Recap: Supervised learning

Statistical learning reading group

Alexander Ly

Psychological Methods
University of Amsterdam

Amsterdam, 2 November 2016

Overview

1. Last year: Supervised learning
2. This year: More topics and applications
3. Organisation
4. Next meeting

Problem statement: Supervised learning

Based on n pairs of data $\binom{x_1}{y_1}, \ldots, \binom{x_n}{y_n}$, where x_i are features (input) and y_i are outcomes (output), predict future outcomes y_{new} given x_{new}.

Assumptions of mean regression

- There exists a true function f^* such that

$$y = f^*(x) + \epsilon \qquad (1)$$

  where ϵ is an error and $E(\epsilon) = 0$.
- Goal 1: Give a "best guess" $\hat{f}(x)$ of the unknown true f^*.
- Goal 2: Based on the best guess $\hat{f}(x)$, estimate a prediction error.

Regression

- There exists a true function f^* such that $y = f^*(x) + \epsilon$. Goal: Give a *single* best guess $\hat{f}(x)$ of $f^*(x)$ based on finite samples $\binom{x_1}{y_1}, \ldots, \binom{x_n}{y_n}$.
- Step 1: Define "best guess" aka define a loss function

$$E(f^*(x) - \hat{f}(x))^2 \tag{2}$$

- Step 2: Define a candidate collection of functions $\mathcal{F}$
- Step 3: Calculate the (empirical) loss for each single candidate $\tilde{f}$ in $\mathcal{F}$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{3}$$

- Step 4: Minimise: Take as best guess:

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\operatorname{argmin}}\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{4}$$

Example of $\mathcal{F}_2$: Linear regression

- Candidate set

$$\mathcal{F}_2 := \left\{ f(x) = \theta_0 + \theta_1 x \,|\, \theta_0, \theta_1 \in \mathbb{R} \right\} \tag{5}$$

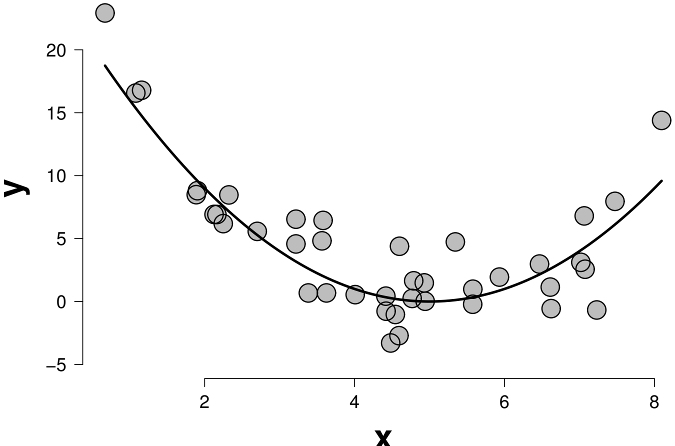
- Trick: Rephrase in terms of matrix algebra:

$$y = X\theta + \epsilon \tag{6}$$

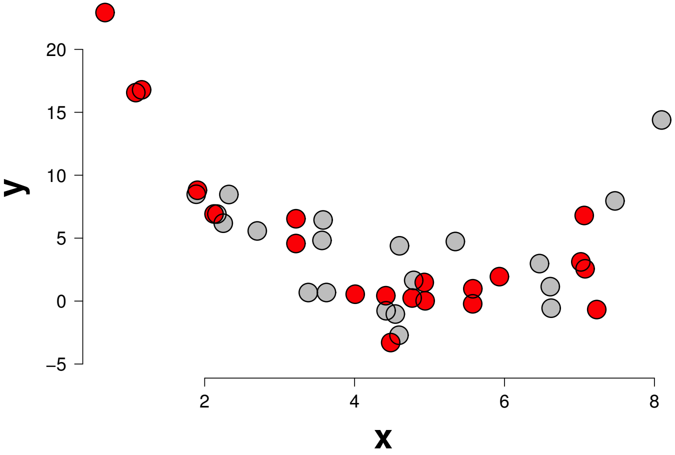
  observed $y \in \mathbb{R}^n$, observed design matrix $X \in \mathbb{R}^{n \times 2}$, parameters $\theta \in \mathbb{R}^2$
- Pro:
  - Computationally: No need to calculate the loss for each $f \in \mathcal{F}$. Solve by matrix algebra $\hat{\theta} = (X^T X)^{-1} X^T y$
  - Unique minimiser: is the plugin $\hat{f}(x_{\text{new}}) = \hat{\theta} x_{\text{new}}$
- Con:
  - Misspecification The true f^* is most likely not linear, thus, $f^* \notin \mathcal{F}_2$

Example: Data generated from $f^*(x) = (x-5)^2 + \epsilon$

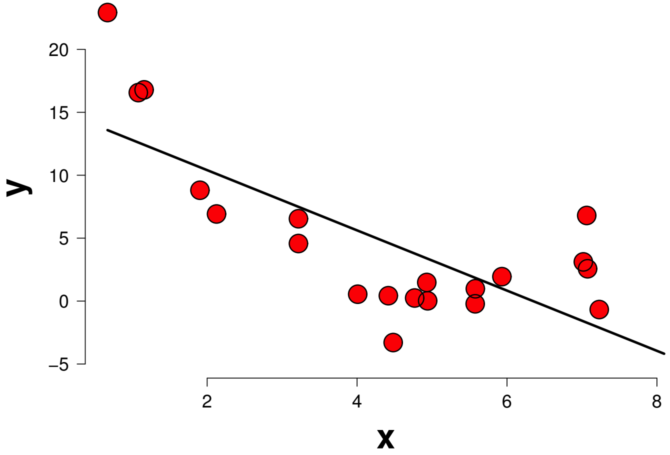
True data generating $f^*(x) = (x-5)^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 2^2)$



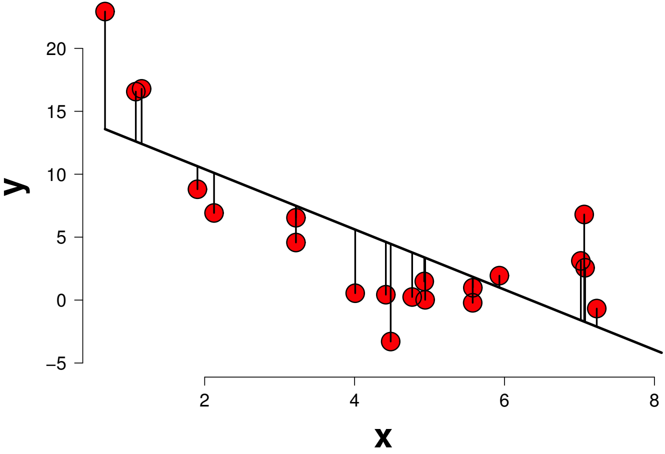
Sample splitting: Training set vs test set



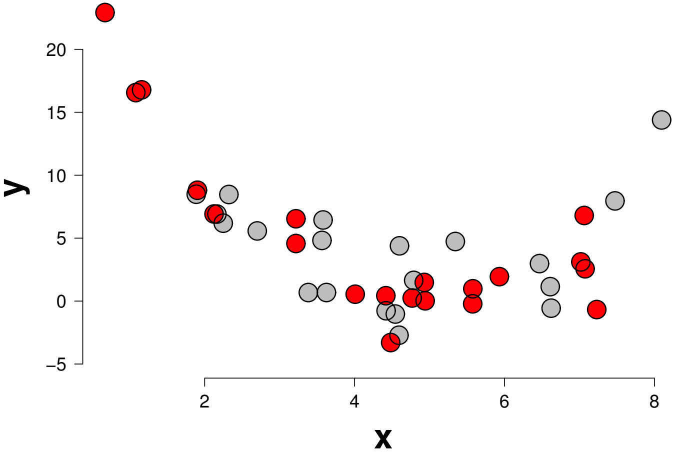
Learning on training set: Estimate the best fit



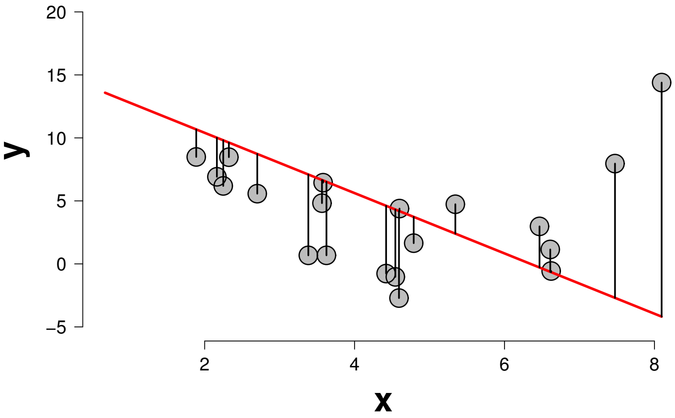
Learning on training set: In-sample error: 4.45



Prediction based on test set



Generalisation: Out-of-sample error: 5.93

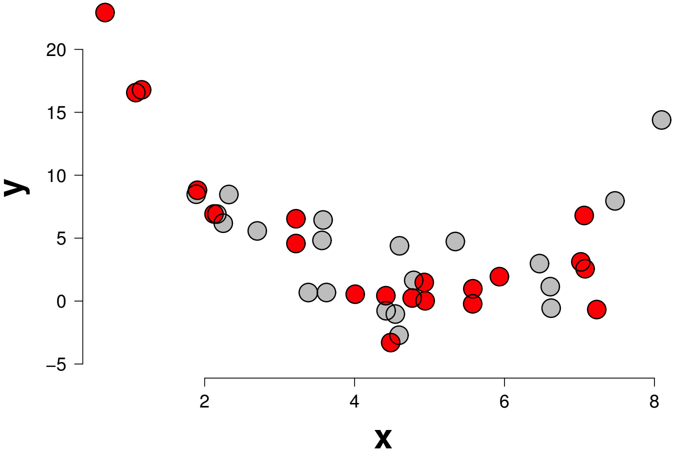


Underfitting

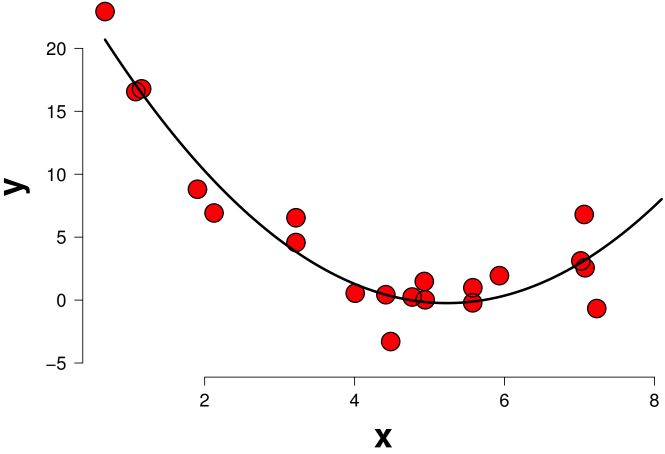
In this case, we know

- True $f^*(x) = (x-5)^2 + \epsilon$
- Thus, misspecification $f^* \notin \mathcal{F}_2 := \left\{ f(x) = \theta_0 + \theta_1 x \,|\, \theta \in \mathbb{R}^2 \right\}$.
- In fact, underfitting
- Why not take a bigger set candidate collection $\mathcal{F}$?
- Try $\mathcal{F}_3 := \left\{ f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \,|\, \theta \in \mathbb{R}^3 \right\}$

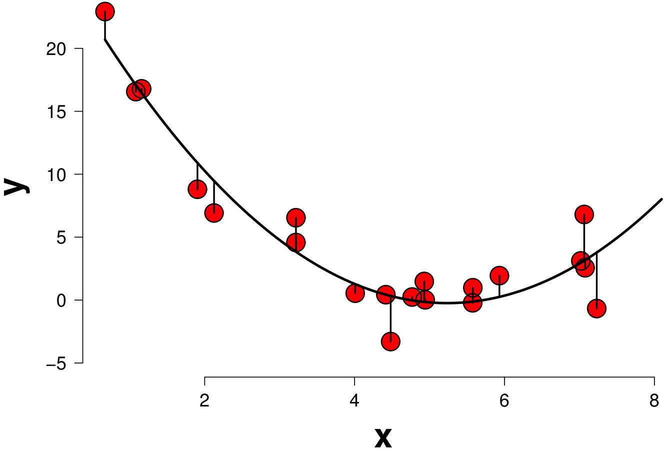
Sample splitting: Training set vs test set



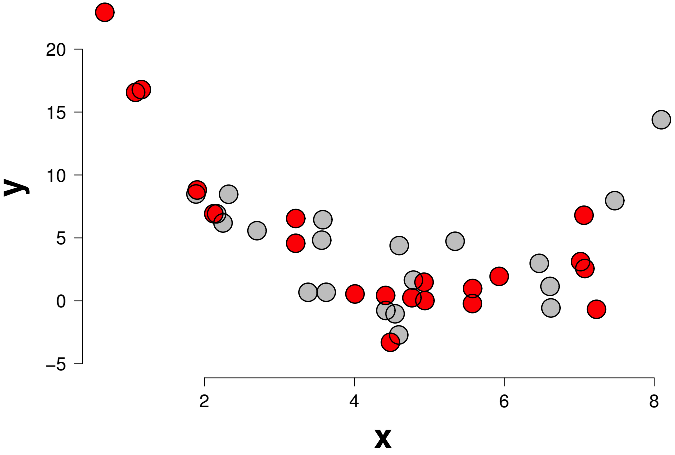
Learning on training set: Estimate the best fit



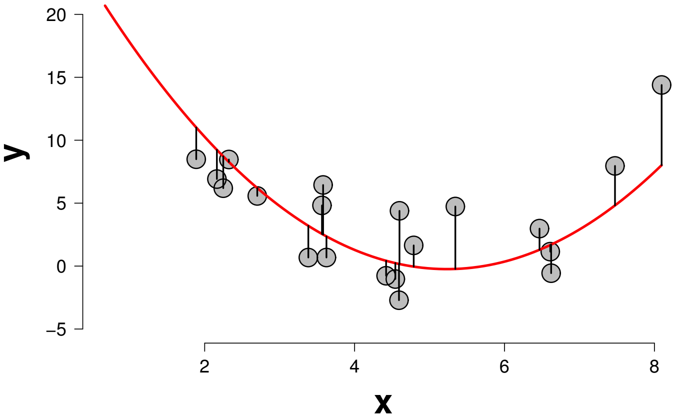
Learning on training set: In-sample error: 1.98



Prediction based on test set



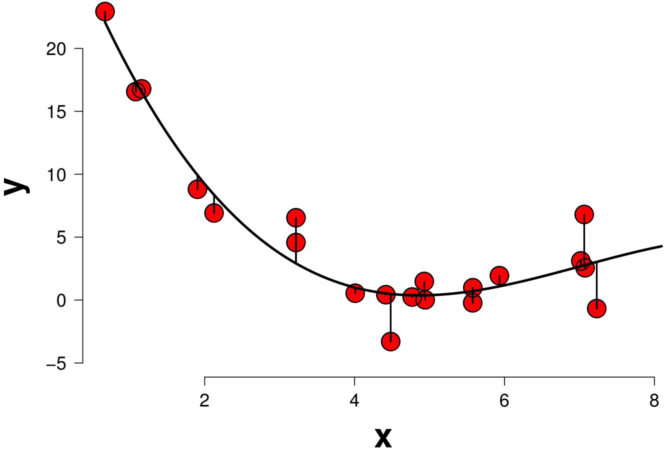
Generalisation: Out-of-sample error: 2.86



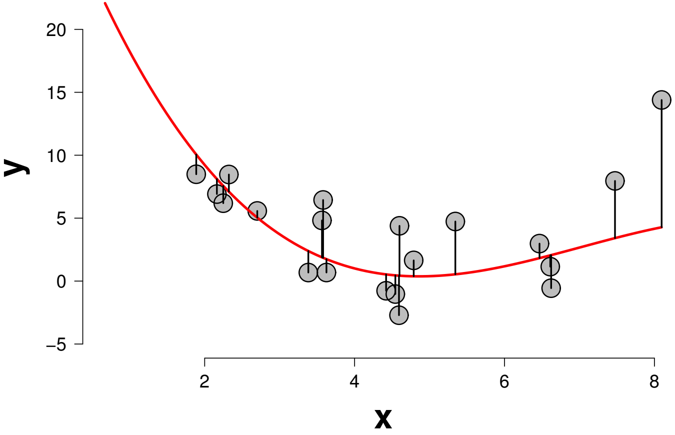
Reality

- We do not know the true f^*
- Why not try to use
  $\mathcal{F}_p := \left\{ \theta_0 + \theta_1 x^1 + \ldots + \theta_{p-1} x^{p-1} \mid \theta \in \mathbb{R}^p \right\}$
- Overfitting :(

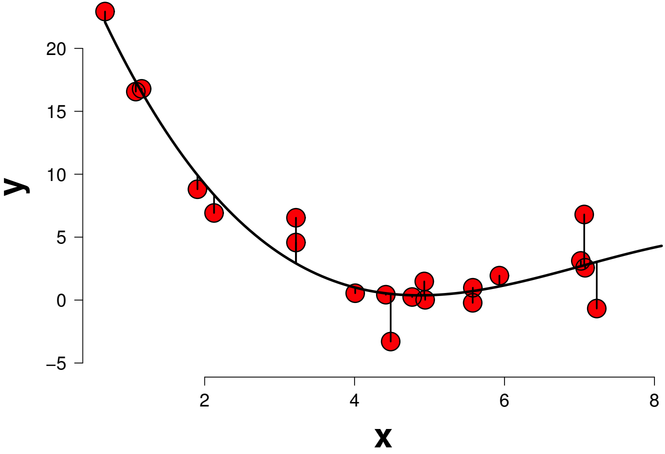
Learning on training set $p = 4$: In-sample error: 1.85



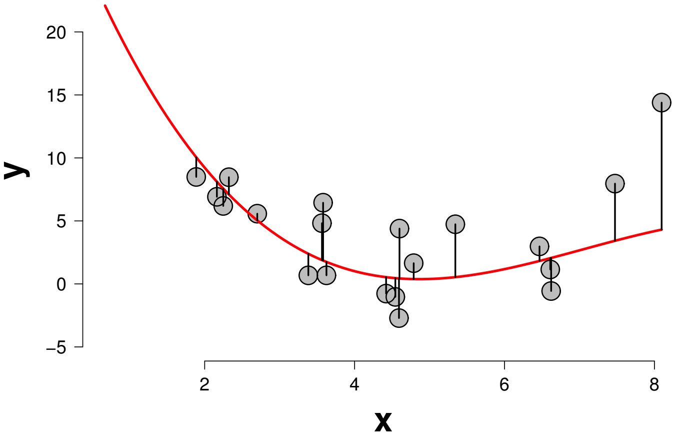
Generalisation: Out-of-sample error $p = 4$: 3.33



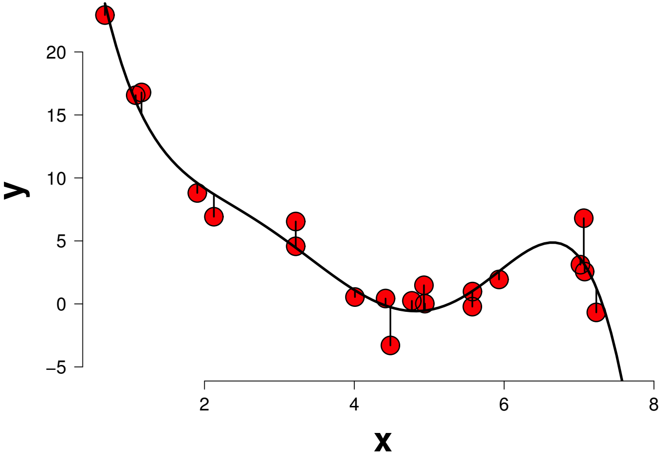
Learning on training set $p = 5$: In-sample error: 1.85



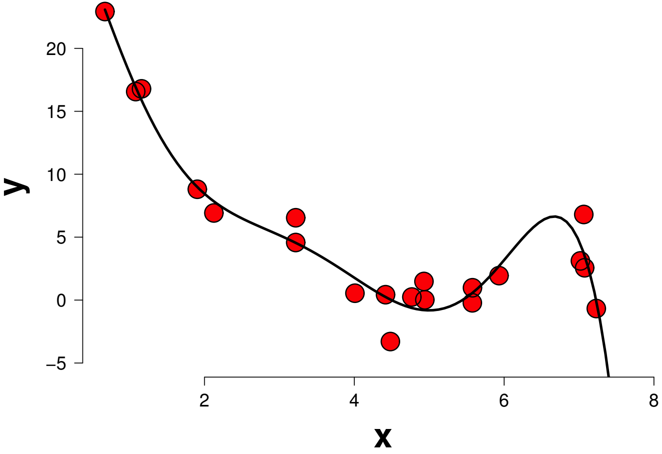
Generalisation: Out-of-sample error $p = 5$: 3.29



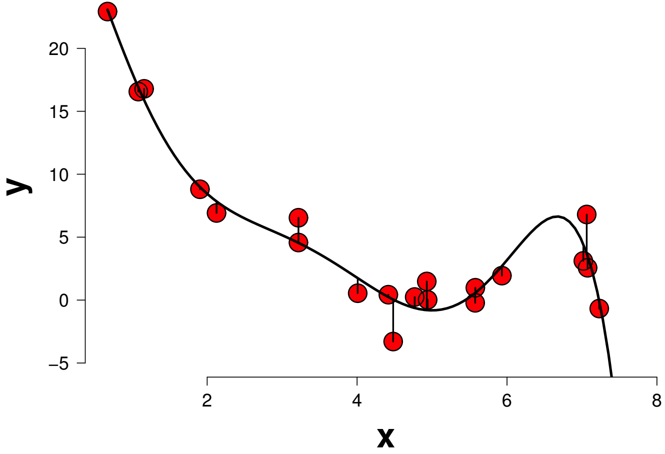
Learning on training set $p = 6$: In-sample error: 1.52



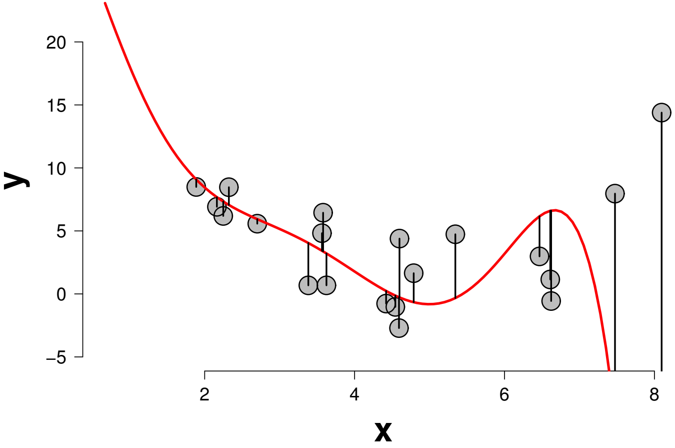
Generalisation: Out-of-sample error $p = 6$: 10.61



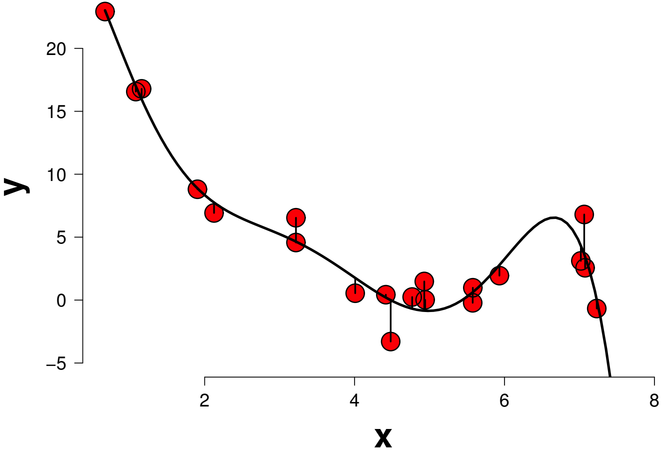
Learning on training set $p = 7$: In-sample error: 1.39



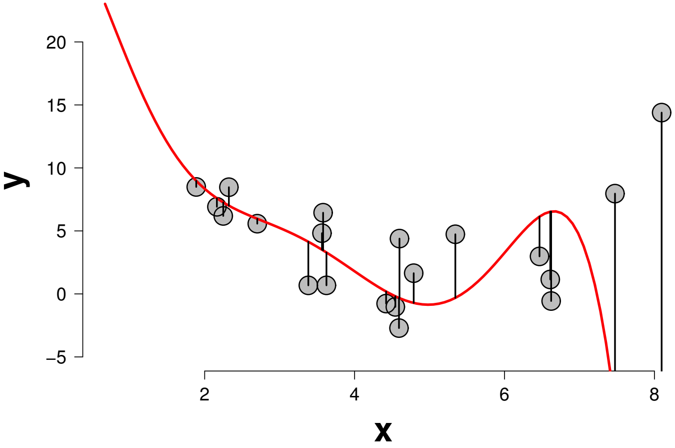
Generalisation: Out-of-sample error $p = 7$: 19.93



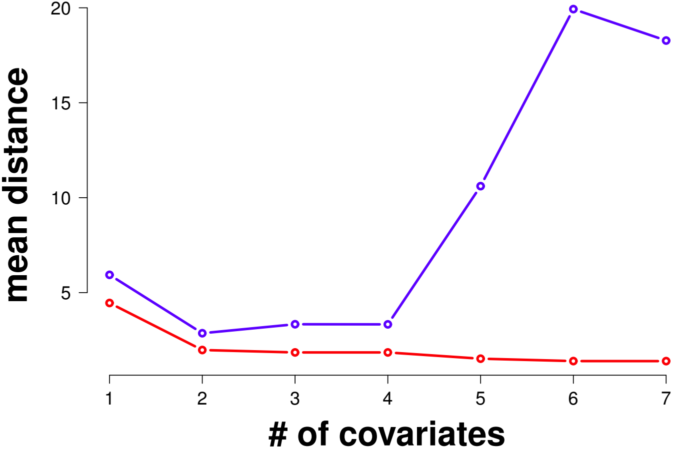
Learning on training set $p = 8$: In-sample error: 1.39



Generalisation: Out-of-sample error $p = 8$: 18.28



In-sample vs out-of-sample error



Summary

- Within sample error vs out-of-sample error
- Generalisation based on a point estimate: best guess
- Uncertainty quantification (within model): Bootstrapping (Quentin)
- Goal 2: Give an estimate prediction error use three-way splitting. Training set, cross validation set, test set. Cross validation (Johnny)
- Small to big model

How far can we go?

- The best guess was based on linear algebra: Write the problem in terms of

$$y = X\theta + \epsilon, \tag{7}$$

  note: resulting best guess $\hat{f}(x)$ was polynomial, though, trick is to write it as a linear combination of basis functions: $x^1, x^2, \ldots, x^p$.
- Minimiser is given by

$$\theta = (X^T X)^{-1} X^T y \tag{8}$$

  where $X \in \mathbb{R}^{n \times p}$.
- Problem $n \ll p$: $X^T X$ is not invertible.

The $n \ll p$ regime: $X^T X$ is not invertible

- This means that there is not a unique solution: If θ_0 is such that $y = X\theta_0 + \epsilon$ then also $y = X(\theta_0 + u) + \epsilon$.
- Solution: Choose amongst all solutions, choose θ "small" by adding a penalty.

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_q \right\} \tag{9}$$

  Lasso: $q = 1$, Ridge: $q = 2$, Elastic net: combination of both.
- Regularisation: We add (Lagrange multiplier) constraints to make the singular matrix regular.

Same theme: Cross-validation

- Solution: Penalised regression

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_q \right\} \tag{10}$$

- Candidate set: $\mathcal{F}_{\lambda,q} = \left\{ f(x) = X\theta \mid \|\theta\|_q = \lambda \right\}$
- Choose λ, use cross validation, see Johnny
- Different forms with $q = 1$, see Lourens.

Regression

- Goal: Give a *single* best guess $\hat{f}(x)$ of $f^*(x)$. Step 1: Define "best guess" aka define a loss function

$$E(f^*(x) - \hat{f}(x))^2 \tag{11}$$

- Step 2: Define a candidate collection of functions $\mathcal{F}$
- Step 3a: Find a good basis for $\mathcal{F}$ use linear algebra trick to find the minimium
- Step 4a: Grow model $\mathcal{F}_{p,q,\lambda}$
- Step 5a: Regularise and cross validate

Regression problem: Y is continuous

- Different types of $\mathcal{F}$s:
  - Linear regression: Tahira
  - Polynomial, local, regression, splines: Alexander
- Choosing basis:
  - Smoothing splines, Gaussian process priors: Alexander
  - Wavelets
  - Gabor patches: Gilles
  - Reproducing Hilbert space kernels
  - Neural networks: Joost (composition basis)
- Multivariate X:
  - Generalised additive models (GAMs)
  - Regression trees: Riet
  - Gaussian graphical models: Lourens
  - B-splines

Classification problems: Y is discrete

- Different types of $\mathcal{F}$s:
  - Logistic regression: Lourens
- Multivariate X:
  - Classification trees: Riet
  - Support vector machines: Udo
  - k-nearest neighbours
  - random forests

Other topics

- Unsupervised learning:
  - Principle component analysis, independent component analysis
  - Clustering: Jonas
  - Topic modelling: Claire/Quentin
  - Mixture models and the EM algorithm
- Reinforcement learning:
  - Act-R: Leendert
  - Decision and game theory: Udo
  - Learning automata: Lourens
  - Bandit problems
  - Markov decision processes
- Internet stuff:
  - Page ranking: Johnny
  - Recommender systems: Don

Other topics

- State-space models:
  - Time series: Sacha
  - Kalman filters
  - Hidden markov models: Ingmar
  - Bayesian time series
  - Kriging: Gaussian processes
- Techniques
  - Boosting
  - Bayesian model averaging
  - More lassos
  - Stochastic gradient descent
  - Wavelets
- Applications:
  - Neural networks for psychological data
  - ...

Set-up

- Same set-up as last year. Do a presentation
- Requires a small peak in preparation of a talk
- Not necessary to understand everything. Can be practical and theoretical
- Still good to read things in advanced. Also website with literature, youtube clips are available.

Clustering

- Wednesday 23rd of November: 12.00 - 13.00
- Wednesday 23rd of November: 13.00 - 14.00
- ???