Non- and semiparametric statistics

Stefan Lang

December 2015

Contents

Introduction

Parametric versus nonparametric inference

Parametric versus nonparametric inference

Example: parametric density estimation

- Assume that the distribution family, e.g. normal distribution, is known.
- Only a few parameters, e.g. the mean and/or the variance, are unknown, i.e.

$$f(x) \in \{f(x \mid \theta), \theta \subset \mathbb{R}^p\}.$$

- Once the unknown parameters θ are estimated through $\hat{\theta}$, the density f is fully specified.
- Advantage: estimators are often unbiased and efficient.
- Disadvantage: distribution family must be known in advance.

Parametric versus nonparametric inference

Example: nonparametric density estimation

- Do not assume a particular distribution family.
- Weak assumptions, e.g. assume only smoothness of the density.
- Advantage: very flexible.
- Disadvantage: loss of statistical efficiency compared to parametric density estimation.
- Examples: histogram or kernel density estimator.

Introduction

Nonparametric density estimators

The Histogram

- Divide the range of values starting from the origin x_0 (e.g. $x_0 = 0, x_0 = x_{min} = x_{(1)}$) in intervals (so called bins) of equal length h (so called binwidth). For the j-th bin we have

$$B_j := [x_0 + (j-1)h, x_0 + jh]$$

and

$$P(X \in B_j) = \int_{x_0+(j-1)h}^{x_0+jh} f(x)\,dx.$$

- An estimator for $P(X \in B_j)$ is the relative frequency of x_i's within B_j, i.e.

$$\widehat{P(X \in B_j)} = \frac{1}{n}(\#x_i \; in \; B_j) = \frac{1}{n}\sum_{i=1}^{n} I_{B_j}(x_i).$$

The Histogram

- For continuous f the mean value theorem for integrals yields

$$\int_{x_0+(j-1)h}^{x_0+jh} f(x)dx = f(\xi) \cdot h$$

for $\xi \in B_j$.

- Approximating f in B_j by a constant value, we obtain

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{B_j}(x_i),$$

for $x \in B_j$.

The Histogram

Definition histogram

Let $x_1, \ldots, x_n$ be an i.i.d. sample of a continuous random variable X with density f. Then the estimator

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j \in Z} I_{B_j}(x_i) I_{B_j}(x)$$

is called histogram with binwidth $h > 0$ and origin x_0.

The Histogram

Advantages of the histogram

- Can be easily computed and presented.
- Available in every statistical package (even in excel).

The Histogram

Disadvantages of the histogram

- Discontinuous estimator for a continuous density.
- Graphical representation is dependent on the origin x_0.
- There are situations, where $\hat{f}_h(x)$ depends more on observations that are far away from x than those that are close, compare the following figure.

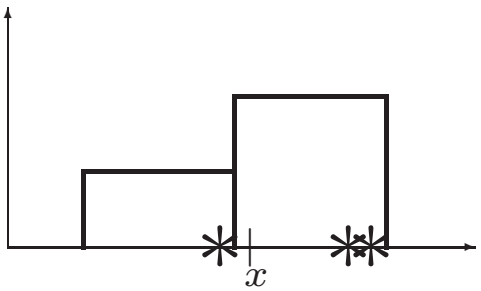


Figure 1: *Illustration of the dependence of the histogram on observations that are far away.*

The histogram

Influence of the bandwidth

$h \to 0$	needleplot
h small	quite wiggly fit
h large	smooth fit
$h \to \infty$	uniform distribution

In many statistical packages the *number of intervals* is specified rather than the binwidth h. The number of intervals induces a certain binwidth.

The histogram

Example: Munich rent index

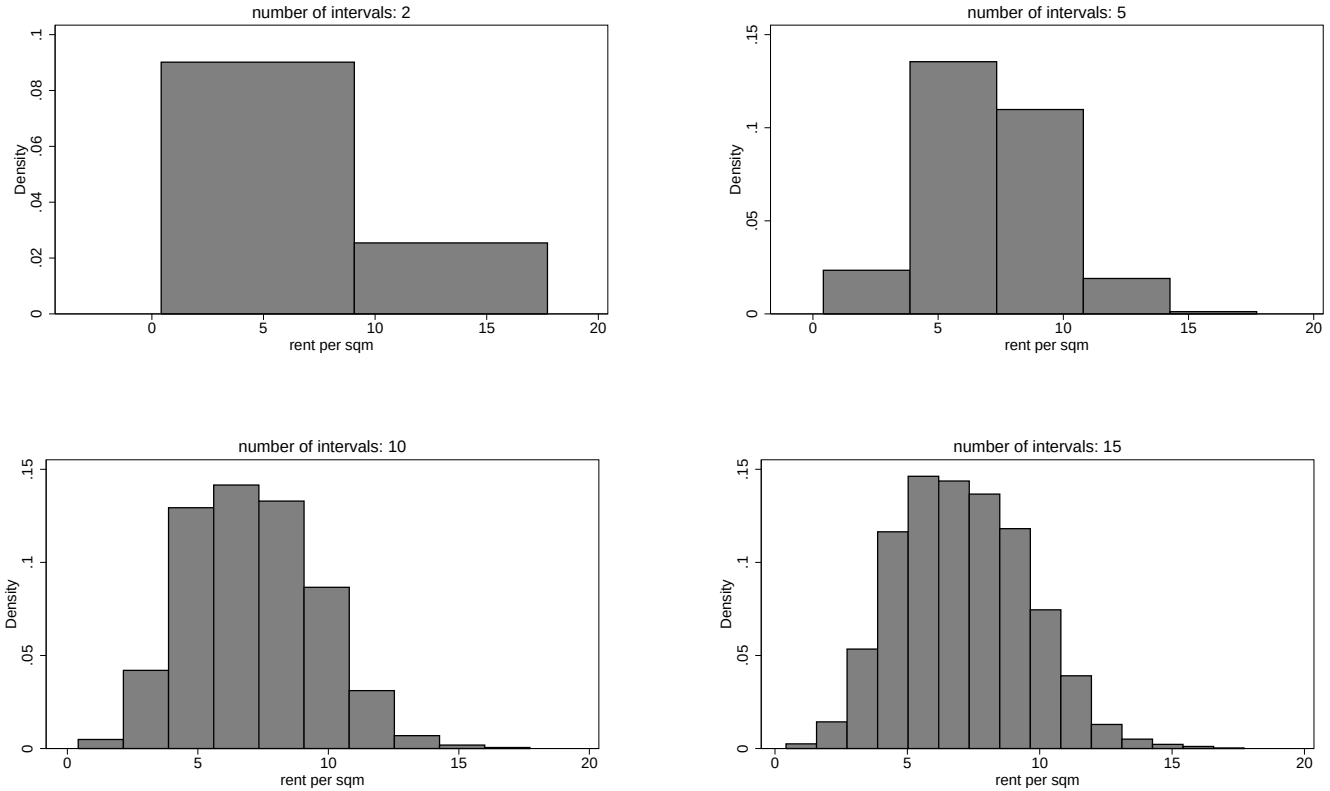


Figure 2: *Dependence of histograms on the number of intervals.*

The histogram

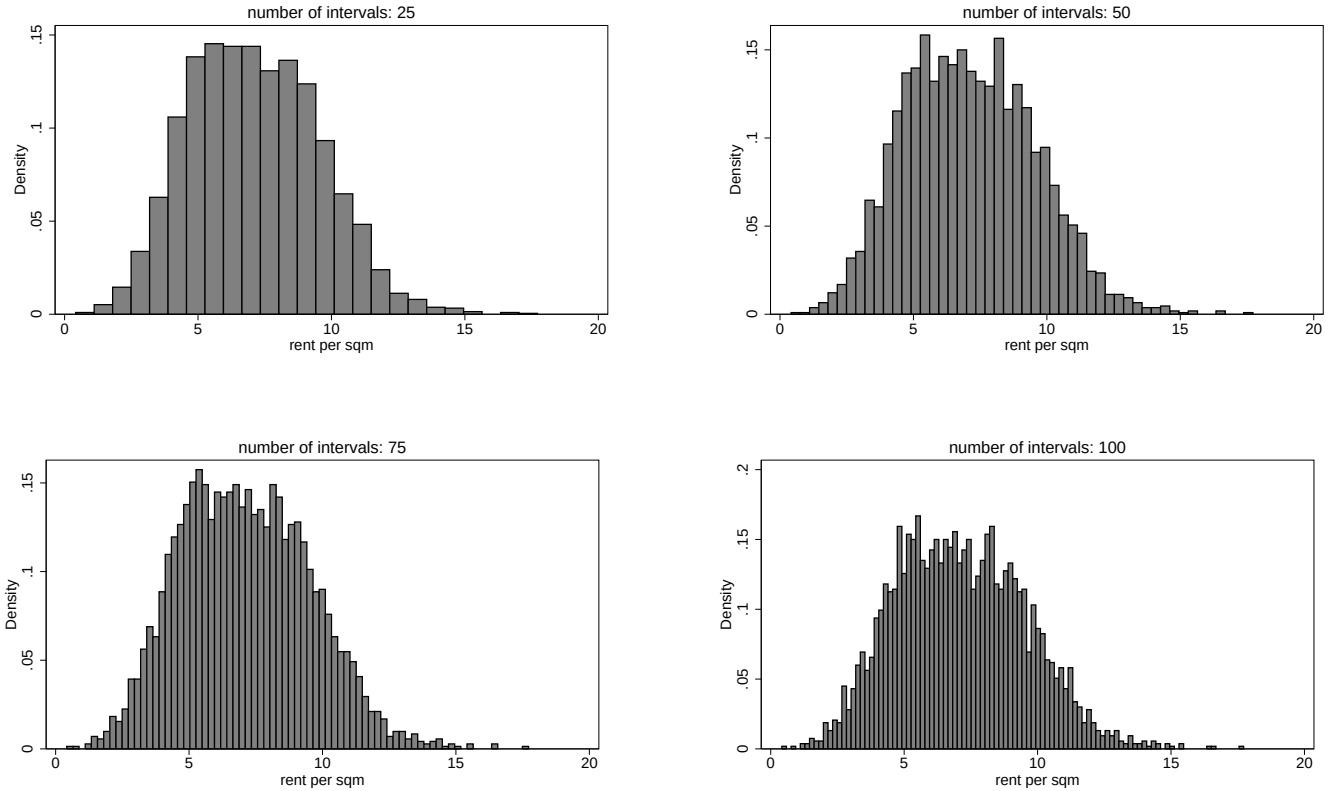


Figure 3: *Dependence of histograms on the number of intervals.*

Kernel density estimators

- Define intervals $[x-h; x+h]$ of width $2h$ and move across the x-axis.
- This yields the estimator

$$\hat{f}_h(x) = \frac{1}{2nh}(\#x_i \text{ within the interval}[x-h; x+h])$$

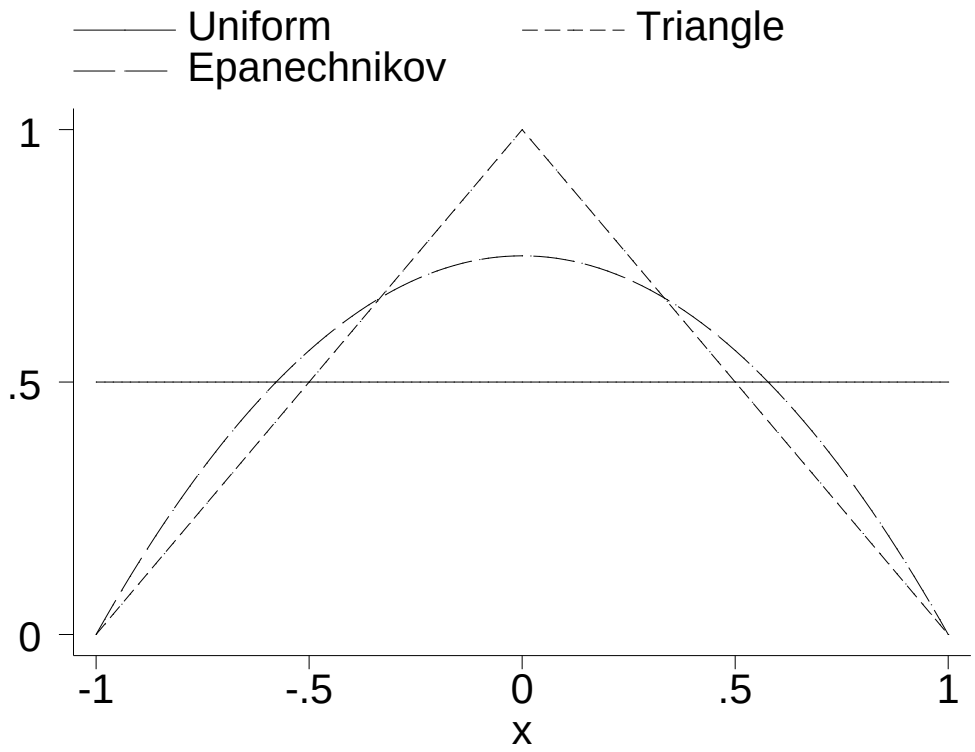
- Using the " kernel"

$$K(u) = \begin{cases} \frac{1}{2} & |u| \leq 1 \\ 0 & \text{else} \end{cases}$$

we obtain

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

Kernels



Kernels

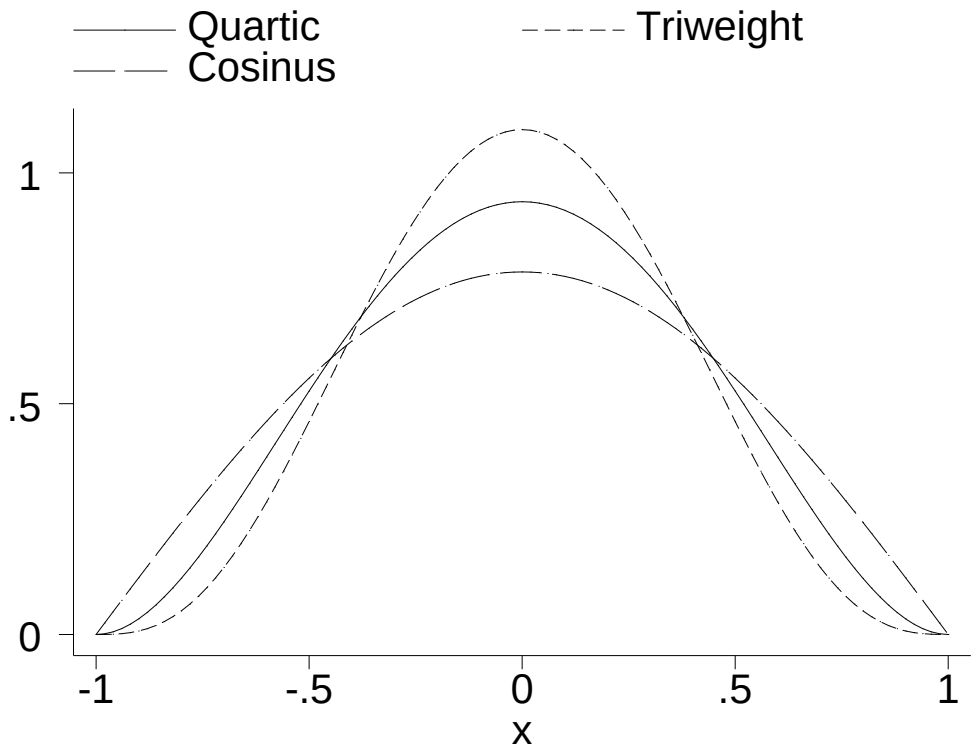
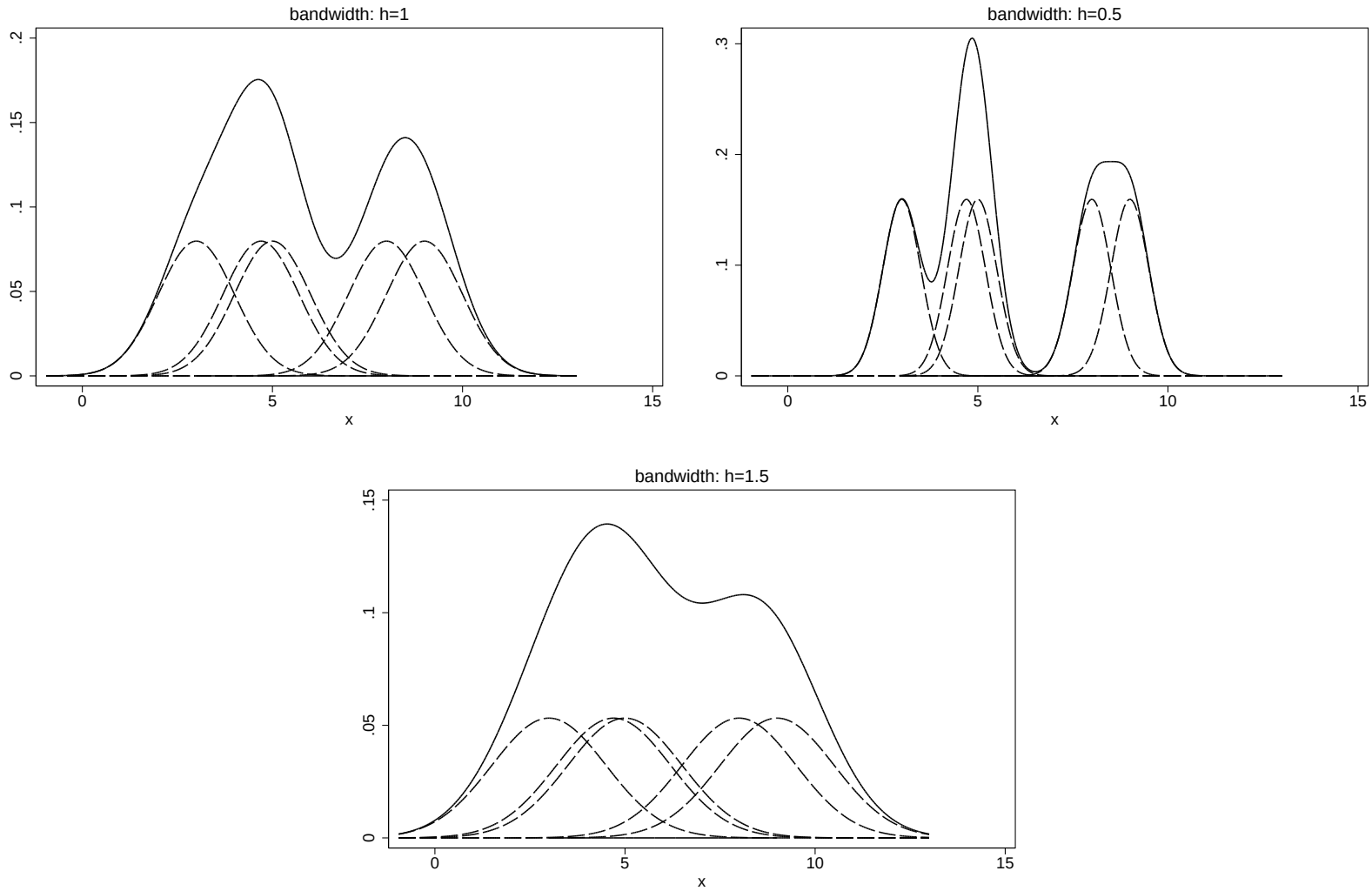


Illustration kernel density estimators



Expected value and variance

For fixed x we obtain for the expected value

$$E(\hat{f}_h(x)) = \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\, dy$$

and the variance

$$Var(\hat{f}_h(x)) = \frac{1}{nh^2} \int_{\mathbb{R}} K^2\left(\frac{x-y}{h}\right) f(y)\, dy - \frac{1}{n} E(\hat{f}_h(x))^2.$$

Bias and MSE

$$Bias(\hat{f}_h(x)) = E(\hat{f}_h(x)) - f(x) = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x).$$

$$\begin{aligned} MSE(\hat{f}_h(x)) &= Var(\hat{f}_h(x)) + Bias^2(\hat{f}_h(x)) \\ &= \frac{1}{nh^2}\int_{\mathbb{R}} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \frac{1}{n}E(\hat{f}_h(x))^2 \\ &\quad + \left(\frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x)\right)^2 \end{aligned}$$

Consistency

Theorem Parzen

Let $R(x)$, $x \in \mathbb{R}$ be a (measurable) function with properties

1. $\sup_{x \in \mathbb{R}} |R(x)| < \infty$ (d.h. $R(x)$ ist beschr"ankt)

2. $\int |R(x)|\, dx < \infty$

3. $|x| R(x) \to 0$.

Consistency

Theorem Parzen

Let $g(x)$, $x \in \mathbb{R}$, be a (measurable) function with $\int |g(x)|\, dx < \infty$. Consider the series

$$g_n(x) = \frac{1}{h_n} \int R\left(\frac{x-y}{h_n}\right) g(y)\, dy$$

where h_n is a series with $\lim_{n\to\infty} h_n = 0$. Then for every continuity point x of g we have

$$g_n(x) \to g(x) \int R(s)\, ds$$

for $n \to \infty$.

Consistency

Let f be continuous. Then

$$E(\hat{f}_{h_n}(x)) \to f(x)$$

provided that the bandwidth h_n converges to zero for $n \to \infty$.

For $nh_n \to \infty$ as $n \to \infty$, we have

$$Var(\hat{f}_{h_n}(x)) \to 0,$$

i.e. $\hat{f}_{h_n}(x)$ is consistent.

Landau Symbols

Assume we are given two real valued sequences $\{a_n\}$ and $\{b_n\}$ with $n \in N$. We write

$$a_n = O(b_n)$$

if

$$\left|\frac{a_n}{b_n}\right|$$

is bounded for $n \to \infty$.

The sequence $\{a_n\}$ is roughly of the same order as $\{b_n\}$.

Obviously $a_n = O(1)$ says that a_n is bounded.

Landau Symbols

We write

$$a_n = o(b_n)$$

if

$$\left|\frac{a_n}{b_n}\right|$$

converges to zero for $n \to \infty$.

The series $\{a_n\}$ converges faster to zero than $\{b_n\}$.

Obviously $a_n = o(1)$ means

$$\lim_{n\to\infty} a_n = 0.$$

Mean integrated squared error

- Let f be at least 2 times continuously differentiable, f'' be bounded, f and f'' square integrable.

- Assume h_n is sequence with $h_n \to 0$.

Using $\int g^2(s)\,ds = ||g||_2^2$ and $\mu_2(g) = \int g(s)s^2\,ds$ for a function g we obtain:

1. $Var(\hat{f}_{h_n}(x)) = \dfrac{1}{nh_n}||K||_2^2 f(x) + o\left(\dfrac{1}{nh_n}\right)$ respectively

$$\int Var(\hat{f}_{h_n}(x))\,dx = \frac{1}{nh_n}||K||_2^2 + o\left(\frac{1}{nh_n}\right).$$

Mean integrated squared error

2. $Bias(\hat{f}_{h_n}(x)) = \frac{{h_n}^2}{2}\mu_2(K)f''(x) + o({h_n}^2)$ respectively

$$\int Bias^2(\hat{f}_{h_n}(x))\,dx = \frac{{h_n}^4}{4}\mu_2^2(K)||f''||_2^2 + o({h_n}^4).$$

3. $MISE(\hat{f}_{h_n}) = \frac{1}{nh_n}||K||_2^2 + \frac{{h_n}^4}{4}\mu_2^2(K)||f''||_2^2 + o(\frac{1}{nh_n} + {h_n}^4)$,

where $MISE(\hat{f}_{h_n}) = \int MSE(\hat{f}_{h_n}(x))dx$

Bias variance trade off

- The bias decreases as h decreases.
- The variance decreases as h increases.
- The bias depends on $f''(x)$ as a measure for the curvature of f. Increased curvature corresponds with increased bias. The bias is positive at local maxima and negative at local minima.
- There is also a dependence of the bias und variance on the chosen kernel K.

AMISE optimal bandwidth

The idea is to optimze the AMISE (Asymptotic Mean Integrated Squared Error), which is obtained from MISE by deleting the o-terms, i.e.

$$AMISE(\hat{f}_h) = \frac{1}{nh}||K||_2^2 + \frac{h^4}{4}\mu_2^2(K)||f''||_2^2$$

The optimal bandwith is given by

$$h_0 = \left(\frac{\|K\|_2^2}{\|f''\|_2^2\mu_2^2(K)n}\right)^{\frac{1}{5}}.$$

AMISE optimal bandwidth

- Obviously, the optimal bandwidth depends on functionals of f(circulus virtuosis).

- In practice a reference density is inserted. Using e.g. a normal distribution and a normal kernel, we obtain

$$\hat{h}_0 = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\,\hat{\sigma}\,n^{-\frac{1}{5}}.$$

Distribution free statistical tests

Distribution free statistical tests
Order statistics and ranks

Order statistics

- Assume we are given at least ordinal data $x_1, \ldots, x_n$.
- $x_1, \ldots, x_n$ are realisations of an i.i.d. random sample $X_1, \ldots, X_n$ with cdf F.
- Ordering the observations yields the vector $(x_{(1)}, \ldots, x_{(n)})$ of the so called *order statistics* $(X_{(1)}, \ldots, X_{(n)})$.
- The component $x_{(j)}$ is the value of the j-th order statistics $X_{(j)}$.

Example: children

Assume 10 children are given a puzzle to solve. The following table contains the time required to solve the puzzle:

child i	1	2	3	4	5	6	7	8	9	10
time x_i in seconds	78	58	60	82	83	85	65	72	70	61

Special cases of order statistics

- Special cases are $x_{(1)}$ (minimum), $x_{(n)}$ (maximum) and the median

$$m = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ uneven} \\ 0.5(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \text{ even.} \end{cases}$$

- The range of the data is given by

$$d = x_{(n)} - x_{(1)}.$$

Example: children

Empirical cumulative distribution function (cdf)

- For unordered observations $x_1, \ldots, x_n$ the empirical cumulative distribution function F_n is defined for a real value x as

$$F_n(x) = \frac{\#(x_i \leq x)}{n}$$

- In terms of the ordered values $x_{(1)}, \ldots, x_{(n)}$ we obtain

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{m}{n} & x_{(m)} \leq x < x_{(m+1)} \\ 1 & x \geq x_{(n)}. \end{cases}$$

Properties of the empirical cdf

- F_n is a monotonic increasing step function with discontinuities at $x_{(1)}, \ldots, x_{(n)}$.
- For unbounded observations F_n increases in $x_{(j)}$ by $1/n$, otherwise by k/n if there are k equal observations $x_{(j)}$.
- $\lim_{x \to -\infty} F_n(x) = 0$ and $\lim_{x \to \infty} F_n(x) = 1$.
- For fixed x $F_n(x)$ is a random variable because $F_n(x)$ depends on the random sample $X_1, \ldots, X_n$.
- $F_n(x)$ is discrete with possibles values m/n for $m = 0, 1, \ldots, n$.

Probability function of the cdf

- The probability function of $F_n(x)$ is given by

$$P(F_n(x) = m/n) = \binom{n}{m} (F(x))^m (1 - F(x))^{n-m}, \qquad m = 0, 1, \ldots, n,$$

  i.e., $nF_n(x)$ is binomial distributed with parameters n and $F(x)$.

  The parameter $F(x)$ depends on the unknown cdf F.

- It follows

$$E(F_n(x)) = F(x) \qquad Var(F_n(x)) = \frac{1}{n} F(x)(1 - F(x)).$$

Asymptotic properties of the cdf

- $F_n(x)$ is unbiased for $F(x)$.
- Since $Var(F_n(x))$ converges to zero as n tends to infinity, $F_n(x)$ is consistent in mean square.
- It immediately follows that $F_n(x)$ converges also in probability to $F(x)$.

Definition ranks

- Let $X_1, \ldots, X_n$ be a random sample of a continuous random variable X.
- If X_i has value $x_{(j)}$ within the order statistics, the rank of X_i is defined as $rank(X_i) = R(X_i) = j$.
- The rank $R_i = R(X_i)$ is a discrete random variable with realisations $1, 2 \ldots, n$.
- Obviously $X_{(R_i)} = X_i$ and

$$R_i = \sum_{j=1}^{n} \chi(X_i - X_j)$$

where

$$\chi(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{else} \end{cases}$$

Example: children

Properties of the rank

1. $P(R_1 = r_1, \ldots, R_n = r_n) = 1/n!$, where $r_1, \ldots, r_n$ is a permutation of $1, \ldots, n$.

2. $P(R_i = j) = 1/n$, $i = 1, \ldots, n$, where $j \in \{1, \ldots, n\}$.

3. $P(R_i = k, R_j = l) = \frac{1}{n(n-1)}$ for $1 \leq i, j, k, l \leq n, i \neq j, k \neq l$.

4. $E(R_i) = \frac{n+1}{2}$, $i = 1, \ldots, n$

5. $Var(R_i) = \frac{n^2-1}{12}, i = 1, \ldots, n$.

6. $Cov(R_i, R_j) = -\frac{n+1}{12}, 1 \leq i, j \leq n, i \neq j$.

7. $Corr(R_i, R_j) = -\frac{1}{n-1}, 1 \leq i, j \leq n, i \neq j$.

Distribution of $F(X)$

- Note that $F(x)$ is a fixed number whereas $F(X)$ is a random variable.
- Assume X is a random variable with continuous cdf F. Then $F(X)$ is uniformly distributed within the unit interval $[0, 1]$.
- For continuous F $F(X_1), \ldots, F(X_n)$ can be seen as a random sample according to the uniformly distributed random variable $F(X)$ $(F(X_1), \ldots, F(X_n))$
- For continuous F $(F(X_{(1)}), \ldots, F(X_{(n)}))$ can be regarded as order statistics from a uniformly distributed population.

Joint distribution of order statistics

- The joint distribution of the random sample $X_1, \ldots, X_n$ is given by

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n),$$

  where f is the density of X.

- Define $Y_i = X_{(i)}$. Then $Y_1 < Y_2 < \cdots < Y_n$ and the order statistics $(Y_1, \ldots, Y_n)$ can be regarded as a transformation of $(X_1, \ldots, X_n)$.

- The joint distribution of the order statistic $(Y_1, \ldots, Y_n)$ is given by

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \ldots f(y_n) & y_1 < \cdots < y_n \\ 0 & \text{else.} \end{cases}$$

Distribution free statistical tests
Goodness of fit tests

Kolomogorov-Smirnow-Test

Data

Cardinal measuring scale.

Assumptions

We assume an independent random sample $X_1, \ldots, X_n$ with continuous cdf F.

Kolomogorov-Smirnow-Test

Test problem

- Case 1: Two-tailed

$$H_0 : F(x) = F_0(x) \qquad \text{versus} \qquad H_1 : F(x) \neq F_0(x)$$

- Case 2: One-tailed

$$H_0 : F(x) \geq F_0(x) \qquad \text{versus} \qquad H_1 : F(x) < F_0(x)$$

- Case 3: One-tailed

$$H_0 : F(x) \leq F_0(x) \qquad \text{versus} \qquad H_1 : F(x) > F_0(x)$$

Kolomogorov-Smirnow-Test

Test statistics

- Case 1: Two-tailed

$$K_n = \sup_{x \in \mathbb{R}} |F_0(x) - F_n(x)|$$

- Case 2: One-tailed

$$K_n^+ = \sup_{x \in \mathbb{R}} (F_0(x) - F_n(x))$$

- Case 3: One-tailed

$$K_n^- = \sup_{x \in \mathbb{R}} (F_n(x) - F_0(x))$$

The test statistic $K_n = \sup_{x \in \mathbb{R}} |F_0(x) - F_n(x)|$ is for all continuous cdfs F_0 distribution free, i.e. its distribution is independent of F_0.

Kolomogorov-Smirnow-Test

Test procedure

- Case 1: Two-tailed

  Reject H_0 if $K_n \geq k_{1-\alpha}$.

- Case 2: One-tailed

  Reject H_0 if $K_n^+ \geq k_{1-\alpha}^+$.

- Case 3: One-tailed

  Reject H_0 if $K_n^- \geq k_{1-\alpha}^-$.

χ^2-test

Data

Any measuring scale. The data have to be grouped into k disjoint groups:

group	1	2	...	k
number of observations	n_1	n_2	...	n_k

Assumptions

The randon sample $X_1, \ldots, X_n$ is independent.

Test problem

Let F be the unknown cdf and F_0 a fully specified cdf.

$$H_0 : F(x) = F_0(x) \qquad \text{for all } x \in \mathbb{R}$$
$$H_1 : F(x) \neq F_0(x) \qquad \text{for at least one } x \in \mathbb{R}$$

χ^2-test

Test statistics

Under H_0, let p_i be the probability that the random variable X takes a value in the i-th group. Then use the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Test procedure

H_0 is rejected, if $\chi^2 \leq \chi^2_{1-\alpha}(k-1)$ where $\chi^2_{1-\alpha}(k-1)$ is the $1-\alpha$ quantile of the χ^2 distribution with $k-1$ degrees of freedom.

Binomial-Test

Data

Any measuring scale. The data are grouped into two disjoint groups.

Assumptions

- The random samples $X_1, \ldots, X_n$ are independent.
- The probability for an observation belonging to class 1 is p for all n observations.

Binomial-Test

Test problem

- Case 1: Two-tailed

$$H_0 : p = p_0 \qquad \text{versus} \qquad H_1 : p \neq p_0$$

- Case 2: One-tailed

$$H_0 : p \geq p_0 \qquad \text{versus} \qquad H_1 : p < p_0$$

- Case 3: One-tailed

$$H_0 : p \leq p_0 \qquad \text{versus} \qquad H_1 : p > p_0$$

Binomial-Test

Test statistics

Use the number of observations T belonging to class 1. Under H_0 we have $T \sim B(n, p_0)$.

Test procedure

- Case 1: Two-tailed

  Reject H_0 if $T \geq t_{1-\alpha_1}$ or $T \leq t_{\alpha_2}$

  where $t_{1-\alpha_1}$ and t_{α_2} are defined through

$$P(T \geq t_{1-\alpha_1}) = \alpha_1 \qquad P(T \leq t_{\alpha_2}) = \alpha_2$$

  with $\alpha_1 + \alpha_2 = \alpha$.

Binomial-Test

- Case 2: One-tailed

  Reject H_0 if $T \geq t_{1-\alpha}$ where

$$P(T \geq t_{1-\alpha}) = \alpha.$$

- Case 3: One-tailed

  Reject H_0 if $T \leq t_{\alpha}$ where

$$P(T \leq t_{\alpha}) = \alpha.$$

Example: Quality control

A manufacturer claims that a component is defect with probability 5 percent or less. A random sample of $n = 20$ yields 3 defect components.

Distribution free statistical tests
Linear rank tests

Initial situation

- Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $X_i \sim F(x - \theta)$ for $i = 1, \ldots, n$ where the cdf F is continuous with density f and symmetric about θ.

- θ is a location parameter, e.g. the median of F.

- We would like to test $H_0 : \theta = \theta_0$ against the Alternatives $\theta < \theta_0$, $\theta > \theta_0$ or $\theta \neq \theta_0$.

Basic idea

- **Idea:** Use as the test statistic a function of the ranks rather than the observations.
- No loss of information if measuring scale is ordinal.
- Some loss of information for cardinal measuring scale, but surprisingly little loss of efficiency when using rank based statistics in hypothesis testing.

Definition of linear rank statistic

- In order to define the general linear rank statistic define the differences $D_i = X_i - \theta_0$ and their absolute values $|D_i| = |X_i - \theta_0|$, $i = 1, \ldots, n$.

- Let $R_i^+ = R(|D_i|)$ be the rank of $|D_i|$ and

$$Z_i = \begin{cases} 1 & D_i > 0 \\ 0 & D_i < 0. \end{cases}$$

- Then the linear rank statistic L_n^+ has the general form

$$L_n^+ = \sum_{i=1}^{n} g(R_i^+) Z_i,$$

where $g(R_i^+) \in \mathbb{R}$ are suitable weights.

Alternative definition of the linear rank statistic

- Let $|D|_{(1)} < \cdots < |D|_{(n)}$ be the order statistics of $|D_1|, \ldots, |D_n|$.
- Define

$$V_i = \begin{cases} 1 & |D|_{(i)} \text{ belongs to a positive difference} \\ 0 & |D|_{(i)} \text{ belongs to a negative difference} \end{cases}$$

- Then

$$L_n^+ = \sum_{i=1}^{n} g(i) V_i$$

- For the expected value and the variance we have

$$E(L_n^+) = \frac{1}{2} \sum_{i=1}^{n} g(i) \qquad Var(L_n^+) \sum_{i=1}^{n} (g(i))^2.$$

Sign test

Data

Cardinal measuring scale.

Assumptions

The random sample $X_1, \ldots, X_n$ is independent with continuous cdf $F(x - \theta)$.

Sign test

Test problem

- Case 1: Two-tailed

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

- Case 2: One-tailed

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0$$

- Case 3: One-tailed

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta < \theta_0$$

Sign test

Test statistics

- The test statistic is given by

$$V_n^+ = \sum_{i=1}^{n} V_i.$$

- V_n^+ is a special case of L_n^+ with $g(i) = 1$.
- Hence V_n^+ counts the observations that are larger than θ_0.
- Under H_0 we have $V_n^+ \sim B(n, 0.5)$.

Sign test

Test procedure

In analogy to the Binomial test.

- Case 1: Two-tailed

  Reject H_0 if $V_n^+ \geq v_{1-\alpha/2}^+$ or $V_n^+ \leq v_{\alpha/2}^+$

- Case 2: One-tailed

  Reject H_0 if $V_n^+ \geq v_{1-\alpha}^+$.

- Case 3: One-tailed

  Reject H_0 if $T \leq v_\alpha^+$.

Example: intelligence test

Wilcoxon test

Data

Cardinal measuring scale.

Assumptions

The random sample $X_1, \ldots, X_n$ is independent with continuous cdf $F(x - \theta)$ that is symmetric around θ.

Wilcoxon test

Test problem

- Case 1: Two-tailed

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

- Case 2: One-tailed

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0$$

- Case 3: One-tailed

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta < \theta_0$$

Wilcoxon test

Test statistics

- The test statistic is given by

$$W_n^+ = \sum_{i=1}^{n} iV_i = \sum_{i=1}^{n} R_i^+ Z_i$$

- W_n^+ is a special case of L_n^+ with $g(i) = i$.
- W_n^+ may be interpreted as the sum of the ranks of the absolute values $|D_i|$ with positive differences. In contrary to the sign test the size of the differences additionally enters the test statistic.
- Under H_0 W_n^+ is symmetric around $E(W_n^+) = n(n+1)/4$ and we have

$$P(W_n^+ = w^+) = a(w^+)/2^n.$$

Wilcoxon test

Test procedure

In analogy to the Binomial test.

- Case 1: Two-tailed

  Reject H_0 if $W_n^+ \geq w_{1-\alpha/2}^+$ or $W_n^+ \leq w_{\alpha/2}^+$

- Case 2: One-tailed

  Reject H_0 if $W_n^+ \geq w_{1-\alpha}^+$.

- Case 3: One-tailed

  Reject H_0 if $W \leq w_\alpha^+$.

Non- and semiparametric regression

Univariate smoothing

Illustration: Malnutrition in Zambia

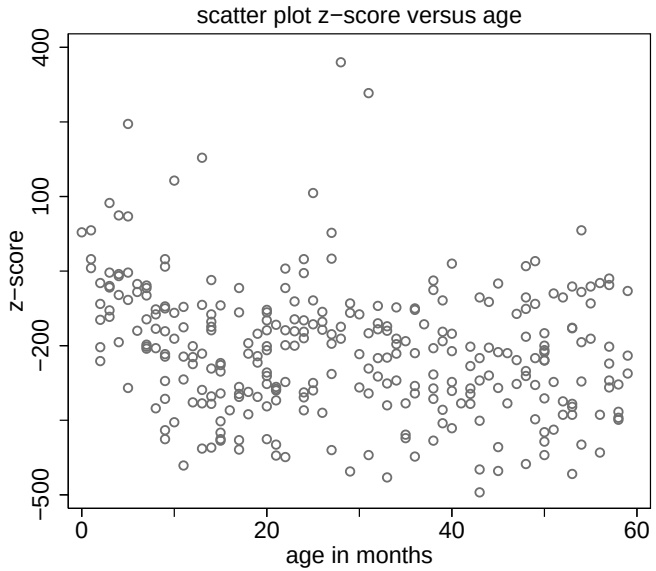


Figure 4: *Malnutrition in Tanzania: scatter plot of the Z-score for chronic malnutrition versus the age of the child in months for one of the districts in Tanzania (Ruvuma).*

Illustration: simulated data

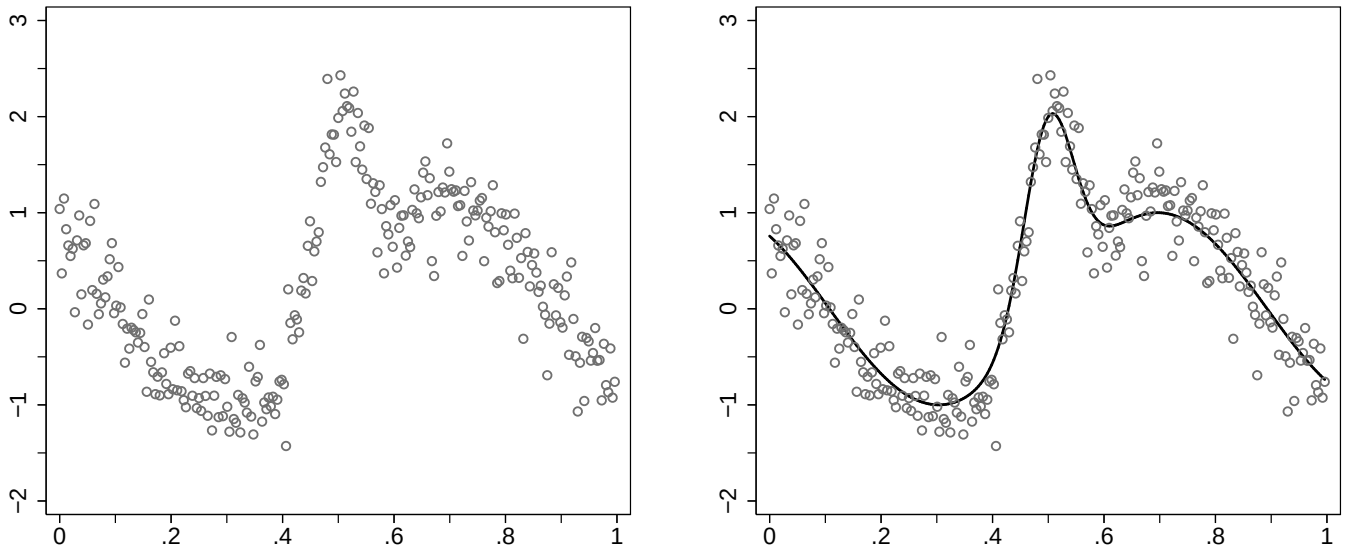


Figure 5: *Scatter plot of a simulated data set with nonlinear effect of the covariate: The right panel additionally shows the true covariate effect. The data have been simulated according to the model* $y = f(x) + \varepsilon$ *where* $f(x) = \sin(2(4x-2)) + 2\exp(-(16^2)(x-0.5)^2)$ *and* $\varepsilon \sim \mathrm{N}(0, 0.3^2)$.

Definition univariate smoothing

Data

Measurements (y_i, z_i), $i = 1, \ldots, n$, for a continuous response variable y and a continuous covariate z.

Model

$$y_i = f(z_i) + \varepsilon_i$$

with independent and identically distributed errors and

$$\mathsf{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathsf{Var}(\varepsilon_i) = \sigma^2.$$

In some cases, we additionally assume that the errors are i.i.d. normally distributed, so that

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2).$$

Polynomial splines

Polynomial regression

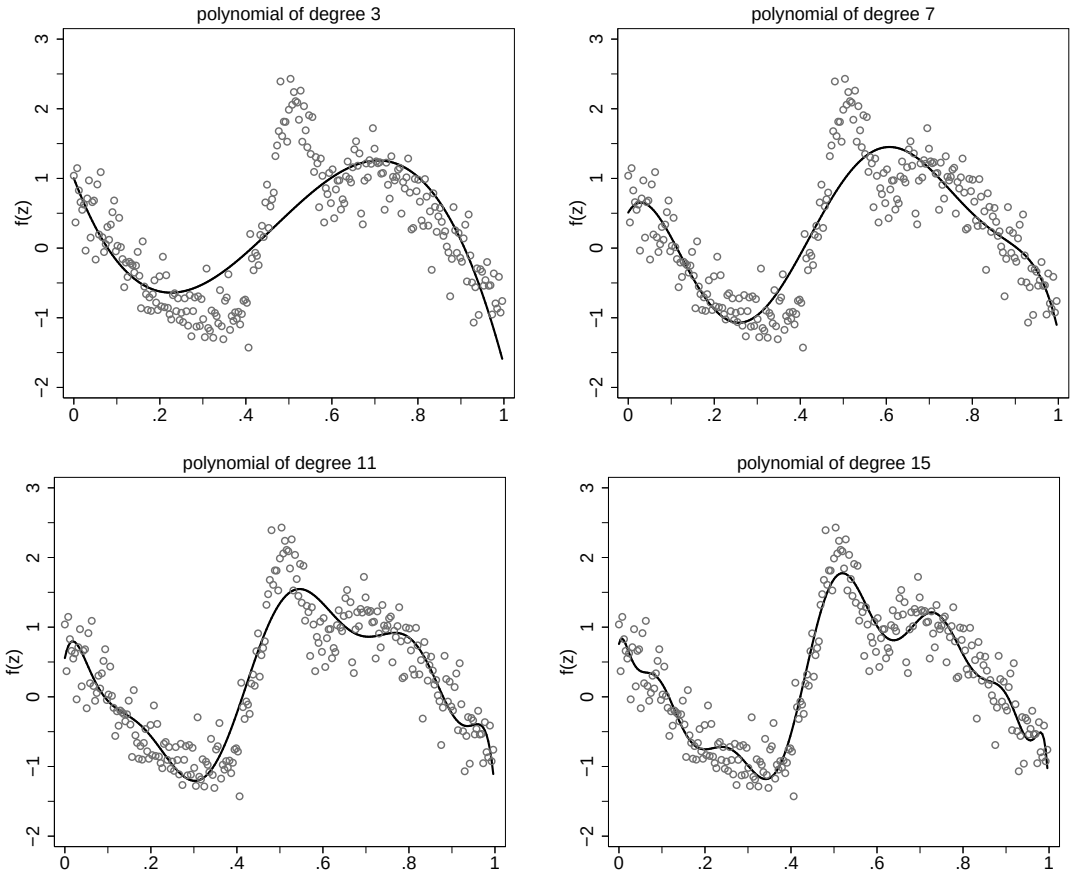


Figure 6: *Polynomial regression models for the simulated data set.*

Polynomial splines

Piecewise polynomial regression

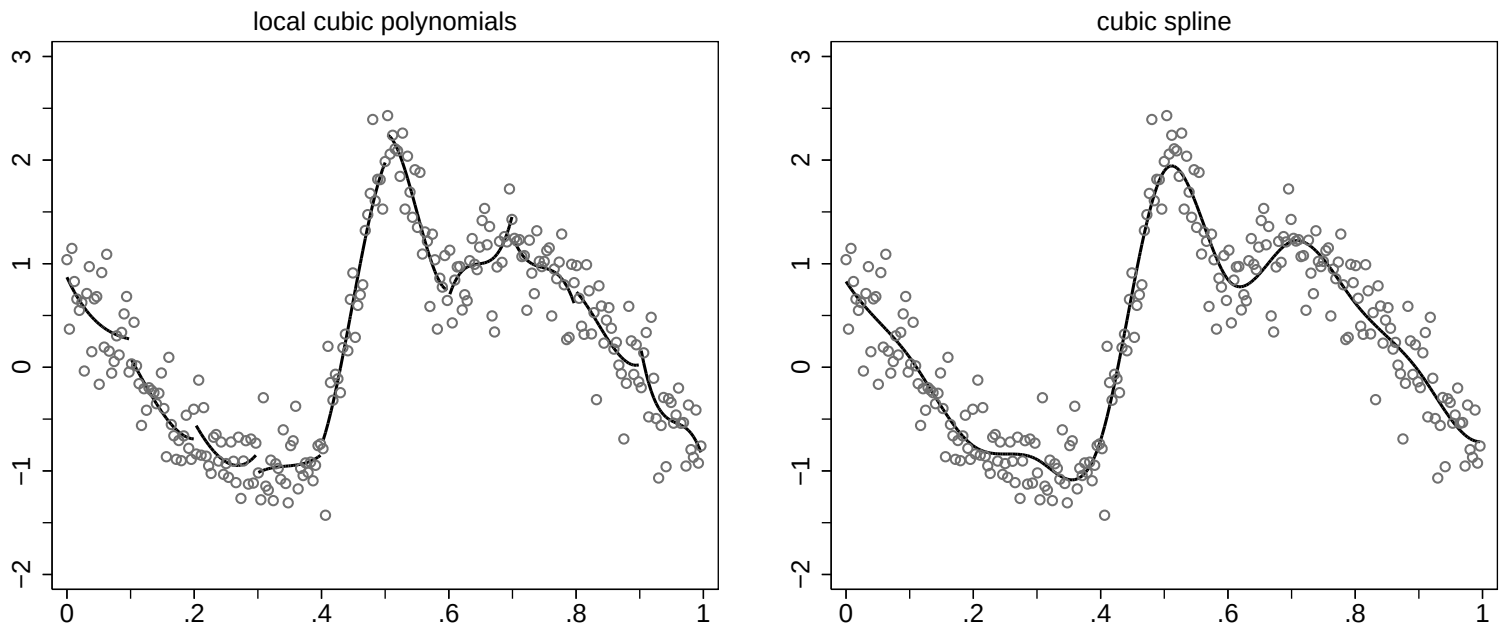


Figure 7: *Piecewise polynomial regression (left) and polynomial splines (right).*

Polynomial splines

Definition

A function $f : [a, b] \to \mathbb{R}$ is called a polynomial spline of degree $l \geq 0$ with knots $a = \kappa_1 < \ldots < \kappa_m = b$, if it fulfills the following conditions:

1. $f(z)$ is $(l-1)$-times continuously differentiable. The special case of $l = 1$ corresponds to $f(z)$ being continuous (but not differentiable). We do not state any smoothness requirements for $f(z)$ when $l = 0$.

2. $f(z)$ is a polynomial of degree l on the intervals $[\kappa_j, \kappa_{j+1})$ defined by the knots.

Polynomial splines

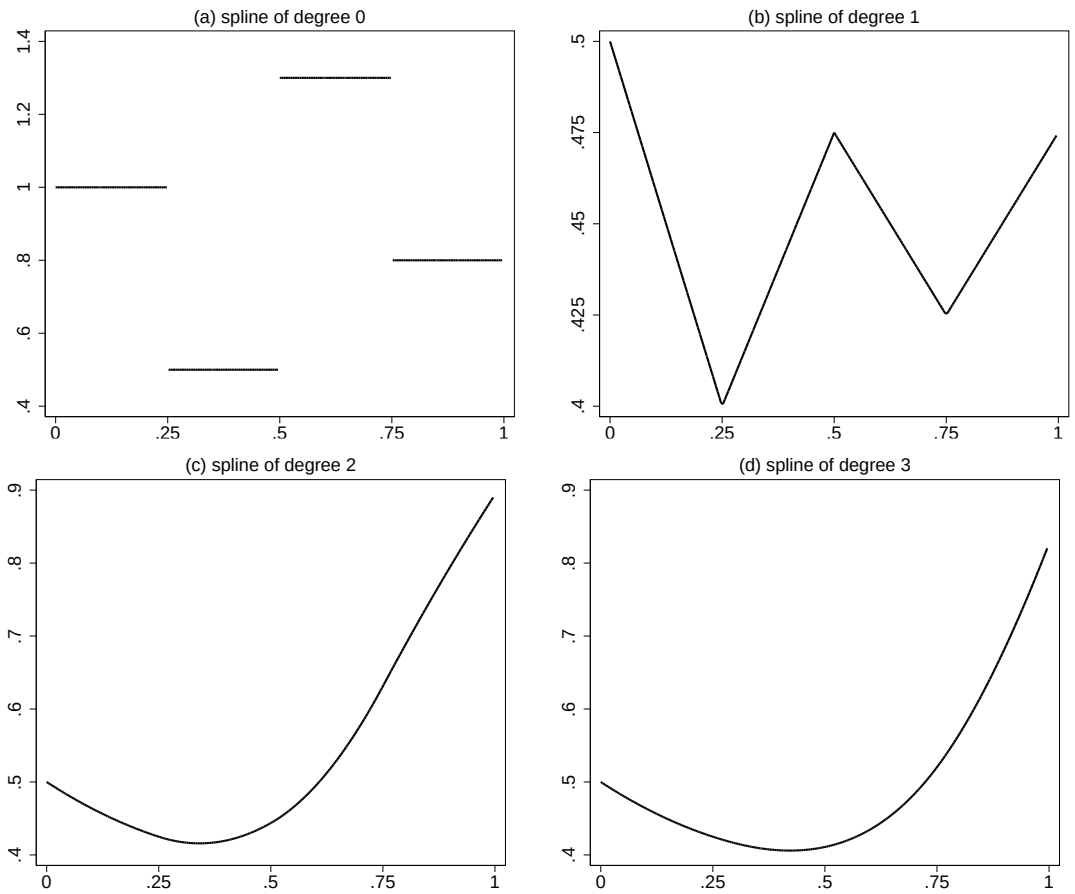


Figure 8: *Examples of polynomial splines of degree 0, 1, 2 and 3 with knots* $\kappa_1 = 0$, $\kappa_2 = 0.25$, $\kappa_3 = 0.5$, $\kappa_4 = 0.75$ *and* $\kappa_5 = 1$.

Polynomial splines

Truncated power series

$$y_i = \gamma_1 + \gamma_2 z_i + \ldots + \gamma_{l+1} z_i^l + \gamma_{l+2}(z_i - \kappa_2)_+^l + \ldots + \gamma_{l+m-1}(z_i - \kappa_{m-1})_+^l + \varepsilon_i$$

with

$$(z - \kappa_j)_+^l = \begin{cases} (z - \kappa_j)^l & z \geq \kappa_j, \\ 0 & \text{otherwise.} \end{cases}$$

Polynomial splines

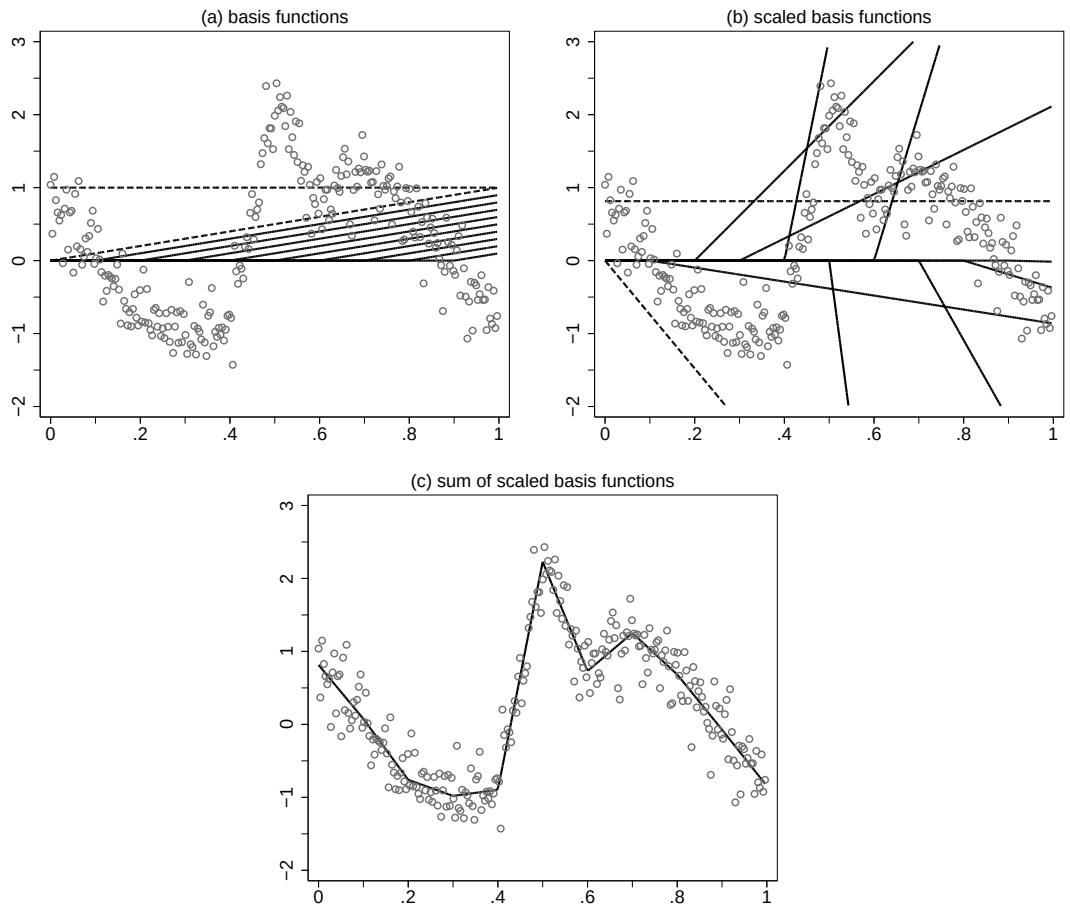


Figure 9: *Polynomial spline fit with linear truncated polynomials.*

Polynomial splines

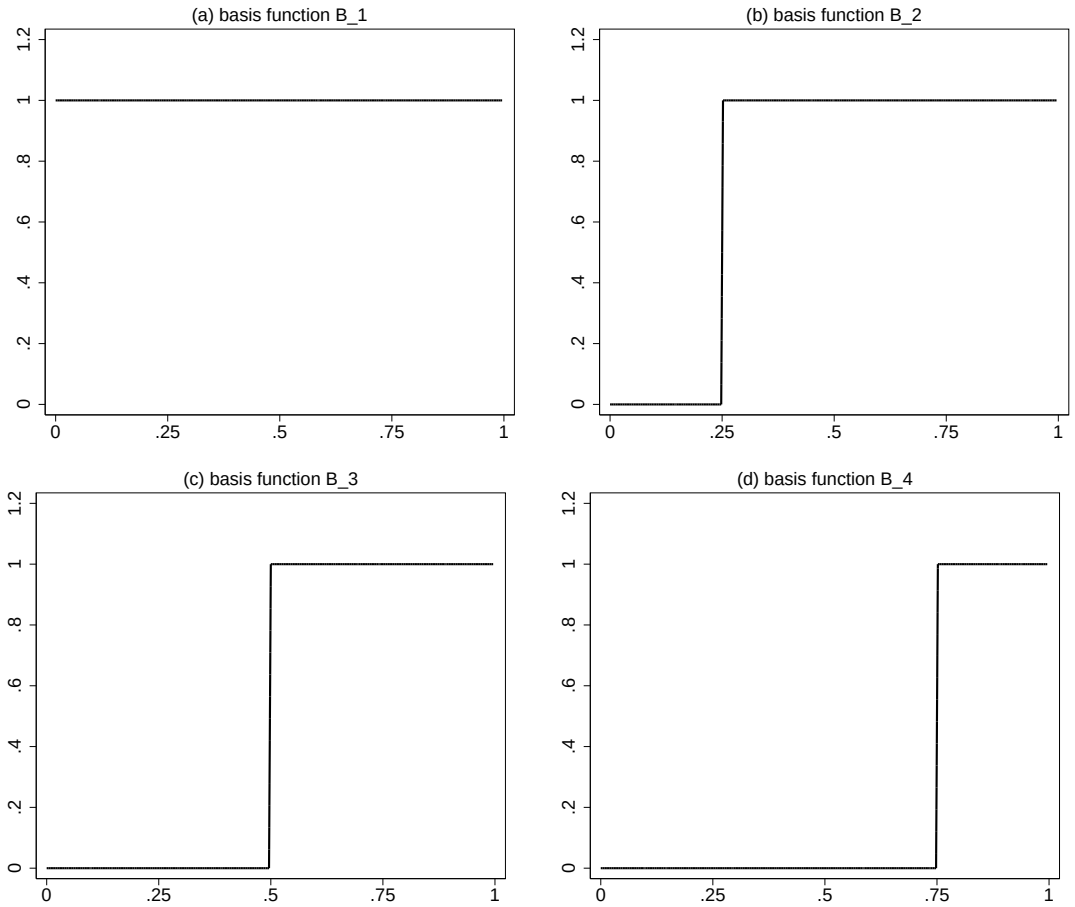


Figure 10: *TP basis for splines of degree 0 based on the knots* $\{0,0.25,0.5,0.75,1\}$.

Polynomial splines

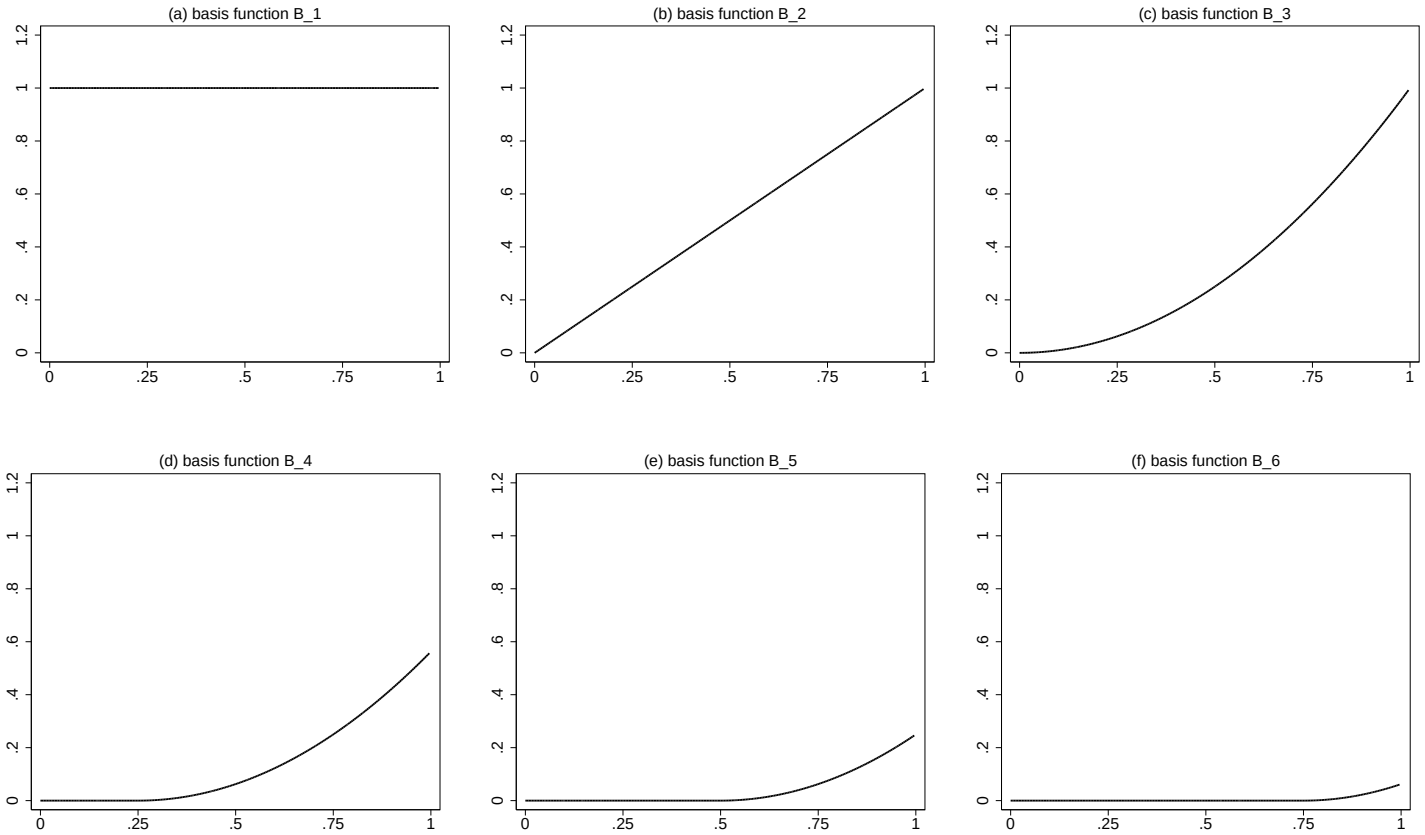


Figure 11: *TP basis for splines of degree 2 based on the knots* $\{0,0.25,0.5,0.75,1\}$.

Polynomial splines

Defining the vectors of the observed response variables $\mathbf{y}$ and the errors ε, as well as the design matrix

$$\mathbf{Z} = \begin{pmatrix} B_1(z_1) & \dots & B_d(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \dots & B_d(z_n) \end{pmatrix} = \begin{pmatrix} 1 & z_1 & \dots & z_1^l & (z_1-\kappa_2)_+^l & \dots & (z_1-\kappa_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & z_n & \dots & z_n^l & (z_n-\kappa_2)_+^l & \dots & (z_n-\kappa_{m-1})_+^l \end{pmatrix},$$

we obtain the equation

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \varepsilon,$$

with the coefficient vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)'$. The usual least squares estimate is thus

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

and

$$\hat{f}(z) = \mathbf{z}'\hat{\boldsymbol{\gamma}}$$

with $\mathbf{z} = (\mathbf{B_1(z)}, \dots, \mathbf{B_d(z)})'$ depending on the chosen covariate value z.

Polynomial splines

Example: Malnutrition in Tanzania

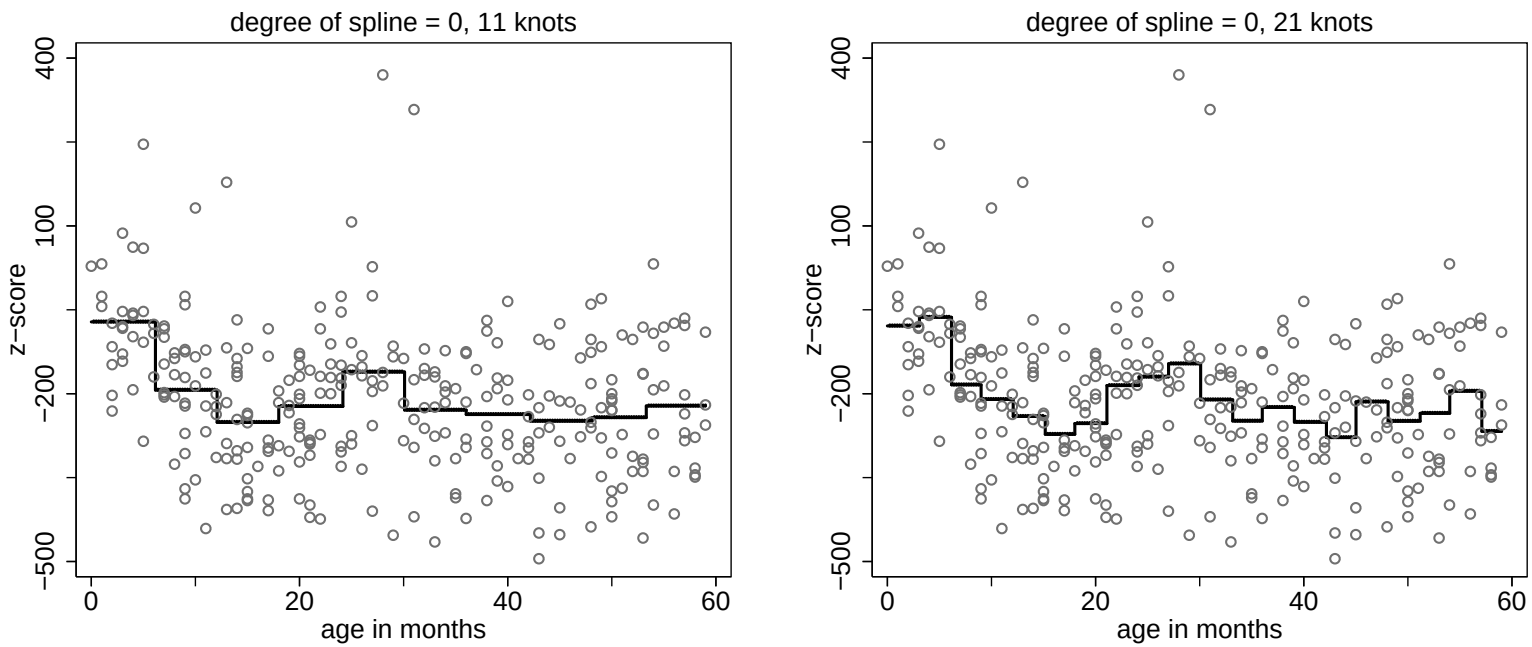


Figure 12: *Nonparametric estimates for the age effect based on polynomial splines.*

Polynomial splines

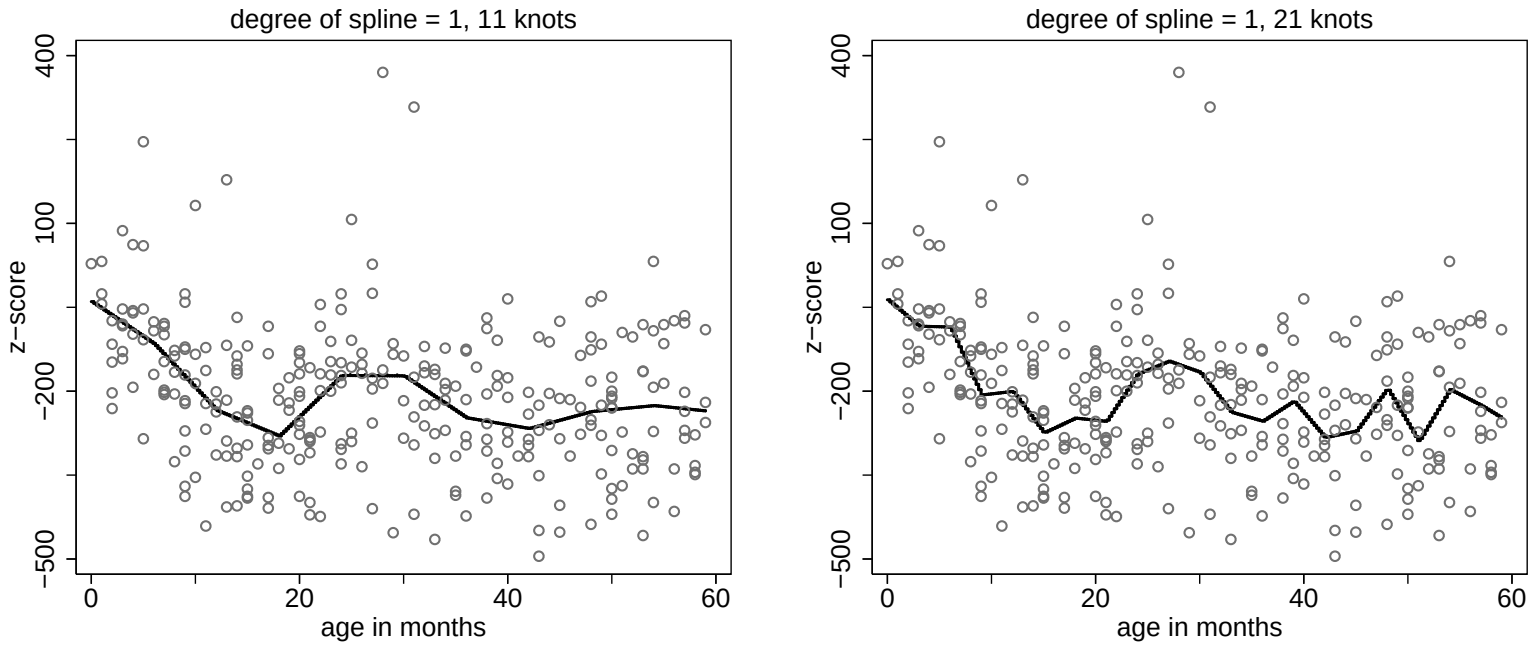


Figure 13: *Nonparametric estimates for the age effect based on polynomial splines.*

Polynomial splines

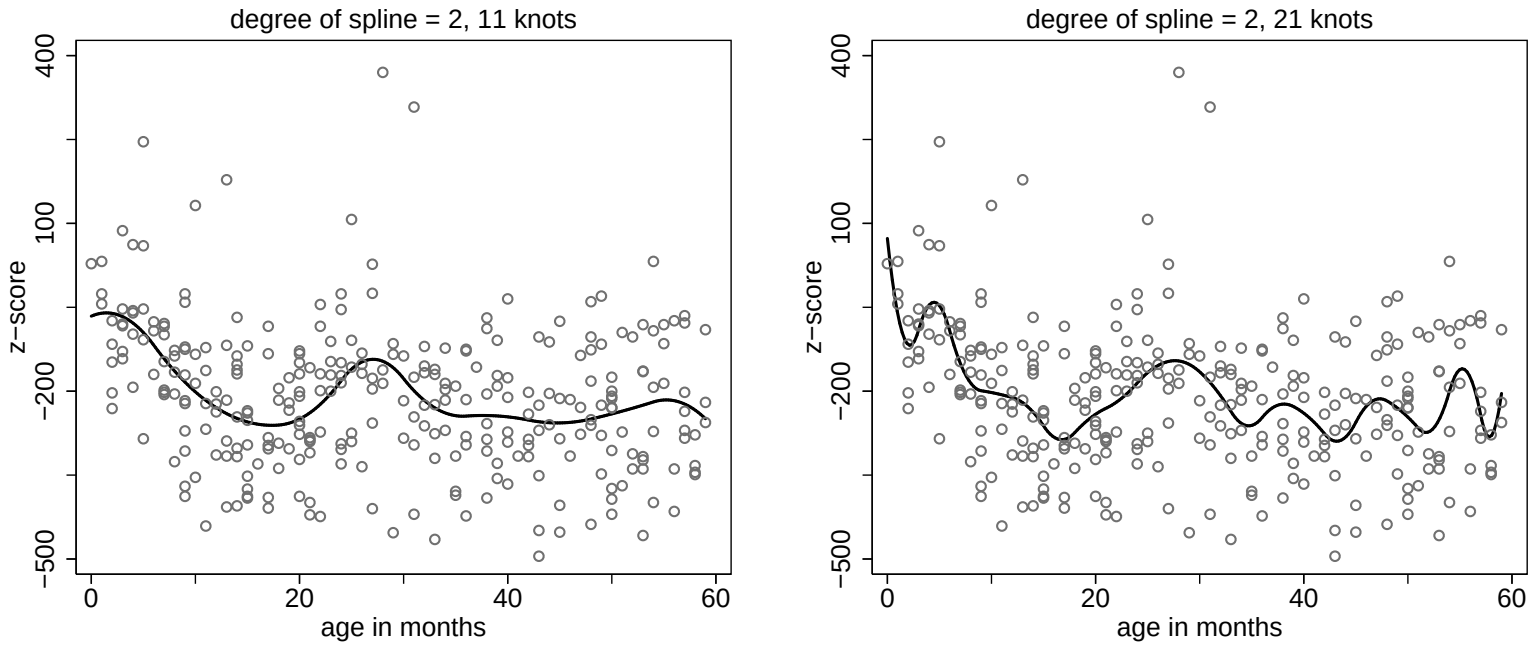


Figure 14: *Nonparametric estimates for the age effect based on polynomial splines.*

Polynomial splines

Influence of the knots

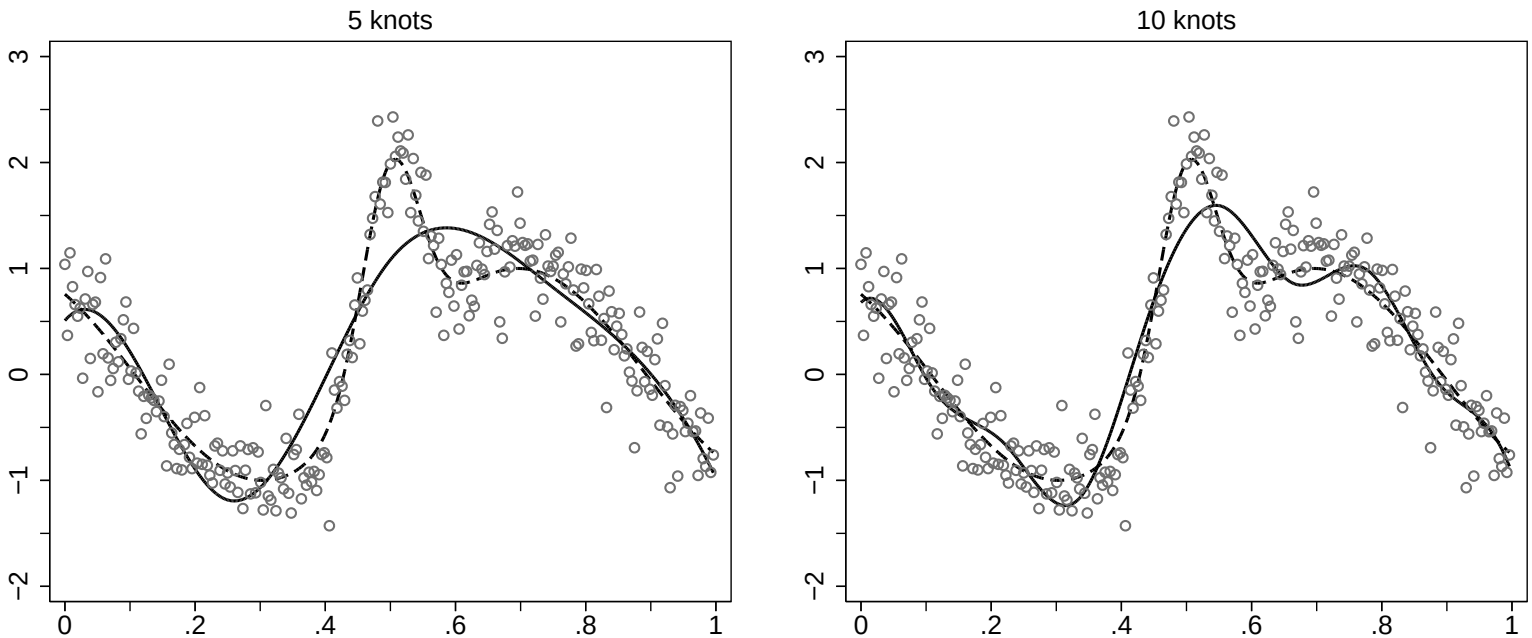


Figure 15: *Impact of the number of knots on cubic spline fits: The estimated function is represented as a solid line while the true function is superimposed as a dashed line.*

Polynomial splines

Influence of the knots

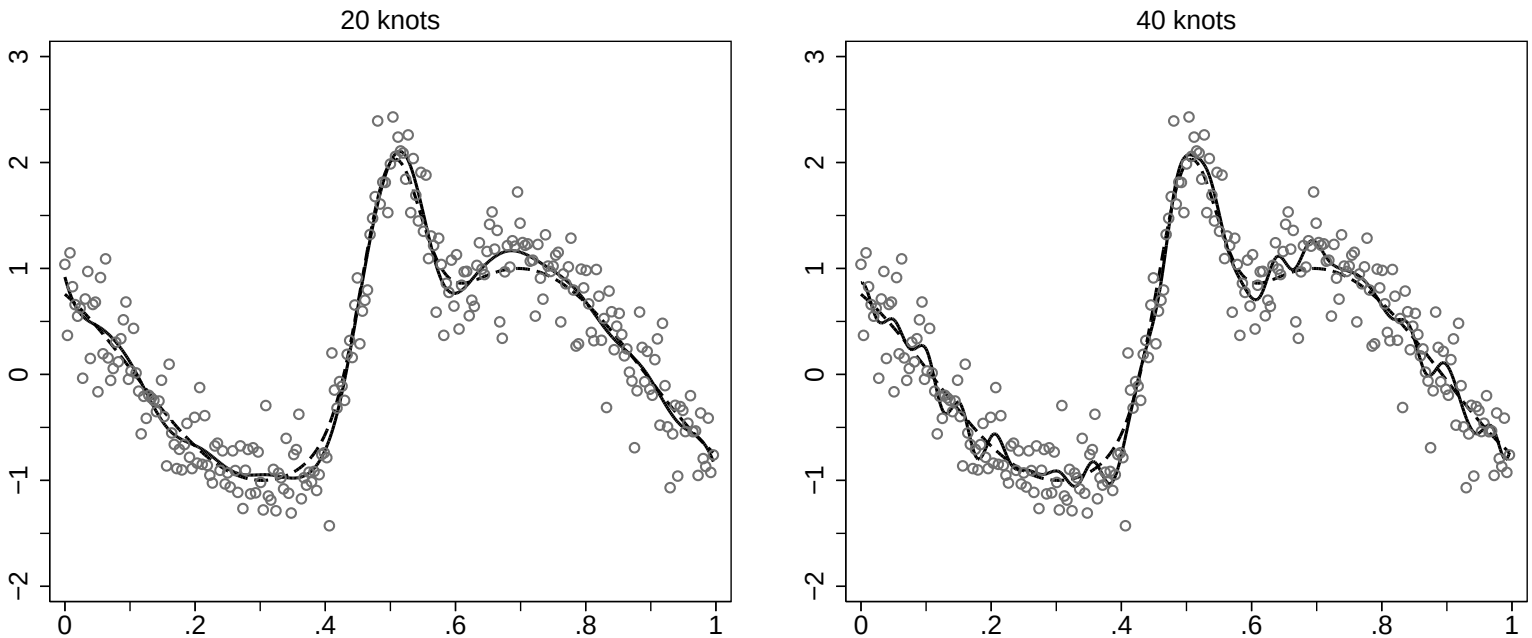


Figure 16: *Impact of the number of knots on cubic spline fits: The estimated function is represented as a solid line while the true function is superimposed as a dashed line.*

Polynomial splines

- *Equidistant knots:* The domain $[a, b]$ of z is split into $m - 1$ intervals of width

$$h = \frac{b - a}{m - 1}$$

in order to obtain the knots

$$\kappa_j = a + (j - 1) \cdot h, \quad j = 1, \ldots, m,$$

In all examples considered so far, we have always tacitly assumed equidistant knots.

- *Quantile-based knots:* Use the $(j - 1)/(m - 1)$-quantiles $(j = 1, \ldots, m)$ of the observed covariate values $z_1, \ldots, z_n$ as knots.

- *Visual knot choice based on a scatter plot*

Polynomial splines

B-splines

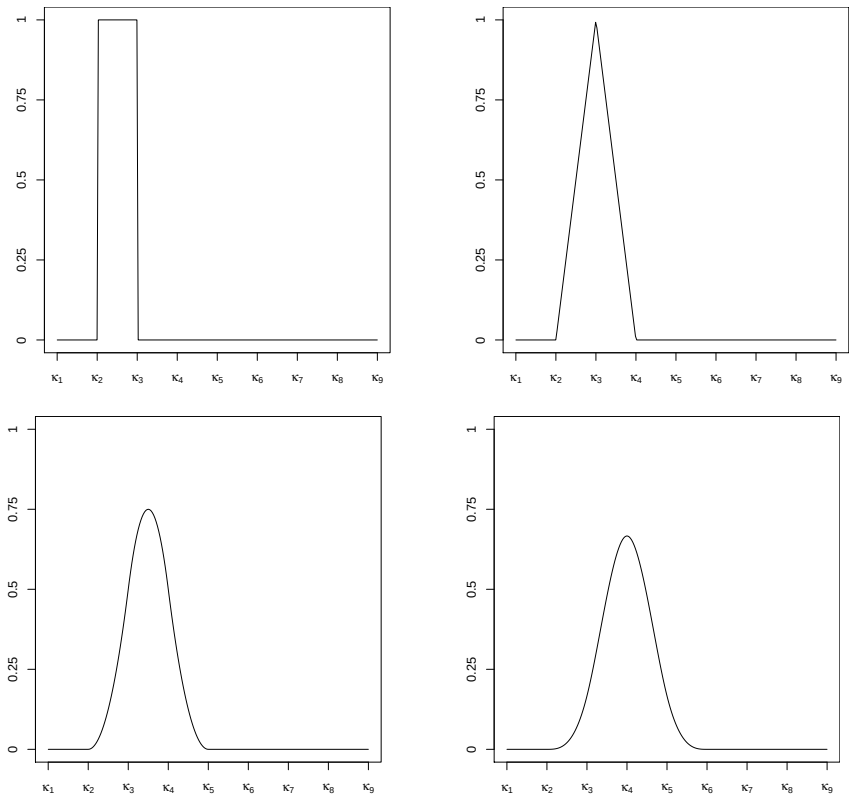


Figure 17: *Single B-spline basis functions for degrees* $l = 0, 1, 2, 3$ *and equidistant knots.*

Polynomial splines

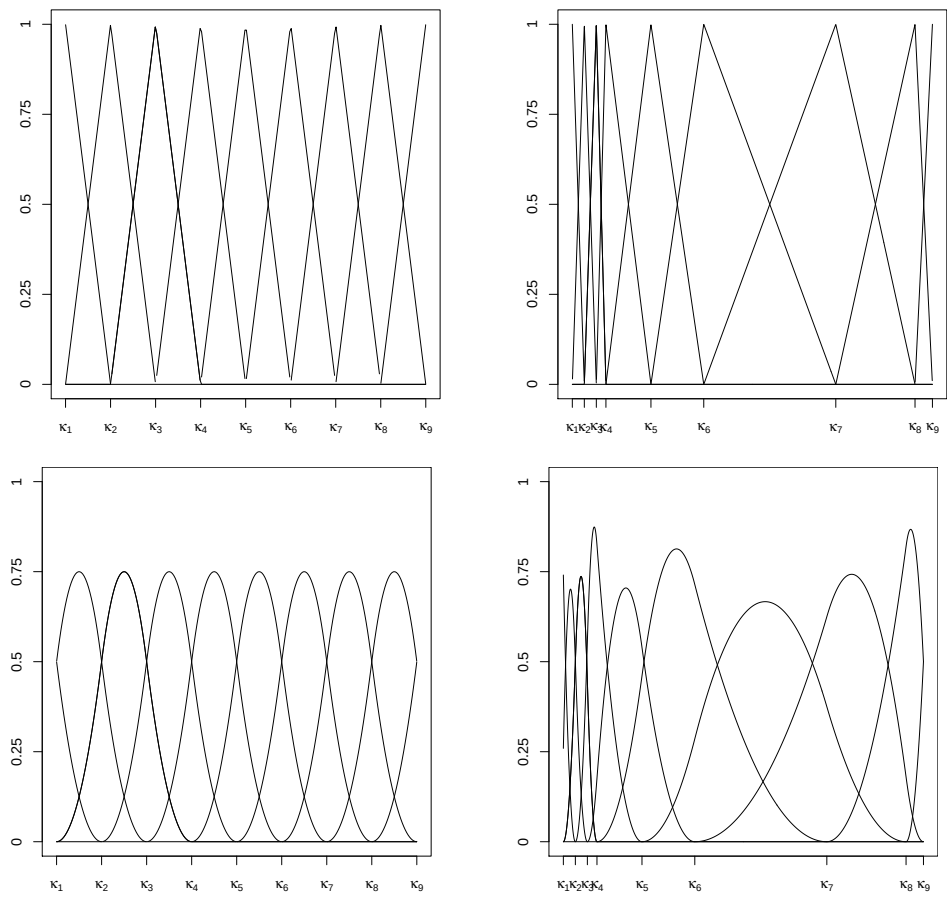


Figure 18: *B-spline bases of degree* $l = 1, 2, 3$ *with equidistant knots (left panel) and unevenly distributed knots (right panel).*

Polynomial splines

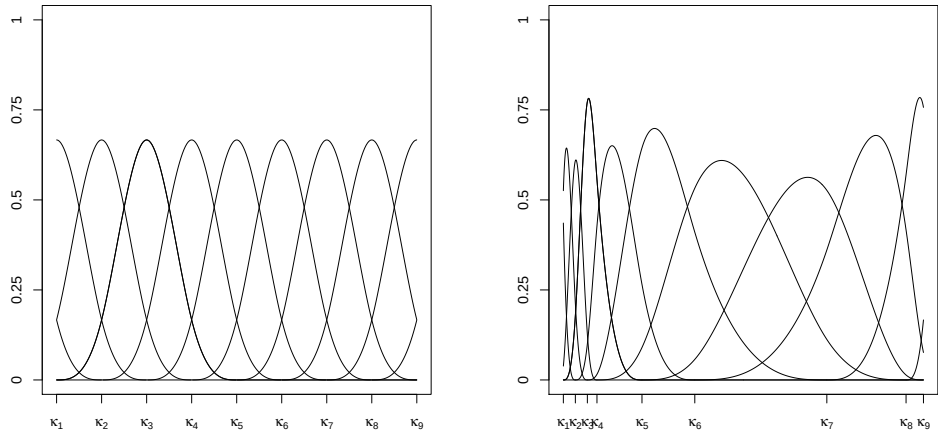


Figure 19: *B-spline bases of degree* $l = 1, 2, 3$ *with equidistant knots (left panel) and unevenly distributed knots (right panel).*

Polynomial splines

Definition B-splines

The function $f(z)$ can again be represented through a linear combination of $d = m + l - 1$ basis functions, i.e.

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z).$$

B-splines of order $l = 0$ are defined as

$$B_j^0(z) = I(\kappa_j \leq z < \kappa_{j+1}) = \begin{cases} 1 & \kappa_j \leq z < \kappa_{j+1} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, d-1,$$

where $I(\cdot)$ denotes the indicator function.

Polynomial splines

We obtain the basis functions of degree $l = 1$ as

$$B_j^1(z) = \frac{z - \kappa_{j-1}}{\kappa_j - \kappa_{j-1}} I(\kappa_{j-1} \leq z < \kappa_j) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_j} I(\kappa_j \leq z < \kappa_{j+1}),$$

i.e. each basis function is defined by two linear segments on the intervals $[\kappa_{j-1}, \kappa_j)$ and $[\kappa_j, \kappa_{j+1})$, which are continuously combined at the knot κ_j.

In general, higher order B-splines are defined recursively:

$$B_j^l(z) = \frac{z - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(z).$$

Polynomial splines

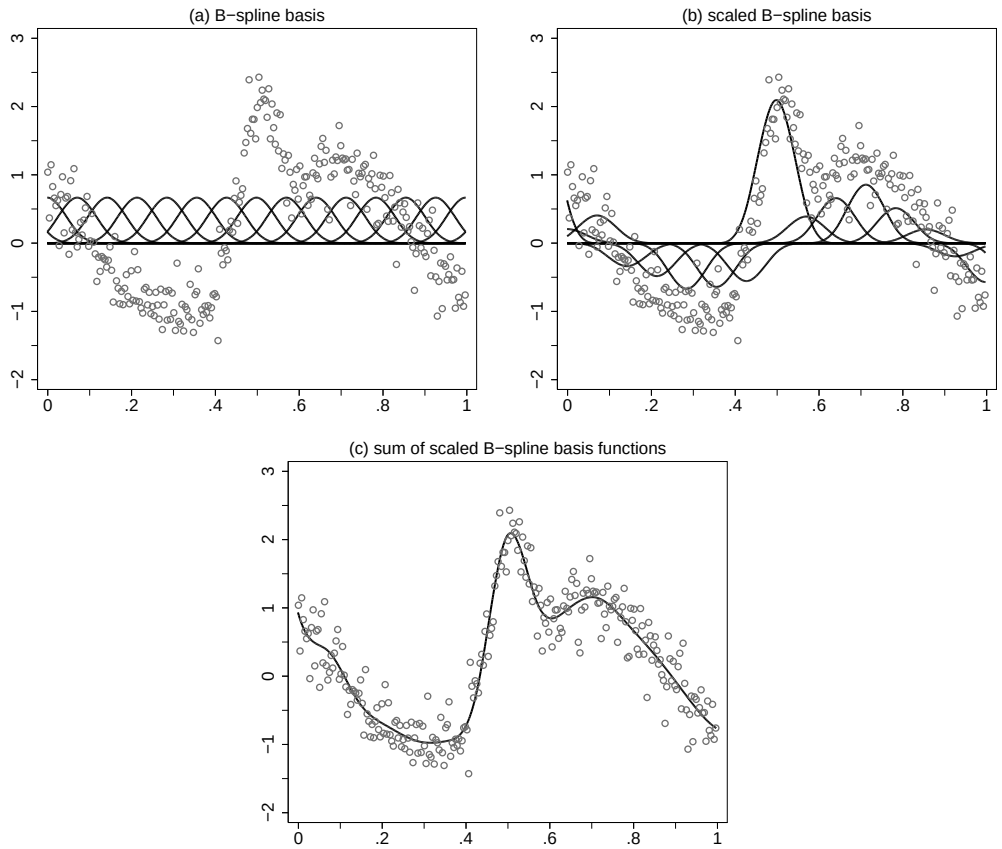


Figure 20: *Schematic representation of a nonparametric fit with cubic B-splines.*

Polynomial splines

Design matrix and fit

$$\mathbf{Z} = \begin{pmatrix} B_1^l(z_1) & \dots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \dots & B_d^l(z_n) \end{pmatrix}.$$

The usual least squares estimate is thus

$$\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

and

$$\hat{f}(z) = \mathbf{z}'\hat{\gamma}$$

with $\mathbf{z} = (\mathbf{B_1(z)}, \dots, \mathbf{B_d(z)})'$ depending on the chosen covariate value z.

P- splines

The main idea of *penalized splines (P-splines)* can be summarized as follows:

- Approximate the function $f(z)$ with a polynomial spline that uses a generous number of knots (usually about 20 to 40). This ensures that $f(z)$ can be approximated with enough flexibility to represent even highly complex functions.

- Introduce an additional penalty term that prevents overfitting and minimize a *penalized least squares criterion* instead of the usual least squares criterion.

P- splines

P-splines based on TP basis

$$f(z) = \gamma_1 + \gamma_2 z + \ldots + \gamma_{l+1} z^l + \gamma_{l+2}(z - \kappa_2)_+^l + \ldots + \gamma_d (z - \kappa_{m-1})_+^l.$$

Minimize the penalized residual sum of squares

$$\mathrm{PLS}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2.$$

P- splines

P-Splines based on B-Splines

Minimize the penalized residual sum of squares

$$\mathrm{PLS}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=r+1}^{d} (\Delta^r \gamma_j)^2,$$

where Δ^r denotes rth order differences, which are recursively defined by

$$\begin{aligned}
\Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1}, \\
\Delta^2 \gamma_j &= \Delta^1 \Delta^1 \gamma_j = \Delta^1 \gamma_j - \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}, \\
&\vdots \\
\Delta^r \gamma_j &= \Delta^{r-1} \gamma_j - \Delta^{r-1} \gamma_{j-1}.
\end{aligned}$$

P- splines

Influence of the number of knots

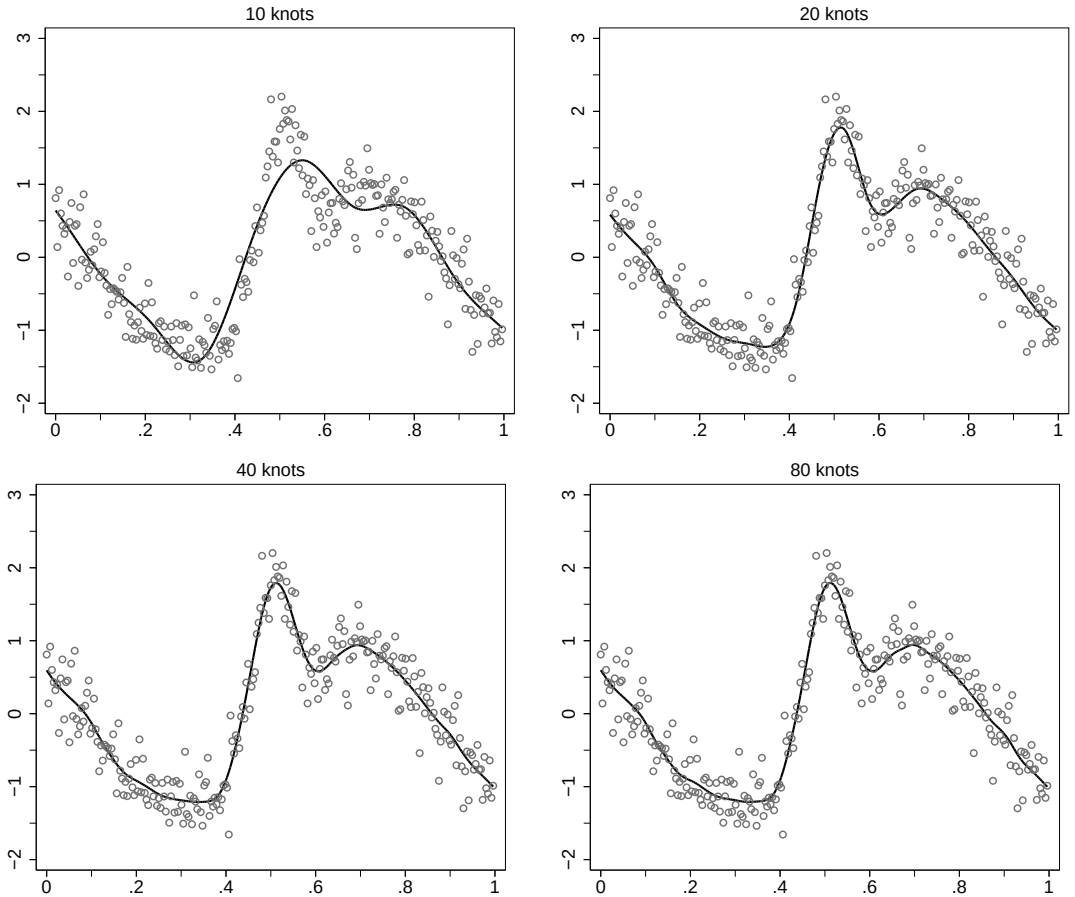


Figure 21: *Influence of the number of knots on estimated P-splines.*

P- splines

Impact of the smoothing parameter

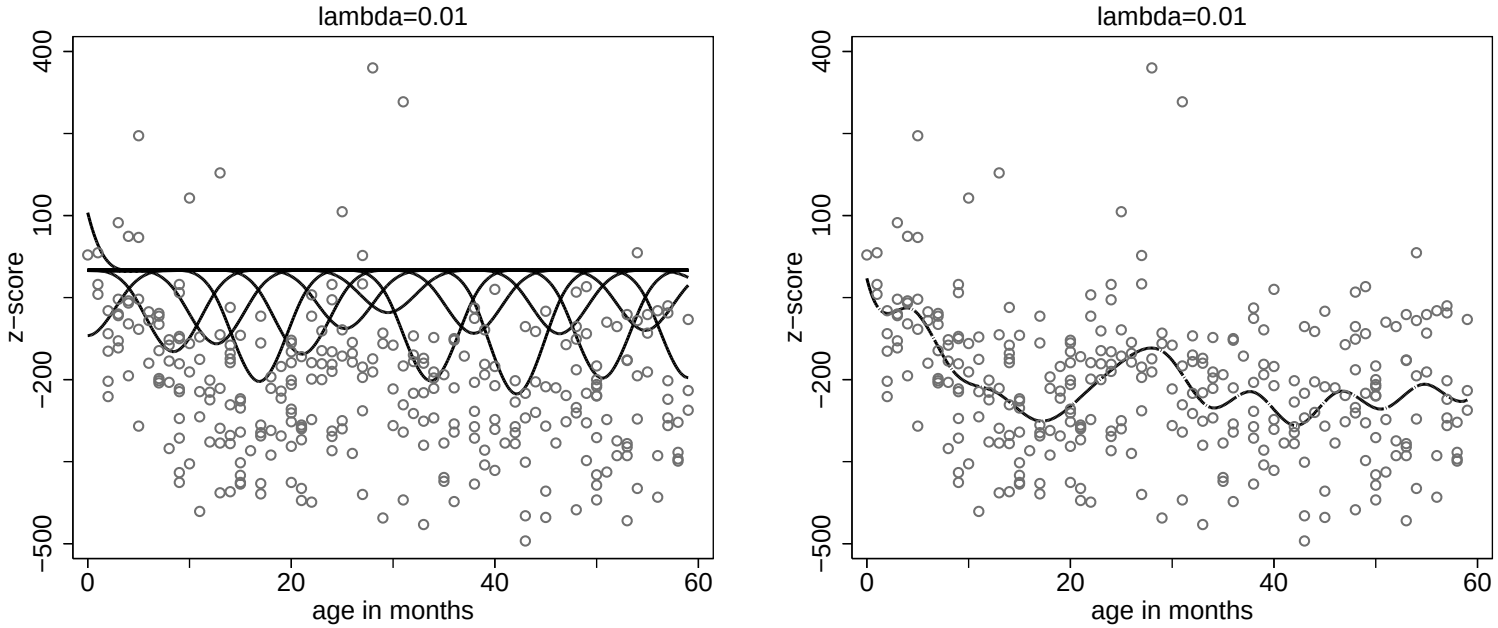


Figure 22: *Malnutrition in Tanzania: impact of the smoothing parameter on estimated P-splines with second order difference penalty.*

P- splines

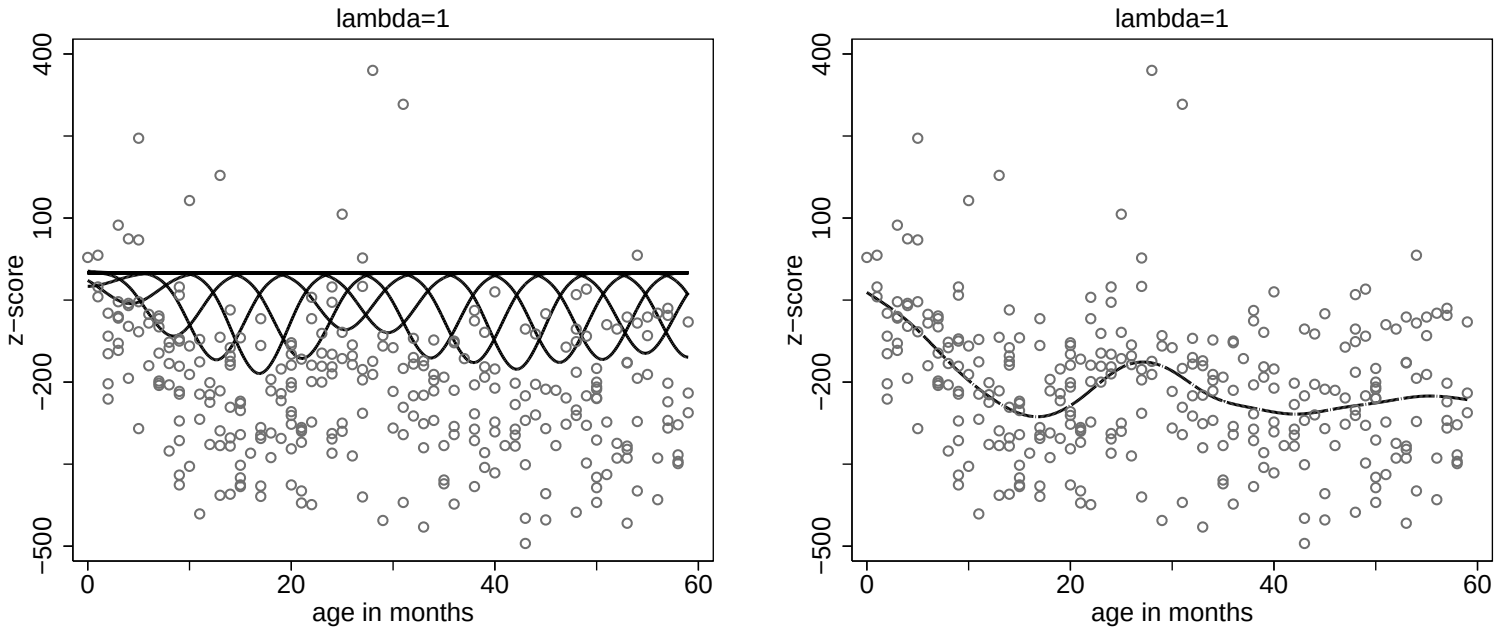


Figure 23: *Malnutrition in Tanzania: impact of the smoothing parameter on estimated P-splines with second order difference penalty.*

P- splines

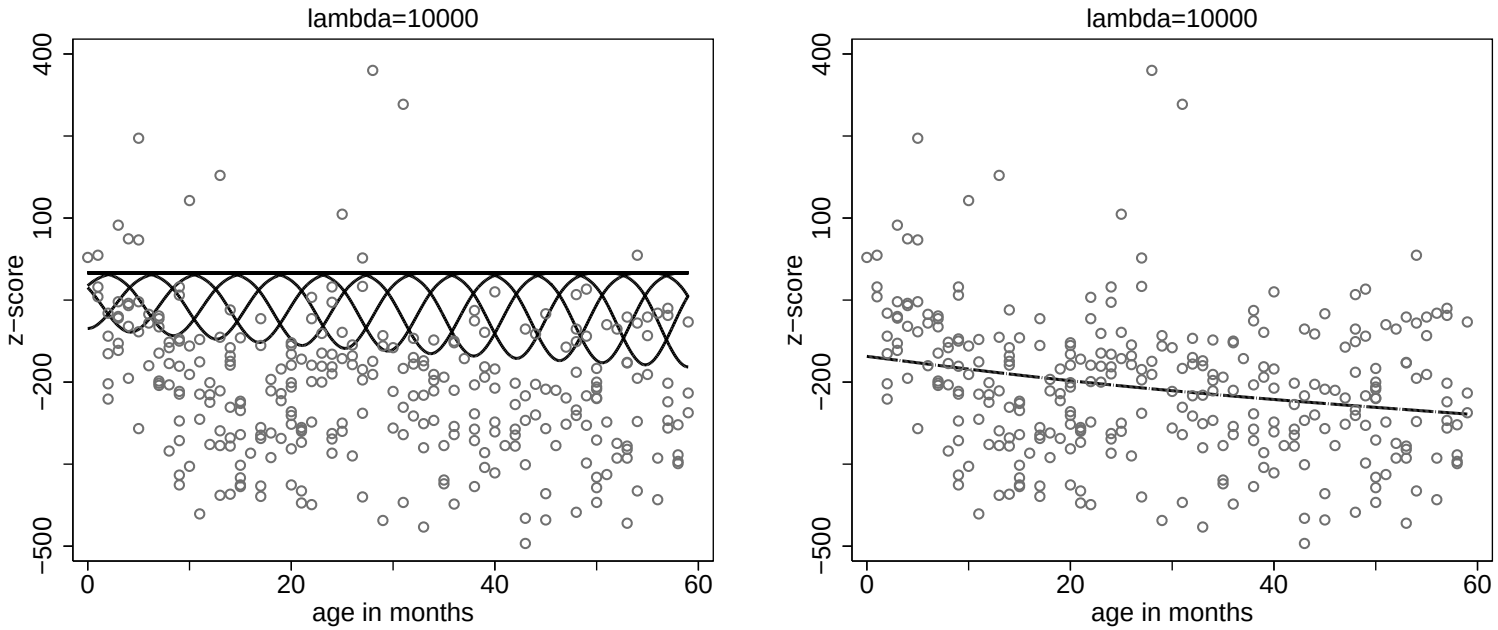


Figure 24: *Malnutrition in Tanzania: impact of the smoothing parameter on estimated P-splines with second order difference penalty.*

P- splines

Summary

Model

We approximate function f using polynomial splines so that we are able to write the nonparametric regression model as a linear model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Penalized Least Squares Criterion

Estimate $\boldsymbol{\gamma}$ by minimizing the penalized least squares criterion

$$\mathrm{PLS}(\lambda) = (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})'(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\mathbf{K}\boldsymbol{\gamma}.$$

P- splines

Smoothing parameter

The smoothing parameter $\lambda \geq 0$ controls the compromise between fidelity to the data and smoothness of the resulting function estimate. For splines in a TP basis representation, we penalize the sum of squared coefficients of the truncated powers. For B-splines, we construct the penalty based on the sum of squared differences of neighboring coefficients or based on the integral of the function's squared second derivative.

Penalized Least Squares Estimation

In either case, the penalized least squares estimate has the form

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{K})^{-1}\mathbf{Z}'\mathbf{y}.$$

Local smoothing procedures

Nearest neighbor estimates

For a time series y_t, $t = 1, \ldots, T$, running means of order 3 are, for example, defined by

$$\hat{y}_t = \frac{1}{3}(y_{t-1} + y_t + y_{t+1}),$$

with appropriate modifications at the boundaries.

Nearest neighbor estimates extend the concept of running means into a more general framework. In general, a nearest neighbor estimate is defined by

$$\hat{f}(z) = \underset{j \in N(z)}{\text{Ave}}\, y_j,$$

where Ave defines some averaging operator and $N(z)$ is an appropriate neighborhood of z.

Local smoothing procedures

The following averaging operators are often used for the determination of nearest neighbor estimates:

1. Arithmetic mean (running mean): Determine the arithmetic mean of the response variable in the neighborhood of z, i.e.:

$$\hat{f}(z) = \frac{1}{|N(z)|} \sum_{j \in N(z)} y_j,$$

   where $|N(z)|$ is the number of neighbors of z.

2. Median (running median): Determine the median of the response variables in the neighborhood of z, i.e.

$$\hat{f}(z) = \mathrm{Median}\{y_j, j \in N(z)\}.$$

Local smoothing procedures

3. Linear regression (running line): Estimate a linear regression based on the observations in the neighborhood of z and use the prediction from this model as the estimate, i.e.
$$\hat{f}(z) = \hat{\gamma}_{0,z} + \hat{\gamma}_{1,z} z,$$
where $\hat{\gamma}_{0,z}$ and $\hat{\gamma}_{1,z}$ are the least squares estimates using the data $\{(y_j, z_j), j \in N(z)\}$.

Commonly used neighborhood definitions are:

1. symmetric neighborhoods of order k or

2. neighborhoods that consist of k-nearest neighbors.

Local smoothing procedures

Example: Malnutrition in Zambia

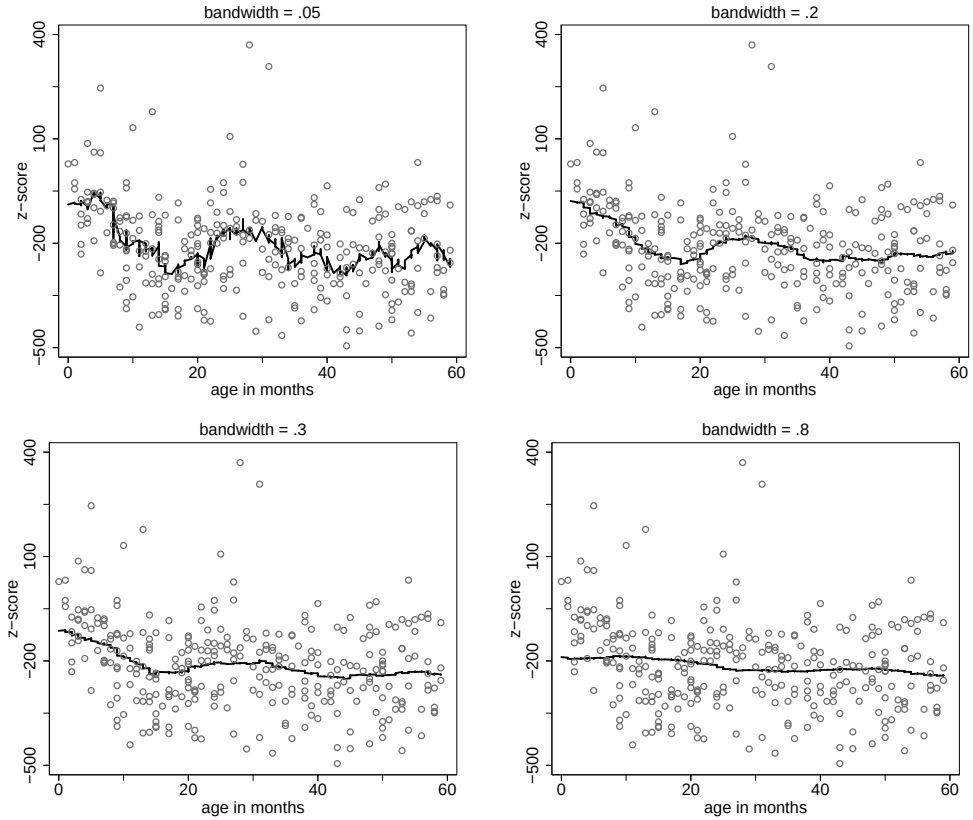


Figure 25: *Malnutrition in Tanzania: running mean with different bandwidths.*

Local smoothing procedures

Local Polynomial Regression and the Nadaraya-Watson Estimator

We consider the (local) approximation of an l-times continuously differentiable function $f(z_i)$ using a *Taylor series expansion* around z, yielding

$$f(z_i) \approx f(z) + (z_i - z)f'(z) + (z_i - z)^2 \frac{f''(z)}{2!} + \ldots + (z_i - z)^l \frac{f^{(l)}(z)}{l!}.$$

Hence, we approximate the function $f(z_i)$ with polynomials of the form $(z_i - z)^j$ in a neighborhood of z_i. The polynomials are weighted by the derivatives $f^{(j)}(z)/j!$ evaluated at the expansion point z.

Local smoothing procedures

We obtain

$$\begin{aligned} y_i &= f(z_i) + \varepsilon_i \\ &\approx f(z) + (z_i - z)f'(z) + (z_i - z)^2 \frac{f''(z)}{2!} + \ldots + (z_i - z)^l \frac{f^{(l)}(z)}{l!} + \varepsilon_i \\ &= \gamma_0 + (z_i - z)\gamma_1 + (z_i - z)^2\gamma_2 + \ldots + (z_i - z)^l\gamma_l + \varepsilon_i, \end{aligned}$$

for each observation (y_i, z_i) with expansion point z.

We obtain an implicit estimate for the function value $f(z)$ through $\hat{\gamma}_0 = \hat{f}(z)$, and more generally, we even obtain estimates for the derivatives through $j!\hat{\gamma}_j = \hat{f}^{(j)}(z)$.

Local smoothing procedures

Since the Taylor series approximation is only valid locally, i.e. close to the expansion point z, estimation is based on a weighted version of the residual sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{l} \gamma_j (z_i - z)^j \right)^2 w_\lambda(z, z_i)$$

with weights $w_\lambda(z, z_i)$. These are typically constructed based on the distances $|z_i - z|$ such that larger weights result for observations with a small distance.

Local smoothing procedures

A general class of such weights results with the use of *kernel functions* K in

$$w_\lambda(z, z_i) = K\left(\frac{z_i - z}{\lambda}\right)$$

Typical examples of kernel functions include

$$K(u) = \begin{cases} \frac{1}{2} & -1 \leq u \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{Uniform kernel,}$$

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & -1 \leq u \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{Epanechnikov kernel,}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \qquad \text{Gaussian kernel, ,}$$

Local smoothing procedures

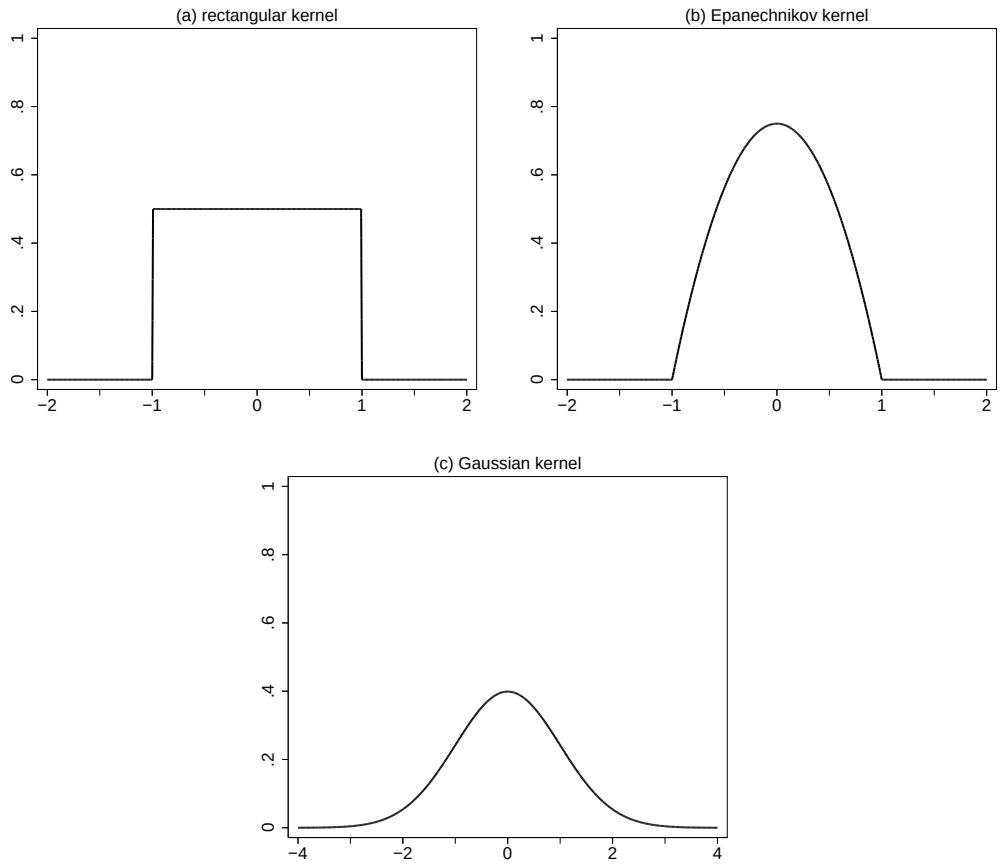


Figure 26: *Examples of kernel functions.*

Choosing the smoothing parameter

Choice Based on Optimality Criteria

E.g. cross validation criterion

$$\mathsf{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{(-i)}(z_i))^2$$

or

$$\mathsf{AIC} = n\log(\hat{\sigma}^2) + 2(\mathsf{df} + 1).$$

Mixed Model Representation of Penalization Approaches

Bayesian approach

Choosing the smoothing parameter

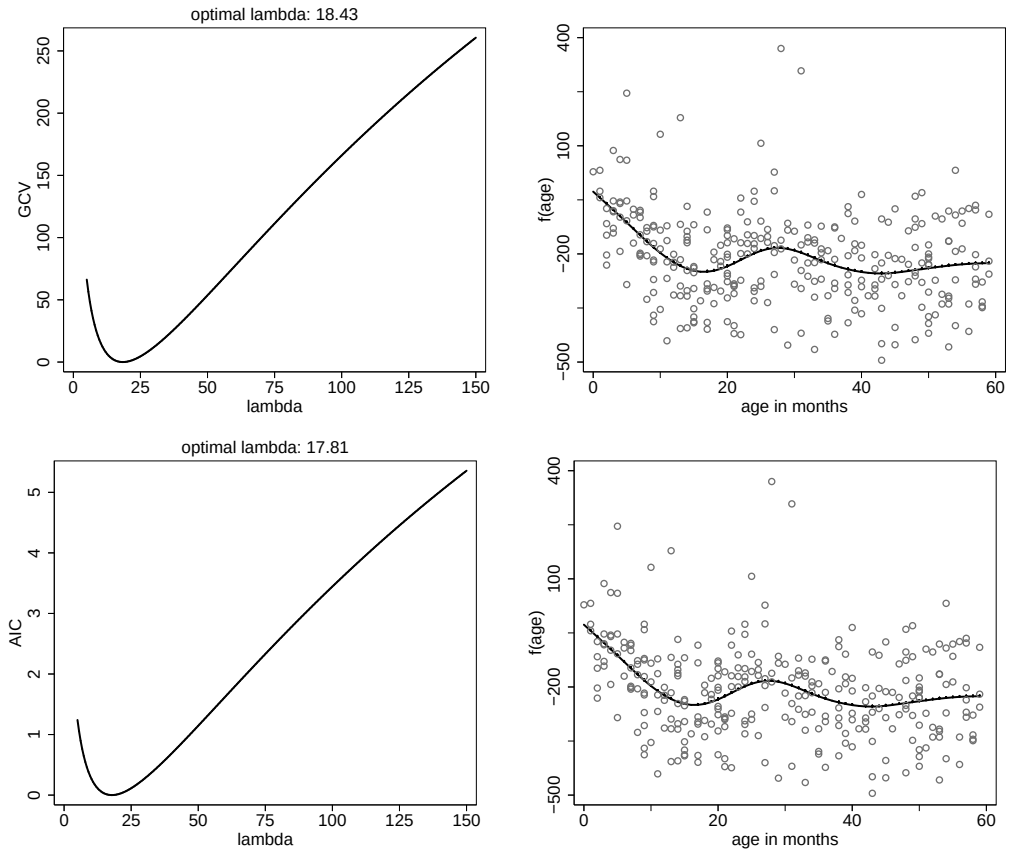


Figure 27: *Malnutrition in Tanzania: GCV and AIC (left panel) and cubic P-splines fits resulting with the corresponding optimal smoothing parameters (right panel).*

Non- and semiparametric regression
Additive models

Additive Models

Data

$(y_i, z_{i1}, \ldots, z_{iq}, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, with y and $x_1, \ldots, x_k$ as in the linear regression model and additional continuous covariates $z_1, \ldots, z_q$.

Model

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

The functions $f_1(z_1)$, ..., $f_q(z_q)$ are assumed to be smooth nonlinear effects of the continuous covariates $z_1, \ldots, z_q$. The same assumptions are made for the errors ε_i as with the classical linear model.

Additive Models

Modeling nonlinear effects

The functions f_j, $j = 1, \ldots, q$, will be approximated by

$$f_j(z_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(z_j),$$

where the basis functions B_l represent TP- or B-spline bases of polynomial splines.

The vector $\mathbf{f_j} = (\mathbf{f_j}(\mathbf{z_{1j}}), \ldots, \mathbf{f_j}(\mathbf{z_{nj}}))'$ of function values evaluated at the observed covariate values $z_{1j}, \ldots, z_{nj}$ can then be expressed as

$$\mathbf{f_j} = \mathbf{Z_j}\boldsymbol{\gamma}_\mathbf{j},$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd_j})'$ is the vector of regression coefficients.

Additive Models

The design matrix $\mathbf{Z_j}$ consists of the basis functions evaluated at the observed covariate values, i.e. $\mathbf{Z_j[i,l] = B_l(z_{ij})}$.

The additive model can then be written in matrix notation in the form

$$\mathbf{y} = \mathbf{Z_1}\boldsymbol{\gamma_1} + \ldots + \mathbf{Z_q}\boldsymbol{\gamma_q} + \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

where the design matrices $\mathbf{Z_1}, \ldots, \mathbf{Z_q}$ consist of the basis functions evaluated at the given covariate values. The design matrix $\mathbf{X}$ is constructed as in the linear model.

Additive Models

Example: Munich rent index

Assume the additive model

$$rentsqm = f_1(area) + f_2(yearc) + \beta_0 + \beta_1 glocation + \beta_2 tlocation + \varepsilon,$$

where the functions f_1 and f_2 are modeled by P-splines with 20 interior knots and second order difference penalties.

For the location, the model includes the dummy variables $glocation$ for good locations, and $tlocation$ for top location. The average location serves as the reference category.

Additive Models

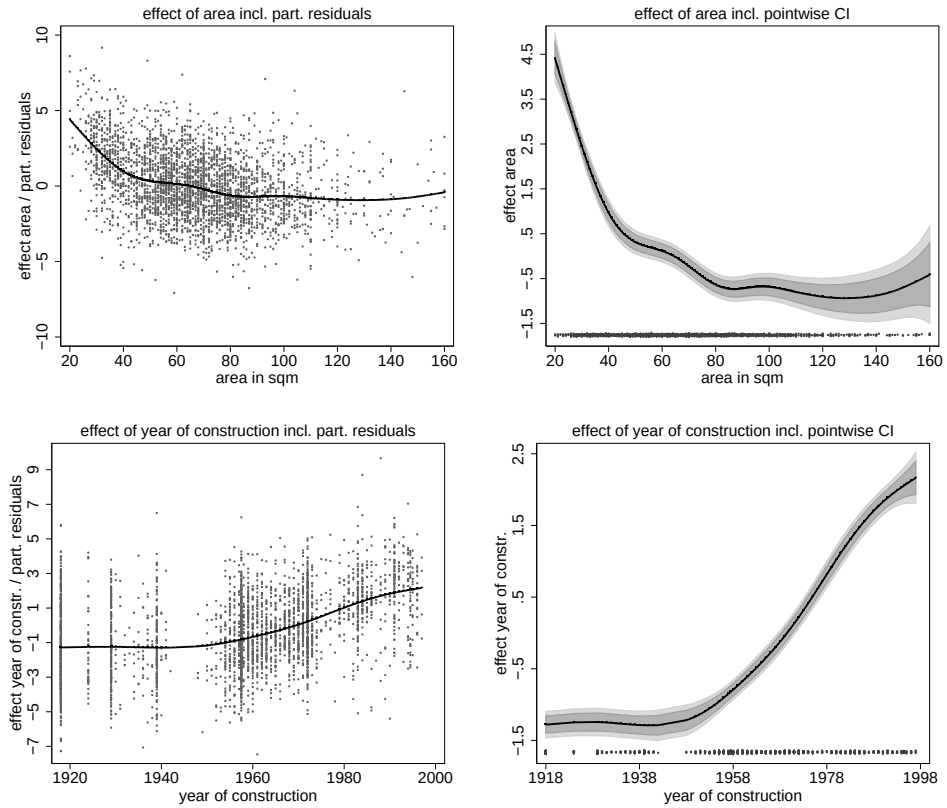


Figure 28: *Munich rent index: estimated nonlinear effects of area and year of construction.*

Quantile regression

Introduction

Basic Idea

Example: Munich rent index

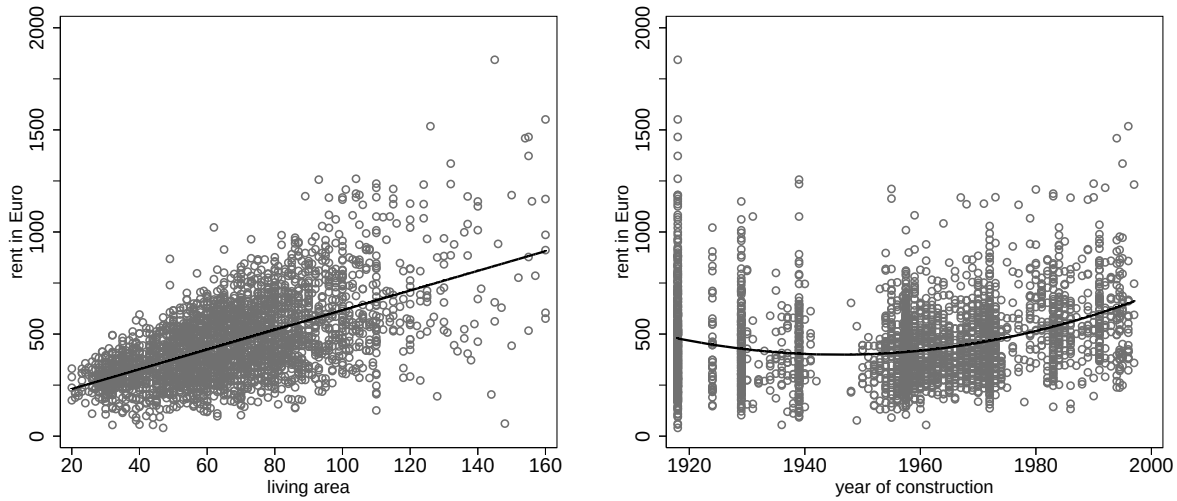


Figure 29: *Munich rent index: scatter plots of rents in Euro versus living area (left panel) and year of construction (right panel) together with a linear (left panel) and a quadratic (right panel) least squares fit.*

Basis Idea

- Focus on the quantiles of the response distribution and relate these quantiles to covariate effects.
- The basic idea is that a dense set of quantiles completely describes any given distribution.

Advantages

- Quantile regression allows investigation of covariate effects, not only on the mean of a response variable, but on the complete conditional distribution of the response given covariates.

- Quantile regression avoids some of the restrictive assumptions of mean regression models. More specifically, we will not require homoscedasticity or a specific type of distribution for the responses (or equivalently the error terms).

- In applications, there often is a genuine interest in regression quantiles that describe "extreme" observations in terms of covariate.

Theoretical Quantiles

The theoretical quantiles q_τ, $\tau \in (0,1)$, of a random variable y are commonly and implicitly defined by the equations

$$\mathsf{P}(y \leq q_\tau) \geq \tau \quad \text{and} \quad \mathsf{P}(y \geq q_\tau) \geq 1 - \tau,$$

i.e. the probability of observing a value below (or equal to) q_τ should be (at least) τ while the probability of observing a value above (or equal to) q_τ should be (at least) $1 - \tau$.

Theoretical Quantiles

Reformulate this implicit definition as the optimization problem

$$q_\tau = \arg\min_q \mathsf{E}\left(w_\tau(y,q)|y-q|\right)$$

with weights

$$w_\tau(y,q) = \begin{cases} 1-\tau & y < q \\ 0 & y = q \\ \tau & y > q. \end{cases}$$

Theoretical Quantiles

If y is continuous with strictly increasing cumulative distribution function $F(y)$ and density $f(y)$, the theoretical quantile is unique and is given by the inverse of the cumulative distribution function evaluated at τ, i.e.

$$q_\tau = F^{-1}(\tau) \qquad \text{and} \qquad F(q_\tau) = \tau.$$

The function $Q(\tau) = F^{-1}(\tau) = q_\tau$ is also called the quantile function of the distribution of y.

Empirical Quantiles

Empirical quantiles correspond to the estimated quantiles $\hat{q}_\tau$ determined from an i.i.d. sample $y_1, \ldots, y_n$ of observations from the corresponding distribution.

At least a fraction of τ observations should be smaller or equal than $\hat{q}_\tau$ and at least a fraction of $1-\tau$ observations should be larger or equal than $\hat{q}_\tau$, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \leq \hat{q}_\tau) \geq \tau \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} I(y_i \geq \hat{q}_\tau) \geq 1-\tau,$$

where $I(\cdot)$ denotes the indicator function.

An equivalent definition is given as the solution of an optimization criterion

$$\hat{q}_\tau = \arg\min_q \sum_{i=1}^{n} w_\tau(y_i, q)|y_i - q|.$$

Quantile regression

Linear quantile regression

Definition

Model

$$y_i = \mathbf{x_i'}\beta_\tau + \varepsilon_{\mathbf{i}\tau}, \quad \mathbf{i = 1, \ldots, n},$$

with assumptions

1. $F_{\varepsilon_{i\tau}}(0) = \tau$.

2. $\varepsilon_{1\tau}, \ldots, \varepsilon_{n\tau}$ are independent.

Estimation of Regression Coefficients

The regression coefficients β_τ are determined by minimizing

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})|y_i - \eta_{i\tau}|,$$

where $\eta_{i\tau} = \mathbf{x_i'}\beta_\tau$.

Software

- R package `quantreg`: Implements linear programming for determining $\hat{\beta}_\tau$.
- R package `mboost`: Implements functional gradient descent boosting for determining $\hat{\beta}_\tau$. for details.
- Software package `BayesX` (see also the R interface `R2BayesX`).

Example: Munich rent index

Models

- Quantile regression: $y = \mathbf{x}'\beta_\tau + \varepsilon_\tau$ as introduced in this chapter;
- Homoscedastic linear model with i.i.d. Gaussian error terms: $y = \mathbf{x}'\beta + \varepsilon$, $\varepsilon \sim \mathrm{N}(0, \sigma^2)$;
- Heteroscedastic linear model with independent Gaussian error terms: $y = \mathbf{x}'\beta + \exp(\mathbf{x}'\tilde{\alpha})\varepsilon$, $\varepsilon \sim \mathrm{N}(0, 1)$.

In each case determine the 11 quantiles $\tau = 0.01, 0.1, 0.2, \ldots, 0.8, 0.9, 0.99$ of the net rent distribution.

Example: Munich rent index

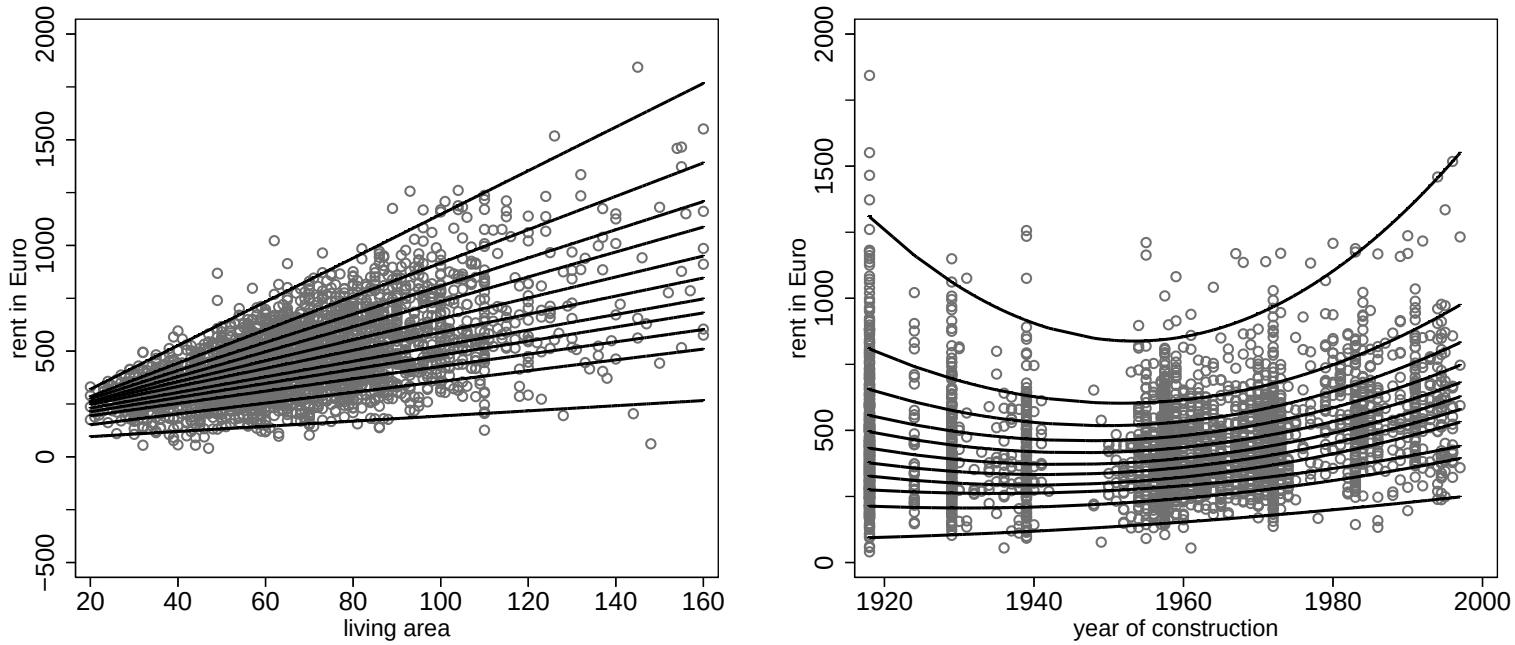


Figure 30: *Munich rent index: scatter plots of the rents in Euro versus living area (left column) and year of construction (right panel) together with linear/quadratic quantile regression fits for 11 quantiles.*

Example: Munich rent index

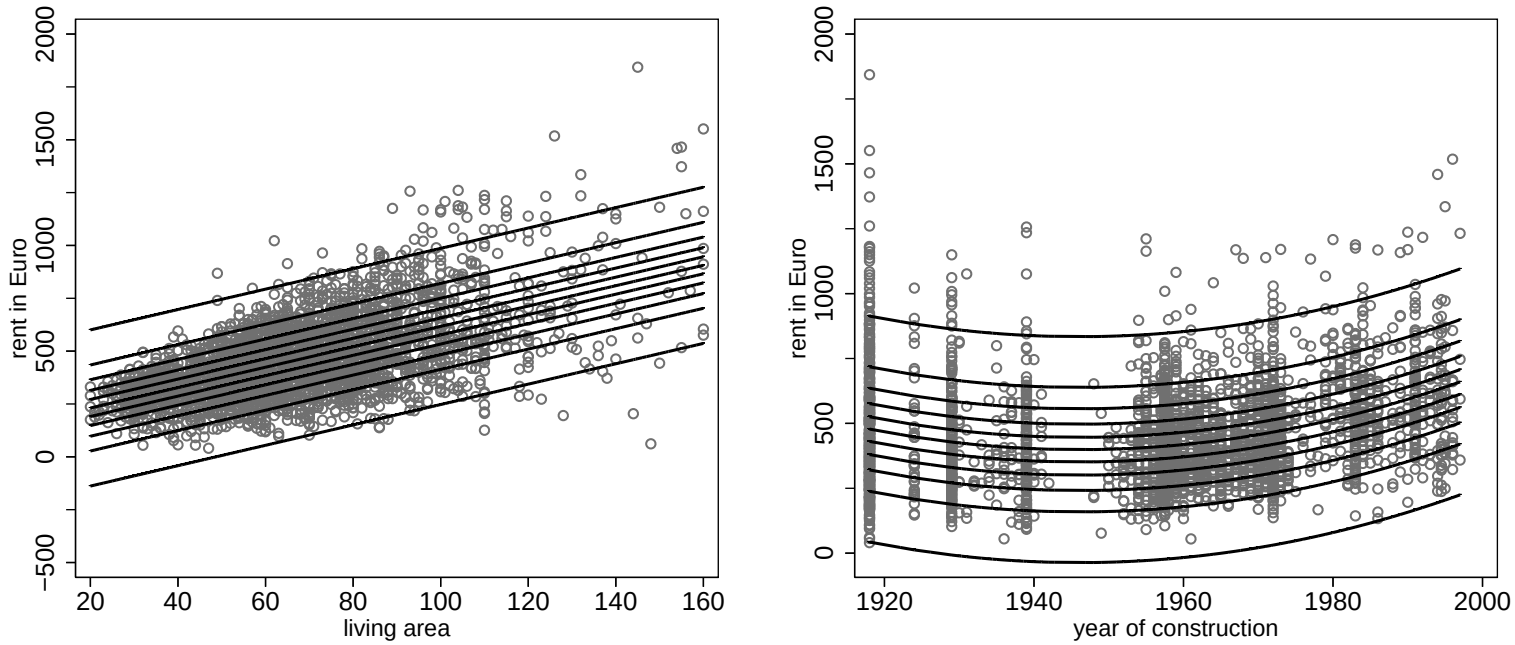


Figure 31: *Munich rent index: scatter plots of the rents in Euro versus living area (left column) and year of construction (right panel) together with quantiles determined from a homoscedastic linear model.*

Example: Munich rent index

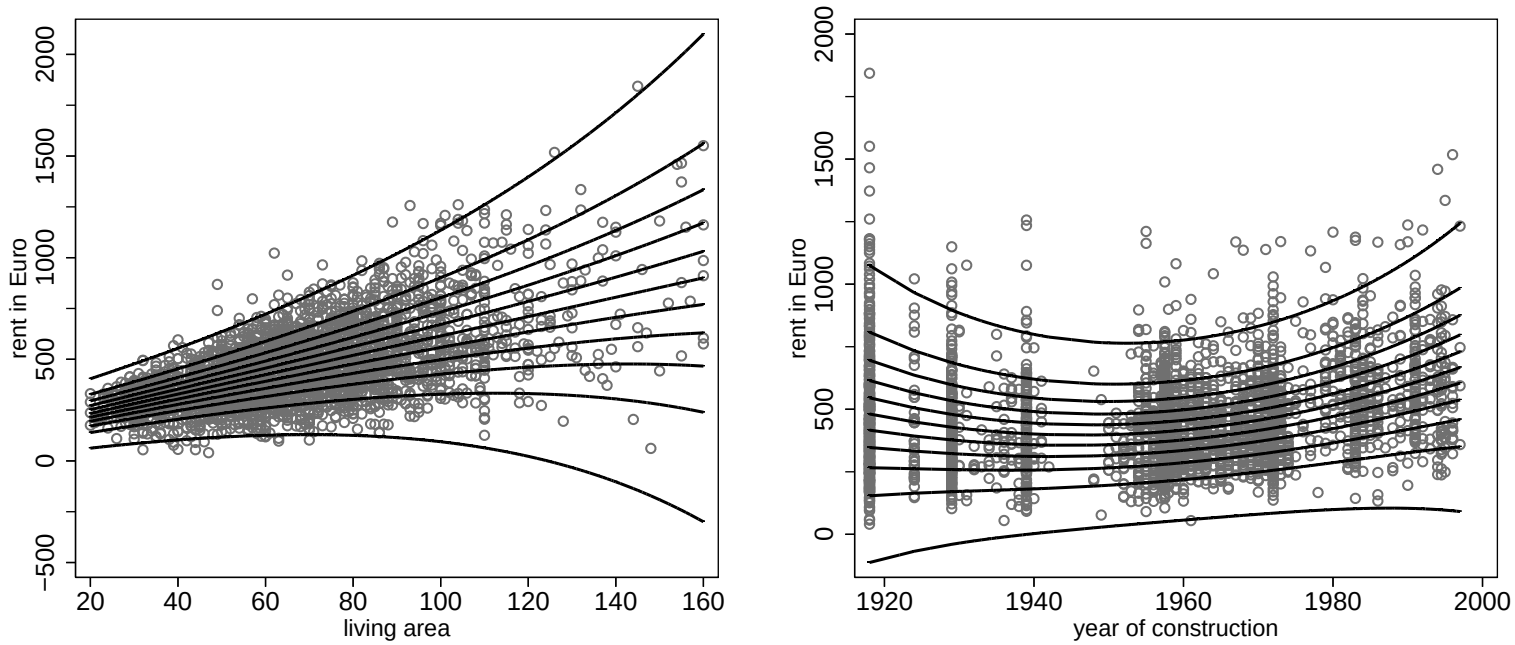


Figure 32: *Munich rent index: scatter plots of the rents in Euro versus living area (left column) and year of construction (right panel) together with quantiles determined from a heteroscedastic linear model.*

Example: Munich rent index

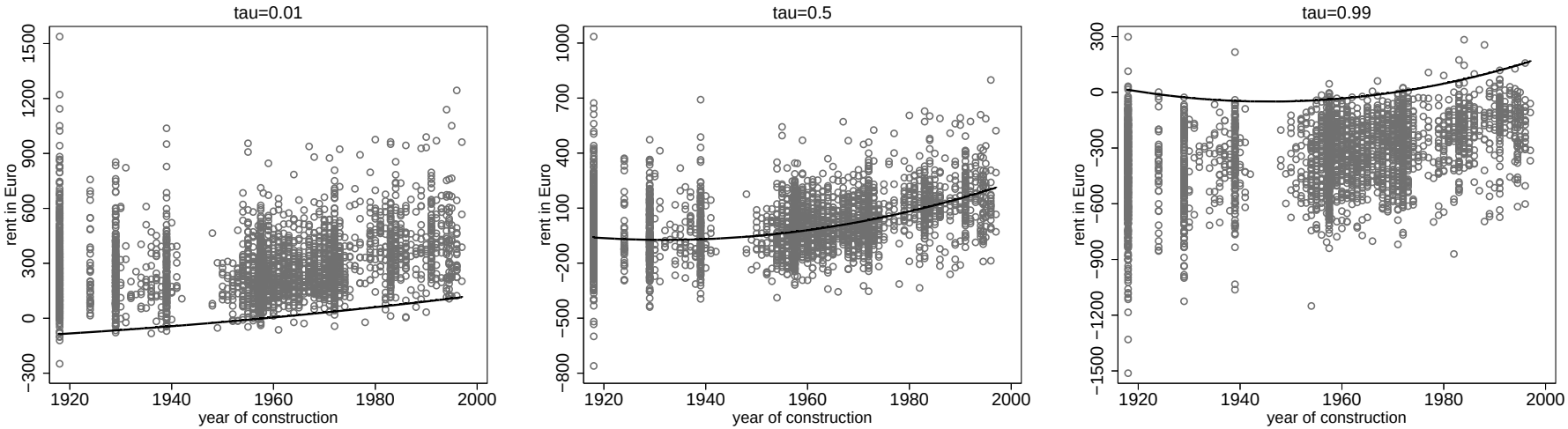


Figure 33: *Munich rent index: estimated effects of year of construction together with partial residuals for different quantiles.*

Example: Munich rent index

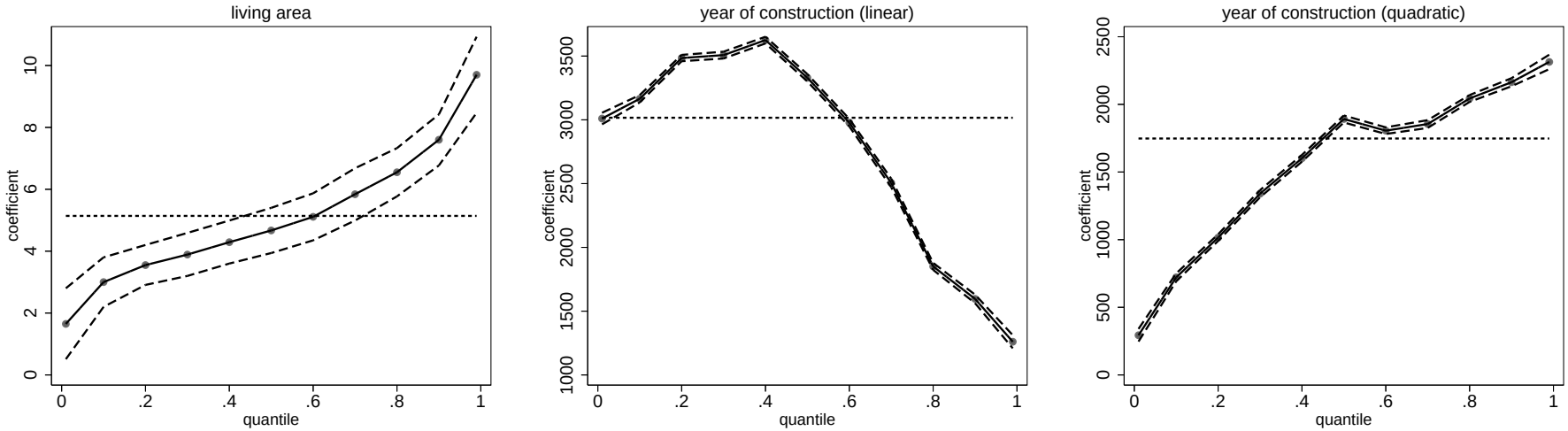


Figure 34: *Munich rent index: paths of estimated coefficients (solid line) together with 95% confidence intervals (dashed lines) obtained from inverting a rank-based test for various quantiles τ. The horizontal dotted line corresponds to the least squares estimate.*

Properties

- *Invariance under monotonic transformations:* If $\mathbf{x}'\hat{\beta}_\tau$ is an estimate for the τ-quantile of the distribution of the response y, given covariates $\mathbf{x}$, then for any monotonically increasing transformation h the transformed estimate $h(\mathbf{x}'\hat{\beta}_\tau)$ is an estimate for the τ-quantile of the distribution of $h(y)$.

- *Asymptotic distribution:* In case of i.i.d. errors, the asymptotic distribution of $\hat{\beta}_\tau$ is given by

$$\hat{\beta}_\tau \overset{a}{\sim} \mathrm{N}\left(\beta_\tau, \frac{\tau(1-\tau)}{\mathbf{f}_{\varepsilon_\tau}(0)^2}(\mathbf{X}'\mathbf{X})^{-1}\right).$$

Properties

- In theory, the quantiles of the distribution of a response should be ordered such that
$$\mathbf{x}'\beta_{\tau_1} \leq \mathbf{x}'\beta_{\tau_2} \quad \text{for} \quad \tau_1 \leq \tau_2,$$
holds for any covariate vector $\mathbf{x}$. It can be shown that the ordering is preserved for the average covariate vector
$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i},$$
when replacing the theoretical quantiles with estimated quantiles, i.e.
$$\bar{\mathbf{x}}'\hat{\beta}_{\tau_1} \leq \bar{\mathbf{x}}'\hat{\beta}_{\tau_2} \quad \text{for} \quad \tau_1 \leq \tau_2.$$

- In general this results will not transfer to arbitrary vectors $\mathbf{x}$ and will not even hold for all observed covariate vectors.

Quantile regression

Bayesian quantile regression

Model

Observation model

The observations y_i, $i = 1, \ldots, n$, are conditionally independent following an asymmetric Laplace distribution, i.e. y_i i.i.d. $\mathrm{ALD}(\mathbf{x_i'}\beta_\tau, \sigma^{\mathbf{2}}, \tau)$. The scale mixture representation of the asymmetric Laplace distribution yields

$$y_i \,|\, z_i, \beta_\tau, \sigma^{\mathbf{2}} \sim \mathbf{N}(\mathbf{x_i'}\beta_\tau + \xi \mathbf{z_i}, \sigma^{\mathbf{2}}/\mathbf{w_i}),$$

where

$$\xi = \frac{1-2\tau}{\tau(1-\tau)}, \qquad w_i = \frac{1}{\delta^2 z_i}, \qquad \delta^2 = \frac{2}{\tau(1-\tau)}.$$

Priors

$$\begin{aligned} \beta_\tau &\propto \mathrm{const} \\ z_i \,|\, \sigma^2 &\sim \mathrm{Expo}(1/\sigma^2) \\ \sigma^2 &\sim \mathrm{IG}(a, b) \end{aligned}$$

Gibbs sampler

- *Full conditional for the regression coefficients:* $\beta_\tau \mid \cdot \, \mathrm{N}(\mu_{\beta_\tau}, \mathbf{\Sigma}_{\beta_\tau})$ with

$$\mathbf{\Sigma}_{\beta_\tau} = \sigma^{\mathbf{2}}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}, \qquad \mu_{\beta_\tau} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{W}(\mathbf{y} - \xi\mathbf{z}),$$

  where $\mathbf{W} = \mathrm{diag}(\mathbf{w_1}, \ldots, \mathbf{w_n})$ and $\mathbf{z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})'$.

- *Full conditional for the scale parameters:*

$$z_i^{-1} \mid \cdot \sim \mathrm{InvGauss}\left(\sqrt{\frac{\xi^2 + 2\delta^2}{(y_i - \mathbf{x_i'}\beta_\tau)^{\mathbf{2}}}}, \frac{\xi^2 + 2\delta^2}{\sigma^2\delta^2}\right),$$

- Full conditional for the error variance:

$$\sigma^2 \mid \cdot \sim \mathrm{IG}\left(a + \frac{3n}{2}, b + \frac{1}{2}\sum_{i=1}^{n} w_i(y_i - \mathbf{x_i'}\beta_\tau - \xi\mathbf{z_i})^{\mathbf{2}} + \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{z_i}\right).$$

Quantile regression

Additive quantile regression

Models

- An approach for estimating nonlinear quantile functions $f_\tau(z_i)$ of continuous covariates z_i in the scatter plot smoothing model

$$y_i = f_\tau(z_i) + \varepsilon_{i\tau}$$

relies on the fitting criterion

$$\arg\min_{f_\tau} \sum_{i=1}^{n} w_\tau(y_i, f_\tau(z_i))|y_i - f_\tau(z_i)| + \lambda V(f'_\tau).$$

Models

- $V(f'_\tau)$ denotes the total variation of the derivative f'_τ defined as

$$V(f'_\tau) = \sup \sum_{i=1}^{n} |f'_\tau(z_{i+1}) - f'_\tau(z_i)|,$$

where the $\sup$ is taken over all partitions $a \leq z_1 < \ldots < z_n < b$. For twice continuously differentiable functions f_τ, the total variation penalty can be written as

$$V(f'_\tau) = \int |f''_\tau(z)| dz$$

- The approach can be extended to additive models. However, it is typically difficult to determine the smoothing parameters along with the estimated functions in an automatic and data-driven way.

Models

- For Bayesian additive quantile regression, we can easily extend the Gibbs sampler outlined in the previous section.

- Most importantly, the full conditionals for nonparametric effects represented as $\mathbf{V_j}\boldsymbol{\gamma}_\mathbf{j}$ are now given by

$$\gamma_j \,|\, \cdot \;\sim \mathrm{N}(\mathbf{m_j}, \boldsymbol{\Sigma}_\mathbf{j})$$

with expectation and covariance matrix

$$\begin{aligned}
\mathbf{m_j} &= \mathsf{E}(\gamma_j \,|\, \cdot) = \left(\mathbf{V_j'WV_j} + \frac{\sigma^\mathbf{2}}{\tau_\mathbf{j}^\mathbf{2}}\mathbf{K_j}\right)^{-1} \mathbf{V_j W'}(\mathbf{y} - \eta_{-\mathbf{j}}^{struct} - \xi\mathbf{z}) \\
\boldsymbol{\Sigma}_\mathbf{j} &= \mathsf{Cov}(\gamma_j \,|\, \cdot) = \sigma^2 \left(\mathbf{V_j'WV_j} + \frac{\sigma^\mathbf{2}}{\tau_\mathbf{j}^\mathbf{2}}\mathbf{K_j}\right)^{-1}
\end{aligned}$$

where $\mathbf{W} = \mathsf{diag}(\mathbf{w_1}, \ldots, \mathbf{w_n})$ and $\mathbf{z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})'$.

- Similarly, the full conditional for the error variance has to be adjusted while the full conditionals for the smoothing variances remain unchanged.

References

Nonparametric density estimation

- Lang, S., (2004): Lecture notes Computerintensive Verfahren (in German).
- Pruscha, H., 2000: Vorlesungen über Mathematische Statistik. Teubner, Stuttgart (in German).

Semiparametric regression

- Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013): Regression: models, methods and applications. Springer

- Wood, S. N. (2006): Generalized Additive Models: An Introduction with R, Chapman & Hall / CRC.

Quantile regression

- Klein, N., Kneib, T., Lang, S. and Sohn, A. (2015): Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany. The Annals of Applied Statistics, 9, 1024-1052.

- Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013): Regression: models, methods and applications. Springer

- Koenker, R. (2005): Quantile Regression. New York, Cambridge University Press.