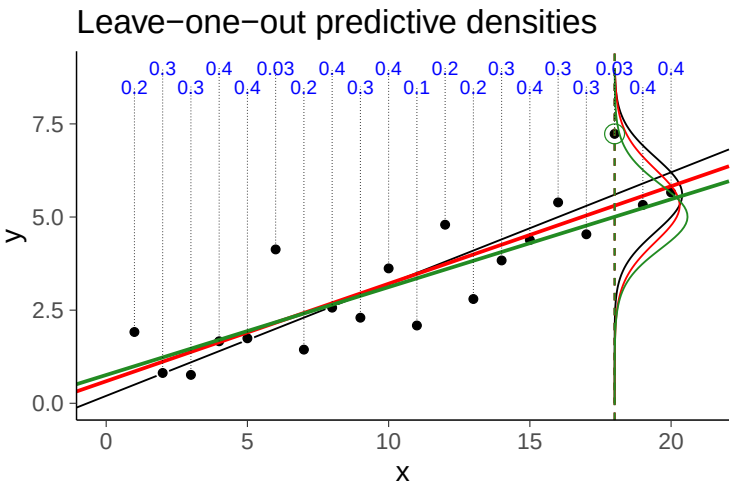
Example: Bayesian predictive accuracy



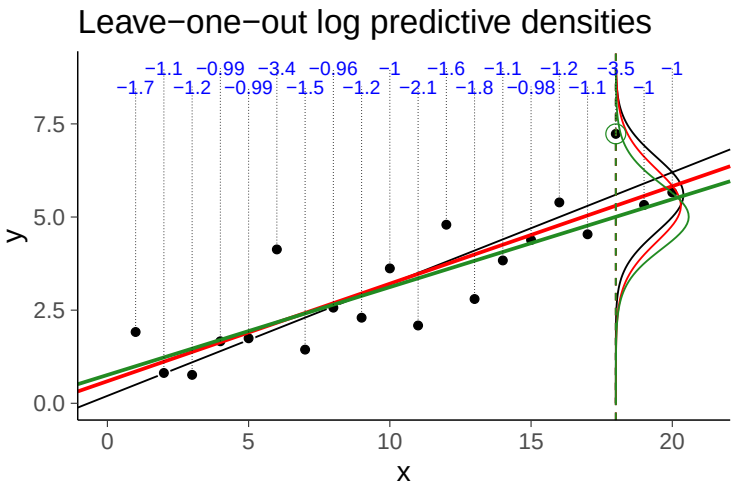
Example: Bayesian predictive accuracy



Example: Bayesian predictive accuracy



Example: Bayesian predictive accuracy



Estimating predictive accuracy: Cross-validation

- The many variations in the cross-validation approach are given by the way in which data is split into training and evaluation
- With large n and a complicated model, it is likely unfeasible to fit all possible n posteriors

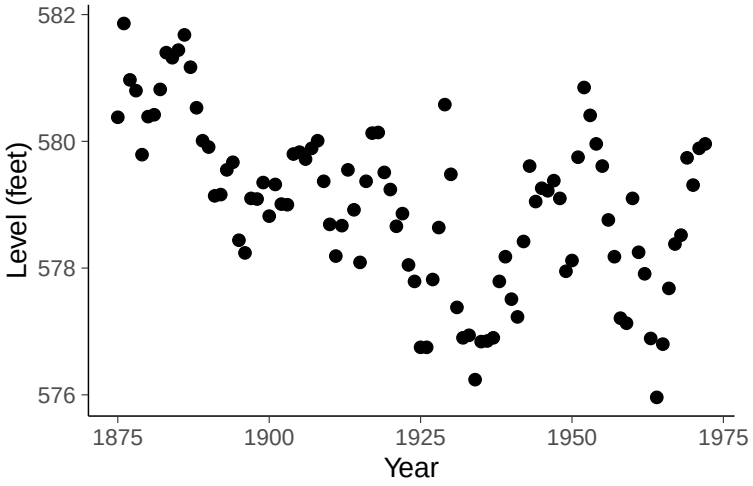
Estimating predictive accuracy: Cross-validation

- The many variations in the cross-validation approach are given by the way in which data is split into training and evaluation
- With large n and a complicated model, it is likely unfeasible to fit all possible n posteriors
- You might instead choose to separate data into K equal parts $(y_1, \ldots, y_K)$ and choose $y_{test} = y_k$, $y_{train} = y_{-k}$ for $k = 1, \ldots, K$
- This is known as K-fold cross-validation. In practice, $K = 10$ is standard as it balances computation time with accuracy

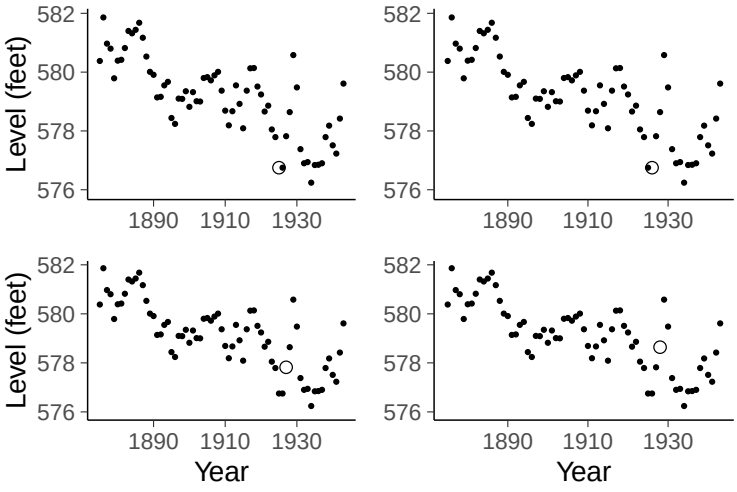
Estimating predictive accuracy: Cross-validation

- The many variations in the cross-validation approach are given by the way in which data is split into training and evaluation
- With large n and a complicated model, it is likely unfeasible to fit all possible n posteriors
- You might instead choose to separate data into K equal parts $(y_1, \ldots, y_K)$ and choose $y_{test} = y_k$, $y_{train} = y_{-k}$ for $k = 1, \ldots, K$
- This is known as K-fold cross-validation. In practice, $K = 10$ is standard as it balances computation time with accuracy
- Cross-validation can be used with structured data: time series or hierarchical
- Be careful about the grouping in the data and the objective of prediction

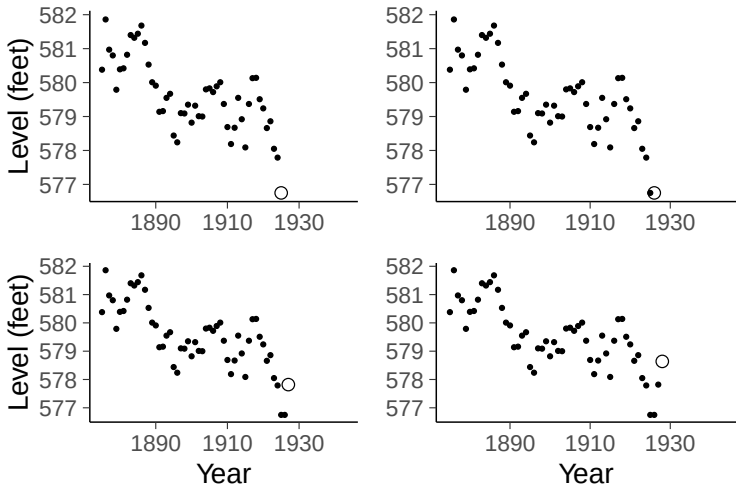
Example: Cross-validation for time series data



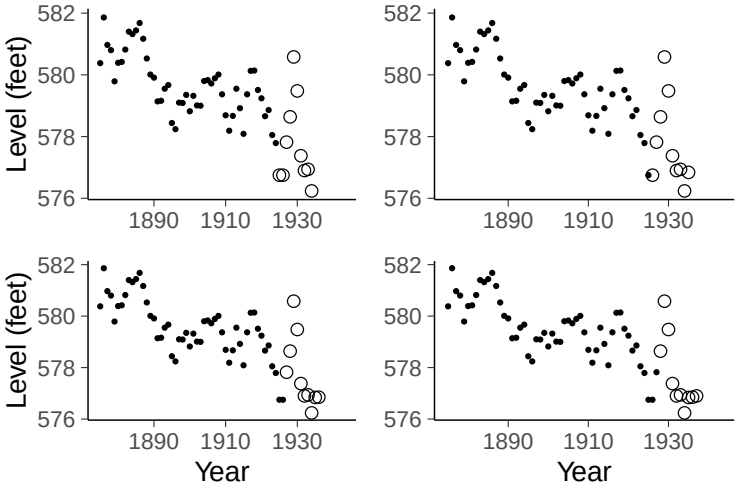
Example: Cross-validation for time series data



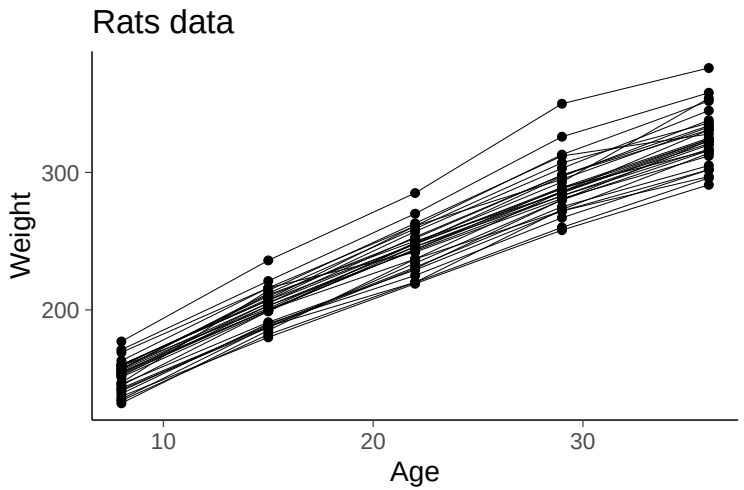
Example: Cross-validation for time series data



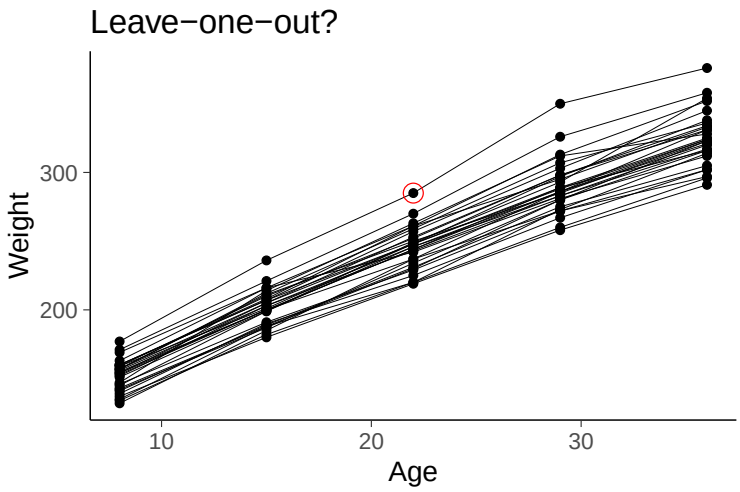
Example: Cross-validation for time series data



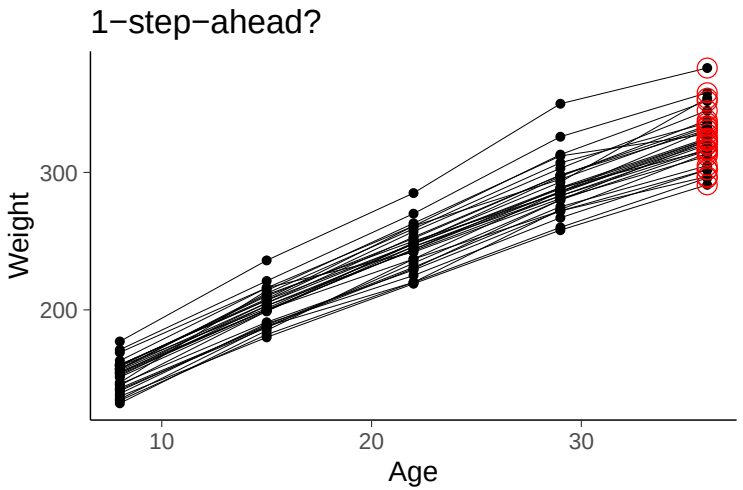
Example: Cross-validation for hierarchical data



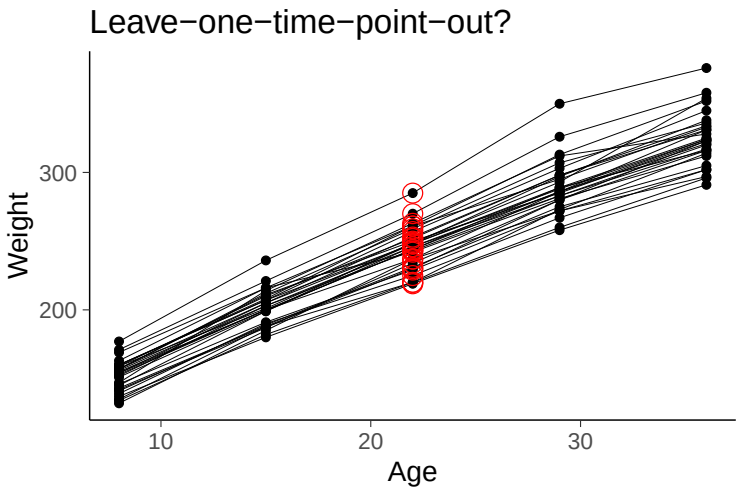
Example: Cross-validation for hierarchical data



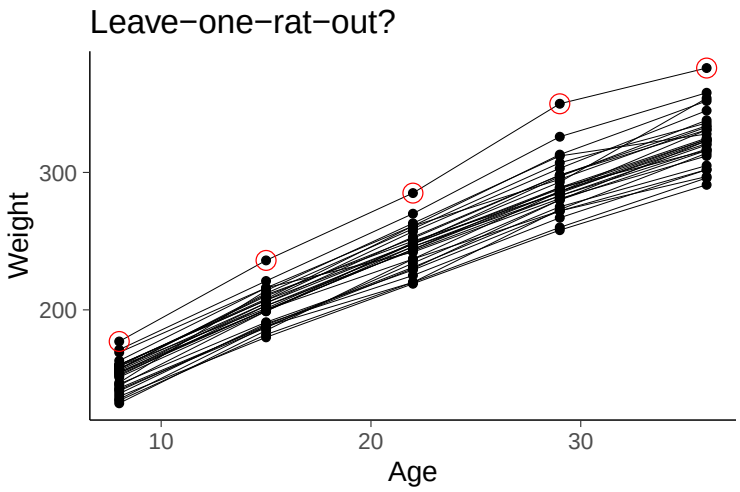
Example: Cross-validation for hierarchical data



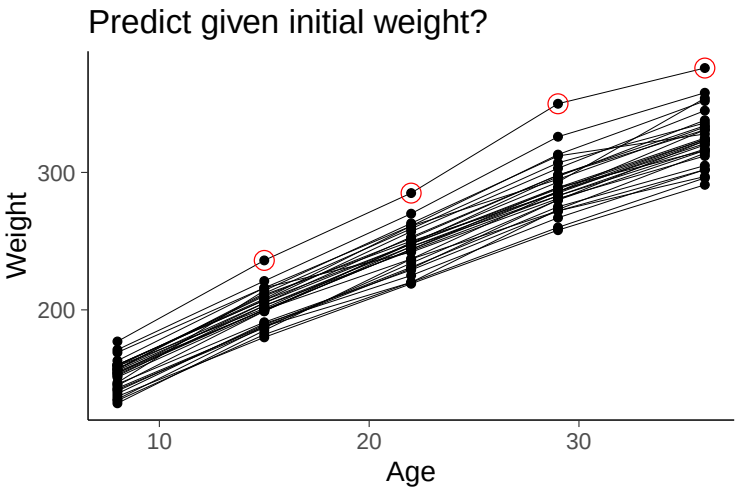
Example: Cross-validation for hierarchical data



Example: Cross-validation for hierarchical data



Example: Cross-validation for hierarchical data



Model comparison: Bayes factors

- Assume R possible models, M_r for $r = 1, \ldots, R$, each depending on parameters θ^r
- For each model, we can obtain

$$p(\theta^r|y, M_r) \propto p(y|\theta^r, M_i)p(\theta^r|M_i)$$

- Similarly, all knowledge about model M_i contained in the data is in

$$p(M_r|y) = \frac{p(y|M_r)p(M_r)}{p(y)}$$

- Note that $p(M_r)$ is a prior model probability, $p(y|M_r)$ is model r's marginal likelihood and $p(M_r|y)$ are *posterior model probabilities*
- We also have

$$p(y|M_r) = \int p(y|\theta^r, M_r)p(\theta^r|M_r)d\theta^r$$

Model comparison: Bayes factors

- To compare models, we usually use *posterior odd ratios* between models M_r and M_t

$$PO_{rt} = \frac{p(M_r|y)}{p(M_t|y)} = \frac{p(y|M_r)p(M_r)}{p(y|M_t)p(M_t)}$$

- These odds can be written as a product of what is known as the *Bayes factor* and the prior odds ratio between models r and t

$$BF_{rt} = \frac{p(y|M_r)}{p(y|M_t)}$$

- This Bayes factor can also be used to test hypotheses about models

Model comparison: Bayes factors

- There are two interesting special cases that arise when testing a restricted vs. an unrestricted model
- Let $\theta = (\theta_1, \theta_2)$ where M_1 assumes $\theta_2 = \theta^*$ but model M_2 lets $\theta_2 \neq \theta^*$
  1. Savage-Dickey ratio: If the prior is such that $p(\theta_1|\theta_2 = \theta^*, M_2) = p(\theta_1|M_1)$, then
  $$BF_{12} = \frac{p(\theta_2 = \theta^*|y, M_2)}{p(\theta_2 = \theta^*|M_2)}$$
  2. Generalized Savage-Dickey ratio: If no restriction is imposed in the prior $p(\theta_1|M_1)$, then
  $$BF_{12} = \frac{p(\theta_2 = \theta^*|y, M_2)}{p(\theta_2 = \theta^*|M_2)} \cdot \mathbb{E}_{p(\theta_1|\theta_2=\theta^*, y, M_2)} \left[\frac{p(\theta_1|M_1)}{p(\theta_1|\theta_2 = \theta^*, M_2)} \right]$$

Model comparison: Bayes factors

- In general, Bayes factors have no closed-form solution
- A big issue is that to compute them, you must compute a quantity we have avoided throughout the lectures: the marginal likelihood $p(y|M)$
- A few methods to compute $p(y|M)$ exist in the literature:
  - Gelfand and Dey (1994)
  - Chib (1995) for Gibbs sampling output
  - Chib and Jeliazkov (2001) for Metropolis-Hastings sampling output
  - Bridge sampling (Bennett, 1976; Meng & Wong, 1996)

Alternatives to Bayes factors

- Bayes factors can be misleading in some circumstances:
  - Using improper prior distributions
  - Some non-informative but proper priors
  - Continuous model spaces
- In light of some of the pitfalls associated to using Bayes factors, there are several alternatives in the literature
- We will explore a few of these in the lectures and lab:
  1. Bayesian model averaging (BMA)
  2. Stochastic search variable selection (also known as spike-and-slab priors)
  3. LASSO, elastic nets and other variants using different penalties