Motivation for splines

Statistical learning reading group

Alexander Ly

Psychological Methods
University of Amsterdam

Amsterdam, 23 February 2016

Overview

The regression problem

- The regression assumption: There exists a true function f^* such that $y = f^*(x) + \epsilon$. Give a *single* best guess $\hat{f}(x)$ of $f^*(x)$ based on finite samples $\binom{x_1}{y_1}, \ldots, \binom{x_n}{y_n}$.
- Linear regression: The best guess is of the form $\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$.
- Goal: Choose $\hat{f}$ that yields good predictions.

The regression problem

- The regression assumption: There exists a true function f^* such that $y = f^*(x) + \epsilon$. Give a *single* best guess $\hat{f}(x)$ of $f^*(x)$ based on finite samples $\binom{x_1}{y_1}, \dots, \binom{x_n}{y_n}$.
- Linear regression: The best guess is of the form $\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$.
- Goal: Choose $\hat{f}$ that yields good predictions.
- Give a definition of good predictions: Here, minimise squared error loss.

The regression problem

- The regression assumption: There exists a true function f^* such that $y = f^*(x) + \epsilon$. Give a *single* best guess $\hat{f}(x)$ of $f^*(x)$ based on finite samples $\binom{x_1}{y_1}, \ldots, \binom{x_n}{y_n}$.
- Linear regression: The best guess is of the form $\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$.
- Goal: Choose $\hat{f}$ that yields good predictions.
- Give a definition of good predictions: Here, minimise squared error loss.
- For each estimate $\hat{f}(x)$ the *risk* (expected mean squared error loss) is given by

$$E(f^*(x) - \hat{f}(x))^2 = \mathrm{Bias}^2(\hat{f}(x)) + \mathrm{Var}(\hat{f}) \tag{1}$$

(Population) average mean squared error wrt *any* x, thus, also not observed ones.

True risk versus empirical risk

- For each estimate $\hat{f}(x)$ the *risk* (expected mean squared error loss) is given by

$$E(f^*(x) - \hat{f}(x))^2 = \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}) \qquad (2)$$

(Population) average mean squared error wrt *any* x, thus, also not observed ones.

True risk versus empirical risk

- For each estimate $\hat{f}(x)$ the *risk* (expected mean squared error loss) is given by

$$E(f^*(x) - \hat{f}(x))^2 = \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}) \tag{2}$$

(Population) average mean squared error wrt *any* x, thus, also not observed ones.

- For any x, plugin $f^*(x) = y$. Linear regression: $\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$

$$E(y - \hat{\theta}_0 - \hat{\theta}_1 x)^2 \tag{3}$$

Problem: cannot evaluate this population mean, thus, find $\hat{\theta}$ for which this is smallest.

True risk versus empirical risk

- For each estimate $\hat{f}(x)$ the *risk* (expected mean squared error loss) is given by

$$E(f^*(x) - \hat{f}(x))^2 = \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}) \tag{2}$$

  (Population) average mean squared error wrt *any* x, thus, also not observed ones.
- For any x, plugin $f^*(x) = y$. Linear regression: $\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$

$$E(y - \hat{\theta}_0 - \hat{\theta}_1 x)^2 \tag{3}$$

  Problem: cannot evaluate this population mean, thus, find $\hat{\theta}$ for which this is smallest.
- Replace: Population mean by sample mean

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)^2 \tag{4}$$

The regression estimation procedure

- Define a prediction criterion at the population level:

$$E(y - \hat{f}(x))^2, \tag{5}$$

This risk is unknown in practice.

The regression estimation procedure

- Define a prediction criterion at the population level:
$$E(y - \hat{f}(x))^2, \tag{5}$$
This risk is unknown in practice.
- For each candidate $\tilde{f}$ from a collection of $\mathcal{F}$ approximation this risk by its empirical version
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2, \tag{6}$$

The regression estimation procedure

- Define a prediction criterion at the population level:

$$E(y - \hat{f}(x))^2, \tag{5}$$

This risk is unknown in practice.

- For each candidate $\tilde{f}$ from a collection of $\mathcal{F}$ approximation this risk by its empirical version

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2, \tag{6}$$

- Define the best guess $\hat{f}$ (i.e., point estimate of f^*) as

$$\hat{f}(x) = \operatorname*{argmin}_{\tilde{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{7}$$

where the minimisation is over the collection $\mathcal{F}$.

Beyond linearity = Changing the collection $\mathcal{F}$

- Define the best guess $\hat{f}$ (i.e., point estimate of f^*) as

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{8}$$

  where the minimisation is over the collection $\mathcal{F}$.
- Example: linear regression: the collection of functions is given by $\mathcal{F} := \{f : X \to Y \mid f(x) = \theta_0 + \theta_1 x\}$

Beyond linearity = Changing the collection $\mathcal{F}$

- Define the best guess $\hat{f}$ (i.e., point estimate of f^*) as

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{8}$$

  where the minimisation is over the collection $\mathcal{F}$.
- Example: linear regression: the collection of functions is given by $\mathcal{F} := \{f : X \to Y \mid f(x) = \theta_0 + \theta_1 x\}$
- Later: The collection $\mathcal{F}$ consists of polynomials,

Beyond linearity = Changing the collection $\mathcal{F}$

- Define the best guess $\hat{f}$ (i.e., point estimate of f^*) as

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{8}$$

  where the minimisation is over the collection $\mathcal{F}$.
- Example: linear regression: the collection of functions is given by $\mathcal{F} := \{f : X \to Y \mid f(x) = \theta_0 + \theta_1 x\}$
- Later: The collection $\mathcal{F}$ consists of polynomials,
- Later: The collection $\mathcal{F}$ consists of natural/regression splines,

Beyond linearity = Changing the collection $\mathcal{F}$

- Define the best guess $\hat{f}$ (i.e., point estimate of f^*) as

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{8}$$

  where the minimisation is over the collection $\mathcal{F}$.
- Example: linear regression: the collection of functions is given by $\mathcal{F} := \{f : X \to Y \mid f(x) = \theta_0 + \theta_1 x\}$
- Later: The collection $\mathcal{F}$ consists of polynomials,
- Later: The collection $\mathcal{F}$ consists of natural/regression splines,
- Later: The collection $\mathcal{F}$ consists of smoothing splines

Beyond linearity = Changing the collection $\mathcal{F}$

- Define the best guess $\hat{f}$ (i.e., point estimate of f^*) as

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{8}$$

  where the minimisation is over the collection $\mathcal{F}$.
- Example: linear regression: the collection of functions is given by $\mathcal{F} := \{f : X \to Y \mid f(x) = \theta_0 + \theta_1 x\}$
- Later: The collection $\mathcal{F}$ consists of polynomials,
- Later: The collection $\mathcal{F}$ consists of natural/regression splines,
- Later: The collection $\mathcal{F}$ consists of smoothing splines
- Later: The collection $\mathcal{F}$ consists of classification and regression trees, neural networks, support vector machines, etc, etc

Reasons to broaden the collection $\mathcal{F}$

- Misspecification: linear regression works best if $f^*(x)$ is indeed linear.
- More general, $\mathcal{F}$ works best if true $f^* \in \mathcal{F}$, but this is not known in practice.

Reasons to broaden the collection $\mathcal{F}$

- Q: Why not choose $\mathcal{F}$ just the collection of all possible functions?

Reasons to broaden the collection $\mathcal{F}$

- Q: Why not choose $\mathcal{F}$ just the collection of all possible functions?
- Computationally intensive. Minimisation is then over an uncountable set of functions.

Reasons to broaden the collection $\mathcal{F}$

- Q: Why not choose $\mathcal{F}$ just the collection of all possible functions?
- Computationally intensive. Minimisation is then over an uncountable set of functions.
- No unique minimiser

Reasons to broaden the collection $\mathcal{F}$

- Q: Why not choose $\mathcal{F}$ just the collection of all possible functions?
- Computationally intensive. Minimisation is then over an uncountable set of functions.
- No unique minimiser
- Overfitting: *If* we can evaluate the (theoretical) average

$$E(y - \tilde{f}(x))^2 = \int (y - \tilde{f}(x))^2 dx \tag{9}$$

for every possible function $\tilde{f}$ at "every" (uncountably many) x, thus, also $y = f^*(x)$. The minimiser over all possible functions is then $f^*(x)$.

Reasons to broaden the collection $\mathcal{F}$

- Q: Why not choose $\mathcal{F}$ just the collection of all possible functions?
- Computationally intensive. Minimisation is then over an uncountable set of functions.
- No unique minimiser
- Overfitting: *If* we can evaluate the (theoretical) average

$$E(y - \tilde{f}(x))^2 = \int (y - \tilde{f}(x))^2 \mathrm{d}x \tag{9}$$

  for every possible function $\tilde{f}$ at "every" (uncountably many) x, thus, also $y = f^*(x)$. The minimiser over all possible functions is then $f^*(x)$.
- Overfitting comes from not being able to evaluate Eq. (9).

How to broaden $\mathcal{F}$:

- Make it bigger, lower misspecification, but not let the variance grow too much

How to broaden $\mathcal{F}$:

- Make it bigger, lower misspecification, but not let the variance grow too much
- 1. Practical: Retain computational efficiency.

How to broaden $\mathcal{F}$:

- Make it bigger, lower misspecification, but not let the variance grow too much
- 1. Practical: Retain computational efficiency.
- 2.a. Have $\mathcal{F}$ big, but have a unique minimiser

How to broaden $\mathcal{F}$:

- Make it bigger, lower misspecification, but not let the variance grow too much
- 1. Practical: Retain computational efficiency.
- 2.a. Have $\mathcal{F}$ big, but have a unique minimiser
- 2.b. In case of multiple minimisers, choose the "smallest" solution (i.e., regularisation).

Computations: The general regression solution

- In general computationally heavy: For each candidate $\tilde{f}$ from $\mathcal{F}$ approximation the risk by $\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2$, Consequently, select the minimiser:

$$\hat{f}(x) = \operatorname*{argmin}_{\tilde{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{10}$$

Computations: The general regression solution

- In general computationally heavy: For each candidate $\tilde{f}$ from $\mathcal{F}$ approximation the risk by $\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2$, Consequently, select the minimiser:

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{10}$$

- Alternatively, Linear regression: $y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n\times p}$ the ordinary least square is

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{11}$$

Computations: The general regression solution

- In general computationally heavy: For each candidate $\tilde{f}$ from $\mathcal{F}$ approximation the risk by $\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2$, Consequently, select the minimiser:

$$\hat{f}(x) = \underset{\tilde{f}\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 \tag{10}$$

- Alternatively, Linear regression: $y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n\times p}$ the ordinary least square is

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{11}$$

- The minimiser is basically derived from simple matrix algebra, i.e., $\hat{\theta} = Ay$. This idea is exploited in polynomial and regression splines models.

Uniqueness and regularisation

- Linear regression: $Y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$.

Uniqueness and regularisation

- Linear regression: $Y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$.
- Problem: when $p > n$ then also have $Y = X(\theta_{(0)} + u) + \epsilon$, where $Xu = 0$. There are many u s.t. $Xu = 0$, non-uniqueness.

Uniqueness and regularisation

- Linear regression: $Y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$.
- Problem: when $p > n$ then also have $Y = X(\theta_{(0)} + u) + \epsilon$, where $Xu = 0$. There are many u s.t. $Xu = 0$, non-uniqueness.
- Solution: Choose the solution s.t. $\theta_{(0)} + u$ is small. In other words, instead of minimising $\sum_{i=1}^n (y_i - \tilde{f}(x_i))^2$ minimise the following instead

$$\hat{f}(x) = \operatorname*{argmin}_{\tilde{f} \in \mathcal{F}} \sum_{i=1}^n (y_i - \tilde{f}(x_i))^2 + \lambda \, \text{penalty}(\tilde{f}). \tag{12}$$

for some fixed $\lambda > 0$.

Uniqueness and regularisation

- Linear regression: $Y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$.
- Problem: when $p > n$ then also have $Y = X(\theta_{(0)} + u) + \epsilon$, where $Xu = 0$. There are many u s.t. $Xu = 0$, non-uniqueness.
- Solution: Choose the solution s.t. $\theta_{(0)} + u$ is small. In other words, instead of minimising $\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2$ minimise the following instead

$$\hat{f}(x) = \underset{\tilde{f} \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \tilde{f}(x_i))^2 + \lambda \, \text{penalty}(\tilde{f}). \tag{12}$$

  for some fixed $\lambda > 0$.
- Example: Lasso/ridge/elastic nets and remarkably: smoothing splines.

Polynomial regression

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_m$ the family of order-m polynomials:

$$\mathcal{F}_m = \{f(x) = \theta_0 x^0 + \theta_1 x^1 + \ldots + \theta_{m-1} x^{m-1} = \sum_{j=0}^{m-1} \theta_j x^j\} \tag{13}$$

Polynomial regression

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_m$ the family of order-m polynomials:

$$\mathcal{F}_m = \{f(x) = \theta_0 x^0 + \theta_1 x^1 + \ldots + \theta_{m-1} x^{m-1} = \sum_{j=0}^{m-1} \theta_j x^j\} \tag{13}$$

- Example: $m = 2$: linear regression: $\{f(x) = \theta_0 + \theta_1 x\}$

Polynomial regression

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_m$ the family of order-m polynomials:

$$\mathcal{F}_m = \{f(x) = \theta_0 x^0 + \theta_1 x^1 + \ldots + \theta_{m-1} x^{m-1} = \sum_{j=0}^{m-1} \theta_j x^j\} \tag{13}$$

- Example: $m = 2$: linear regression: $\{f(x) = \theta_0 + \theta_1 x\}$
- Idea: Let the order m grow.

Polynomial regression

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_m$ the family of order-m polynomials:

$$\mathcal{F}_m = \{f(x) = \theta_0 x^0 + \theta_1 x^1 + \ldots + \theta_{m-1} x^{m-1} = \sum_{j=0}^{m-1} \theta_j x^j\} \tag{13}$$

- Example: $m = 2$: linear regression: $\{f(x) = \theta_0 + \theta_1 x\}$
- Idea: Let the order m grow.
- Problem: overfitting. Q: How far can we go? How bad is this problem?

Computationally: Exploit matrix algebra

- Given a chosen m: Frame the problem as Linear regression: $y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^m$, $X \in \mathbb{R}^{n \times m}$, where

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^{m-1} \\ 1 & x_2^1 & x_2^2 & \dots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & x_n^2 & \dots & x_n^{m-1} \end{pmatrix}. \tag{14}$$

Computationally: Exploit matrix algebra

- Given a chosen m: Frame the problem as Linear regression: $y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^m$, $X \in \mathbb{R}^{n \times m}$, where

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^{m-1} \\ 1 & x_2^1 & x_2^2 & \dots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & x_n^2 & \dots & x_n^{m-1} \end{pmatrix}. \tag{14}$$

- Again, the ordinary least square is

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{15}$$

exploit this linear structure (matrix algebra). If $m < n$ there is a unique solution (thus, minimiser).

Computationally: Exploit matrix algebra

- Given a chosen m: Frame the problem as Linear regression: $y = X\theta + \epsilon$ with $y \in \mathbb{R}^n, \theta \in \mathbb{R}^m$, $X \in \mathbb{R}^{n \times m}$, where

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^{m-1} \\ 1 & x_2^1 & x_2^2 & \dots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & x_n^2 & \dots & x_n^{m-1} \end{pmatrix}. \tag{14}$$

- Again, the ordinary least square is

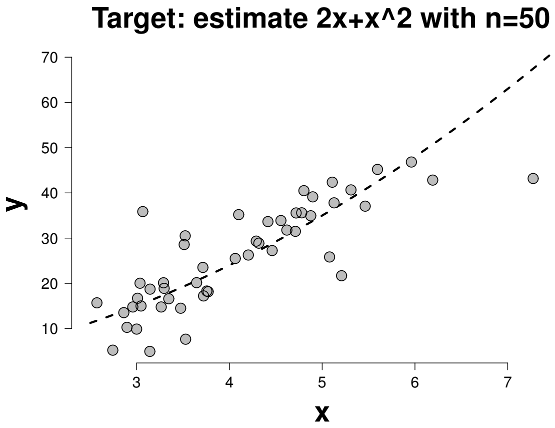
$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{15}$$

exploit this linear structure (matrix algebra). If $m < n$ there is a unique solution (thus, minimiser).

- Of course, how to choose the additional parameter m? Cross validation, etc etc.

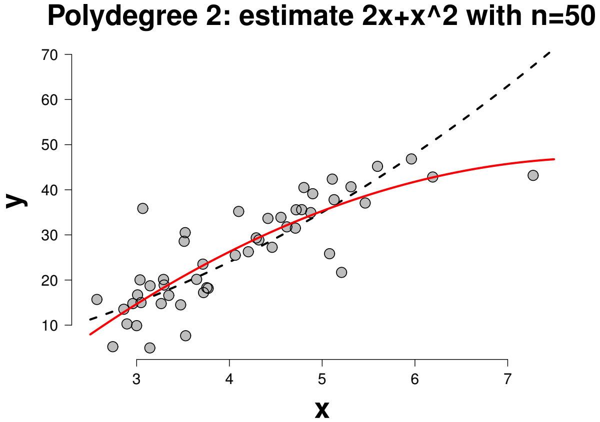
Growing model and how bad is bad?

Suppose the true is $f^*(x) = 2x + x^2$. Data sampled as $Y = f^*(x) + \epsilon$.



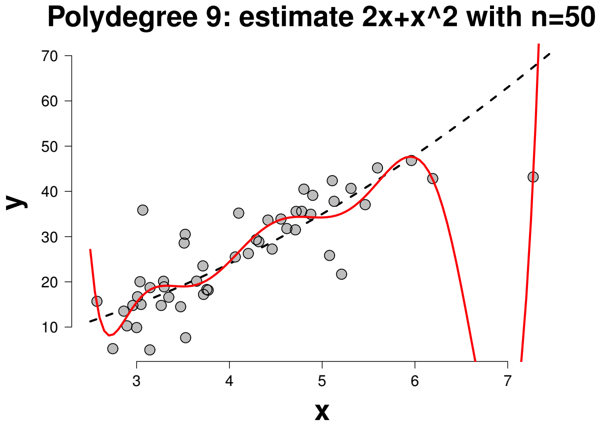
Well-specified, right order

Note $f^* \in \mathcal{F}_m$ when $m = 3$. Thus, well-specified:



Well-specified, order too large

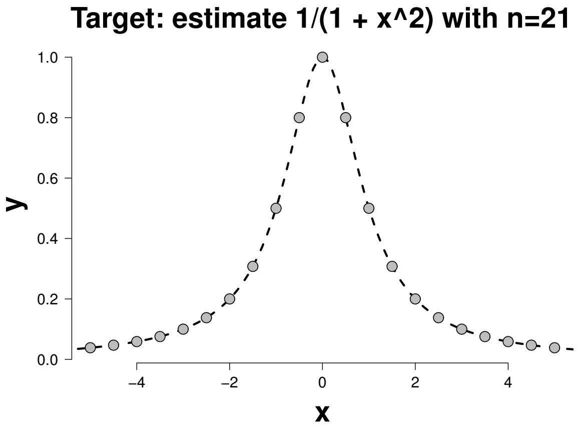
Note still $f^* \in \mathcal{F}_m$ when $m = 9$. Thus, well-specified, but $m > 3$:



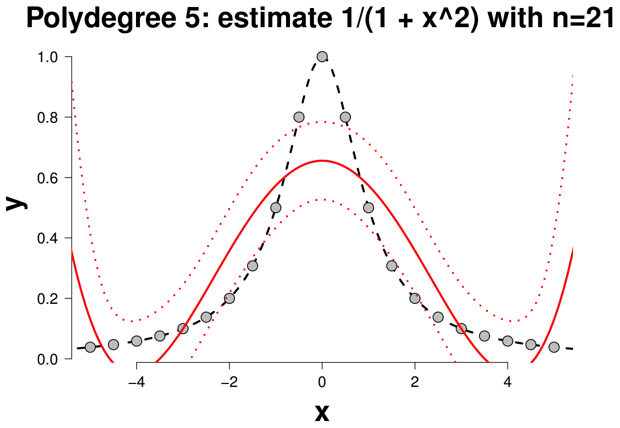
Overfit: Random error is seen as structural.

How far can we go with polynomial regression?

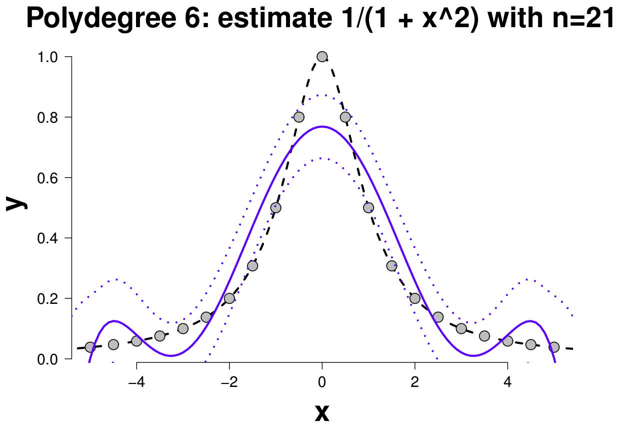
Suppose the true is $f^*(x) = 1/(1+x^2)$. Data sampled as $Y = f^*(X)$, uniform $X \in [-5, 5]$. Note: no error



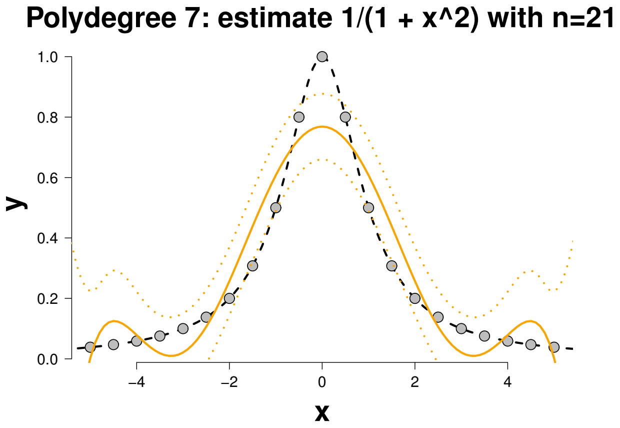
Misspecified, order?



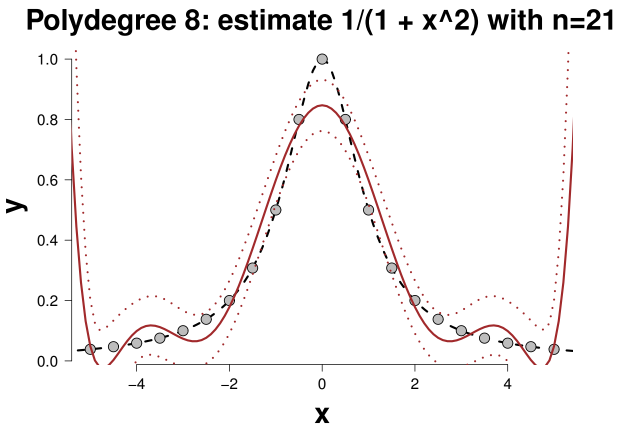
Misspecified, order?



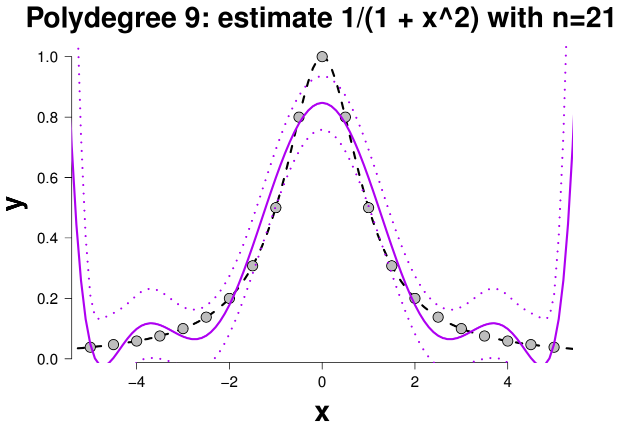
Misspecified, order?



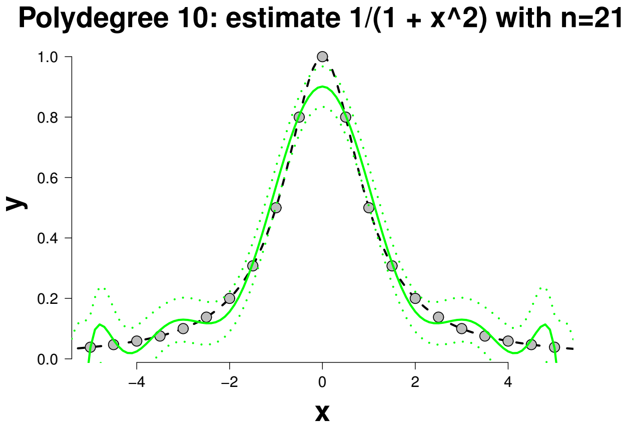
Misspecified, order?



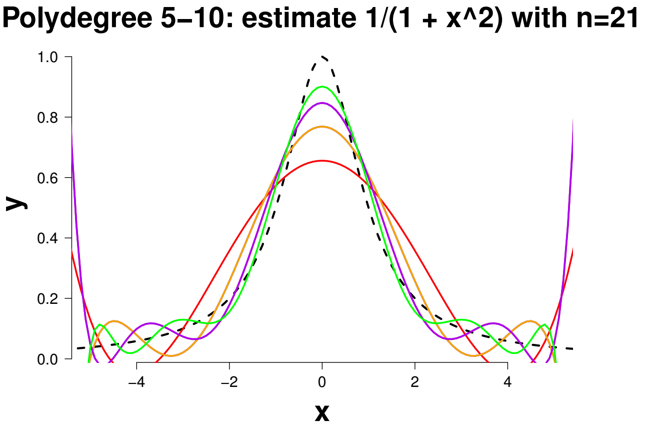
Misspecified, order?



Misspecified, order?

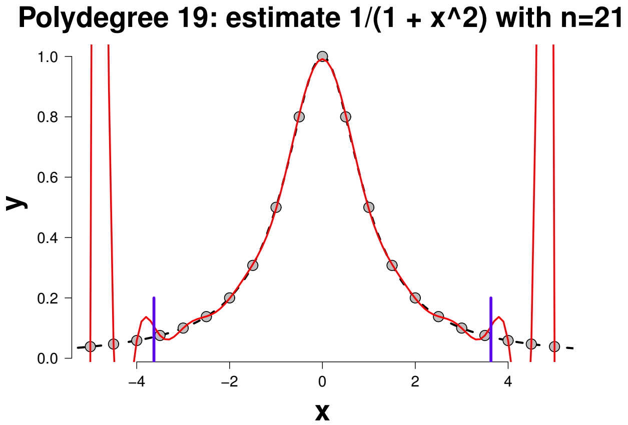


Misspecified, order?



How far can we go

Here, the OLS solution works until $m = n - 1$



Interpolation.

Conclusions

- Models have limits, even if there is no error term here.
- The target f^* is known as Runge example or Runge phenomenon for interpolation (polynomial regression with $m = n - 1$.

Conclusions

- Models have limits, even if there is no error term here.
- The target f^* is known as Runge example or Runge phenomenon for interpolation (polynomial regression with $m = n - 1$.
- Runge phenomenon: When interpolating $f^*(x$ in $[-5, 5]$ with x having equal step-size in $[-5, 5]$ *impossible* to interpolate $f^*(x)$ well within $|x| < 3.63$ and $|x| > 3.63$ at the same time.

Conclusions

- Models have limits, even if there is no error term here.
- The target f^* is known as Runge example or Runge phenomenon for interpolation (polynomial regression with $m = n - 1$.
- Runge phenomenon: When interpolating $f^*(x$ in $[-5, 5]$ with x having equal step-size in $[-5, 5]$ *impossible* to interpolate $f^*(x)$ well within $|x| < 3.63$ and $|x| > 3.63$ at the same time.
- Cautious when designing experiments with polynomial interpolation. (Chebyshev polynomials)

Conclusions

- Models have limits, even if there is no error term here.
- The target f^* is known as Runge example or Runge phenomenon for interpolation (polynomial regression with $m = n - 1$.
- Runge phenomenon: When interpolating $f^*(x$ in $[-5, 5]$ with x having equal step-size in $[-5, 5]$ *impossible* to interpolate $f^*(x)$ well within $|x| < 3.63$ and $|x| > 3.63$ at the same time.
- Cautious when designing experiments with polynomial interpolation. (Chebyshev polynomials)
- Global (over whole $[-5, 5]$) versus local fits (within $|x| < 3.63$) and at the tails.

Piecewise polynomials

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_{m,K}$ consisting of polynomials of order-m with K knots $\xi_1, \ldots, \xi_K$.

Piecewise polynomials

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_{m,K}$ consisting of polynomials of order-m with K knots $\xi_1, \ldots, \xi_K$.
- Knots split the domain in $K + 1$ area allowing for global and local behaviour.

$$\mathcal{F}_{m,k} = f(x) = \begin{cases} \sum_{j=0}^{m-1} \theta_{j,1} x^j & \text{if } x \leq \xi_1 \\ \sum_{j=0}^{m-1} \theta_{j,2} x^j & \text{if } \xi_1 < x \leq \xi_2 \\ \sum_{j=0}^{m-1} \theta_{j,k} x^j & \text{if } \xi_{k-1} < x \leq \xi_k \\ \sum_{j=0}^{m-1} \theta_{j,K} x^j & \text{if } \xi_{K-1} < x \leq \xi_K \end{cases} \quad (16)$$

Piecewise polynomials

- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_{m,K}$ consisting of polynomials of order-m with K knots $\xi_1, \dots, \xi_K$.
- Knots split the domain in $K+1$ area allowing for global and local behaviour.

$$\mathcal{F}_{m,k} = f(x) = \begin{cases} \sum_{j=0}^{m-1} \theta_{j,1} x^j & \text{if } x \leq \xi_1 \\ \sum_{j=0}^{m-1} \theta_{j,2} x^j & \text{if } \xi_1 < x \leq \xi_2 \\ \sum_{j=0}^{m-1} \theta_{j,k} x^j & \text{if } \xi_{k-1} < x \leq \xi_k \\ \sum_{j=0}^{m-1} \theta_{j,K} x^j & \text{if } \xi_{K-1} < x \leq \xi_K \end{cases} \tag{16}$$

- Note: there are $(K+1)m$ parameters

Piecewise polynomials and basis functions

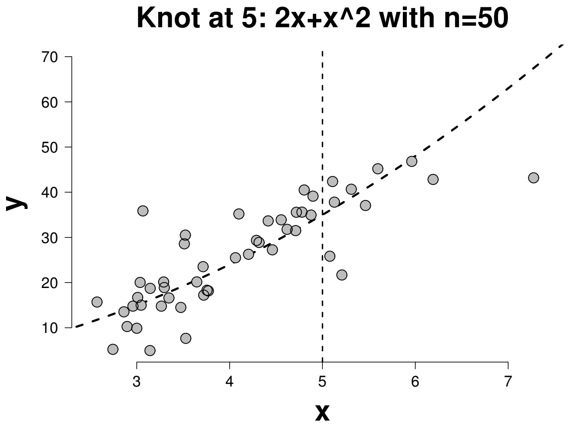
- Assumption $y = f^*(x) + \epsilon$, take the candidate collection $\mathcal{F}_{m,K}$ consisting of polynomials of order-m with K knots $\xi_1, \dots, \xi_K$.
- Knots split the domain in $K + 1$ area allowing for global and local behaviour.

$$\mathcal{F}_{m,k} = f(x) = \{ \sum_{j=0,k=1}^{m-1,K} \theta_{j,k} g_{j,k}(x) \} \tag{17}$$

  where $g_{j,k}(x) = x^j \, 1_{(\xi_{k-1},\xi_k]}(x)$.
- Note: there are $(K+1)m$ parameters

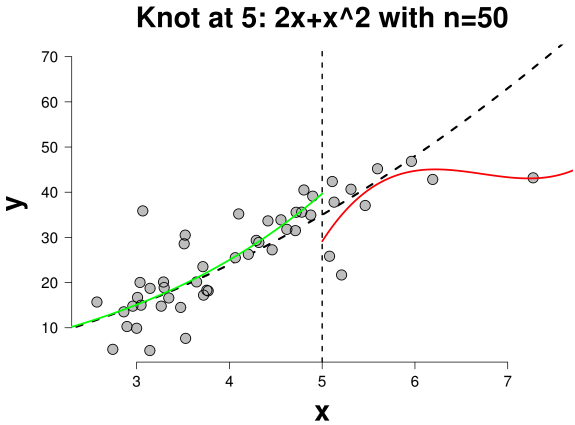
Piecewise polynomials with one knot

Knot at $\xi_1 = 5$



Piecewise cubic polynomials with one knot

Knot at $\xi_1 = 5$ and $M = 4$ on each domain.



Computationally: Exploit matrix algebra

- At each domain do linear regression
- Simpler: write it as basis functions. Recall polynomial regression with $X \in \mathbb{R}^{n \times m}$, where

$$X = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^{m-1} \\ 1 & x_2^1 & \dots & x_2^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \dots & x_n^{m-1} \end{pmatrix} = \begin{pmatrix} g_0(x_1) & g_1(x_1) & \dots & g_{m-1}(x_1) \\ g_0(x_2) & g_1(x_2) & \dots & g_{m-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_0(x_n) & g_1(x_n) & \dots & g_{m-1}(x_n) \end{pmatrix}$$

Computationally: Exploit matrix algebra

- At each domain do linear regression
- Simpler: write it as basis functions. Recall polynomial regression with $X \in \mathbb{R}^{n \times m}$, where

$$X = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^{m-1} \\ 1 & x_2^1 & \dots & x_2^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \dots & x_n^{m-1} \end{pmatrix} = \begin{pmatrix} g_0(x_1) & g_1(x_1) & \dots & g_{m-1}(x_1) \\ g_0(x_2) & g_1(x_2) & \dots & g_{m-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_0(x_n) & g_1(x_n) & \dots & g_{m-1}(x_n) \end{pmatrix}$$

- Again, the ordinary least square is

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{18}$$

exploit this linear structure (matrix algebra). If $m < n$ there is a unique solution (thus, minimiser).

Computationally: Exploit matrix algebra

- At each domain do linear regression
- Simpler: write it as basis functions. Recall polynomial regression with $X \in \mathbb{R}^{n \times m}$, where

$$X = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^{m-1} \\ 1 & x_2^1 & \dots & x_2^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \dots & x_n^{m-1} \end{pmatrix} = \begin{pmatrix} g_0(x_1) & g_1(x_1) & \dots & g_{m-1}(x_1) \\ g_0(x_2) & g_1(x_2) & \dots & g_{m-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_0(x_n) & g_1(x_n) & \dots & g_{m-1}(x_n) \end{pmatrix}$$

- Again, the ordinary least square is

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{18}$$

  exploit this linear structure (matrix algebra). If $m < n$ there is a unique solution (thus, minimiser).
- Of course, how to choose the additional parameter m and K and where?

How far can we go with piecewise polynomials?

Note with $M = 1$ and some number K that this method leads to functions that look like histograms with the height of the bar given by the empirical mean of the samples in each domain. As K increases, say, K is the number of elements in the domain, you get the space of all functions.

Next time

- Choosing knots and "continuous" piecewise regressions: splines
- Parameters and smoothing
- Smoothing and degrees of freedom
- Reproducing kernel Hilbert spaces.

References

- Bousquet, O, Boucheron, S, Lugosi, G (2004). Introduction to statistical learning theory
- Epperson, JF (1987). On the Runge example
- de Boor, C (2001). A practical guide to splines