Advanced Regression Analysis

Stefan Lang

June 2019

Contents

1. Overview

2. Linear models revisited

3. Distributional regression

4. Scatterplot smoothers and additive models

5. 2D and spatial smoothing

6. Mixed models

7. Quantileregression

Advanced Regression Analysis

Overview

Pioneer of regression analysis: Francis Galton

- Sir Francis Galton (1822–1911) is regarded as the pioneer of regression analysis.
- At the end of the nineteenth century, he examined the *relationship* between body heights of children and the average body height of both parents, see Table 1 and Figure 1.
- To adjust for the natural height differences across gender, the body height of women was multiplied by a factor of 1.08.

Pioneer of regression analysis: Francis Galton

height of children	average height of parents											
	64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73.0	Total
73.7	0	0	0	0	0	0	5	3	2	4	0	14
73.2	0	0	0	0	0	3	4	3	2	2	3	17
72.2	0	0	1	0	4	4	11	4	9	7	1	41
71.2	0	0	2	0	11	18	20	7	4	2	0	64
70.2	0	0	5	4	19	21	25	14	10	1	0	99
69.2	1	2	7	13	38	48	33	18	5	2	0	167
68.2	1	0	7	14	28	34	20	12	3	1	0	120
67.2	2	5	11	17	38	31	27	3	4	0	0	138
66.2	2	5	11	17	36	25	17	1	3	0	0	117
65.2	1	1	7	2	15	16	4	1	1	0	0	48
64.2	4	4	5	5	14	11	16	0	0	0	0	59
63.2	2	4	9	3	5	7	1	1	0	0	0	32
62.2	-	1	0	3	3	0	0	0	0	0	0	7
61.7	1	1	1	0	0	1	0	1	0	0	0	5
Total	14	23	66	78	211	219	183	68	43	19	4	928

Table 1: *Galton heredity data: contingency table between the height of 928 adult children and the average height of their 205 set of parents.*

Pioneer of regression analysis: Francis Galton

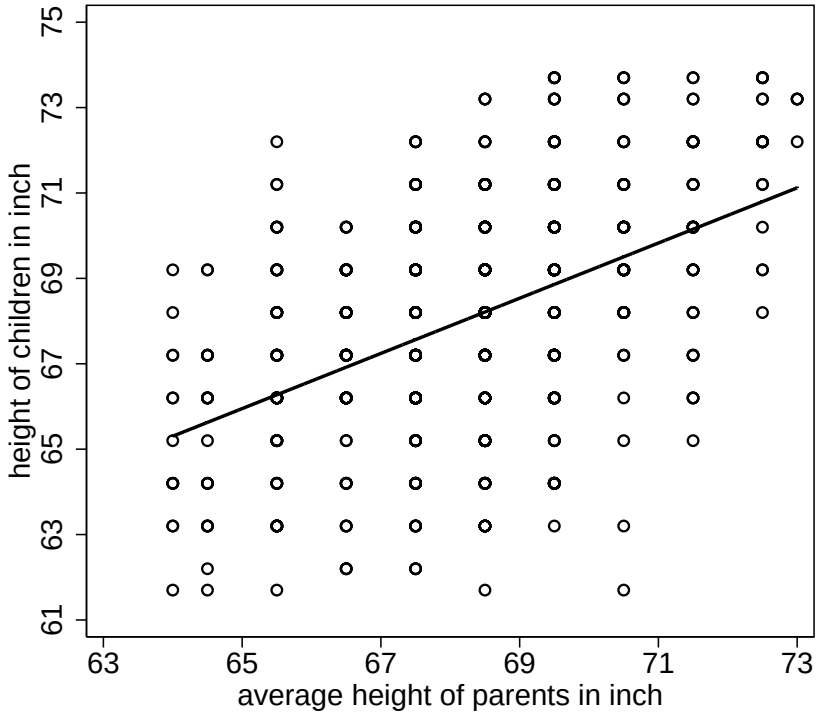


Figure 1: *Galton heredity data: scatter plot including a regression line between the height of children and the average height of their parents.*

Pioneer of regression analysis: Francis Galton

- Column-wise, i.e. for given average heights of the parents, the heights of the adolescents approximately follow a normal distribution.
- The normal distributions in each column have a common variance.
- When examining the relationship between the height of the children and the average height of the parents, an approximate linear trend was found with a slope of 2/3.
- There is a tendency towards the population average, and Galton referred to this as *regression* towards the mean.

Pioneer of regression analysis: Francis Galton

- In modern terms, Galton studied the systematic influence of the *explanatory variable* $x =$ "average size of the parents" on the *response variable* $y =$ "height of grown-up children".

- The fact that the linear relationship is not exact, but rather depends on random errors, is a main characteristic for regression problems.

- Galton assumed the most simple regression model,

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

  where the systematic component $\beta_0 + \beta_1 x$ is linear and ε constitutes the random error.

First steps

- Univariate distributions of the variables
  - Goals: summary and exploration of the distribution of the variables, identification of extreme values and outliers, identification of incorrect variable coding.
  - Use summary statistics and visualize data using histograms, box plots etc.

- Graphical Association Analysis
  - Use scatterplots and clustered scatter plots for continuous covariates
  - Use e.g. boxplots for categorical covariates

Example: Munich rental guide

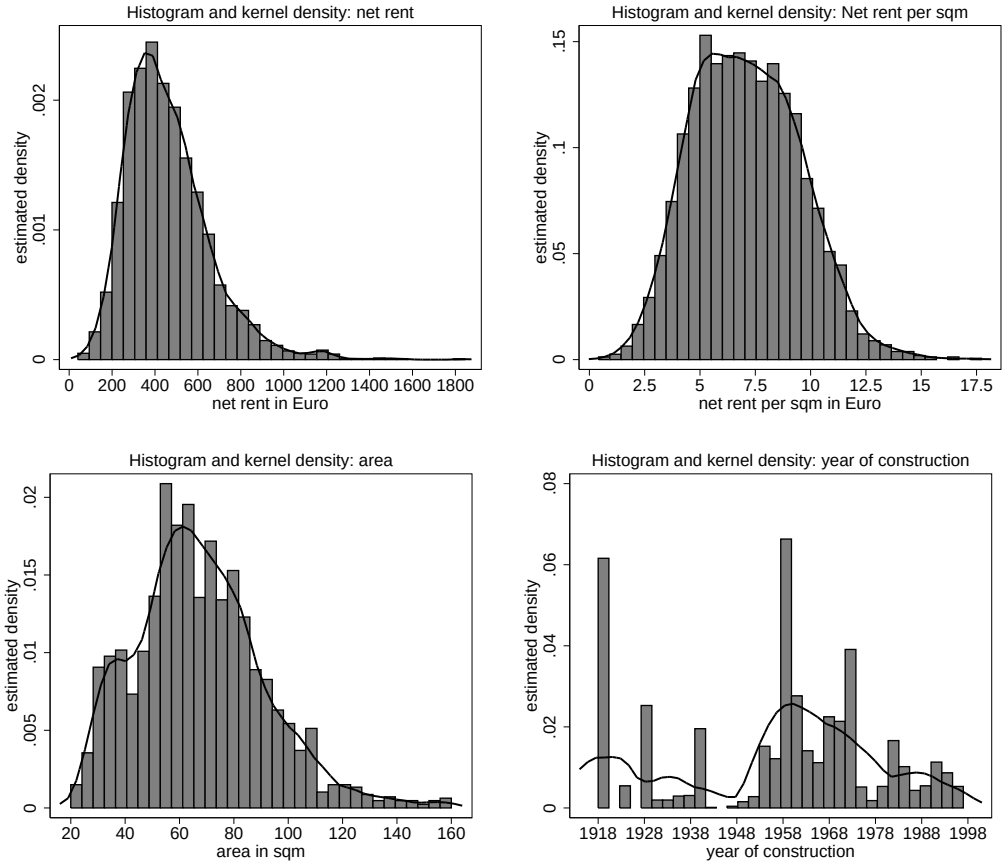


Figure 2: *Munich rent index: histograms and kernel density estimators for the continuous variables* $rent$, $rentsqm$, $area$ *and* $yearc$.

Example: Munich rental guide

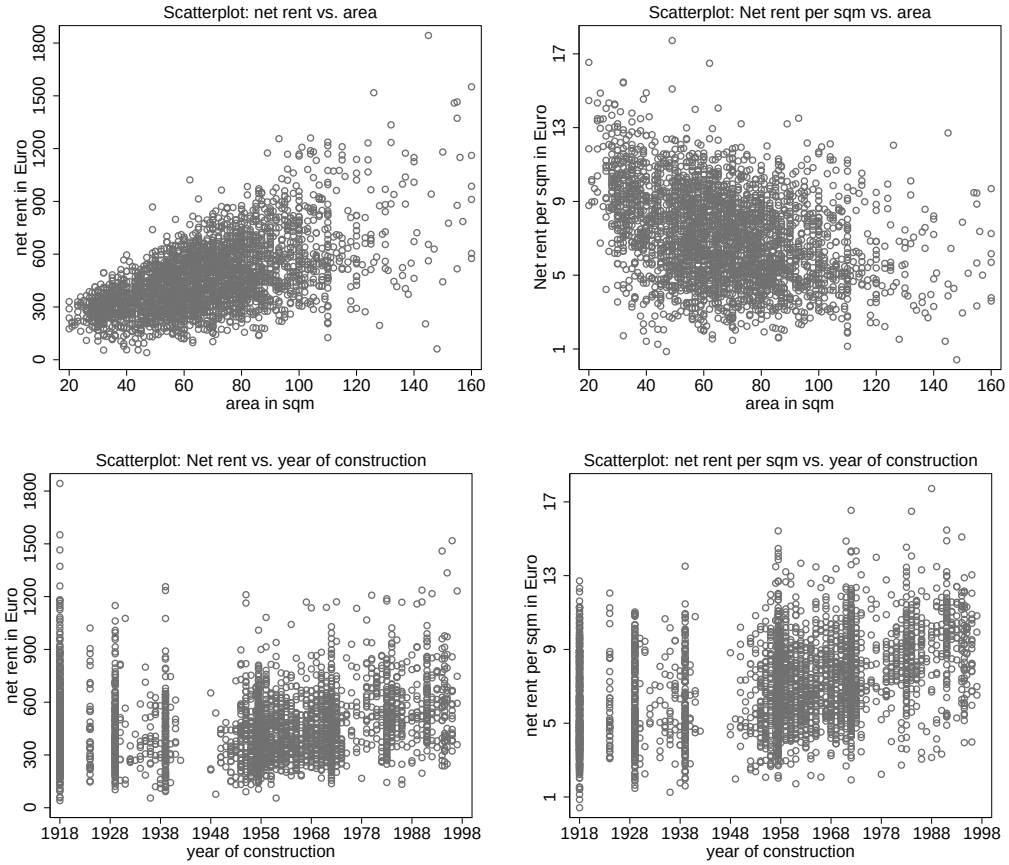


Figure 3: *Munich rent index: scatter plots between net rent (left) / net rent per sqm (right) and the covariates area and year of construction.*

Example: Munich rental guide

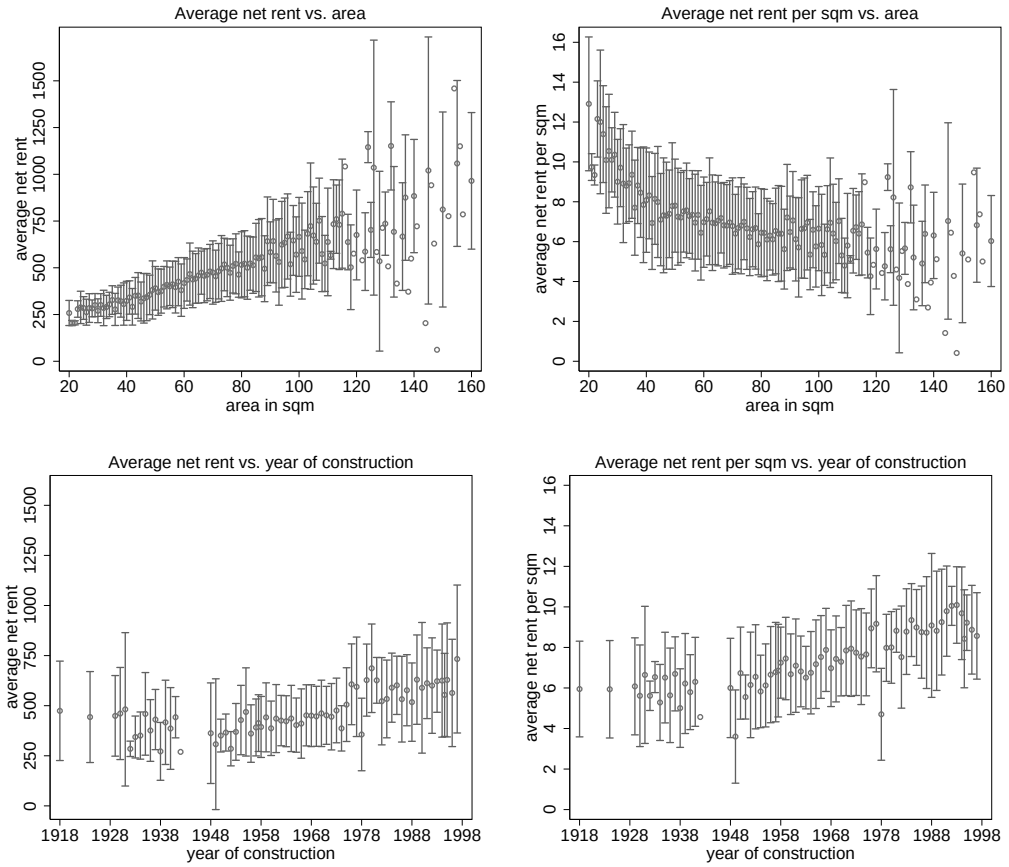


Figure 4: *Munich rent index: average net rent (left) and net rent per sqm (right) plus/minus one standard deviation versus area and year of construction.*

Simple Linear Regression Model

- **Data:** (y_i, x_i), $i = 1, \ldots, n$, with continuous variables y and x.

- **Model:**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \quad i = 1, \ldots, n\,.$$

The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) with

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

- We can interpret the estimated regression line $\hat{f}(x) = \hat{\beta}_\mathbf{0} + \hat{\beta}_\mathbf{1}\mathbf{x}$ as an estimate $\widehat{\mathsf{E}(y|x)}$ for the conditional expected value of y given the covariate value x. We can, thus, predict y through $\hat{y} = \hat{\beta}_\mathbf{0} + \hat{\beta}_\mathbf{1}\mathbf{x}$.

Example: Munich rental guide

- Assume first the relationship

$$rentsqm_i = f(area_i) + \varepsilon_i = \beta_0 + \beta_1 \cdot area_i + \varepsilon_i \tag{1}$$

  or alternatively

$$\mathsf{E}(rentsqm \mid area) = f(area) = \beta_0 + \beta_1 \cdot area.$$

- Using the method of least squares we obtain

$$\hat{f}(area) = 10.5 - 0.041 \cdot area.$$

Example: Munich rental guide

- A better fit can be achieved by defining the new explanatory variable

$$x = \frac{1}{area},$$

that yields a simple linear regression of the form

$$rentsqm_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 \frac{1}{area_i} + \varepsilon_i. \tag{2}$$

- (2) is again a simple linear regression and we can still use the method of least squares to estimate the parameters β_0 and β_1 of the function

$$f(area) = \beta_0 + \beta_1 \cdot \frac{1}{area}.$$

- We obtain $\hat{f}(area) = 5.44 + 138.32 \cdot \frac{1}{area}$.

Example: Munich rental guide

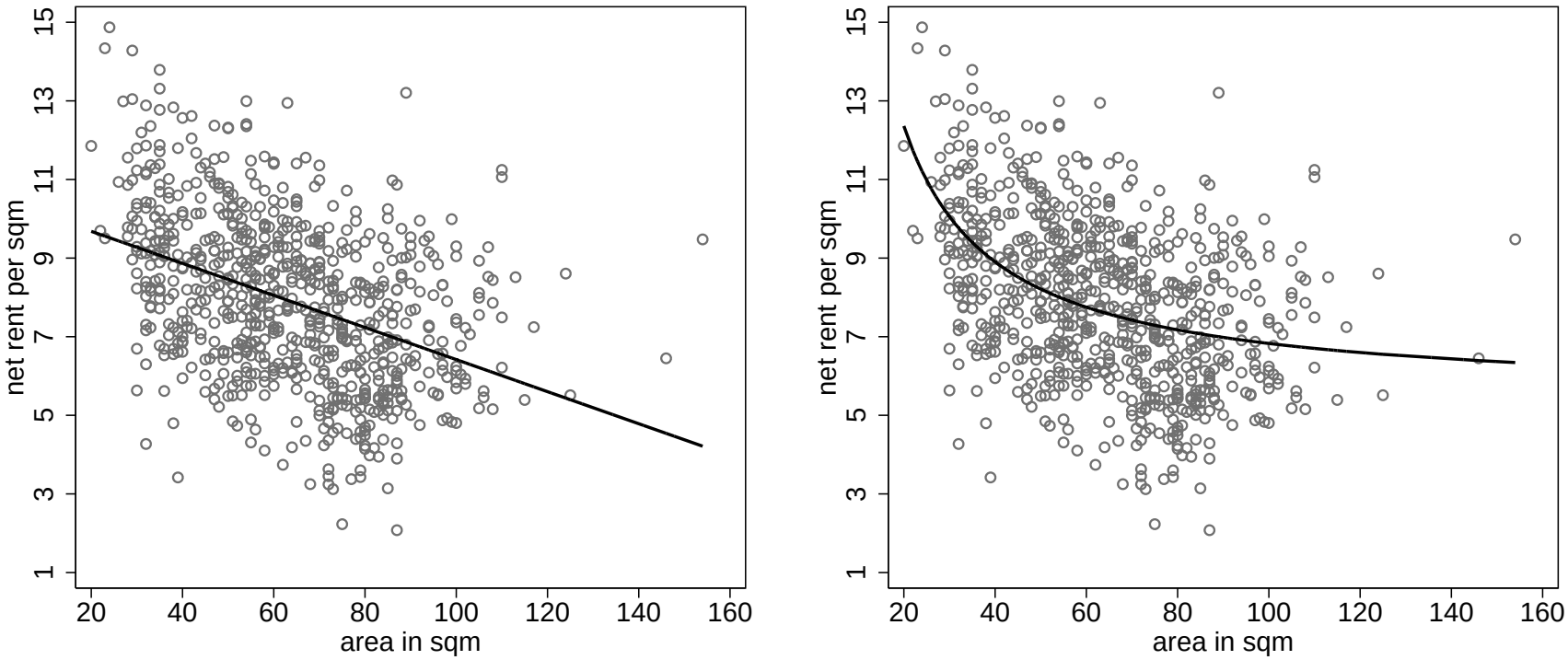


Figure 5: *Munich rent index: scatter plots between net rent per square meter and area. Included is the estimated effect $\hat{f}$ of living area for a linear (left) and reciprocal (right) area effect.*

Classical linear regression model

- **Data** $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, for a continuous variable y and continuous or appropriately coded categorical regressors $x_1, \ldots, x_k$.

- **Model**

$$y_i = f(x_1, \ldots, x_k) + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \,, \quad i = 1, \ldots, n \,.$$

  The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) with

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

- The estimated linear function $\hat{f}(x_1, \ldots, x_k) = \hat{\beta}_\mathbf{0} + \hat{\beta}_\mathbf{1}\mathbf{x_1} + \ldots + \hat{\beta}_\mathbf{k}\mathbf{x_k}$ can be used as an estimator $\widehat{\mathsf{E}}(y|x_1, \ldots, x_k)$ for the conditional expected value of y given the covariates $x_1, \ldots, x_k$. As such it can be used to predict y, denoted as $\hat{y}$.

Example: Munich rental guide

- Estimate the model

$$\begin{aligned} rentsqm_i &= \beta_0 + \beta_1 \cdot (1/area_i) + \beta_2\, yearc_i + \beta_3\, yearc_i^2 + \beta_4\, glocation_i + \\ &\quad \beta_5\, tlocation_i + \beta_6\, bath_i + \beta_7\, kitchen_i + \beta_8\, cheating_i + \varepsilon_i. \end{aligned}$$

- The binary regressors $bath$, $kitchen$, and $cheating$ (central heating system) are dummy-coded.

Example: Munich rental guide

- Figure 6 shows the estimated nonlinear effects of living area and year of construction. The average effect plots (solid lines) are obtained by inserting in the predicted rent per sqm

$$\begin{aligned}\widehat{rentsqm} &= 3684.991 + 137.5044 \cdot (1/area) - 3.8007\, yearc + 0.0098\, yearc_i^2 + \\ &\quad 0.6795\, glocation_i + 1.5187\, tlocation_i + 0.5027\, bath_i + \\ &\quad 0.8664\, kitchen + 1.8704\, cheating\end{aligned}$$

differing values for living area (respectively year of construction) while the other covariates are held constant at their mean values.

Example: Munich rental guide

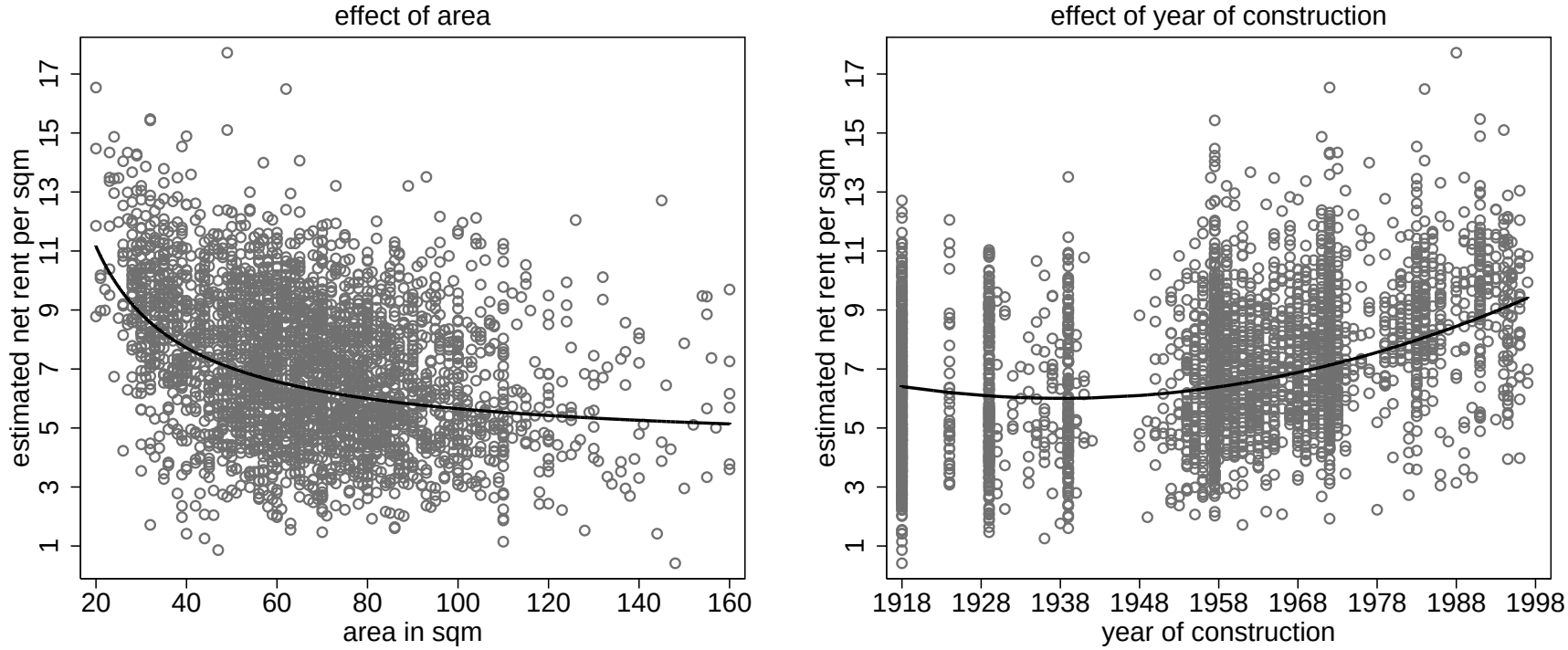


Figure 6: *Munich rent index: effects of area (left) and year of construction (right).*

Simple nonparametric regression: Nearest neighbor

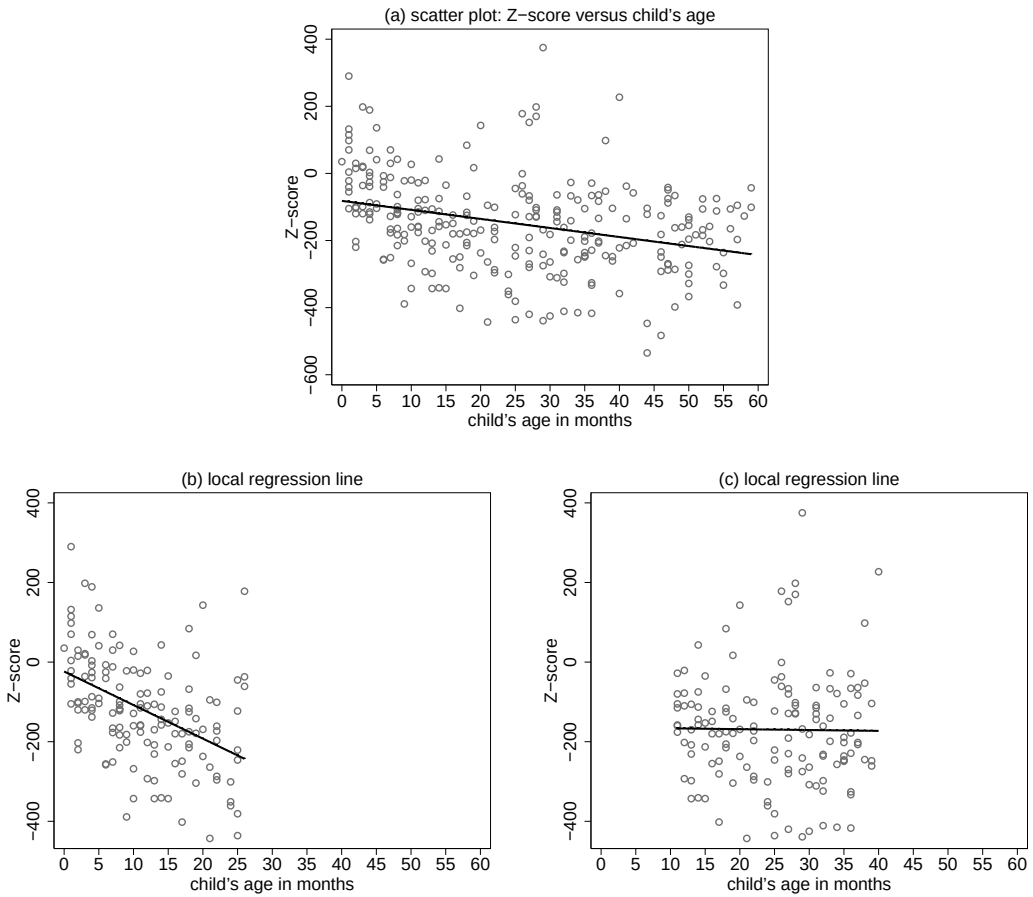


Figure 7: *Malnutrition in Zambia: illustration of a global and local regression.*

Simple nonparametric regression: Nearest neighbor

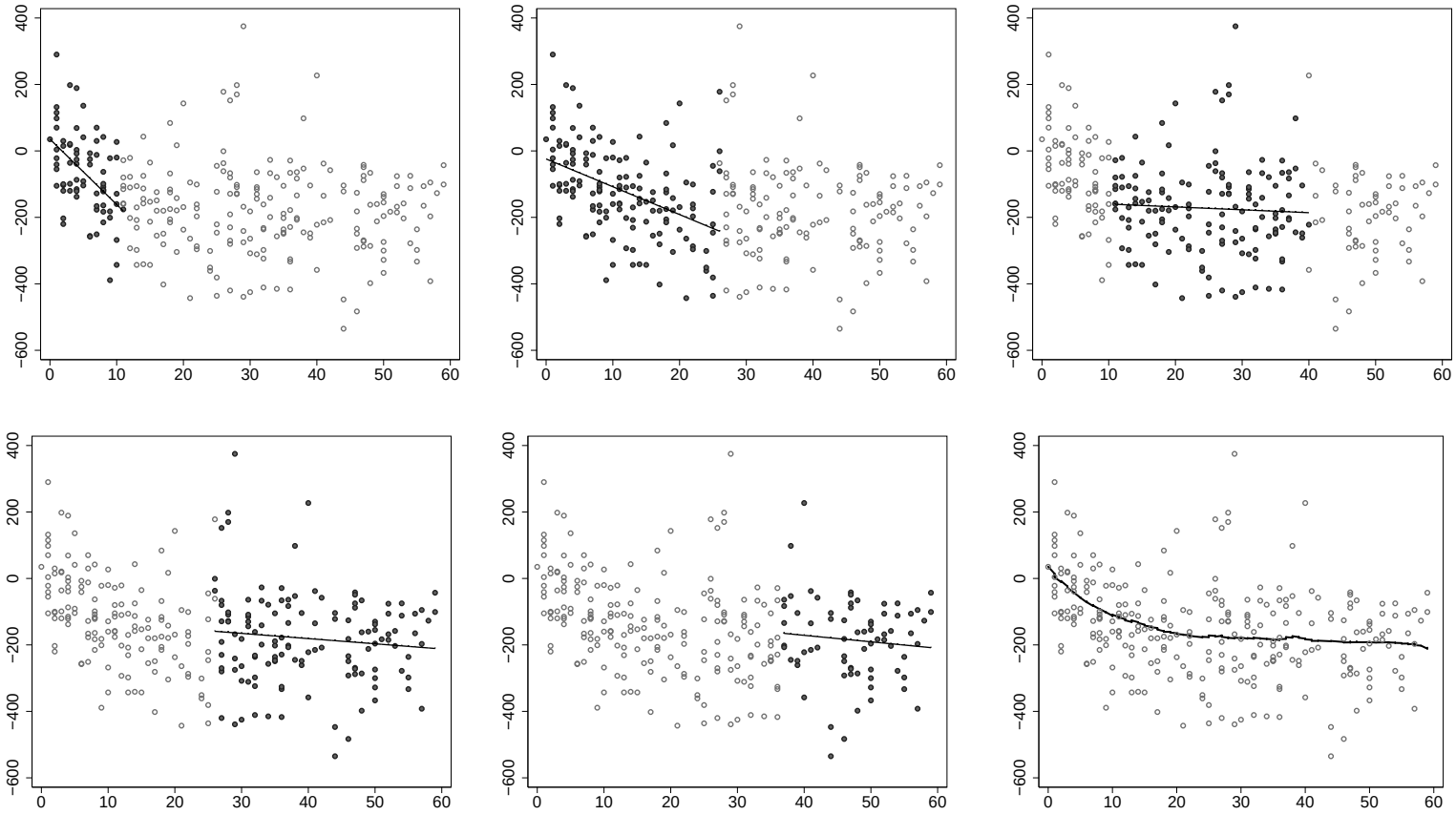


Figure 8: *Malnutrition in Zambia: illustration for a nearest neighbor estimator.*

Simple nonparametric regression: Nearest neighbor

- Determine a number of values

$$c_age_1 < c_age_2 < \ldots < c_age_m$$

  within the support of c_age for which estimates $\hat{f}(c_age_j)$, $j = 1, \ldots, m$, of f are to be computed.

- When estimating f at c_age_j, use a predetermined number of observations in the *neighborhood* of c_age_j. In Figure 7 b) and c) the nearest 70 observations, either to the right or left, of $c_age_j = 11$ (panel b) and $c_age_j = 28$ (panel c) were used.

Simple nonparametric regression: Nearest neighbor

- Estimate a local regression line based on the observations taken in step 2. to obtain the estimate $\hat{f}(c_age_j) = \hat{\beta}_0 + \hat{\beta}_1 c_age_j$. Note that for every value of c_age_j, a separate regression line is estimated. This implies that the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ vary according to c_age.

- Combine the obtained estimates $\hat{f}(c_age_1), \ldots, \hat{f}(c_age_m)$ and visualize the estimated curve.

Simple nonparametric regression: Splines

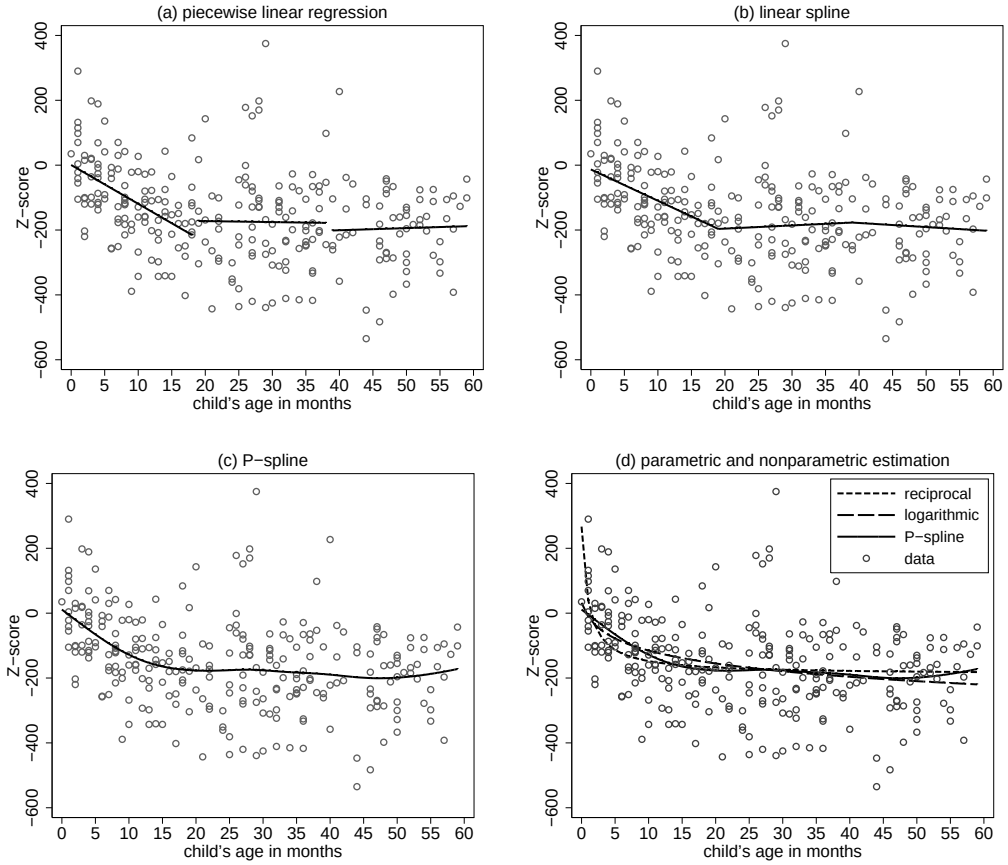


Figure 9: *Malnutrition in Zambia: piecewise linear regression (panel a), linear spline (panel b) and cubic P-spline (panel c).*

Standard nonparametric regression model

- **Data** (y_i, x_i), $i = 1, \ldots, n$, with continuous response variable y and continuous covariate x.

- **Model**

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

  We do not assume a simple parametric form for function f. We rather assume certain smoothness characteristics, for example continuity or differentiability.

- The same assumptions as in the classical linear regression model apply for errors ε_i.

Additive models: Malnutrition in Zambia

- We specify the *additive model*

$$\begin{aligned} zscore &= f_1(c_age) + f_2(m_bmi) + f_3(m_agebirth) + f_4(m_height) \\ &\quad + \beta_0 + \beta_1\, m_education1 + \ldots + \beta_{11}\, region9 + \varepsilon \end{aligned} \tag{3}$$

  rather than a linear model.

- Impose the constraints

$$\sum_{i=1}^{n} f_1\left(c_age_i\right) = \ldots = \sum_{i=1}^{n} f_4\left(m_height_i\right) = 0,$$

  i.e. each nonlinear function is centered around zero, to guarantee identifiability.

Additive models: Malnutrition in Zambia

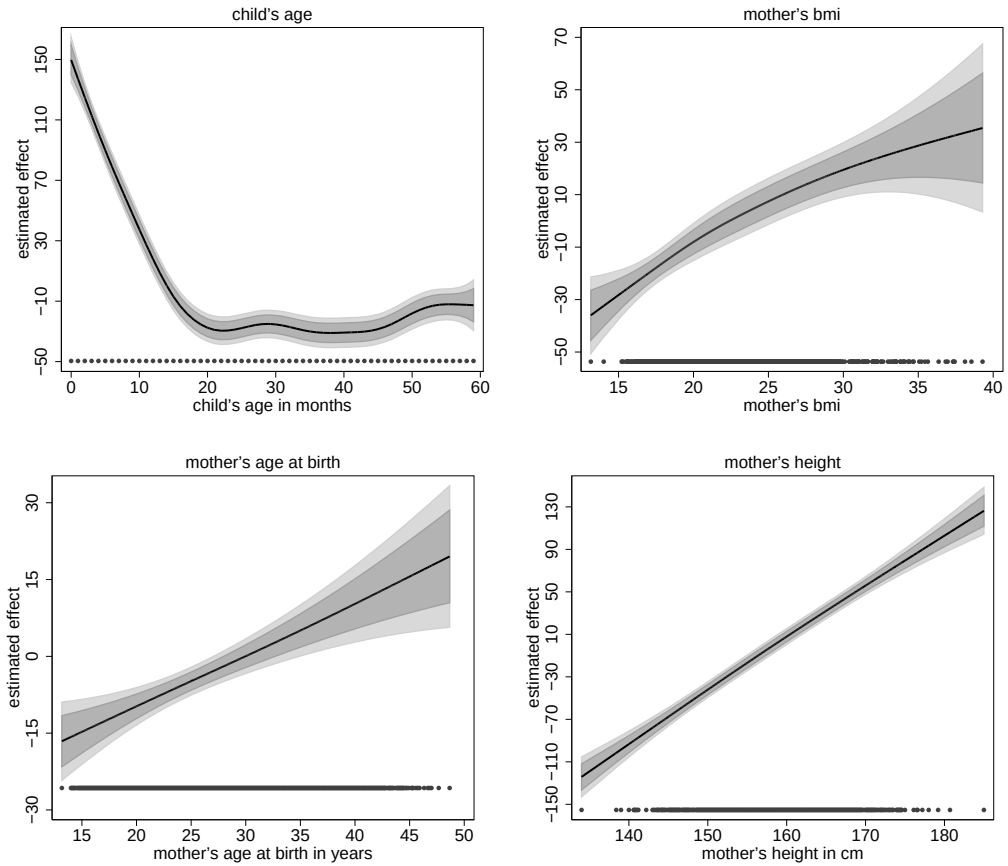


Figure 10: *Malnutrition in Zambia: estimated nonlinear functions including 80% and 95% pointwise confidence intervals.*

Standard Additive Regression Models

- **Data** $(y_i, z_{i1}, \ldots, z_{iq}, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, with y and $x_1, \ldots, x_k$ similar to those in linear regression models and additional continuous covariates $z_1, \ldots, z_q$.

- **Model**

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

- For the errors ε_i the same assumptions as in the classical linear regression model are made.

- The functions $f_1(z_1)$, ..., $f_q(z_q)$ are assumed to be "smooth" and represent nonlinear effects of the continuous covariates $z_1, \ldots, z_q$.

Binary regression

- During the validation of a patent application, it is possible that objections are raised. The response variable patent opposition (opp) is binary and coded by

$$opp_i = \begin{cases} 1 & \text{opposition against patent } i, \\ 0 & \text{otherwise.} \end{cases}$$

  The decision for an opposition against a patent may depend on various covariates.

- The expected value of a binary variable y is given by

$$\mathsf{E}(y) = \mathsf{P}(y = 0) \cdot 0 + \mathsf{P}(y = 1) \cdot 1 = \mathsf{P}(y = 1).$$

Binary regression

- The aim of a regression analysis with binary responses $y \in \{0, 1\}$ is to model the expected value $\mathsf{E}(y)$ or in other words the probability

$$\mathsf{P}(y = 1) = \mathsf{P}(y = 1 \,|\, x_1, \ldots, x_k) = \pi$$

  in the presence of covariates.

- Generally we assume the model

$$\pi_i = \mathsf{P}(y_i = 1) = F(\eta_i),$$

  where the domain of the (monotonic) function F is restricted to the interval $[0, 1]$.

Binary regression

- Choosing the logistic distribution function

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

  yields the logit model

$$\pi_i = \mathsf{P}(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \tag{4}$$

  with the *linear* or *additive predictor* η_i.

- If the value of the predictor η increases to $\eta + 1$, the probability for $y = 1$ increases in a *nonlinear* way from $F(\eta)$ to $F(\eta + 1)$.

Binary regression

- An alternative interpretation is obtained by solving the model equation (4) for η to obtain

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\mathsf{P}(y_i=1)}{1-\mathsf{P}(y_i=1)}\right) = \eta \tag{5}$$

- Alternatively (because of $\exp(a+b) = \exp(a)\cdot\exp(b)$) we obtain

$$\frac{\pi_i}{1-\pi_i} = \frac{\mathsf{P}(y_i=1)}{\mathsf{P}(y_i=0)} = \exp(\beta_0)\exp(\beta_1 x_{i1})\cdot\ldots\cdot\exp(\beta_k x_{ik}) \tag{6}$$

  in case of a linear predictor.

Binary regression: patent opposition

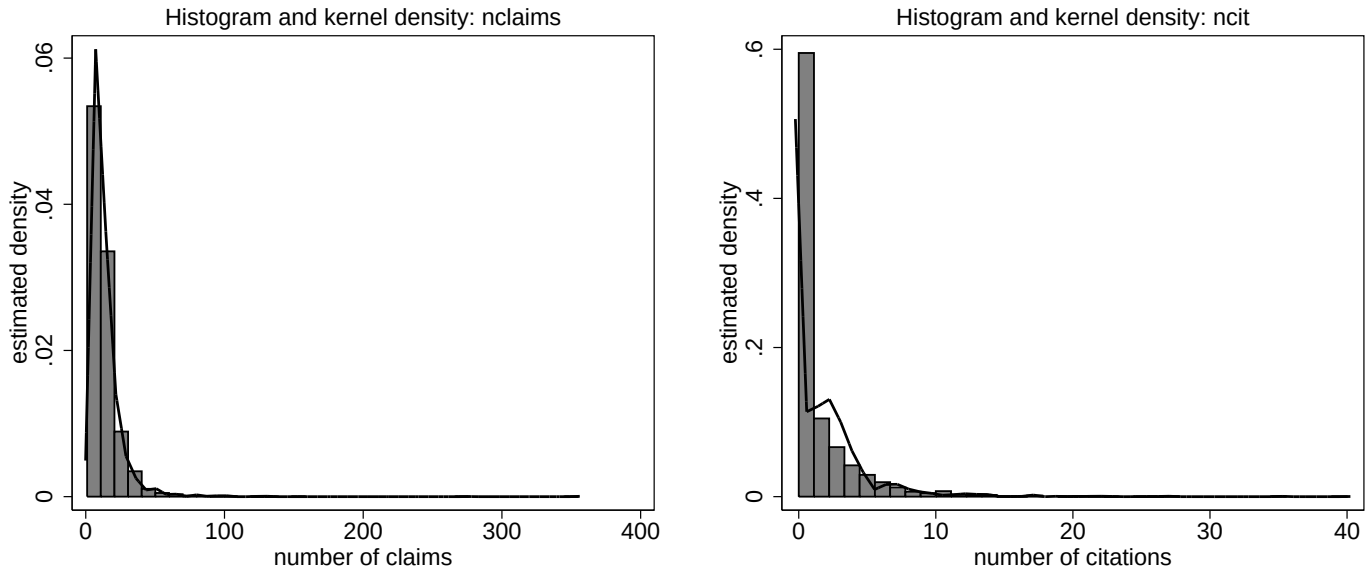


Figure 11: *Patent opposition: histogram and kernel density estimator for the continuous covariates* $nclaims$ *(left) and* $ncit$ *(right).*

Binary regression: patent opposition

- We divide the data into two groups: $biopharm = 0$ and $biopharm = 1$. For the subset $biopharm = 0$, i.e. the patents derived from the semi-conductor/computer industry, a logit model

$$\mathrm{P}(opp_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

  is estimated with the main effects linear predictor

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1\, year_i + \beta_2\, ncit_i + \beta_3\, nclaims_i + \beta_4\, ustwin_i + \beta_5\, patus_i + \\ &\quad + \beta_6\, patgsgr_i + \beta_7\, ncountry_i. \end{aligned}$$

- Table 2 contains the estimated coefficients $\hat{\beta}_\mathrm{j}$, $j = 0, \ldots, 7$, together with the corresponding odds ratios $\exp(\hat{\beta}_\mathrm{j})$.

Binary regression: patent opposition

- In multiplicative form (6) we obtain

$$\frac{\mathrm{P(opposition)}}{\mathrm{P(no\ opposition)}} = \exp(201.74)\cdot\exp(-0.102\cdot year_i)\cdot\ldots\cdot\exp(0.097\cdot ncountry_i).$$

- We observe, for instance, an increase in the odds of opposition against a patent from Germany, Switzerland, or Great Britain ($patgsgr = 1$) by the factor $\exp(0.196) = 1.217$ relative to a patent from the United States with the same values of the other covariates.

- A prediction of the odds $\mathrm{P(opposition)}\,/\,\mathrm{P(no\ opposition)}$ for a new patent is obtained by inserting the observed covariate values into the estimated model.

Binary regression: patent opposition

Variable	Estimated Coefficient	Estimated Odds Ratio
$intercept$	$\hat{\beta}_0 = 201.74$	
$year$	$\hat{\beta}_1 = -0.102$	$\exp(\hat{\beta}_1) = 0.902$
$ncit$	$\hat{\beta}_2 = 0.113$	$\exp(\hat{\beta}_2) = 1.120$
$nclaims$	$\hat{\beta}_3 = 0.026$	$\exp(\hat{\beta}_3) = 1.026$
$ustwin$	$\hat{\beta}_4 = -0.402$	$\exp(\hat{\beta}_4) = 0.668$
$patus$	$\hat{\beta}_5 = -0.526$	$\exp(\hat{\beta}_5) = 0.591$
$patgsgr$	$\hat{\beta}_6 = 0.196$	$\exp(\hat{\beta}_6) = 1.217$
$ncountry$	$\hat{\beta}_7 = 0.097$	$\exp(\hat{\beta}_7) = 1.102$

Table 2: *Patent opposition: estimated coefficients and odds ratios for the logit model.*

Poisson regression: vehicle insurance

- The calculation of vehicle insurance premiums is based on a detailed statistical analysis of the risk structure of the policyholders.

- An important part of the analysis is the modeling of the claim frequency, which generally depends on the characteristics of the policyholder and the vehicle type.

- The claim history is measured with the help of a 23-step bonus-malus-score (bm). The higher the score, the worse is the insurant's claims history.

- The statistical analysis is based on regression models with the claim frequency (within one year) as the response variable.

Poisson regression: vehicle insurance

- For count data, the Poisson distribution is often assumed for the response, i.e. $y \sim \mathrm{Po}(\lambda)$ with $\lambda = \mathrm{E}(y)$ as the expected number of claims.

- We assume $\lambda = \exp(\eta)$ in order to ensure a positive expected claim frequency.

- We model the claim frequencies of the Belgian insurance data using an additive predictor with possibly nonlinear functions of the variables age, age_v, hp, and bm:

$$\eta_i = f_1(age_i) + f_2(age_v_i) + f_3(hp_i) + f_4(bm_i) + \beta_0 + \beta_1 gender_i + \ldots$$

Poisson regression: vehicle insurance

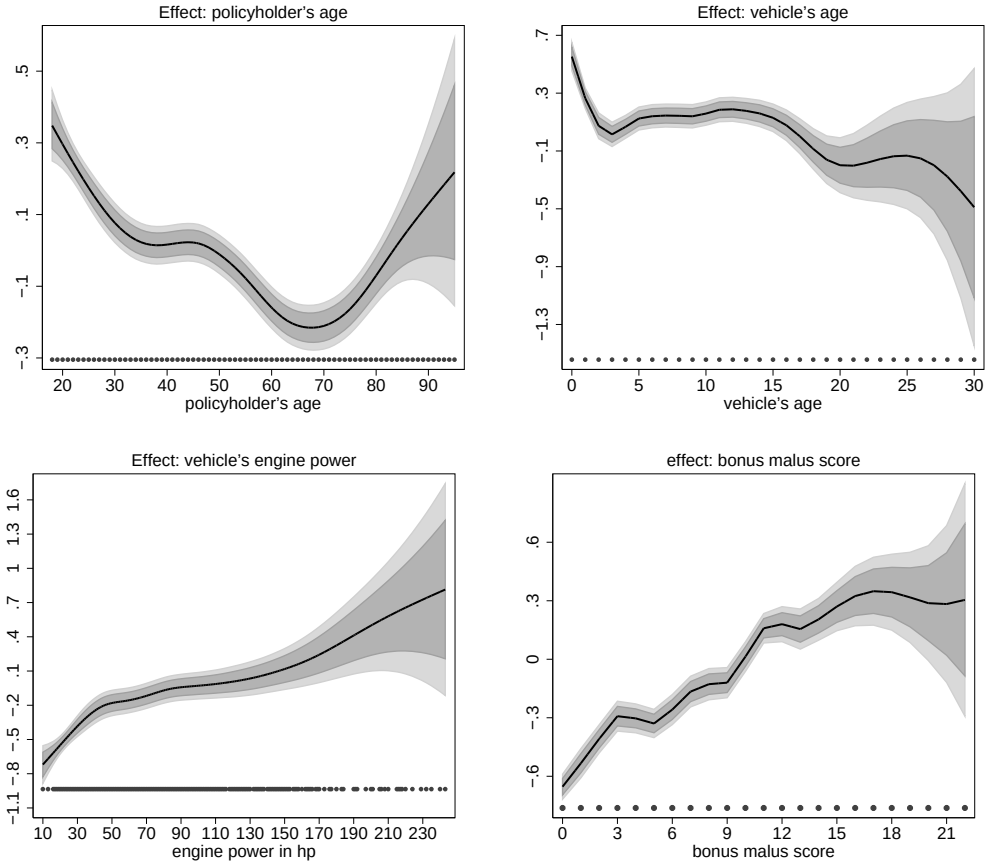


Figure 12: *Vehicle insurance: estimated nonlinear functions including 80% and 95% pointwise confidence intervals.*

Poisson regression: vehicle insurance

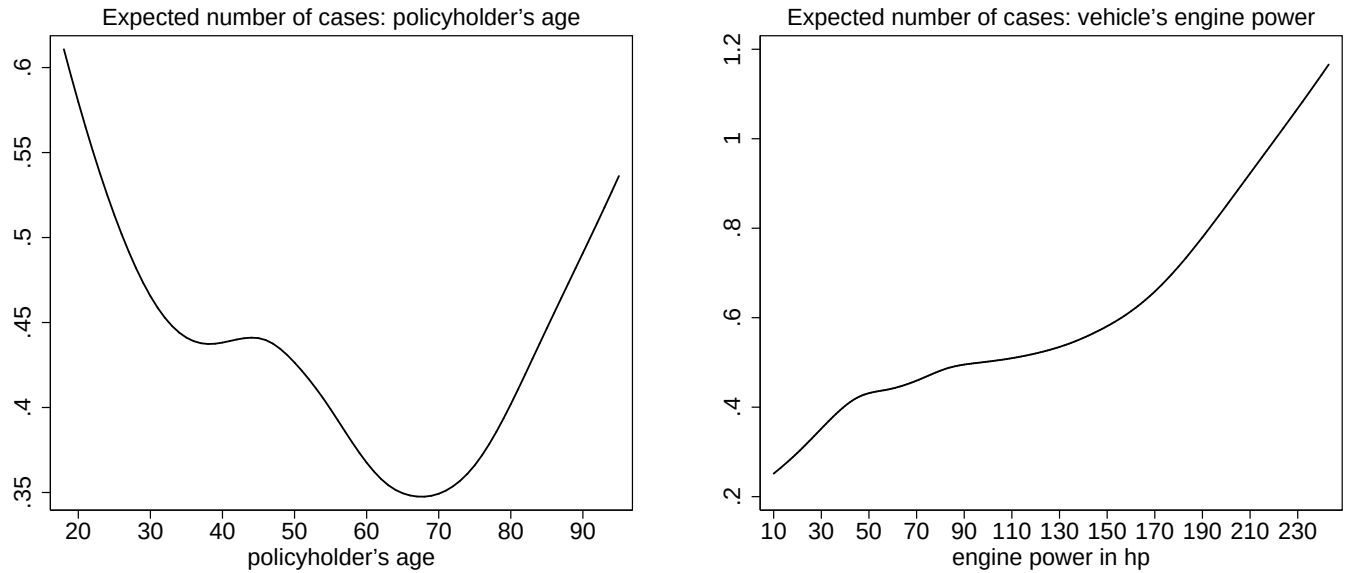


Figure 13: *Vehicle insurance: expected number of cases per year among 1000 policyholders depending on the policyholder's age and vehicle's engine power. The effects of the remaining covariates are held fixed at their mean values.*

Poisson additive model

- A Poisson additive model $y_i \sim \text{Po}(\lambda_i)$ is defined via the rate

$$\lambda_i = \mathsf{E}(y_i) = \exp(\eta_i)$$

and the additive predictor

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

Geoadditive regression: malnutrition in Zambia

- We reanalyze the data with the following *geoadditive model*:

$$\begin{aligned} zscore &= f_1(c_age) + f_2(m_bmi) + f_3(m_agebirth) + f_4(m_height) + \\ & \quad f_{geo}(district) + \beta_0 + \beta_1\, m_education + \ldots + \beta_4\, m_work + \varepsilon \end{aligned}$$

- Figure 14 shows the map of Zambia, which is divided into color-coded district-specific effects. The geographic effect can now be interpreted similar to a nonlinear effect of a continuous covariate. An effect of e.g. 40 implies an average Z-score increase of 40 points relative to a district with a zero effect.

Geoadditive regression: malnutrition in Zambia

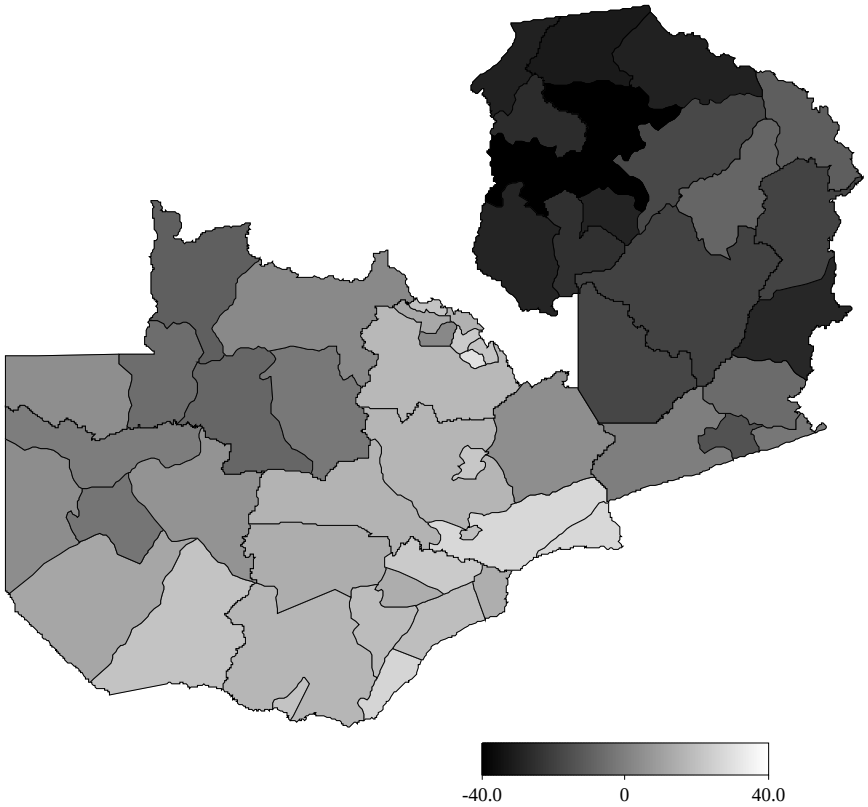


Figure 14: *Malnutrition in Zambia: estimated spatial effect.*

Geoadditive regression: vehicle insurance

- We extend the additive predictor to

$$\eta_i = f_1(age_i) + f_2(age_v) + f_3(hp_i) + f_4(bm_i) + f_{geo}(district_i) + \ldots,$$

  where $f_{geo}(district)$ represents a spatial district-specific effect.

- Figure 15 shows the estimated geographic effect: the darker the color, the higher is the estimated effect. The hatched areas mark districts for which we do not have any observations.

- In comparison to a region with an effect of zero, an effect of approximately 0.3 implies an increase of the expected frequency by a factor of $\exp(0.3) = 1.35$.

Geoadditive regression: vehicle insurance

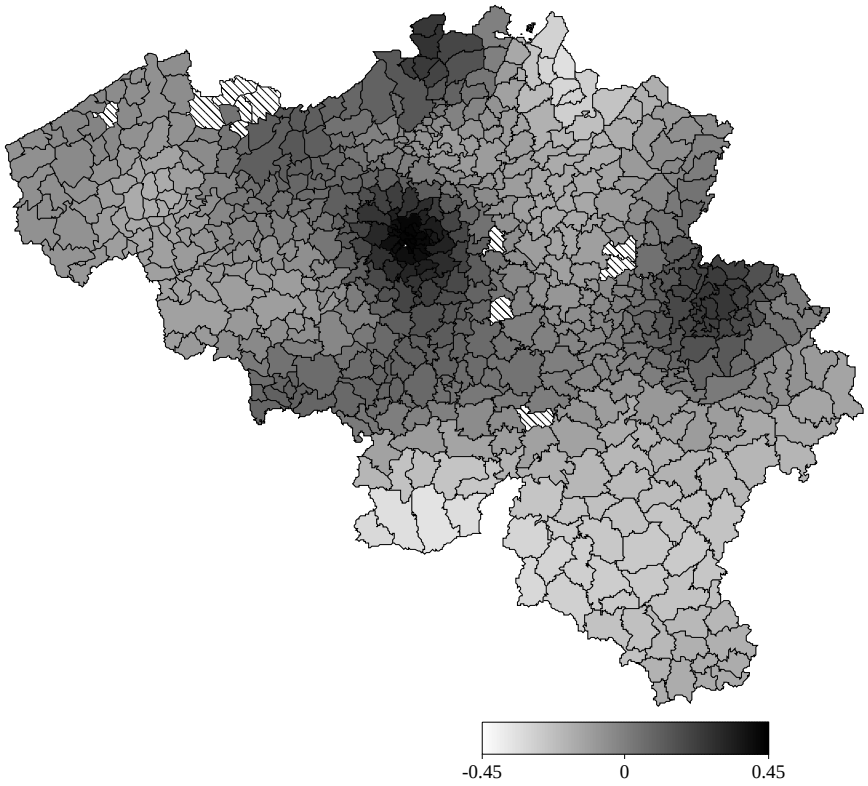


Figure 15: *Vehicle insurance: estimated spatial effect.*

Beyond mean regression: distributional regression

- In distributional regression not only the mean but all parameters of a distribution family is modelled in dependence of a suitable predictor.

- E.g. a straightforward approach to extend the multiple linear regression model to cope with variances depending on covariates is to assume $y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ where additionally to the means

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

the standard deviations (alternatively the variances)

$$\sigma_i = \exp(\alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik}) \tag{7}$$

depend on the covariates.

Distributional regression: munich rent index

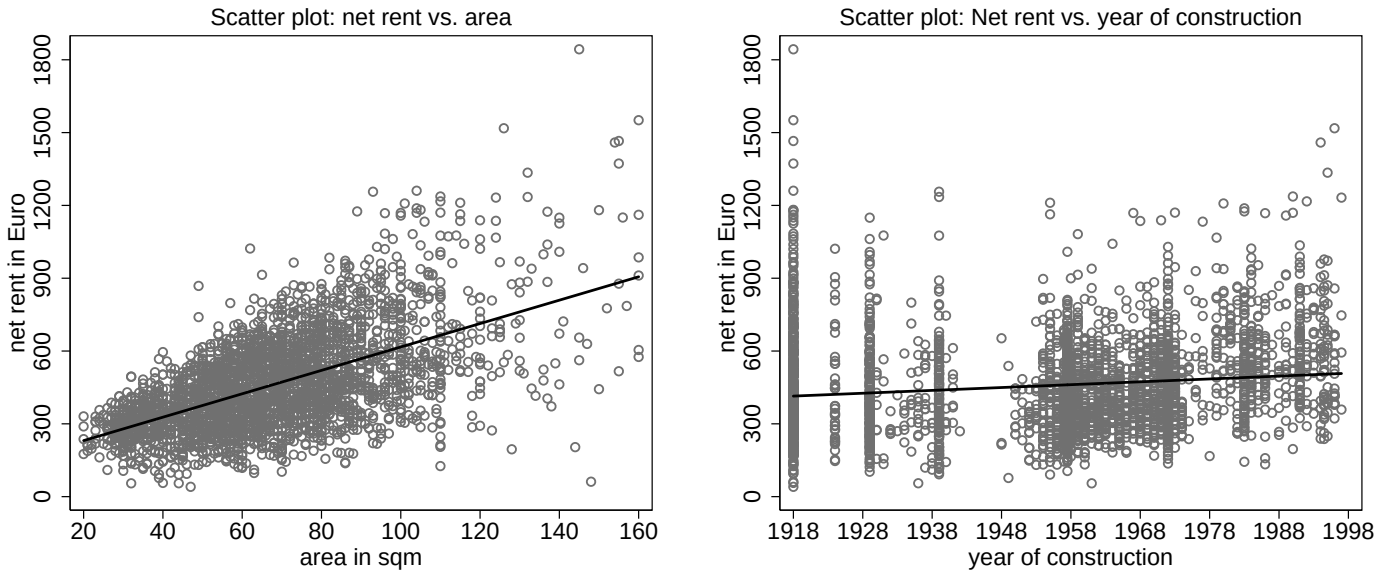


Figure 16: *Munich rent index: scatter plots of rents in Euro versus living area (left panel) and year of construction (right panel) together with estimated regression lines.*

Distributional regression: Munich rent index

- We take the data on the Munich rent index and assume the model $rent_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ with

$$\mu_i = \beta_0 + \beta_1 area_i + \beta_2 yearc_i, \qquad \sigma_i = \exp\left(\alpha_0 + \alpha_1 area_i + \alpha_2 yearc_i\right),$$

  for the expected value and standard deviation of the net rents. For simplicity, we restrict ourselves to the two covariates living area and year of construction.

- Using the R package `GAMLSS` we obtain the estimates

$$\hat{\mu}_i = -4617.6889 + 5.1847 \cdot area_i + 2.4162 \cdot yearc_i$$

  and

$$\hat{\sigma}_i = \exp\left(8.5235 + 0.0141 area_i - 0.0023 yearc_i\right).$$

Distributional regression: munich rent index

- Interpretation for the mean:
  - Increasing the living area by 1 square meter leads to an average increase of the net rent of about 5.18 Euro.
  - Modern flats are on average more expensive than older flats. Every year increases the average net rent by 2.42 Euro.

- Interpretation for the variance:
  - A unit increase of the living area increases the standard deviation by a (small) factor of $\exp(0.014094) = 1.0141938$.
  - A unit increase of the year of construction decreases the standard deviation of net rents by the factor $\exp(-0.002347) = 0.99765575$ which is again close to unity.

Distributional regression: munich rent index

- Assume now possibly nonlinear effects of living area and year of construction in the equations for the mean and the standard deviation:

$$\mu_i = f_1(area_i) + f_2(yearc_i) + \beta_0 \qquad \sigma_i = \exp\left(g_1(area_i) + g_2(yearc_i) + \alpha_0\right)$$

- Results are displayed in Figure 17.

Distributional regression: munich rent index

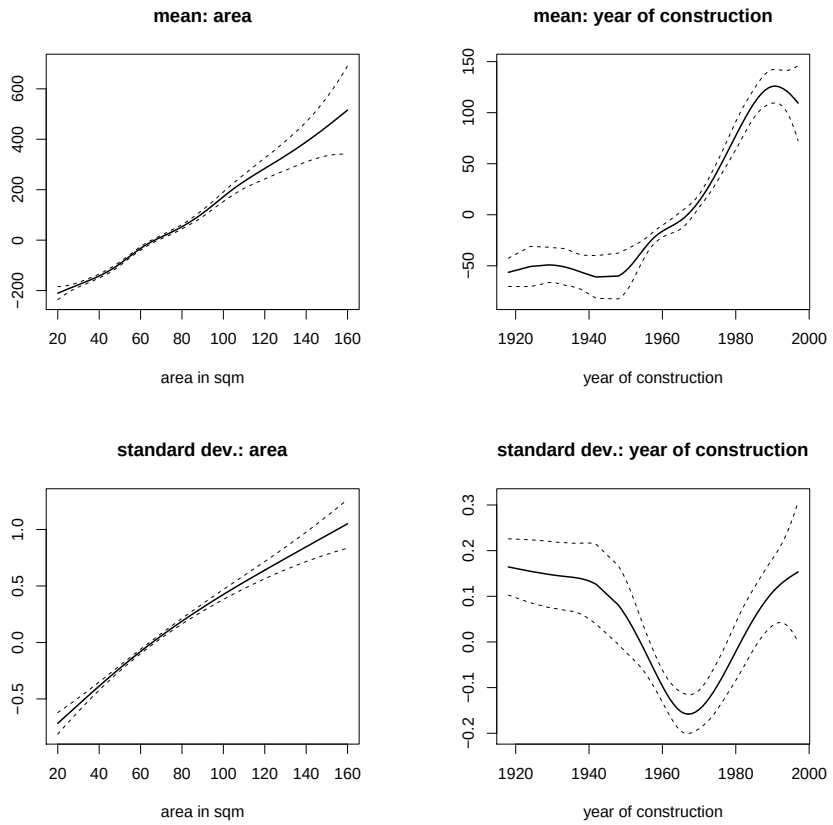


Figure 17: *Munich rent index: estimated effects of living area and year of construction for the mean and standard deviation.*

Beyond mean regression: Quantile regression

- Quantile regression aims at directly modeling the *quantiles* of the response distribution in dependence of covariates without resorting to a specific parametric distribution family.

- For $0 < \tau < 1$ let q_τ be the τ-quantile of the response distribution, e.g. $q_{0.75}$ is the 75% quantile.

- Then in linear quantile regression we assume

$$q_{\tau,i} = \beta_{\tau,0} + \beta_{\tau,1} x_{i1} + \ldots + \beta_{\tau,k} x_{ik},$$

  i.e. the quantile q_τ of the response distribution is a linear combination of the covariates.

Quantile regression: Munich rent index

- We estimate a linear effect of living area on 11 quantiles q_τ, $\tau = 0.05, 0.1, \ldots, 0.9, 0.95$, of the net rent in Euro, i.e.

$$q_{\tau,i} = \beta_{\tau,0} + \beta_{\tau,1} \cdot area_i,$$

  see Figure 18.

- We observe a clear change of the slope (and to a lesser extent also the intercept) of the regression lines with the quantile τ.

- Estimates for the quantiles of the response distribution can be obtained also on the basis of the linear models based on normally distributed responses.

Quantile regression: Munich rent index

- Assuming the simple linear model $rent_i = \beta_0 + \beta_1 area_i + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ implies $rent_i \sim \mathrm{N}(\beta_0 + \beta_1 area_i, \sigma^2)$ and the quantiles $q_{\tau,i} = \beta_0 + \beta_1 area_i + \sigma \cdot z_\tau$ where z_τ is the τ-quantile of the $\mathrm{N}(0, 1)$ distribution.

- Assuming a model with linear predictors for mean and standard deviation, i.e.

$$rent_i \sim N(\beta_0 + \beta_1 area_i, \sigma_i^2), \qquad \sigma_i = \exp(\alpha_0 + \alpha_1 area_i),$$

  results in the quantiles

$$q_{\tau,i} = \mu_i + \sigma_i z_\tau = \beta_0 + \beta_1 area_i + \exp(\alpha_0 + \alpha_1 area_i) z_\tau,$$

  which are no longer parallel to each other because the standard deviations of the rents depend on the living area.

Quantile regression: Munich rent index

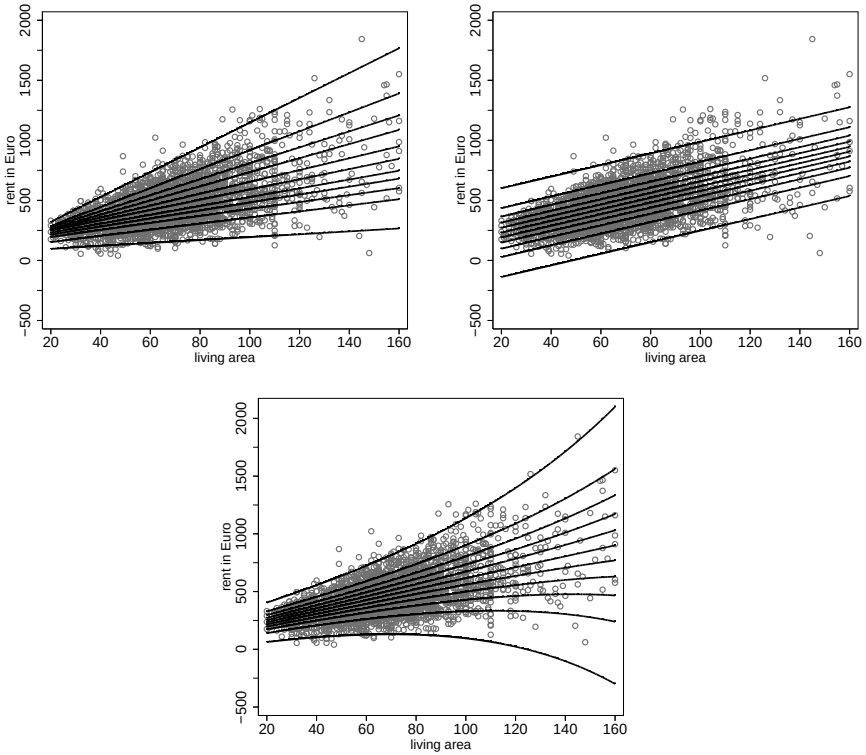


Figure 18: *Munich rent index: scatter plots of the rents in Euro versus living area together with linear quantile regression fits for 11 quantiles (top left panel), quantiles determined from a classical linear model (top right panel) and quantiles determined from distributional regression (bottom panel).*

Advanced Regression Analysis

Linear models revisited

The classical linear model

The model

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

is called the classical linear regression model, if the following assumptions are true:

1. $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$.

2. $\mathsf{Cov}(\boldsymbol{\varepsilon}) = \mathsf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$.

3. The design matrix $\mathbf{X}$ has full column rank, i.e. $\mathsf{rk}(\mathbf{X}) = \mathbf{k} + \mathbf{1} = \mathbf{p}$.

The classical normal regression model is obtained if additionally

4. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$

holds. For stochastic covariates these assumptions are to be understood conditionally on $\mathbf{X}$.

The classical linear model

- From our model assumptions it immediately follows that

$$\begin{aligned} \mathsf{E}(y_i) &= \mathsf{E}(\mathbf{x_i'}\beta + \varepsilon_\mathbf{i}) = \mathbf{x_i'}\beta = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x_{i1}} + \ldots + \beta_\mathbf{k}\mathbf{x_{ik}}, \\ \mathsf{Var}(y_i) &= \mathsf{Var}(\mathbf{x_i'}\beta + \varepsilon_\mathbf{i}) = \mathsf{Var}(\varepsilon_\mathbf{i}) = \sigma^\mathbf{2}, \\ \mathsf{Cov}(y_i, y_j) &= \mathsf{Cov}(\varepsilon_i, \varepsilon_j) = 0, \end{aligned}$$

  for the mean and the variance of y_i, as well as the covariance between y_i and y_j, respectively.

- In matrix notation we obtain

$$\mathsf{E}(\mathbf{y}) = \mathbf{X}\beta \quad \text{and} \quad \mathsf{Cov}(\mathbf{y}) = \sigma^\mathbf{2}\mathbf{I}.$$

- If we additionally assume normally distributed errors, we have

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^\mathbf{2}\mathbf{I}).$$

Model Parameters, Estimation, and Residuals

- Estimates of β and σ^2 are denoted by $\hat{\beta}$ and $\hat{\sigma}^2$, respectively. Note $\hat{\beta} \neq \beta$ in general.

- Based on an estimator $\hat{\beta}$ for β, a straightforward estimator of the mean $\mathsf{E}(y_i)$ of y_i is given by

$$\widehat{\mathsf{E}(y_i)} = \hat{\beta}_{\mathbf{0}} + \hat{\beta}_{\mathbf{1}}\mathbf{x_{i1}} + \cdots + \hat{\beta}_{\mathbf{k}}\mathbf{x_{ik}} = \mathbf{x_i'}\hat{\beta}.$$

  We usually refer to this estimator as $\hat{y}_i$, i.e. $\hat{y}_i = \widehat{\mathsf{E}(y_i)}$.

Model Parameters, Estimation, and Residuals

- The estimated error, i.e. the deviation $y_i - \mathbf{x_i'}\hat{\beta}$ between the true value y_i and the estimated value $\hat{y}_i$ is called residual and denoted by $\hat{\varepsilon}_i$. We have

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \mathbf{x_i'}\hat{\beta}.$$

- Defining the vector of residuals $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)'$, we obtain

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\beta}.$$

  Again the residuals $\hat{\varepsilon}_i$ are not identical to the errors ε_i.

Model Parameters, Estimation, and Residuals

- Especially for model diagnostics, we compute *partial residuals*, which quantify the removal of the effect of most, but not of all, covariates.

- Partial residuals with respect to the jth covariate x_j are defined as follows:

$$\begin{aligned}
\hat{\varepsilon}_{x_j,i} &= y_i - \hat{\beta}_0 - \ldots - \hat{\beta}_{j-1} x_{i,j-1} - \hat{\beta}_{j+1} x_{i,j+1} - \ldots - \hat{\beta}_k x_{ik} \\
&= y_i - \mathbf{x_i'}\hat{\beta} + \hat{\beta}_\mathbf{j}\mathbf{x_{ij}} \\
&= \hat{\varepsilon}_i + \hat{\beta}_j x_{ij}.
\end{aligned}$$

Discussion of model assumptions: Linearity of effects

- Within the scope of linear models, nonlinear relations are possible.
- Consider for instance the following model, in which the effect of the explanatory variable z_i is logarithmic:

$$y_i = \beta_0 + \beta_1 \log(z_i) + \varepsilon_i.$$

- Defining a new variable $x_i = \log(z_i)$ yields the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
- In general, nonlinear relationships can be connected to linear models provided that they are *linear in the parameters*.
- An example for a model, which is nonlinear in the parameter β_2, is given by

$$y_i = \beta_0 + \beta_1 \sin(\beta_2 z_i) + \varepsilon_i.$$

Discussion of model assumptions: Homoscedastic variances

- Figure 19 illustrates the difference between homo- and heteroscedastic variances with simulated data.

- The data are based on the model $y_i \sim \mathrm{N}(-1 + 2x_i, 1)$ (panel a) and b)) and $y_i \sim \mathrm{N}(-1 + 2x_i, (0.1 + 0.3(x_i + 3))^2)$ (panel c) and d))

- Panel a) displays the typical ideal setting of homoscedastic errors: All observations fluctuate with a stable variability around the regression line. Panel b) shows the associated errors centered at zero.

- Figures 19 c) and d) provide a typical example for heteroscedastic variances. With an increasing x, the variation around the regression line also increases. The error values in panel d) still fluctuate around zero, but the variances increase with increasing x.

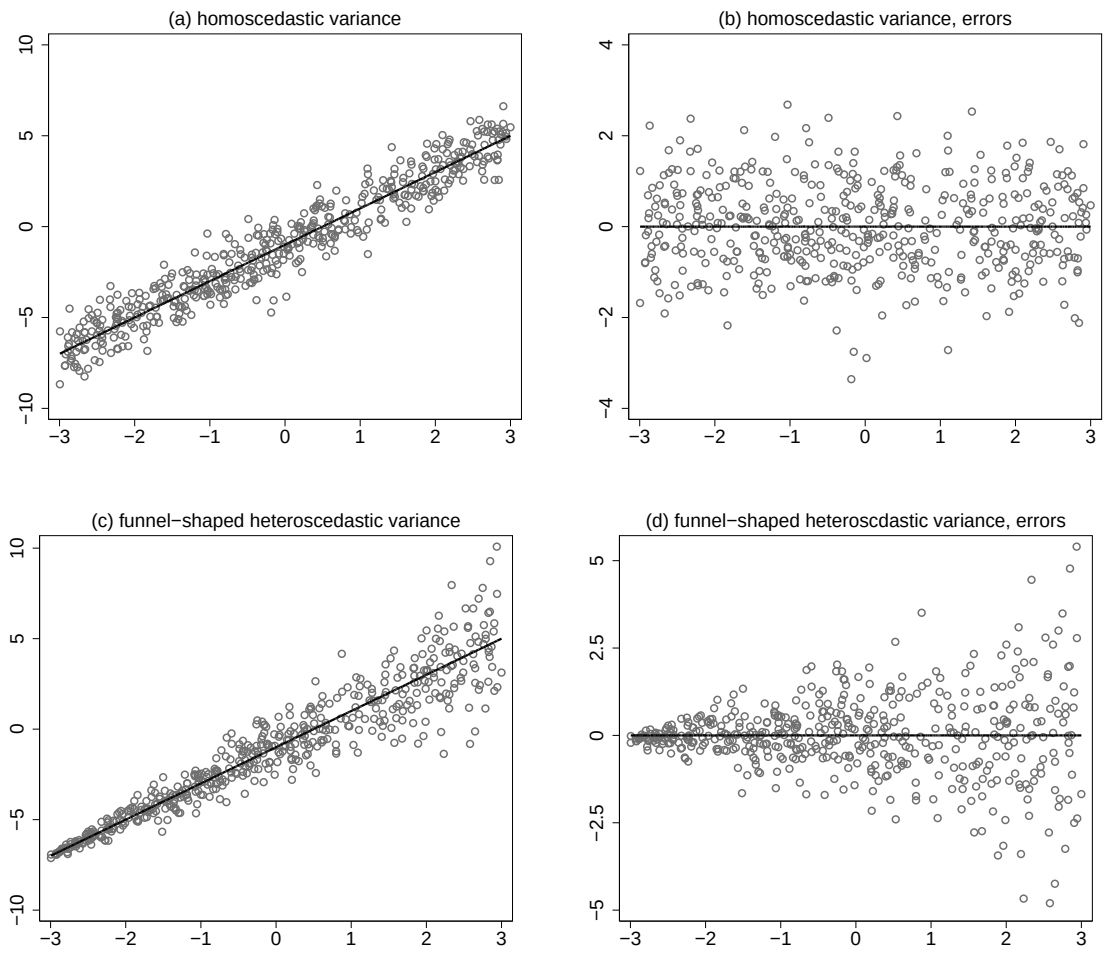


Figure 19: *Illustration for homo- and heteroscedastic variances*

Discussion of model assumptions: Homoscedastic variances

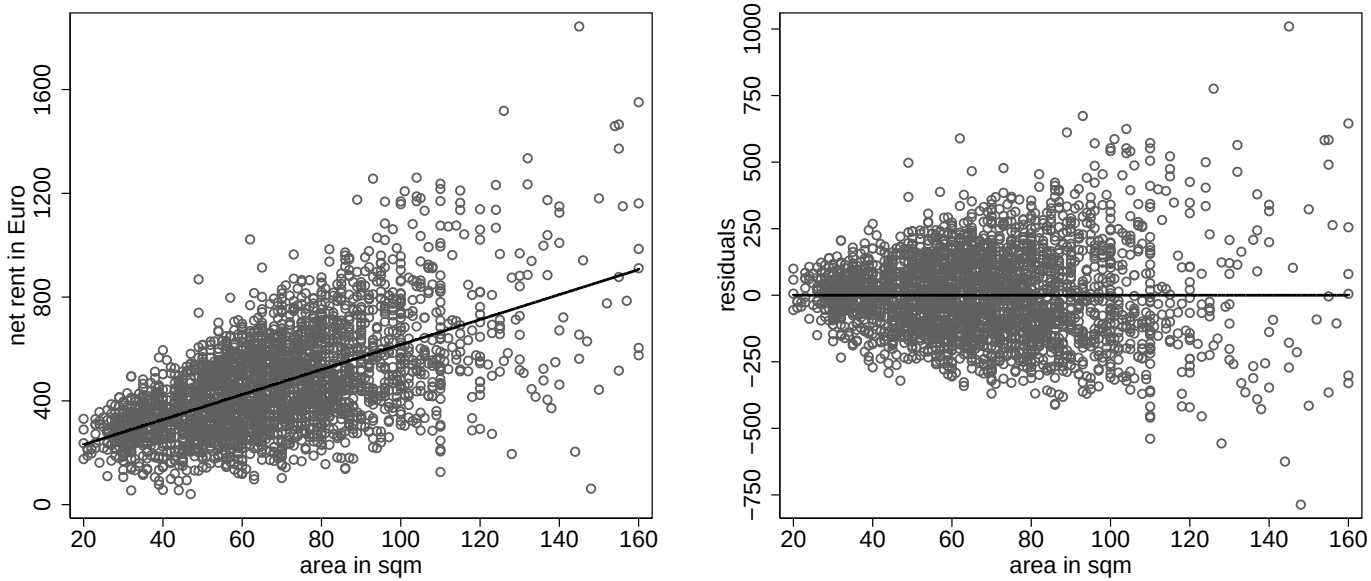


Figure 20: *Munich rent index: illustration of heteroscedastic variances. The left panel shows a scatter plot between net rent and area together with the estimated regression line. The right panel displays the corresponding residuals versus area.*

Discussion of model assumptions: Uncorrelated errors

- Figure 21 illustrates autocorrelated errors.
- Panels a) and b) shows errors with positive autocorrelation, panels c) and d) correspond to negative autocorrelation. In case of negative autocorrelation, observations are connected in order to emphasize the changing algebraic sign.
- The data with positive correlation are simulated according to the model $y_i = -1 + 2x_i + \varepsilon_i$ where $\varepsilon_i = 0.9\varepsilon_{i-1} + u_i$ and $u_i \sim \mathrm{N}(0, 0.5^2)$.
- The data with negative correlation in the errors are simulated according to $y_i = -1 + 2x_i + \varepsilon_i$ where $\varepsilon_i = -0.9\varepsilon_{i-1} + u_i$ and $u_i \sim \mathrm{N}(0, 0.5^2)$.

Discussion of model assumptions: Uncorrelated errors

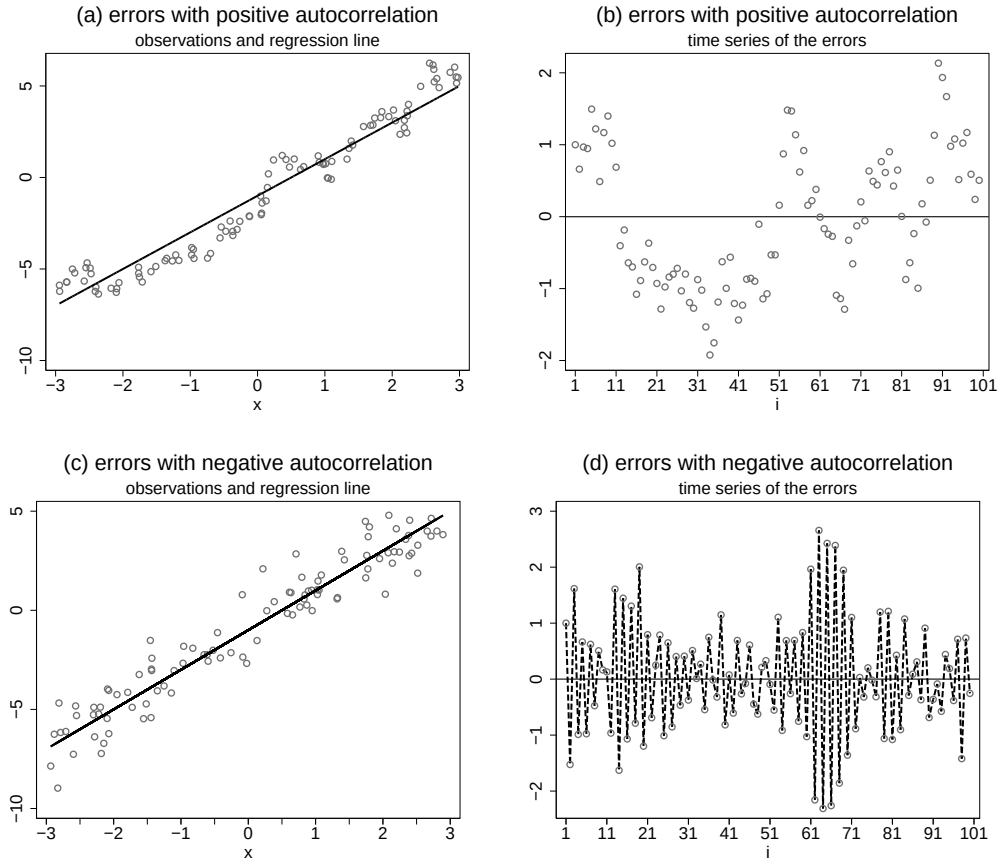


Figure 21: *Illustration for autocorrelated errors*

Discussion of model assumptions: Uncorrelated errors

Autocorrelated errors usually appear when the regression model is misspecified, e.g. a covariate is missing or the effect of a continuous covariate is nonlinear rather than linear.

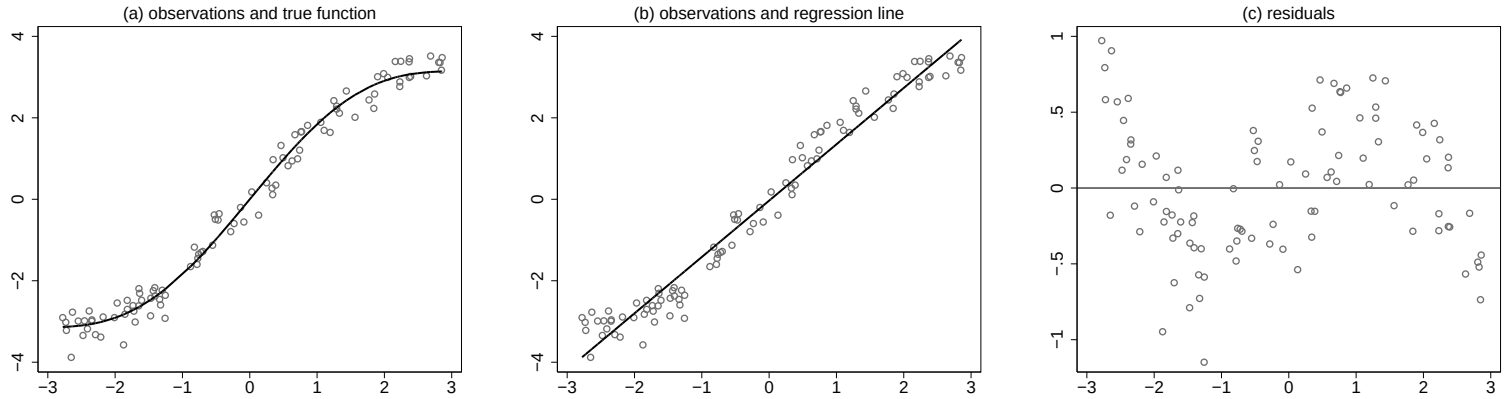


Figure 22: *Illustration for correlated residuals when the model is misspecified: Panel a) displays (simulated) data based on the function* $E(y_i \mid x_i) = \sin(x_i) + x_i$ *and* $\varepsilon_i \sim \mathrm{N}(0, 0.3^2)$*. Panel b) shows the estimated regression line, i.e. the nonlinear relationship is ignored. The corresponding residuals can be found in panel c).*

Discussion of model assumptions: Uncorrelated errors

- Autocorrelated errors are most often encountered in time series or longitudinal data. In many cases, relevant covariates can not be included in the model because they can not be observed.

- If the unobserved, but relevant covariates show a temporal or seasonal trend, correlated errors are induced.

- Panels a) and b) of Figure 23 show a time series plot of variables x_1 and x_2. Whereas x_1 is apparently subject to a clear temporal trend, x_2 fluctuates randomly around zero.

Discussion of model assumptions: Uncorrelated errors

- Now consider the regression model $y_i = -1 + x_1 - 0.6\, x_2 + \varepsilon_i$, $i = 1, \ldots, 100$, with independent and identically distributed normal errors $\varepsilon_i \sim \mathrm{N}(0, 0.5^2)$.

- We obtain the estimates $\hat{y}_i = -1.03 + 0.98 x_{i1} - 0.61 x_{i2}$. Figure 23 c) shows the corresponding residuals, which does not reveal any conspicuous behavior.

Discussion of model assumptions: Uncorrelated errors

- If we estimate a regression model that does not include covariate x_2, we obtain $\hat{y}_i = -1.01 + 0.95 \cdot x_{i1}$. The corresponding time series of residuals is presented in Figure 23 d).

- Even though covariate x_2 is missing, there does not appear to be any anomalies in the fit.

- In case that we neglect to model the variable x_1 fluctuating about some trend, we obtain the estimates $\hat{y}_i = -2.22 - 0.60 \cdot x_{i2}$.

- Figure 23 e) shows the residuals, which now suggest presence of autocorrelation.

Discussion of model assumptions: Uncorrelated errors

- The omitted effects $\beta_1 x_1$ or $\beta_2 x_2$ can be absorbed into the error term. If x_1 (x_2) is omitted in the model, we can denote the corresponding errors as $\tilde{\varepsilon} = \beta_1 x_1 + \varepsilon$ ($\tilde{\varepsilon} = \beta_2 x_2 + \varepsilon$).

- Since x_2 essentially shows no time trend, the residuals $\tilde{\varepsilon}$ are still uncorrelated.

- However, if the effect of x_1 is absorbed into $\tilde{\varepsilon}$, the trend in x_1 is reflected in the errors and autocorrelated residuals are obtained.

Discussion of model assumptions: Uncorrelated errors

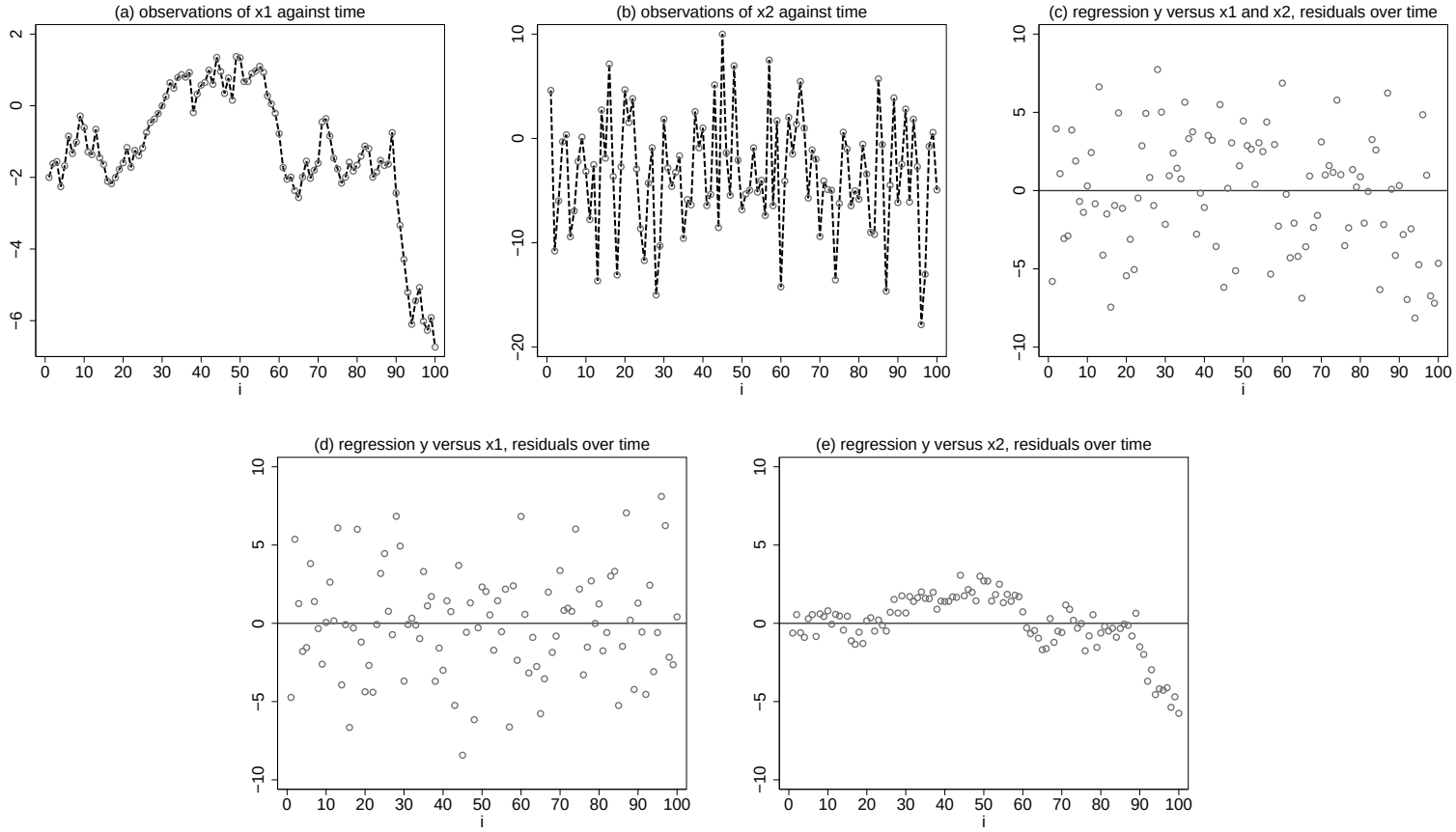


Figure 23: *Illustration for autocorrelated errors if relevant covariates showing a temporal trend are ignored.*

Discussion of model assumptions: Additivity of errors

- In the vast majority of cases, two alternative error structures are assumed: additive errors and multiplicative errors.

- An example for multiplicative errors is the exponential model

$$\begin{aligned} y_i &= \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i) \\ &= \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})\exp(\varepsilon_i). \end{aligned} \tag{8}$$

  with multiplicative errors $\tilde{\varepsilon}_i = \exp(\varepsilon_i)$.

- Models with multiplicative error structure are more plausible for exponential relationships since the errors are proportional to the mean value of y.

Discussion of model assumptions: Additivity of errors

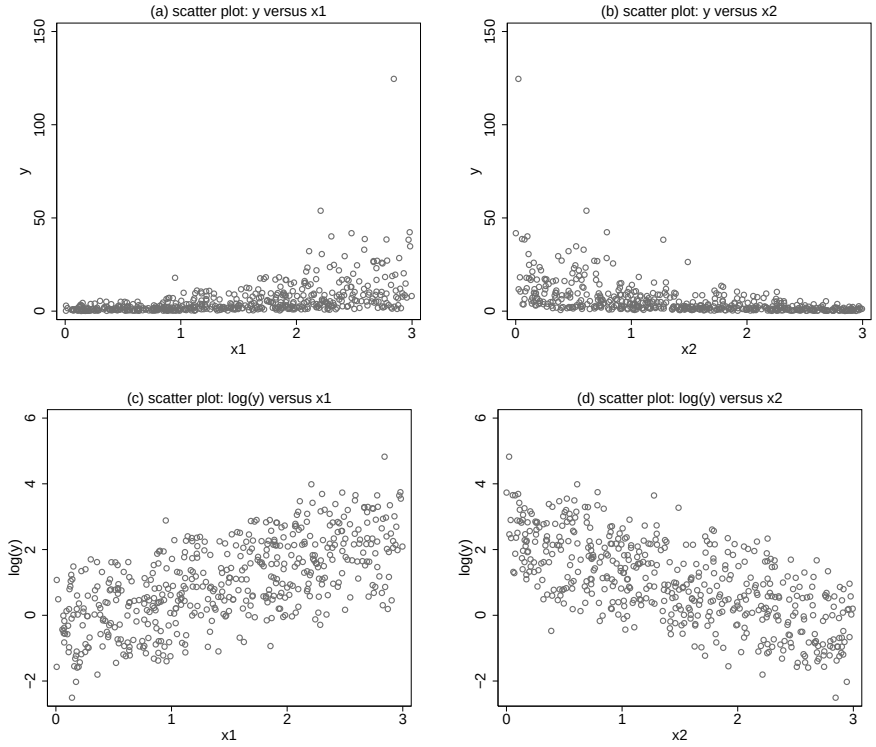


Figure 24: *Example for a multiplicative model: Panels a) and b) show scatter plots between simulated data* y *and* x_1 *respectively* x_2 *based on the model* $y_i = \exp(1 + x_{i1} - x_{i2} + \varepsilon_i)$ *with* $\varepsilon_i \sim \mathrm{N}(0, 0.4^2)$*. Panels c) and d) display scatter plots of* $\log(y)$ *versus* x_1 *and* x_2*, respectively.*

Discussion of model assumptions: Additivity of errors

- Many popular models with multiplicative errors can be expressed as linear models with additive errors through a simple transformation of variables.

- A logarithmic transformation of (8) results in the linear model

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

- Provided that the errors are normally distributed, the response y is *log-normally distributed* resulting in

$$\mathsf{E}(y_i) = \exp(\mathbf{x_i'}\beta + \sigma^{\mathbf{2}}/\mathbf{2})$$

  for the (conditional) mean of y_i.

Discussion of model assumptions: Additivity of errors

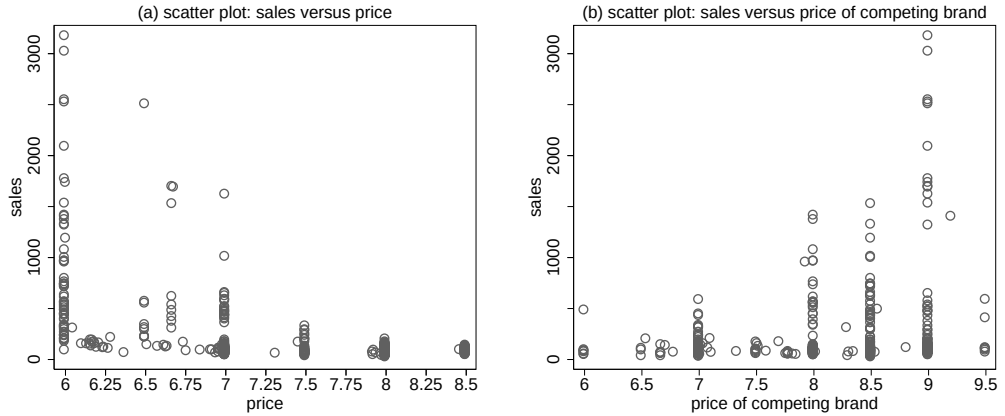


Figure 25: *Supermarket scanner data: scatter plot between the sales of a particular brand and its price (panel a) and the price of a competing brand (panel b), respectively.*

Continuous covariates: variable transformation

- If the continuous covariate z has an approximately nonlinear effect $\beta_1 f(z)$ with known transformation f, then the model

$$y_i = \beta_0 + \beta_1 f(z_i) + \ldots + \varepsilon_i,$$

can be transformed into the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \varepsilon_i,$$

where $x_i = f(z_i) - \bar{f}$.

Continuous covariates: variable transformation

- By subtracting

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(z_i),$$

  the estimated effect $\hat{\beta}_1\mathbf{x}$ is automatically centered around zero.

- The estimated curve is best interpreted by plotting $\hat{\beta}_1\mathbf{x}$ against z (instead of x).

Continuous covariates: polynomials

- If the continuous covariate z has an approximately polynomial effect $\beta_1 z + \beta_2 z^2 + \ldots + \beta_l z^l$ of degree l, then the model

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \ldots + \beta_l z_i^l + \ldots + \varepsilon_i$$

  can be transformed into the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots + \beta_l x_{il} + \ldots + \varepsilon_i,$$

  where $x_{i1} = z_i$, $x_{i2} = z_i^2$, ..., $x_{il} = z_i^l$.

Continuous covariates: polynomials

- The centering (and possibly orthogonalization) of the vectors $\mathbf{x^j} = (\mathbf{x_{1j}}, \ldots, \mathbf{x_{nj}})'$, $j = 1, \ldots, l$, to $\mathbf{x^1} - \bar{\mathbf{x}}_\mathbf{1}, \ldots, \mathbf{x^l} - \bar{\mathbf{x}}_\mathbf{l}$ with the mean vector $\bar{\mathbf{x}}_j = (\bar{x}_j, \ldots, \bar{x}_j)'$ facilitates interpretation of the estimated effects.

- A graphical illustration of the estimated polynomial is a useful way to interpret the estimated effect of z.

Munich rent index

- Using the orthogonal variables $yearco$, $yearco2$, $yearco3$, and the centered variable $areainvc$, we obtain the model fit

$$\widehat{rentsqm}_i = 7.11 + 129.57 \cdot areainvc_i + 0.79 \cdot yearco_i + 0.50 \cdot yearco2_i - 0.032 \cdot yearco3_i$$

- Figure 26 displays the corresponding effects $\hat{f}_1(area)$ and $\hat{f}_2(yearc)$ together with the partial residuals $\hat{\varepsilon}_{area,i}$ and $\hat{\varepsilon}_{yearc,i}$, which are defined by

$$\hat{\varepsilon}_{area,i} = rentsqm_i - \hat{\beta}_{\mathbf{0}} - \hat{\mathbf{f}}_{\mathbf{2}}(yearc_{\mathbf{i}}) = \hat{\varepsilon}_{\mathbf{i}} + \hat{\mathbf{f}}_{\mathbf{1}}(area_{\mathbf{i}})$$

and

$$\hat{\varepsilon}_{yearc,i} = rentsqm_i - \hat{\beta}_{\mathbf{0}} - \hat{\mathbf{f}}_{\mathbf{1}}(area_{\mathbf{i}}) = \hat{\varepsilon}_{\mathbf{i}} + \hat{\mathbf{f}}_{\mathbf{2}}(yearc_{\mathbf{i}}),$$

respectively.

Munich rent index

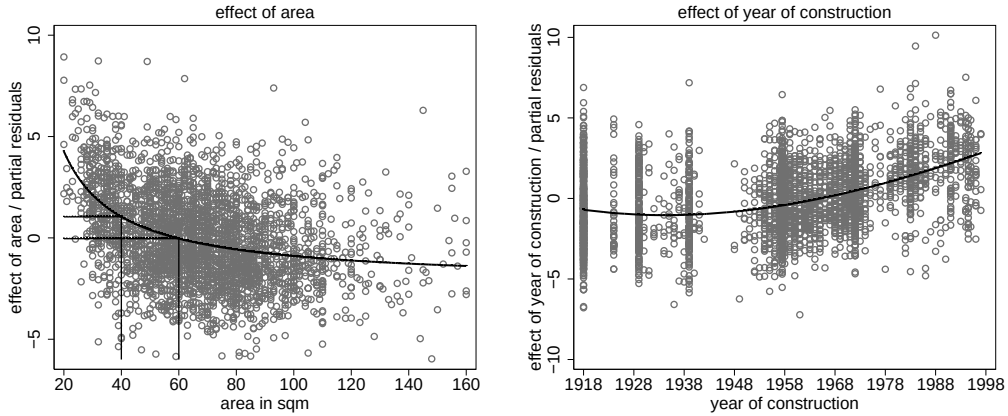


Figure 26: *Munich rent index: plots of the nonlinear and centered effects of living area and year of construction including partial residuals.*

Coding categorical covariates

- For modeling the effect of a covariate $x \in \{1, \ldots, c\}$ with c categories using dummy coding, we define the $c-1$ dummy variables

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c-1 \\ 0 & \text{otherwise} \end{cases}$$

  for $i = 1, \ldots, n$, and include them as explanatory variables in the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \varepsilon_i.$$

- For reasons of identifiability, we omit one of the dummy variables, in this case the dummy variable for category c. This category is called reference category.

- The estimated effects can be interpreted by direct comparison with the (omitted) reference category.

Munich rent index

- Define the 3 dummy variables

$$alocation = \begin{cases} 1 & location = 1 \text{ (average location)} \\ 0 & \text{otherwise} \end{cases}$$

$$glocation = \begin{cases} 1 & location = 2 \text{ (good location)} \\ 0 & \text{otherwise} \end{cases}$$

$$tlocation = \begin{cases} 1 & location = 3 \text{ (top location)} \\ 0 & \text{otherwise} \end{cases}$$

- If we do not consider the variable $glocation$ as the reference category, we obtain the model

$$rentsqm_i = \beta_0 + \beta_1 \cdot alocation_i + \beta_2 \cdot tlocation_i + \varepsilon_i.$$

Munich rent index

- The effects are now given by $\beta_0 + \beta_1$ for an average location, β_0 for good location, and $\beta_0 + \beta_2$ for top location.

- We obtain the fit

$$\widehat{rentsqm}_i = 7.27 - 0.31 \cdot alocation_i + 0.90 \cdot tlocation_i.$$

- We obtain an estimated effect of $\hat{\beta}_0 + \hat{\beta}_1 = \mathbf{7.27 - 0.31 = 6.96}$ for average location, $\hat{\beta}_0 = \mathbf{7.27}$ for good location, and $\hat{\beta}_0 + \hat{\beta}_2 = \mathbf{7.27 + 0.90 = 8.17}$ for the top location.

- *Compared to* apartments in good location, apartments in average location are 0.31 Euro less expensive and apartments in top location are 0.90 Euro more expensive.

Parameter estimation

- Assuming normally distributed errors, i.e. $\varepsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, the maximum likelihood (ML) estimator for the unknown parameters can be computed.

- We have $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$, which yields the likelihood

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right). \tag{9}$$

- The log-likelihood is thus given by

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta). \tag{10}$$

Parameter estimation

- When maximizing the log-likelihood with respect to β, we can ignore the first two terms of the sum in (10) because they are independent of β.

- Maximizing $-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ is equivalent to minimizing $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$, which is the least squares criterion.

- Thus the ML estimator for β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Parameter estimation

- Differentiation of the log-likelihood (10) with respect to σ^2, and setting to zero, provides

$$\frac{\partial l(\beta, \sigma^{\mathbf{2}})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \stackrel{!}{=} \mathbf{0}.$$

- Substituting the ML or least squares estimator $\hat{\beta}$ for β results in

$$-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta}) = -\frac{\mathbf{n}}{\mathbf{2}\sigma^{\mathbf{2}}}+\frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{4}}}(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}}) = -\frac{\mathbf{n}}{\mathbf{2}\sigma^{\mathbf{2}}}+\frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{4}}}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \stackrel{!}{=} \mathbf{0}$$

  which yields

$$\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}.$$

Parameter estimation

- The mean of the sum of squared residuals is

$$\mathsf{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = (n-p)\cdot\sigma^2 \tag{11}$$

  and thus

$$\mathsf{E}(\hat{\sigma}^2_{ML}) = \frac{n-p}{n}\sigma^2.$$

- This implies that the ML estimator for σ^2 is biased.

- An unbiased estimator is given by

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}, \tag{12}$$

  which is the commonly used estimator for σ^2.

Parameter estimation

- Estimator (12) possesses the feature of being the *restricted maximum likelihood estimator* (REML).

- It can be shown that (12) maximizes the *marginal likelihood*

$$L(\sigma^2) = \int L(\beta, \sigma^{\mathbf{2}})\, \mathbf{d}\beta,$$

  which is obtained by integration over the vector of regression coefficients β.

- In general, the REML estimator for variance parameters is less biased than the ML estimator, and thus is generally preferred. In the present case the REML estimator is even unbiased, in contrast to the ML estimator.

Statistical Properties of the LS Estimator

- *Expectation:* $\mathsf{E}(\hat{\beta}) = \beta$, implying that the least squares estimator is unbiased.

- *Covariance Matrix:* $\mathsf{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. The diagonal elements can be expressed as

$$\mathsf{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2},$$

  where R_j^2 is the coefficient of determination of the regression between x_j as the response variable, and the remaining columns of $\mathbf{X}$ as regressors.

  An estimator for the covariance matrix is given by

$$\widehat{\mathsf{Cov}\,(\hat{\beta})} = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathbf{1}}{\mathbf{n - p}} \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} (\mathbf{X}'\mathbf{X})^{-1}.$$

Statistical Properties of the LS Estimator

- *Gauß-Markov Theorem:* Among all linear and unbiased estimators $\hat{\beta}^{\mathbf{L}}$, the least squares estimator has minimal variances, implying

$$\mathrm{Var}(\hat{\beta}_{\mathbf{j}}) \leq \mathrm{Var}(\hat{\beta}_{\mathbf{j}}^{\mathbf{L}}), \quad \mathbf{j} = \mathbf{0}, \ldots, \mathbf{k}.$$

  For any linear combination $\mathbf{b}'\beta$ it holds

$$\mathrm{Var}(\mathbf{b}'\hat{\beta}) \leq \mathrm{Var}(\mathbf{b}'\hat{\beta}^{\mathbf{L}}).$$

Statistical Properties of the LS Estimator under Normality

- *Distribution of the response variable:*

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

- *Distribution of the least squares estimator:*

$$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

- *Distribution of weighted distance:*

$$\frac{(\hat{\beta}-\beta)'(\mathbf{X}'\mathbf{X})(\hat{\beta}-\beta)}{\sigma^2} \sim \chi^2_p.$$

Asymptotic Properties of the Least Squares Estimator

- The least squares estimator $\hat{\beta}_{\mathbf{n}}$ for β and the ML or REML estimator $\hat{\sigma}_n^2$ for the variance σ^2 are consistent.

- The least squares estimator asymptotically follows a normal distribution, specifically

$$\sqrt{n}(\hat{\beta}_{\mathbf{n}} - \beta) \xrightarrow{\mathbf{d}} \mathrm{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{V}^{-\mathbf{1}}).$$

  That is the difference $\hat{\beta}_{\mathbf{n}} - \beta$ normalized with $\sqrt{n}$ converges in distribution to the normal distribution on the right hand side.

Hypothesis Testing

- We revisit the data from the Munich rent index to illustrate hypothesis testing. We use the data for the 1999 rent index, in combination with the follow-up data from 2001.

- Consider the regression model

$$\begin{aligned} rentsqm_i &= \beta_0 + \beta_1\, areainvc_i + \beta_2\, yearco_i + \beta_3\, yearco2_i + \\ &\quad \beta_4\, yearco3_i + \beta_5\, nkitchen + \beta_6\, pkitchen + \beta_7\, year01 + \varepsilon_i. \end{aligned} \tag{13}$$

- The following estimated model results:

$$\begin{aligned} \widehat{rentsqm}_i &= 6.94 + 124.34\, areainvc_i + 0.73\, yearco_i + 0.44\, yearco2_i - \\ &\quad 0.01\, yearco3_i + 1.04\, nkitchen + 1.31\, pkitchen - 0.19\, year01\,. \end{aligned}$$

Hypothesis Testing

We would like to test the following hypotheses:

- $$H_0 : \beta_7 = 0 \qquad \text{against} \qquad H_1 : \beta_7, \neq 0,$$
  i.e. whether the decrease of 19 Cent in 2001 is significant.

- $$H_0 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{against} \qquad H_1 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$
  i.e. whether the variable $kitchen$ with the three categories "substandard kitchen" (reference category), "standard kitchen" (dummy variable $nkitchen$), and "premium kitchen" (dummy variable $pkitchen$) is significant.

Hypothesis Testing

- $$H_0 : \beta_5 = \beta_6 \qquad \text{against} \qquad H_1 : \beta_5 \neq \beta_6$$

  or equivalently

  $$H_0 : \beta_5 - \beta_6 = 0 \qquad \text{against} \qquad H_1 : \beta_5 - \beta_6 \neq 0,$$

  i.e. we ask whether it is really necessary to differentiate between standard and premium kitchens.

Hypothesis Testing

The discussed problems are exemplary for the most common statistical hypotheses in linear models:

1. *Test of significance*

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0.$$

2. *Composite test of a subvector* $\beta_1 = (\beta_1, \ldots, \beta_r)'$

$$H_0 : \beta_1 = \mathbf{0} \qquad \text{against} \qquad \mathbf{H_1} : \beta_1 \neq \mathbf{0}.$$

3. *Test of equality*

$$H_0 : \beta_j - \beta_r = 0 \qquad \text{against} \qquad H_1 : \beta_j - \beta_r \neq 0.$$

Hypothesis Testing

- We can treat these three test problems as special cases of tests for *general linear hypotheses*

$$H_0 : \mathbf{C}\beta = \mathbf{d} \qquad \text{against} \qquad \mathbf{H_1} : \mathbf{C}\beta \neq \mathbf{d}, \tag{14}$$

  where $\mathbf{C}$ is a $r \times p$-matrix with $\text{rk}(\mathbf{C}) = \mathbf{r} \leq \mathbf{p}$.

- When testing the significance of an influential variable, $\mathbf{d} = \mathbf{0}$ is a scalar and $\mathbf{C}$ is a $1 \times p$-matrix given by $\mathbf{C} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{1}, \mathbf{0}, \ldots, \mathbf{0})$, where the one is at the $(j+1)$-th column of the matrix.

Hypothesis Testing

- In the case of testing the first r components, we obtain the r-dimensional vector $\mathbf{d} = \mathbf{0}$ and the $r \times p$-matrix

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix}.$$

- when testing equality of two coefficients, we obtain the scalar $\mathbf{d} = \mathbf{0}$ and the $1 \times p$-matrix $\mathbf{C} = (\mathbf{0}, \ldots, \mathbf{1}, \ldots, -\mathbf{1}, \ldots, \mathbf{0})$, where the one is at the $(j+1)$-th position and the minus one at the $(r+1)$-th position.

Hypothesis Testing

Assuming Gaussian errors the test statistic is based on the following idea:

1. Compute the residual sum of squares $\text{SSE} = \hat{\varepsilon}'\hat{\varepsilon}$ for the full model.

2. Compute the residual sum of squares $\text{SSE}_{H_0} = \hat{\varepsilon}'_{H_0}\hat{\varepsilon}_{H_0}$ for the model under the null hypothesis, i.e. under the restriction $\mathbf{C}\beta = \mathbf{d}$.

3. Calculate the statistic
$$\frac{\Delta\text{SSE}}{\text{SSE}} = \frac{\text{SSE}_{H_0} - \text{SSE}}{\text{SSE}}, \tag{15}$$
which is the relative distance in residual sum of squares between the restricted model and the full model.

Hypothesis Testing

- For the test statistic we obtain

$$F = 1/r \left(\mathbf{C}\hat{\beta} - \mathbf{d}\right)' \left(\hat{\sigma}^{\mathbf{2}}\mathbf{C}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{C}'\right)^{-\mathbf{1}} \left(\mathbf{C}\hat{\beta} - \mathbf{d}\right) \sim \mathsf{F}_{\mathbf{r},\mathbf{n}-\mathbf{p}}$$

  .

- Reject H_0 in the case of $F > F_{r,n-p}(1-\alpha)$.

- The test is relatively robust against moderate departures from normality.

- In addition, the test can be applied for large sample size, even with non-normal errors.

Hypothesis Testing - Munich rent index

Variable	Coefficient	Standard-error	t-statistic	p-value	95% Confidence-interval	
intercept	6.936	0.037	184.50	<0.001	6.862	7.009
areainv	124.341	4.445	27.97	<0.001	115.625	133.057
yearco	0.729	0.030	24.29	<0.001	0.670	0.788
yearco2	0.436	0.030	14.53	<0.001	0.377	0.495
yearco3	-0.011	0.029	-0.38	0.701	-0.068	0.046
nkitchen	1.044	0.101	10.28	<0.001	0.845	1.243
pkitchen	1.306	0.152	8.57	<0.001	1.007	1.604
year01	-0.192	0.062	-3.08	0.002	-0.314	-0.069

Table 3: *Munich rent index: output for the regression model (13).*

Hypothesis Testing - Munich rent index

- We notice that the decrease of the average net rent by 0.19 Euro in the year 2001 is significant when comparing it to the year of 1999. We can reject the hypothesis $H_0 : \beta_7 = 0$ for every significance level $\alpha > 0.002$.

- For the hypotheses

$$H_0 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{versus} \qquad H_1 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

  the test statistic computed as $F = 82.22$. With a significance level of $\alpha = 0.05$, the corresponding quantile of the F-distribution is given by

$$F_{2,4559-8}(0.95) = 3.00.$$

  Since $F = 82.22 > 3.00 = F_{2,4551}(0.95)$, we can reject the null hypothesis at this level.

Hypothesis Testing - Munich rent index

- For the hypothesis

$$H_0 : \beta_5 - \beta_6 = 0 \quad \text{against} \quad H_1 : \beta_5 - \beta_6 \neq 0.$$

we obtain the test statistic $F = 2.23$. For $\alpha = 0.05$, the corresponding quantile of the F-distribution is given by $F_{1,4551}(0.95) = 3.84$. Since $F = 2.23 < 3.84 = F_{1,4551}(0.95)$, we cannot reject the null hypothesis. This test has a p-value of $p = 0.13$.

Confidence intervals

- The duality between two-sided tests and confidence intervals allows us to construct a confidence interval for a single parameter β_j, $j = 0, \ldots, k$.

- To construct a confidence interval for β_j under normality, we use the test-statistic $t_j = (\hat{\beta}_\mathbf{j} - \mathbf{d_j})/\mathrm{se}_\mathbf{j}$ corresponding to the test $H_0 : \beta_j = d_j$.

- Recall that we reject the null hypothesis when $|t_j| > t_{n-p}(1 - \alpha/2)$. The test is constructed such that the probability of rejecting H_0 when H_0 is true equals α.

- Therefore, under H_0 we have

$$\mathrm{P}(|t_j| > t_{n-p}(1 - \alpha/2)) = \alpha.$$

Confidence intervals

- Thus the probability of not rejecting H_0 (given H_0 is true) is provided by

$$\mathsf{P}(|t_j| < t_{n-p}(1-\alpha/2)) = \mathsf{P}(|(\hat{\beta}_j - \beta_j)/\mathsf{se}_j| < t_{n-p}(1-\alpha/2)) = 1-\alpha.$$

- This is equivalent to

$$\mathsf{P}(\hat{\beta}_\mathbf{j} - \mathbf{t_{n-p}}(\mathbf{1}-\alpha/\mathbf{2}) \cdot \mathsf{se}_\mathbf{j} < \beta_\mathbf{j} < \hat{\beta}_\mathbf{j} + \mathbf{t_{n-p}}(\mathbf{1}-\alpha/\mathbf{2}) \cdot \mathsf{se}_\mathbf{j}) = \mathbf{1}-\alpha,$$

  and we obtain

$$[\hat{\beta}_\mathbf{j} - \mathbf{t_{n-p}}(\mathbf{1}-\alpha/\mathbf{2}) \cdot \mathsf{se}_\mathbf{j}, \hat{\beta}_\mathbf{j} + \mathbf{t_{n-p}}(\mathbf{1}-\alpha/\mathbf{2}) \cdot \mathsf{se}_\mathbf{j}]$$

  as a $(1-\alpha)$-confidence interval for β_j.

Confidence intervals - Munich rent index

- A 95% confidence interval for the regression coefficient β_7, associated with the time dummy variable, can be calculated as

$$-0.192 \pm 1.96 \cdot 0.062 = [-0.313, -0.070].$$

- Here we used $\text{se}_7 = 0.062$ (see Table 3) and $t_{n-p}(1-\alpha/2) = t_{4551}(0.975) = 1.96$.

- The minor differences found between this confidence interval and the one found in Table 3 result from rounding error.

Model choice and variable selection

- In many applications, a large (potentially enormous) number of candidate predictor variables is available, and we face the challenge and decision as to which of these variables to include in the regression model.

- The following are two naive (but often practiced) approaches to the model selection problem:
  - Strategy 1: Estimate the most complex model which includes all available covariates.
  - Strategy 2: First estimate a model with all variables. Then, remove all insignificant variables from the model.

Model choice and variable selection

- We investigate the simulated data illustrated in Figure 27 a).
- The true model used for simulation is $y_i = -1 + 0.3x_i + 0.4x_i^2 - 0.8x_i^3 + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, 0.07^2)$.
- The scatter plot suggests polynomial modeling of the relationship between y and x resulting in the regression model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_l x_i^l + \varepsilon_i.$$

- Figure 27 c) - e) show the estimated relationship for $l = 1$ (regression line), $l = 2$, and $l = 5$.

Model choice and variable selection

- Figure 27 f) additionally displays the mean squared error

$$\text{MSE}(l) = \frac{1}{50}\sum_{i=1}^{50}(y_i - \hat{y}_i(l))^2$$

  of the fitted models depending on the order of the polynomial (continuous line). Clearly, MSE(l) decreases monotonically with increased l.

- Figure 27 b) shows additionally simulated observations for every design point x_i, $i = 1, \ldots, 50$. We refer to this data set as the validation sample, whereas we refer to the first data set (used for estimation) as the training set.

Model choice and variable selection

- Figure 27 f) shows the mean squared error of $\hat{y}_i^*$ for the data y_i^* (dashed line) in the validation set.

- Apparently, the fit to the new data is initially getting better with an increase of the polynomial order. However, from the polynomial order $l = 3$ onward, the fit is getting worse.

- The more complex the model, the better is the fit to the data that were used for estimation. However with new data resulting from the same data generating process, models that are too complex can cause a poorer fit.

Model choice and variable selection

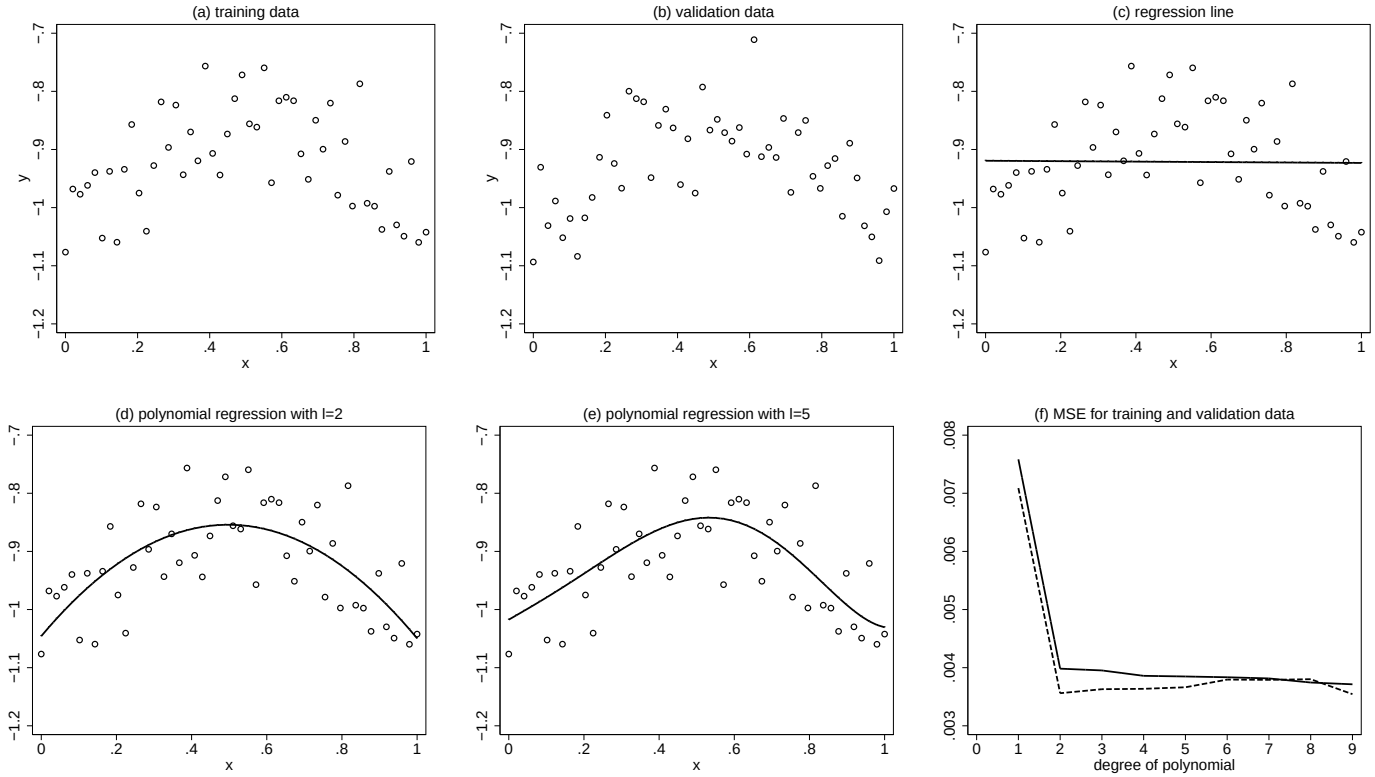


Figure 27: *Simulated training data y_i (panel a) and validation data y_i^* (panel b). Panels c)-e) show estimated polynomials based on the training set. Panel f) displays MSE(l) in relation to the polynomial degree (solid line). The dashed line shows MSE(l), if the estimated polynomials are used to predict the validation data y_i^*.*

Model choice and variable selection

- Consider the $n = 150$ observations $(y_i, x_{i1}, x_{i2}, x_{i3})$, $i = 1, \cdots, 150$, in the scatter plot matrix in Figure 28.

- The variables x_1 and x_3 are independent and uniformly distributed on [0,1]. The variable x_2 is defined as $x_2 = x_1 + u$, where u is also uniformly distributed on [0,1].

- The response variable y is simulated according to the model

$$y \mid x_1, x_2, x_3 \sim \mathrm{N}(-1 + 0.3x_1 + 0.2x_3, 0.2^2).$$

  The conditional mean of y is thus dependent on x_1 and x_3, but not on x_2.

Model choice and variable selection

- We first estimate a regression model with all available covariates x_1, x_2, and x_3, see Table 4.

- Clearly, x_1 and x_2 are nonsignificant. If we followed strategy 2, we would not only eliminate the nonrelevant covariate x_2, but also the relevant variable x_1.

- If we estimate a correctly specified model with true predictor variables x_1 and x_3, we obtain the results shown in Table 5.

- When having a correct model specification, not only is x_3 significant, but so is the previously insignificant variable x_1.

Model choice and variable selection

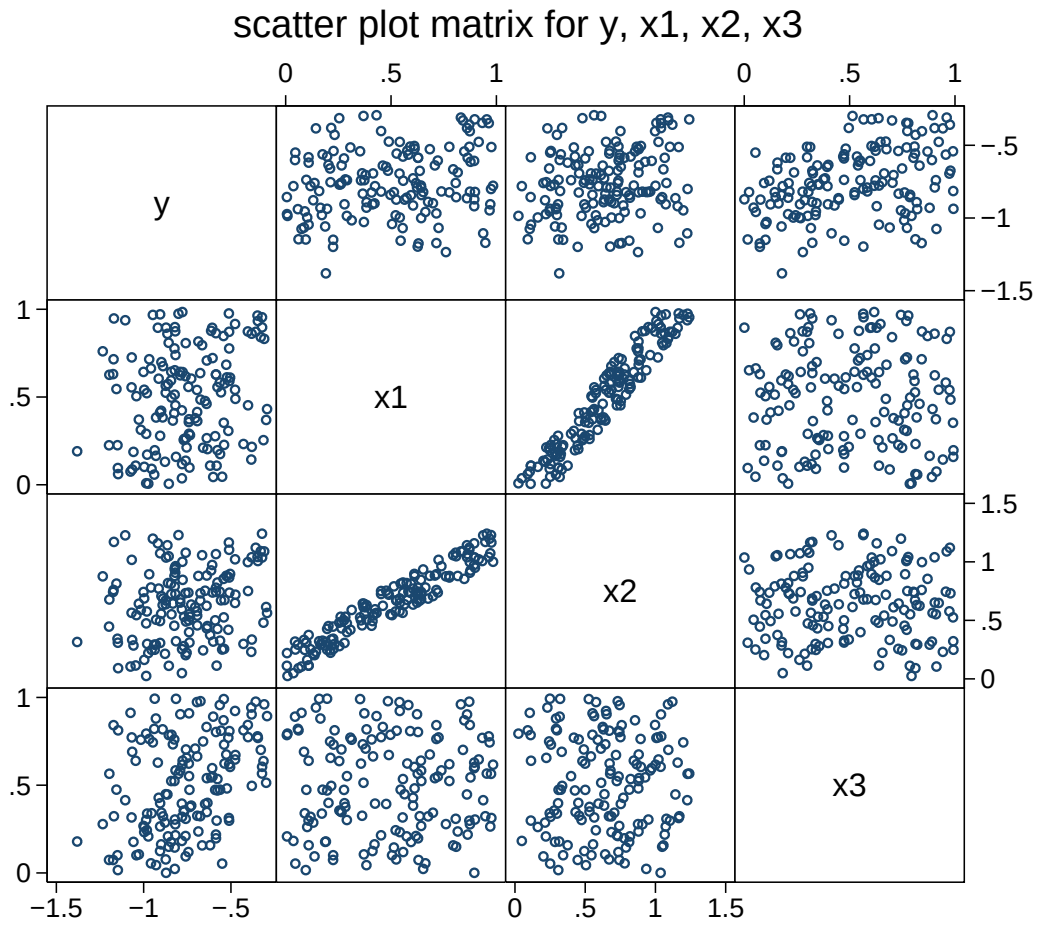


Figure 28: *Scatter plot matrix for the variables* y, x_1, x_2 *and* x_3.

Model choice and variable selection

Variable	Coefficient	Standard-error	t-value	p-value	95% Confidence-interval
intercept	-0.970	0.047	-20.46	$<$0.001	-1.064 -0.877
x_1	0.146	0.187	0.78	0.436	-0.224 0.516
x_2	0.027	0.177	0.15	0.880	-0.323 0.377
x_3	0.227	0.052	4.32	$<$0.001	0.123 0.331

Table 4: *Results for the model based on covariates x_1, x_2 and x_3.*

Model choice and variable selection

Variable	Coefficient	Standard-error	t-value	p-value	95% Confidence-interval
$intercept$	-0.967	0.039	-24.91	$<$0.001	-1.042 -0.889
x_1	0.173	0.055	3.17	0.002	0.065 0.281
x_3	0.226	0.052	4.33	$<$0.001	0.123 0.330

Table 5: *Results for the correctly specified model based on covariates* x_1 *and* x_3*.*

Model choice and variable selection - theoretical insights

We focus on the following questions:

1. *Irrelevant Variables:* What can be said about the bias and the variance of the least squares estimator, in the case that we include irrelevant variables in the model?

2. *Missing Variables:* What can be said about the bias and the variance of the least squares estimator, if we omit relevant variables in the model?

3. *Prediction Quality:* What effect does the model specification, more specifically the selected variables in the model, have on prediction?

Bias and Variance of the Least Squares Estimator

- Consider a partition of the available explanatory variables $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_k})'$ with $x_0 \equiv 1$ into the subsets $\mathbf{x_1} = (\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{k_1}})'$ and $\mathbf{x_2} = (\mathbf{x_{k_1+1}}, \ldots, \mathbf{x_k})'$.

- We look at the two models

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \boldsymbol{\varepsilon}$$

  and

$$\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{u}.$$

- The first model uses all available variables. The second model uses only the subset $\mathbf{x_1}$.

Bias and Variance of the Least Squares Estimator

- For the submodel we obtain the least squares estimators

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad \tilde{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

  respectively.

- For the estimator $\tilde{\beta}_1$ of the submodel, we obtain

$$\mathsf{E}(\tilde{\beta}_1) = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2$$

  and

$$\mathsf{Cov}(\tilde{\beta}_1) == \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

- For the estimator $\hat{\beta}$ in the full model we obtain

$$\mathsf{E}(\hat{\beta}) = \beta \qquad \mathsf{Cov}(\hat{\beta}) = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}.$$

Bias and Variance of the Least Squares Estimator

We now investigate the following two situations:

- *Missing variables:* Even though the complete model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ is correct, we mistakenly estimate the reduced model $\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{u}$. In this case we neglect the relevant variables $\mathbf{x_2}$.

- *Irrelevant Variables:* Even though the reduced model $\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{u}$ is correct, we mistakenly estimate the full model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$. In this case, we included irrelevant variables in the model. The variables in $\mathbf{x_2}$ are redundant.

Bias and Variance of the Least Squares Estimator

In the first case of missing variables the following applies:

- β_1 is biased. An exception is the case when $\mathbf{X_1'X_2} = \mathbf{0}$, i.e. every variable in $\mathbf{X_1}$ is uncorrelated to every variable in $\mathbf{X_2}$.

- It can be shown that the difference $\text{Cov}(\hat{\beta}_1) - \text{Cov}(\tilde{\beta}_1)$ of covariance matrices is positive semi-definite. This implies that the components of the estimator $\tilde{\beta}_1$ based on the submodel $\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{u}$ show a smaller variance than the corresponding components of the estimator $\hat{\beta}_1$ based on the correct model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$. Thus we have $\text{Var}(\hat{\beta}_j) \geq \text{Var}(\tilde{\beta}_j)$.

- It can be shown that situations exist, in which the components in $\tilde{\beta}_1$ based on the misspecified submodel actually show a smaller MSE than the components in $\hat{\beta}_1$, which are based on the full model, i.e. $\text{MSE}(\hat{\beta}_j) \geq \text{MSE}(\tilde{\beta}_j)$.

Bias and Variance of the Least Squares Estimator

In the second case of irrelevant variables, we have:

- Even though irrelevant variables were considered, $\hat{\beta}$ is unbiased. Of course, the estimator $\tilde{\beta}_1$ based on the true model is also unbiased.

- It can be shown that the estimators for the components in β_1 based on $\hat{\beta}$ have larger variance than based on $\tilde{\beta}_1$. Thus we have $\text{Var}(\hat{\beta}_j) \geq \text{Var}(\tilde{\beta}_j)$. If the estimated model contains irrelevant variables, then the precision of the estimators decreases.

We can reach the following conclusion: Preferably, the specified model should not contain irrelevant covariates. Moreover, we should aim for a sparse model so that bias and variance, and thus MSE, are small.

Consequences of the Model Specification on Prediction

- Next we take a look at prediction quality in linear models. Thereby, we do not necessarily assume that the model is correctly specified.

- We assume independent observations y_i, $i = 1, \ldots, n$, with expectation $\mathsf{E}(y_i) = \mu_i$ and variance $\mathsf{Var}(y_i) = \sigma^2$.

- The variables $x_0 = 1, x_1, \ldots, x_k$ are available as potential regressors.

- We assume that a subset $M \subset \{0, 1, 2, \ldots, k\}$ of the available variables will be used for estimation.

- The specified model is defined by the subset of included covariates with corresponding design matrix $\mathbf{X_M}$.

Consequences of the Model Specification on Prediction

- For the least squares estimator we obtain

$$\hat{\beta}_{\mathbf{M}} = (\mathbf{X}'_{\mathbf{M}}\mathbf{X}_{\mathbf{M}})^{-1}\mathbf{X}'_{\mathbf{M}}\mathbf{y}.$$

- An estimator $\hat{\mathbf{y}}_M$ for the vector μ of means $\mu_i = \mathsf{E}(y_i)$ is given by

$$\hat{\mathbf{y}}_M = \mathbf{X}_{\mathbf{M}}\hat{\beta}_{\mathbf{M}}.$$

- We can view the estimator $\hat{y}_{iM}$ also as a prediction for future observations $y_{n+i} = \mu_i + \varepsilon_{n+i}$, $i = 1, \ldots, n$, with given covariates $x_{i1}, \ldots, x_{ik}$.

- In the following, we derive a formula for the sum of the expected squared prediction errors $\sum \mathsf{E}(y_{n+i} - \hat{y}_{iM})^2$.

Consequences of the Model Specification on Prediction

To do so, we need the following, easily verifiable, properties of $\hat{\mathbf{y}}_M$:

- Expectation:

$$\mathsf{E}(\hat{\mathbf{y}}_M) = \mathbf{X_M}(\mathbf{X_M'X_M})^{-1}\mathbf{X_M'}\mathsf{E}(\mathbf{y}).$$

- Covariance matrix:

$$\mathsf{Cov}(\hat{\mathbf{y}}_M) = \sigma^2\mathbf{X_M}(\mathbf{X_M'X_M})^{-1}\mathbf{X_M'}.$$

- Sum of the variances:

$$\sum_{i=1}^{n}\mathsf{Var}(\hat{y}_{iM}) = \sigma^2\mathsf{tr}(\mathbf{X_M}(\mathbf{X_M'X_M})^{-1}\mathbf{X_M'}) = |\mathbf{M}|\,\sigma^2,$$

  where $|M|$ represents the cardinal number of M, i.e. the number of the covariates included in the model. The sum of the variances increases as more covariates are included in the model.

Consequences of the Model Specification on Prediction

- Sum of the mean squared errors (SMSE):

$$\begin{aligned}
\text{SMSE} &= \sum_{i=1}^{n} \mathsf{E}(\hat{y}_{iM} - \mu_i)^2 \\
&= \sum_{i=1}^{n} \mathsf{E}\left((\hat{y}_{iM} - \mu_{iM}) + (\mu_{iM} - \mu_i)\right)^2 \\
&= \sum_{i=1}^{n} \mathsf{Var}(\hat{y}_{iM}) + 2\sum_{i=1}^{n} \mathsf{E}\left((\hat{y}_{iM} - \mu_{iM})(\mu_{iM} - \mu_i)\right) + \sum_{i=1}^{n}(\mu_{iM} - \mu_i)^2 \\
&= |M|\,\sigma^2 + \sum_{i=1}^{n}(\mu_{iM} - \mu_i)^2.
\end{aligned}$$

Here we used $\mu_{iM} = \mathsf{E}(\hat{y}_{iM})$ as an abbreviation for the expectation of the estimator $\hat{y}_{iM}$.

Consequences of the Model Specification on Prediction

These properties provide us with the expected squared prediction error:

$$
\begin{aligned}
\text{SPSE} &= \sum_{i=1}^{n} \mathsf{E}(y_{n+i} - \hat{y}_{iM})^2 \\
&= \sum_{i=1}^{n} \mathsf{E}((y_{n+i} - \mu_i) - (\hat{y}_{iM} - \mu_i))^2 \\
&= \sum_{i=1}^{n} (\mathsf{E}(y_{n+i} - \mu_i)^2 - 2\mathsf{E}((y_{n+i} - \mu_i)(\hat{y}_{iM} - \mu_i)) + \mathsf{E}(\hat{y}_{iM} - \mu_i)^2) \\
&= \sum_{i=1}^{n} \mathsf{E}(y_{n+i} - \mu_i)^2 + \sum_{i=1}^{n} \mathsf{E}(\hat{y}_{iM} - \mu_i)^2 \\
&= n\,\sigma^2 + \text{SMSE} \\
&= n\,\sigma^2 + |M|\,\sigma^2 + \sum_{i=1}^{n} (\mu_{iM} - \mu_i)^2
\end{aligned}
$$

Note that in line 3 of the above derivation, the expectation for the cross product term can be written as the product of expectations due to the independence of $\hat{y}_{iM}$ and y_{n+i}. This way the entire term becomes zero.

Consequences of the Model Specification on Prediction

Thus, the expected squared prediction error can be decomposed into three additive terms:

- *Irreducible Prediction Error:* The first term $n\,\sigma^2$ depends on the error variance. Hence, it cannot be reduced, even by sophisticated inference techniques. This term is therefore referred to as the irreducible prediction error.

- *Variance:* The second term consists of the sum of variances $\text{Var}(\hat{y}_{iM})$ of the estimators $\hat{y}_{iM}$. This term can be manipulated through model choice. It becomes smaller as fewer variables are included in the model.

- *Squared Bias:* The last term $\sum(\mu_{iM} - \mu_i)^2$ can be seen as a bias term. It consists of the squared bias of the estimator $\hat{y}_{iM}$ for the expectation μ_i. This term can also be manipulated through model choice and becomes smaller as more variables are included in the model.

Consequences of the Model Specification on Prediction

- The decomposition of the expected prediction error into an irreducible error, a variance term, and a squared bias term is not limited to linear models, but rather a *fundamental property of prediction* in all statistical models.
- The formula for SPSE shows a classical bias-variance trade-off.
- The more complex the model, the smaller the squared bias and the greater the variance. On the contrary, simpler models show a greater squared bias and in return for that a smaller variance.
- This bias-variance trade-off is not only characteristic for linear models, but for all statistical models.

Model choice: Minimizing SPSE

Estimate SPSE using new and independent data

- If in fact additional observations y_{n+i} are available, we are able to estimate $\text{SPSE} = \sum \text{E}(y_{n+i} - \hat{y}_{iM})^2$ simply by $\widehat{\text{SPSE}} = \sum_{i=1}^{n}(y_{n+i} - \hat{y}_{iM})^2$.

- In practice, it is usually not possible to use this approach, as it rarely happens that additional observations are collected.

- An alternative procedure is the following:
  - Randomly split the data into two parts, i.e. a test and a validation sample.
  - Use the test data set to estimate the specified model.
  - Use the validation set to assess the goodness-of-fit, i.e. for the estimation of SPSE.

Model choice: Minimizing SPSE

Estimate SPSE using existing data

- A naive estimator for SPSE would be the use of the squared sum of residuals $\sum(y_i - \hat{y}_{iM})^2$.
- Note that this sum *underestimates* on average the expected prediction error, as it can be shown that

$$\mathsf{E}\left(\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2\right) = \mathsf{SPSE} - 2|M|\sigma^2.$$

- Thus a better estimate for SPSE is given by

$$\widehat{\mathsf{SPSE}} = \sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2 + 2|M|\hat{\sigma}^2.$$

Model choice: Minimizing SPSE

- Accordingly, we choose a model that minimizes $\widehat{\text{SPSE}}$. In doing so we have to keep in mind that we always use the same estimator for $\hat{\sigma}^2$. Preferably, this estimator should be based on the full model with all available variables, in order to keep the bias in $\hat{\sigma}^2$ small.

- The criterion $\widehat{\text{SPSE}}$ has the typical structure of many model choice criteria. It consists of two terms: The first term, the sum of squared residuals, measures the goodness-of-fit and becomes smaller the more complex the model becomes. The second term $2|M|\hat{\sigma}^2$ measures model complexity and becomes smaller as models become simpler.

Model choice: Corrected coefficient of determination

- When comparing different models, the use of the coefficient of determination is limited, since the coefficient of determination will always increase (never decrease) with the addition of a new covariate into the model.

- The *corrected coefficient of determination* adjusts for this problem, by including a correction term for the number of parameters.

- The corrected coefficient of determination is defined by

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2).$$

  We advice against its usage, since the "penalty" for newly included covariates appears to be too small. It can be shown that $\bar{R}^2$ already increases when a variable with a t-value greater than 1 is included in the model, implying we would include variables with a p-value of about 0.3.

Model choice: Mallow's C_p

- Mallow's C_p ("Complexity parameter") is defined by

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + 2|M|.$$

- C_p can be understood as an estimate of SMSE/σ^2.

- Thus minimizing C_p produces the same optimal model as minimizing $\widehat{\text{SPSE}}$.

Model choice: Akaike information criterion

- AIC is defined by

$$\text{AIC} = -2 \cdot l(\hat{\beta}_{\mathbf{M}}, \hat{\sigma}^{\mathbf{2}}) + \mathbf{2}\,(|\mathbf{M}| + \mathbf{1}),$$

  where $l(\hat{\beta}_{\mathbf{M}}, \hat{\sigma}^{\mathbf{2}})$ is the maximum value of the log-likelihood.

- Smaller values of the AIC correspond to a better model fit.

- In a linear model with Gaussian errors, we obtain

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2\,(|M| + 1).$$

Model choice: Akaike information criterion

- Figure 29 plots AIC for the simulated data from the Example on page 111 as a function of the polynomial degree.

- AIC obtains a minimum for $l = 2$ resulting in a reasonable model, even though we do not obtain the polynomial order of the true model with $l = 3$.

Model choice: Akaike information criterion

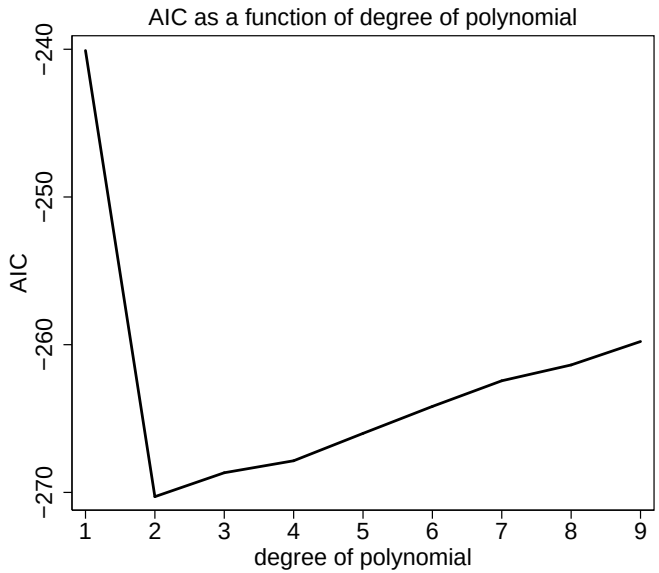


Figure 29: *AIC as a function of the polynomial degree for the simulated data of the example on page 111.*

Model choice: Cross Validation

Cross validation is based on the following general principle:

- Partition the data set into r subsets $1, \ldots, r$, of similar size.
- Start with the first data set as validation set and use the combined remaining $r-1$ data sets for parameter estimation. Based on the estimates, obtain predictions for the validation set and determine the sum of the squared prediction errors.
- Cycle through the partitions using the second, third, up to the rth data set as validation sample, and all other data sets for estimation. Determine the sum of squared prediction errors.
- Use the model with the smallest sum of squared prediction errors, where the final parameter estimates reflect all data. The partition into test and validation samples serves only to estimate the expected squared prediction error.

Model choice: "leave-one-out" Cross Validation

- "leave-one-out" cross validation uses all observations with the exception of *one* for the estimation of model parameters.

- We use this "leave-one-out" estimator to predict the deleted observation and to determine the squared prediction error.

- If we denote the "leave-one-out" estimator with $\hat{y}_{iM}^{-i}$, we obtain the cross validation score

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{iM}^{-i})^2.$$

- It can be shown

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_{iM}}{1 - h_{iiM}}\right)^2,$$

  where h_{iiM} are the diagonal elements of the hat matrix $\mathbf{H_M} = \mathbf{X_M}(\mathbf{X_M'X_M})^{-1}\mathbf{X_M'}$.

Practical Use of Model Choice Criteria

We can use the various model choice criteria to select the most promising models from candidate models. We recommend the following approach:

- On the basis of scientific knowledge, perhaps gained from previous research, we obtain a preselection of potential models. The models can differ in the number of variables but also in model type (e.g. linear vs. nonlinear). The total number of potential models should be as small as possible.

- All potential models can now be assessed with the aid of one of the various model choice criteria. The summary of the results should not be restricted to the "best" model. As a rule, there are a number of competitive models having approximately equal model fit, differing only in small aspects from each other. These differences cause some uncertainty regarding the conclusions.

Practical Use of Model Choice Criteria

If the number of potential models is large, we can use the following partially heuristic methods:

- *Complete Model Selection (All-Subset-Selection)*: In case that the number of covariates is smaller than about 40, we can determine the best model (in the sense of a model choice criterion) with the "leaps and bounds" algorithm. An implementation can be found e.g. in the R package `leaps`.

- *Forward Selection:* Based on a starting model, forward-selection includes one additional variable in every iteration of the algorithm. The variable which offers the greatest reduction of a preselected model choice criteria (C_p, AIC, CV) is chosen. The algorithm terminates if no further reduction is possible.

Practical Use of Model Choice Criteria

- *Backward Elimination:* Backward elimination starts with the full model containing all potential covariates. Subsequently, in every iteration, the covariate which provides the greatest reduction of the model choice criteria (C_p, AIC, CV) is eliminated from the model. The algorithm terminates if no further reduction is possible.

- *Stepwise Selection:* Stepwise selection is a combination of forward selection and backward elimination. In every iteration of the algorithm, the inclusion and the deletion of a variable are both possible.

Practical Use of Model Choice Criteria

- The listed procedures should not be confounded with an algorithm proposed by Efroymson in the 1960s, even though the approach is similar.

- In contrast to what has been proposed above, the Efroymson algorithm includes or excludes those variables in/from the model, which have the highest or lowest t-value.

- The procedure terminates when no variable that potentially needs to be included has a p-value of less than a previously fixed maximal p-value (e.g. 0.05) and when no variable that needs to be excluded has a p-value greater than a minimal p-value (e.g. 0.1).

Practical Use of Model Choice Criteria

This automatic procedure, which is implemented in all major statistical software packages, is often viewed as obscure among statisticians due to the following two reasons:

- Forward, backward, and stepwise selection usually provide different results. This also happens when using a global model choice criterion such as AIC. We can, however, compare the different selected models with the help of the global model choice criterion. When using the Efroymson approach, discrimination between the different models is impossible.

- The repetitive use of the t-test statistic, to assess whether or not a regression coefficient is different from zero, suggests exact tests. However, the t-test statistic does not follow a t-distribution under the null hypothesis, since during the selection process we do not test an arbitrary variable, but rather the variable with the *maximal* t-value.

Case study: prices of used cars

- We illustrate the approaches for model choice using data from the sales price of pre-owned VW Golf automobiles.

- Our goal is to model the relationship between the sales price in 1000 Euro (variable $price$) and the five explanatory variables "age of the car in months" (age), "kilometer reading in 1000 km" ($kilometer$), "number of months until the next appointment with the Technical Inspection Agency" (TIA), "ABS brake yes/no" ($extras1$), and "sunroof yes/no" ($extras2$).

Case study: prices of used cars

The plots of Figure **??** suggest the following effects:

- We can assume a linear or monotonically decreasing nonlinear effect for the variables age and $kilometer$, which could be appropriately modeled using (orthogonal) polynomials of degree three or less.

- The variable TIA appears to either have no effect or a very weak linear effect on the average sales price.

- Cars with ABS ($extras1$) seem to be slightly more expensive than cars without ABS; the effect, however, remains arguable.

- We can attest to no difference in the average sales price for models with or without sunroof ($extras2$).

- All in all, there seems to be a relationship with age and the kilometer reading. The effects of the remaining variables appear doubtful.

Case study: prices of used cars

- We first examine eight regression models (see Table 6), which do not differ in the modeling of the variables age and $kilometer$. For the remaining three regressors, all possible model combinations will be tested.

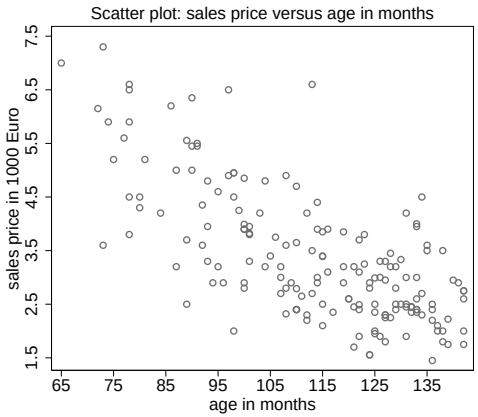
- Using the AIC criterion, we obtain the first model in Table 6 as the preliminary best model. Figure 30 displays the AIC values for the eight models under consideration. In addition, the AIC for a ninth model based on automatic variable selection is provided; see below.

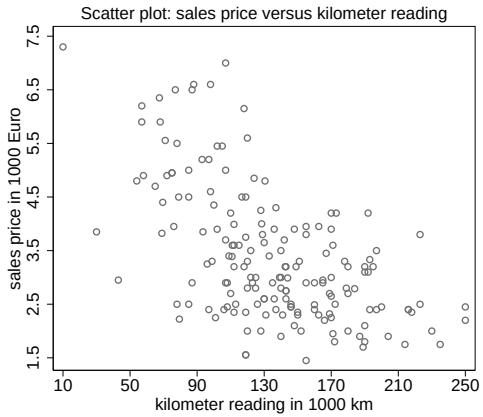
Case study: prices of used cars

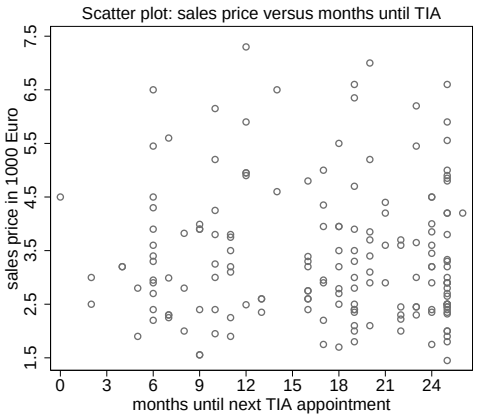
- Since only five explanatory variables are available, we can even determine the AIC best model with the help of the "Leaps and Bounds" algorithm. This model attains an AIC value of 389.35. It differs from the current "best model", in that it only makes use of polynomials of second degree (not third) for the variables age and $kilometer$.

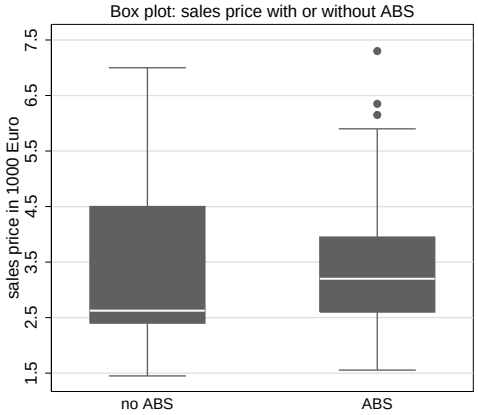
- Figure 30 shows that the obtained AIC for this ninth model is considerably smaller than the best AIC value of all the models that we examined so far.

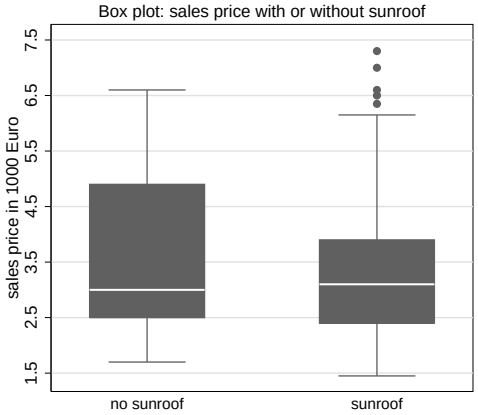
Case study: prices of used cars











Case study: prices of used cars

Model	*kilometer* degree 3	*age* degree 3	*extras1*	*extras2* linear	*TIA*	**AIC**
1	x	x				393.234 (1)
2	x	x	x			394.566 (2)
3	x	x		x		395.119 (4)
4	x	x			x	394.973 (3)
5	x	x	x	x		396.481 (6)
6	x	x	x		x	396.143 (5)
7	x	x		x	x	396.881 (7)
8	x	x	x	x	x	398.085 (8)

Table 6: *Prices of used cars: potential models. The values in brackets indicate the rank of the models according to AIC.*

Case study: prices of used cars

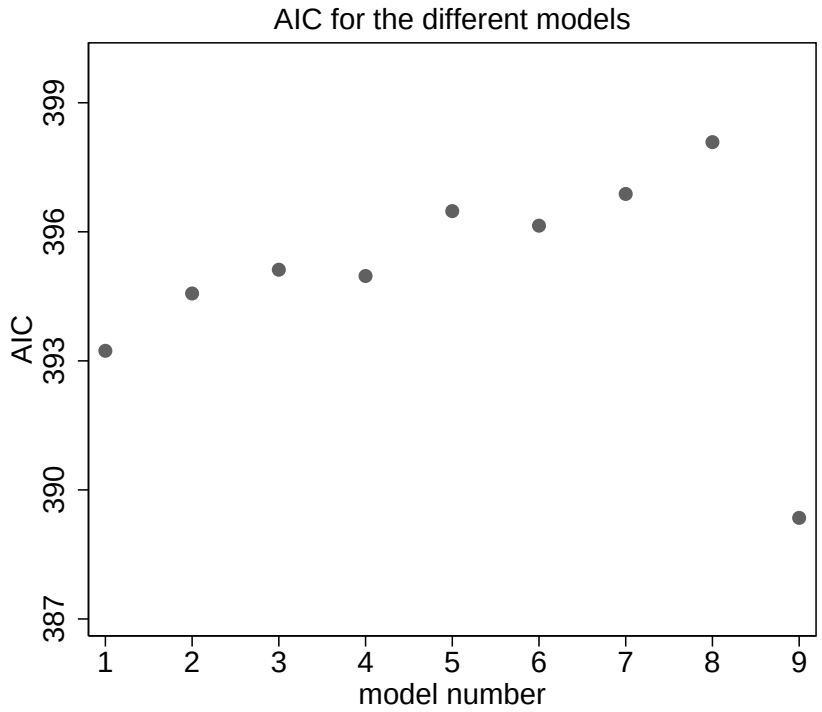


Figure 30: *Prices of used cars: AIC values for the potential models.*

Case study: prices of used cars

Variable	Coefficient	Standard-error	t-value	p-value	95% Confidence-interval	
intercept	3.397	0.056	60.220	<0.001	3.285	3.508
ageop1	-0.705	0.061	-11.470	<0.001	-0.826	-0.584
ageop2	0.187	0.057	3.270	0.001	0.074	0.300
kilometerop1	-0.439	0.061	-7.170	<0.001	-0.560	-0.318
kilometerop2	0.141	0.057	2.460	0.015	0.028	0.254

Table 7: *Prices of used cars: estimation results for the best model according to AIC.*

Case study: prices of used cars

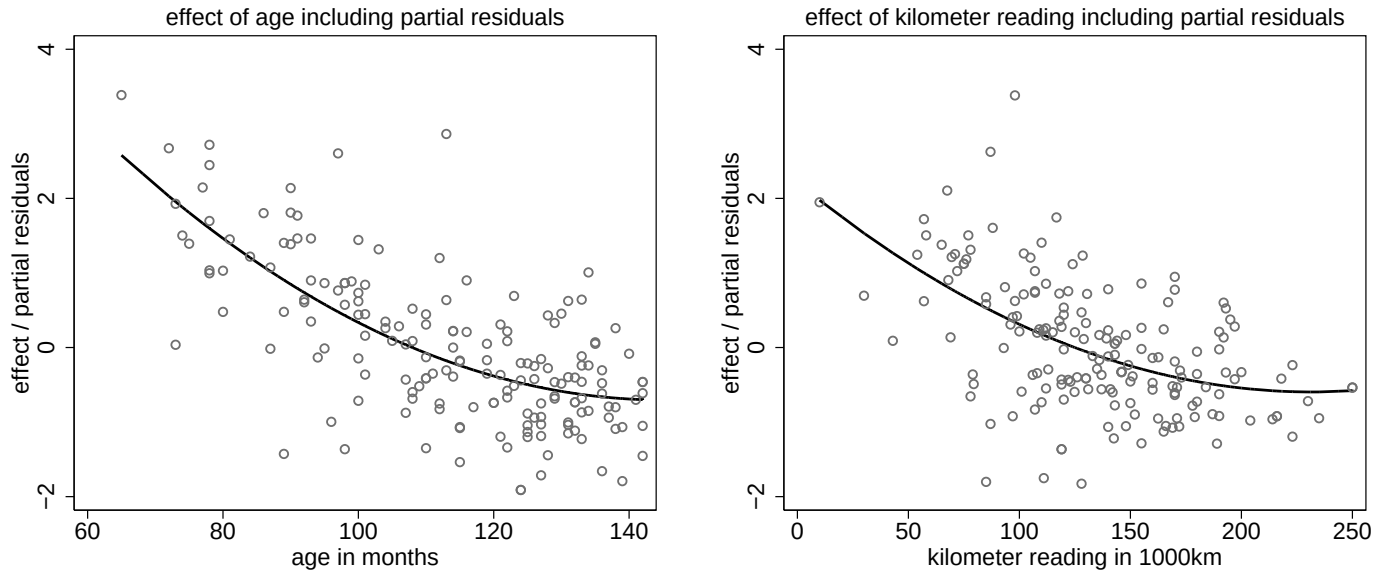


Figure 31: *Prices of used cars: model 9 based on all-subset selection, effects of age and kilometer reading including partial residuals.*

Advanced Regression Analysis

Scatterplot smoothers

Illustrative examples

Figure 32: *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).*

Illustrative examples

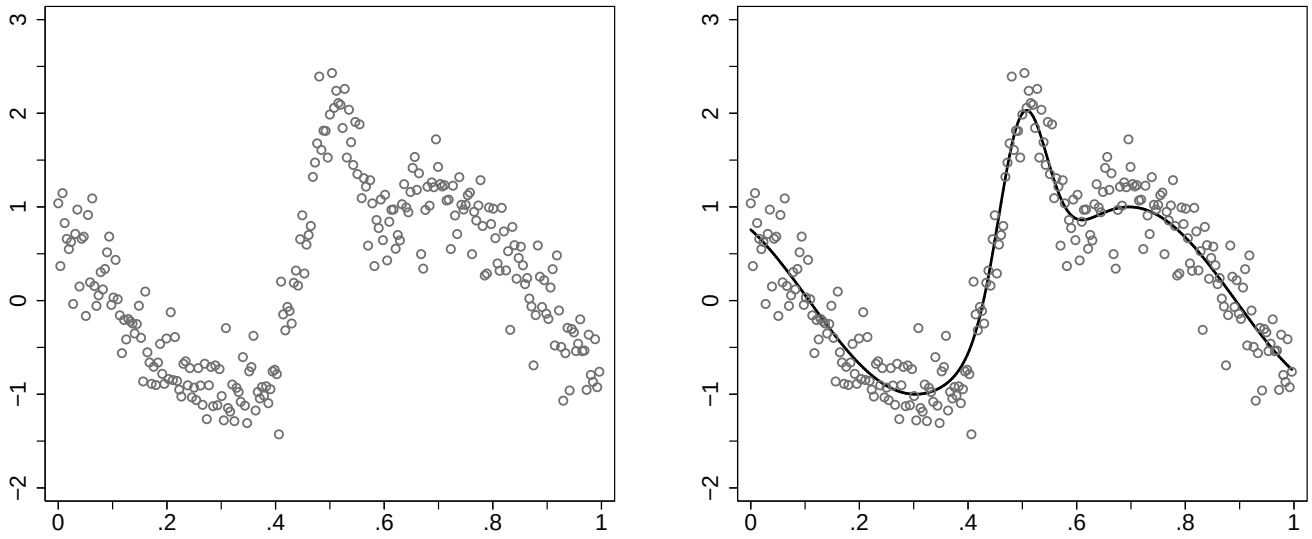


Figure 33: *Scatter plot of a simulated data set with nonlinear effect of the covariate: The right panel additionally shows the true covariate effect. The data have been simulated according to the model* $y = f(x) + \varepsilon$ *where* $f(x) = \sin(2(4x-2)) + 2\exp(-(16^2)(x-0.5)^2)$ *and* $\varepsilon \sim \mathrm{N}(0, 0.3^2)$.

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

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

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

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

Univariate Smoothing: Polynomials

- We revisit the polynomial model

$$f(z_i) = \gamma_0 + \gamma_1 z_i + \ldots + \gamma_l z_i^l.$$

- Figure 34 shows why a purely polynomial model is often insufficient to estimate nonlinear functions $f(z)$.

- We considered polynomials up to a degree of $l = 15$ for estimating f.

- As seen, polynomials are not able to properly reproduce the local maximum at $z = 0.5$, even with (the very high) degree of $l = 11$.

- For $l = 15$ this maximum is nearly obtained, but the estimated relationship is very wiggly and not quite satisfactory in other regions.

Univariate Smoothing: Polynomials

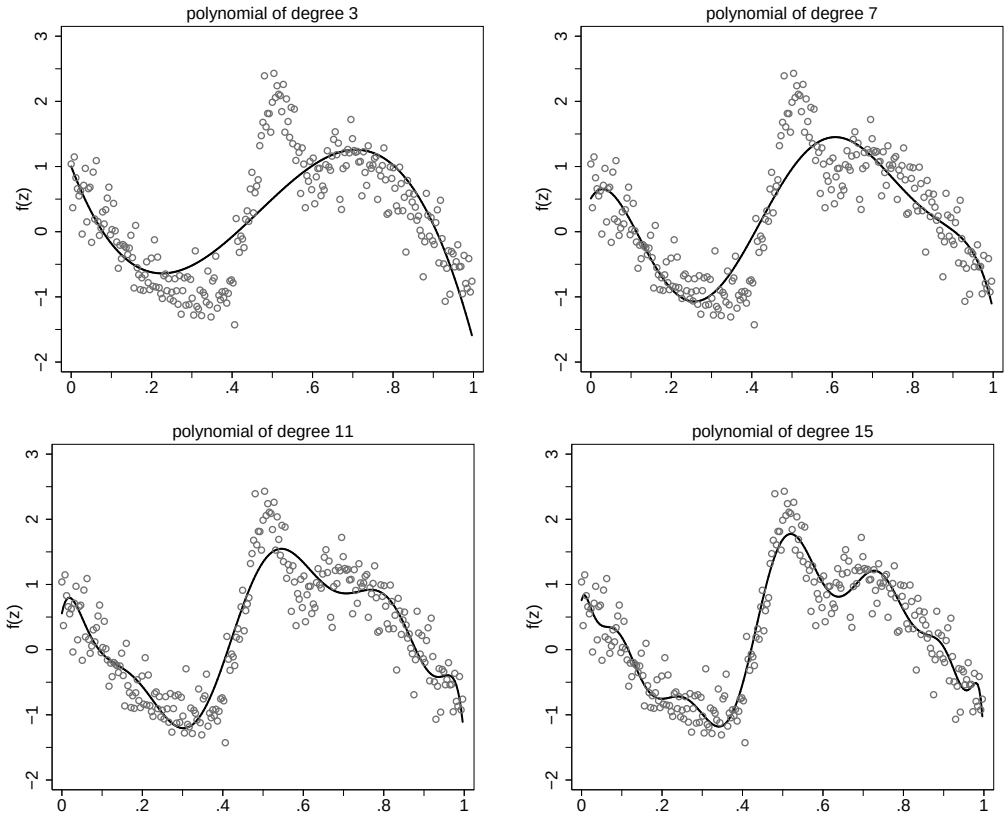


Figure 34: *Polynomial regression models for the simulated data set.*

Univariate Smoothing: Piecewise Polynomials

- To make the polynomial model more flexible, we next partition the domain of the covariate into intervals, and estimate separate polynomials in each interval.
- In the left panel of Figure 35 we split up the domain of z into ten intervals of width 0.1 and fit separate polynomials in each.
- This actually leads to a more flexible estimate of the function, which reflects the form of the true function.
- However, since the polynomials are fit separately on the intervals, the polynomial pieces do not lead to an overall smooth function and, in particular, they show different function values at the interval boundaries.
- Therefore, it would be desirable to impose further smoothness restrictions on the function at the interval boundaries to obtain a function estimate similar to the one shown in the right panel of Figure 35.

Univariate Smoothing: Piecewise Polynomials

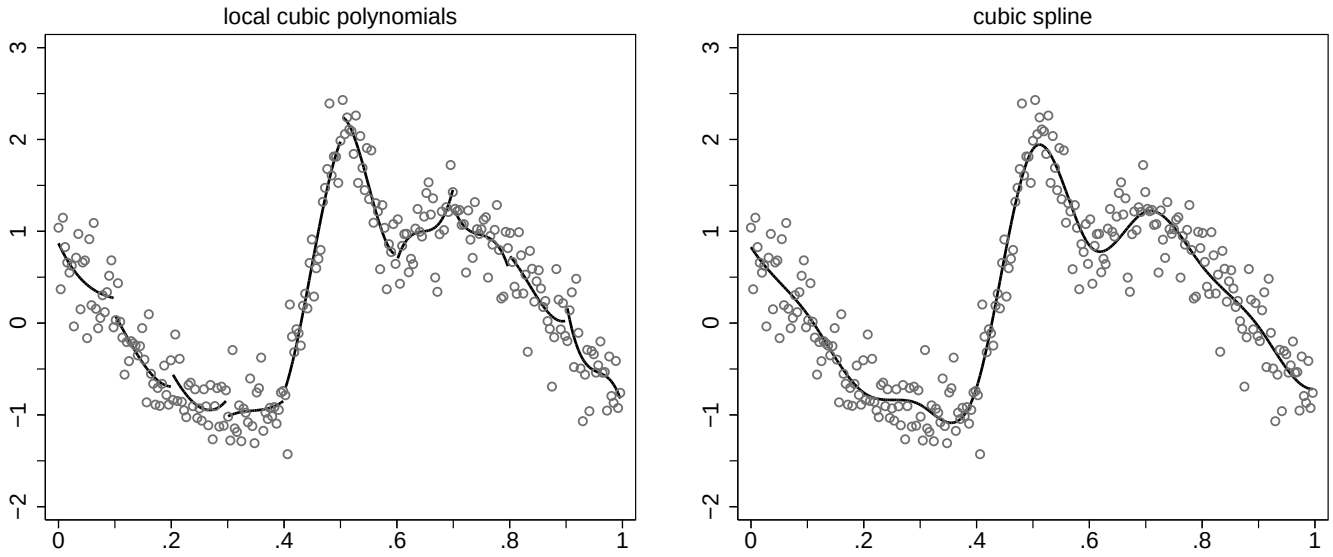


Figure 35: *Piecewise polynomial regression (left) and polynomial splines (right).*

Univariate Smoothing: Polynomial splines

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.

Univariate Smoothing: Polynomial splines

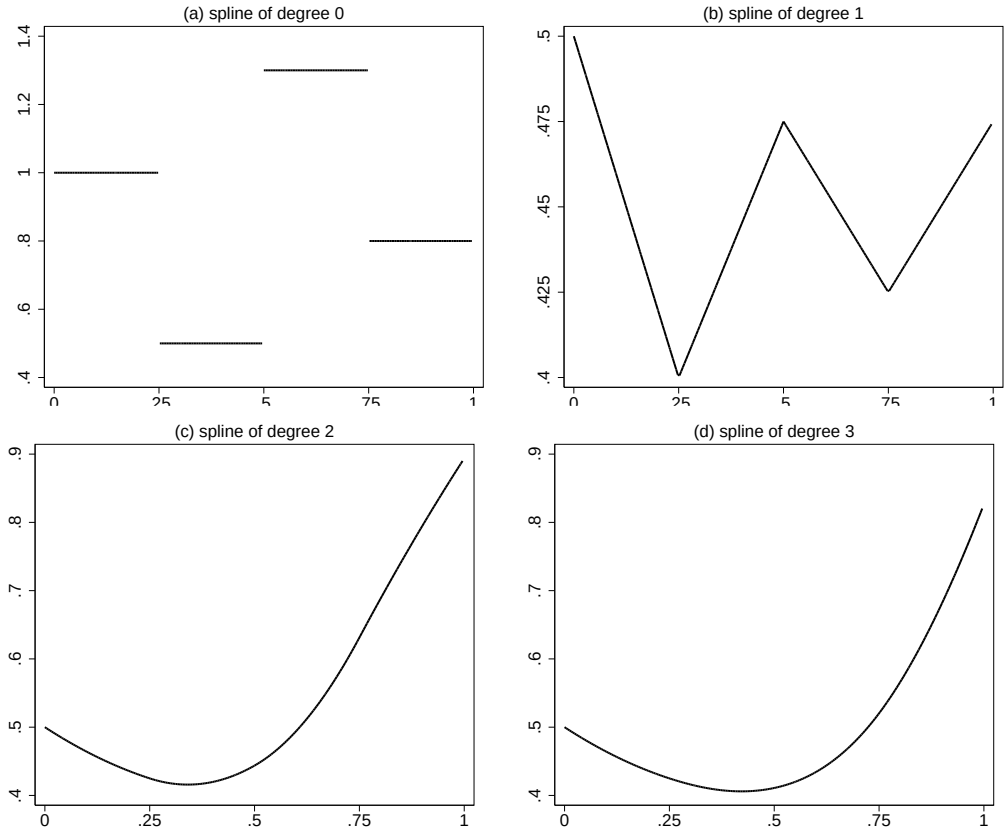


Figure 36: *Examples of polynomial splines of degree 0, 1, 2 and 3 with knots* $\kappa_1 = 0$, $\kappa_2 = 0.25$, $\kappa_3 = 0.5$, $\kappa_4 = 0.75$ *and* $\kappa_5 = 1$.

Univariate Smoothing: Polynomial splines

- It can be shown that each polynomial spline of degree l with knots $\kappa_1 < \ldots < \kappa_m$ can be uniquely determined as a linear combination of the $d = m + l - 1$ functions

$$B_1(z) = 1, \quad B_2(z) = z, \quad \ldots, \quad B_{l+1}(z) = z^l,$$
$$B_{l+2}(z) = (z - \kappa_2)_+^l, \quad \ldots, \quad B_d(z) = (z - \kappa_{m-1})_+^l.$$

- This yields the following representation of the nonparametric regression problem:

$$y_i = f(z_i) + \varepsilon_i = \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i.$$

- We call the functions $B_1, \ldots, B_d$ *basis functions*, since we can uniquely represent all polynomial splines by using these functions.

Univariate Smoothing: Polynomial splines

- Modeling $f(z)$ as a polynomial spline has the advantage that we can still understand the nonparametric regression model as a linear model, but with a possibly large number of parameters.

- If we define the vectors of the observed response variables $\mathbf{y}$ and the errors $\boldsymbol{\varepsilon}$, 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

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

  with the coefficient vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)'$.

Univariate Smoothing: Polynomial splines

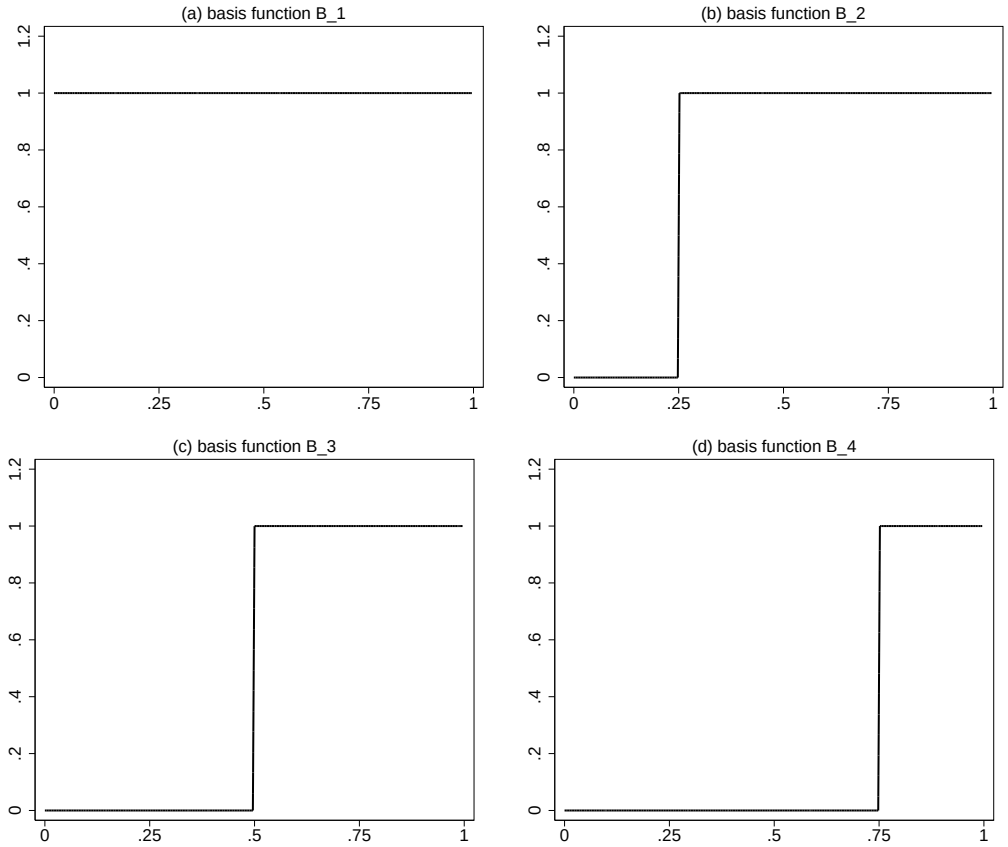


Figure 37: *TP basis for splines of degree 0 based on the knots* $\{0,0.25,0.5,0.75,1\}$.

Univariate Smoothing: Polynomial splines

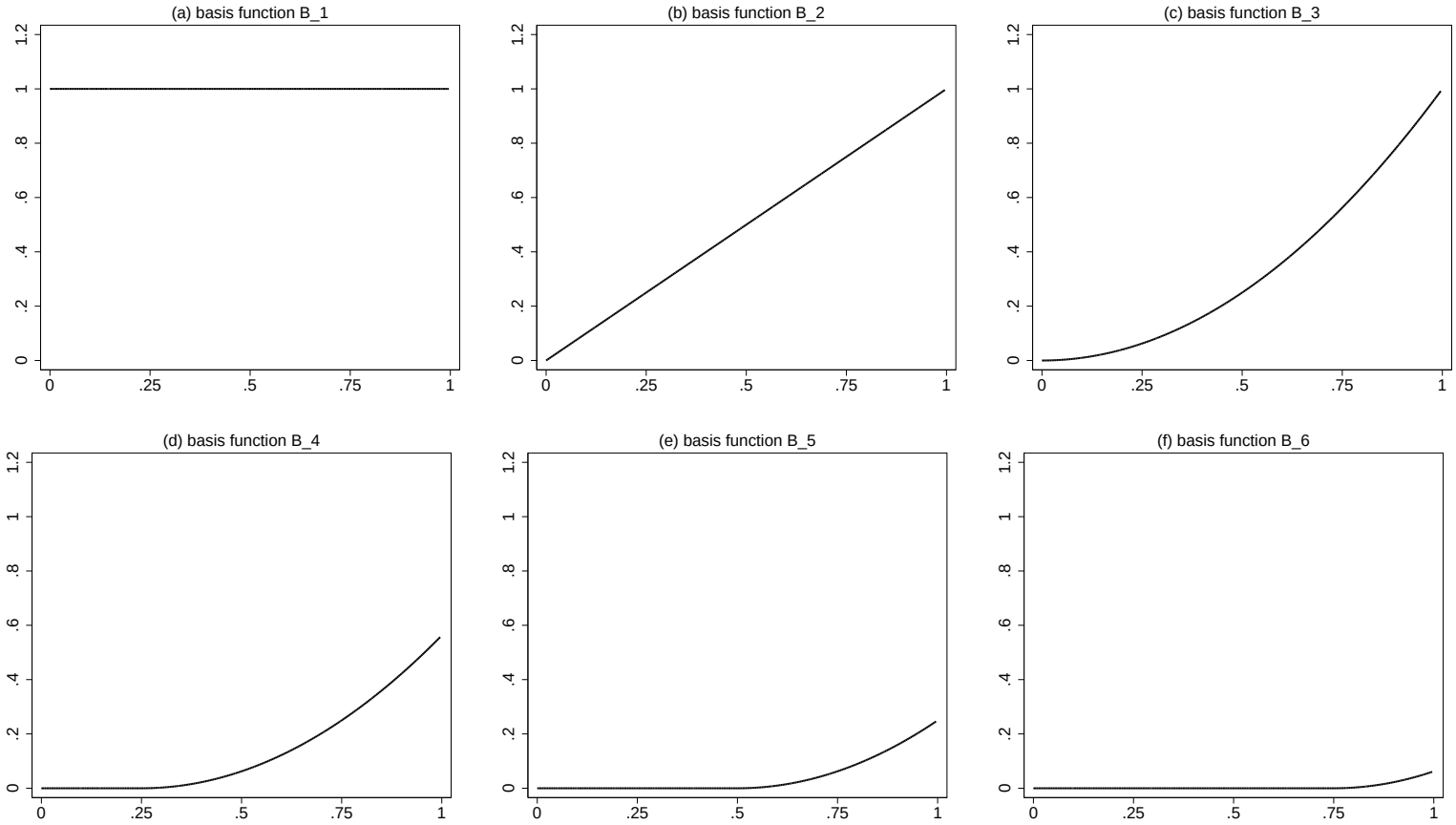


Figure 38: *TP basis for splines of degree 2 based on the knots* $\{0,0.25,0.5,0.75,1\}$.

Univariate Smoothing: Polynomial splines

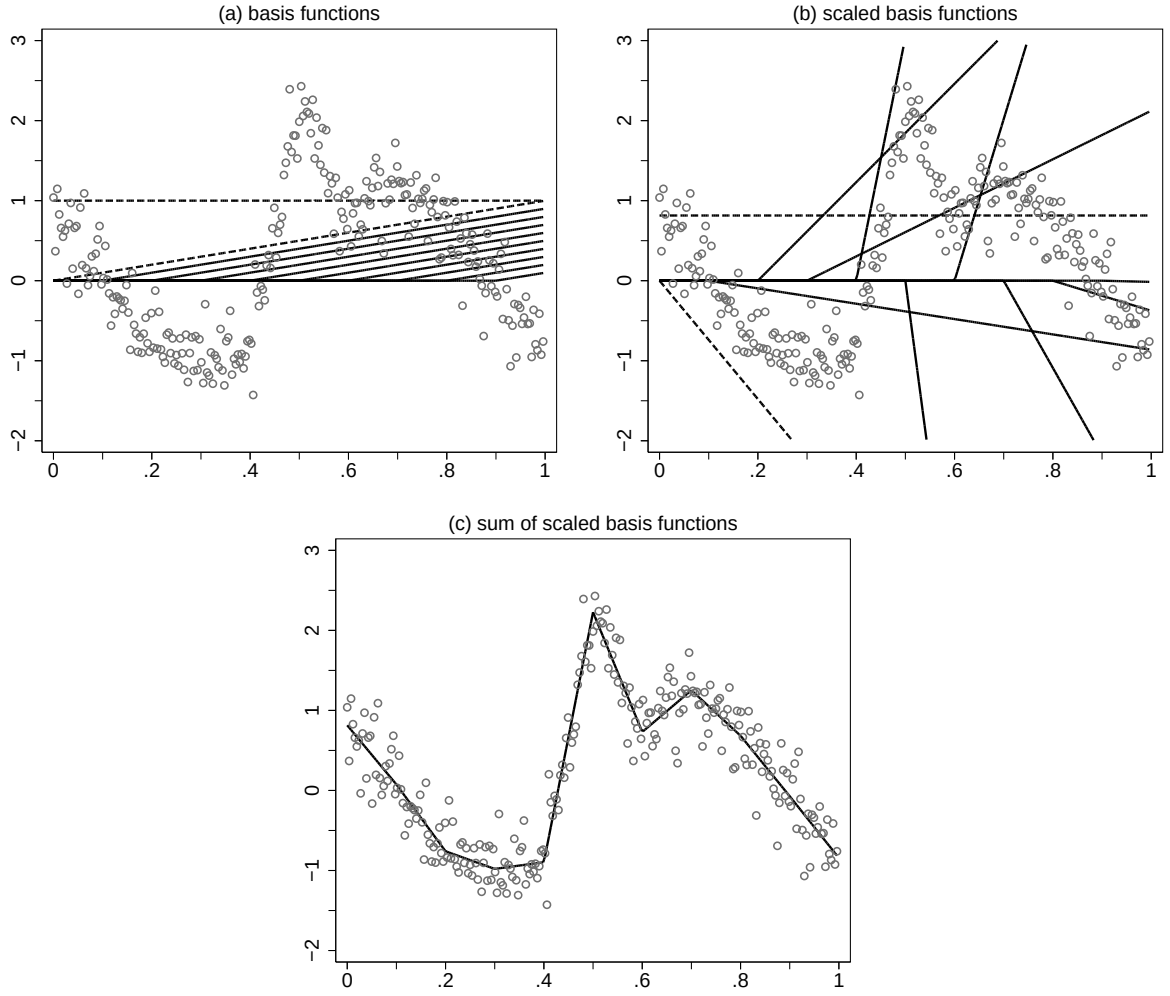


Figure 39: *Polynomial spline fit with linear truncated polynomials.*

Illustration: Malnutrition in Tanzania

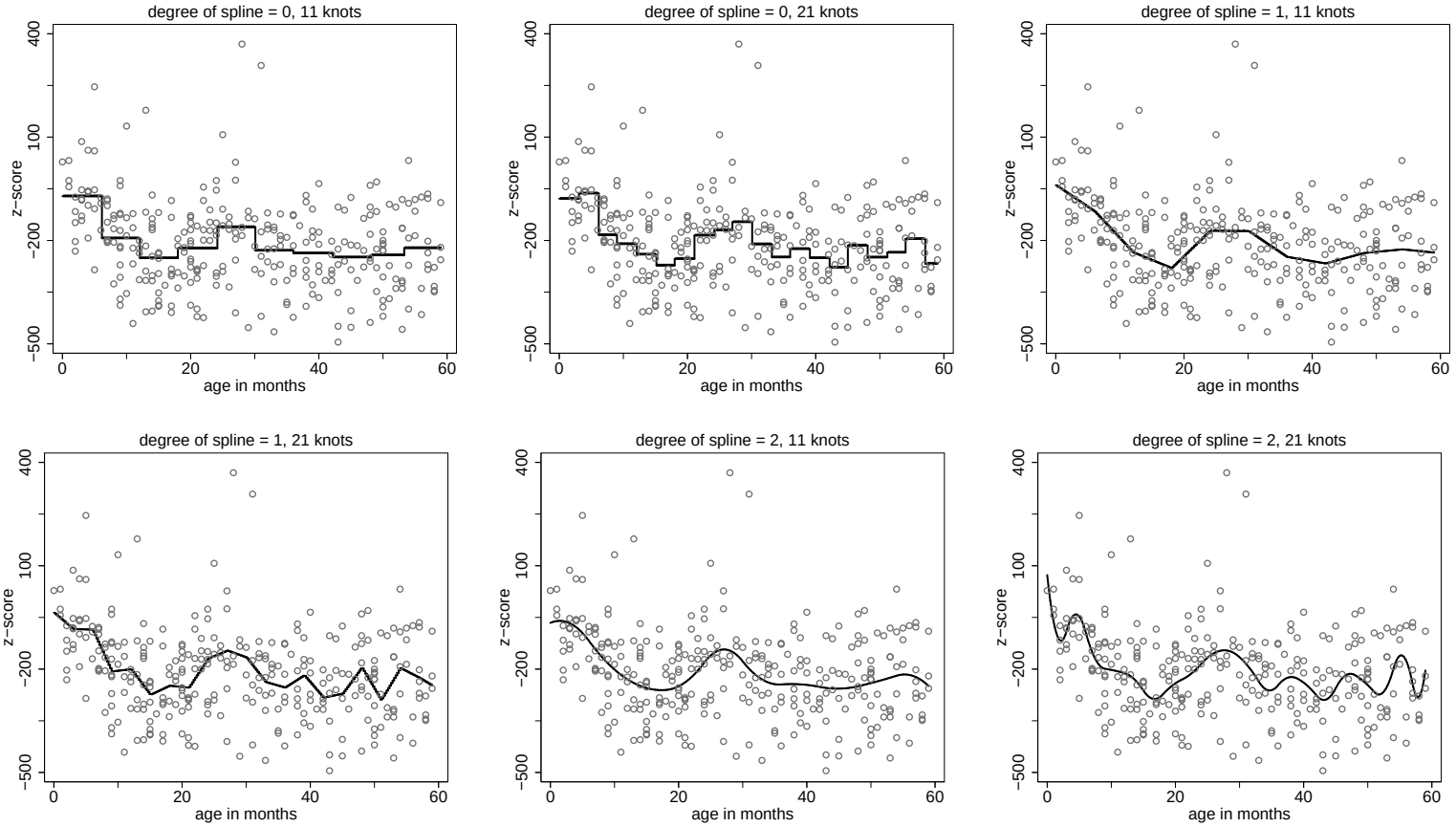


Figure 40: *Nonparametric estimates for the age effect based on polynomial splines.*

Influence of the knots

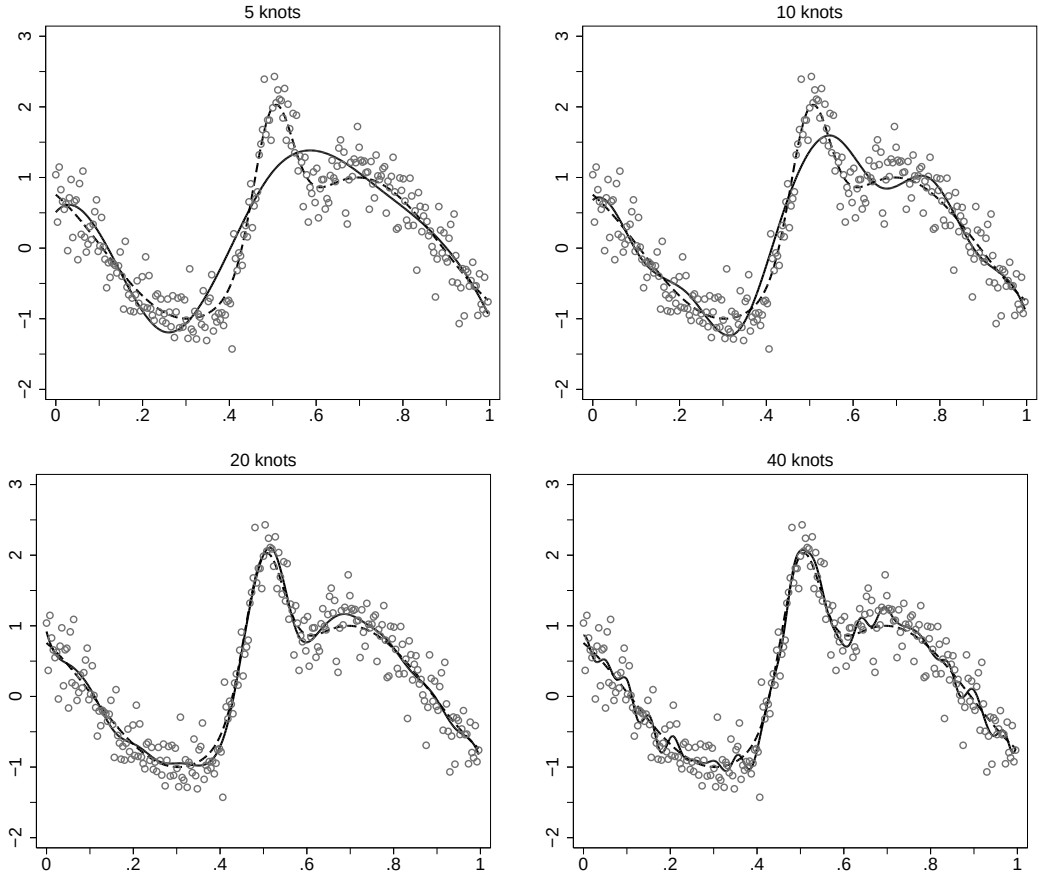


Figure 41: *Impact of the number of knots on cubic spline fits: The estimated function is represented as a solid line while the true function is superimposed as a dashed line.*

Choice of the knots

- *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. By doing so, we place many knots in areas where we have a large number of observations and therefore the distribution of the knots better adapts to the distribution of the explanatory variable.

Choice of the knots

- *Visual knot choice based on a scatter plot:* Studying the scatter plot of the data allows us to (subjectively) target the placement of the knots to either adapt the knot density to the variability of data, or to account for the specific context of the estimation problem.

Another spline basis: B-splines

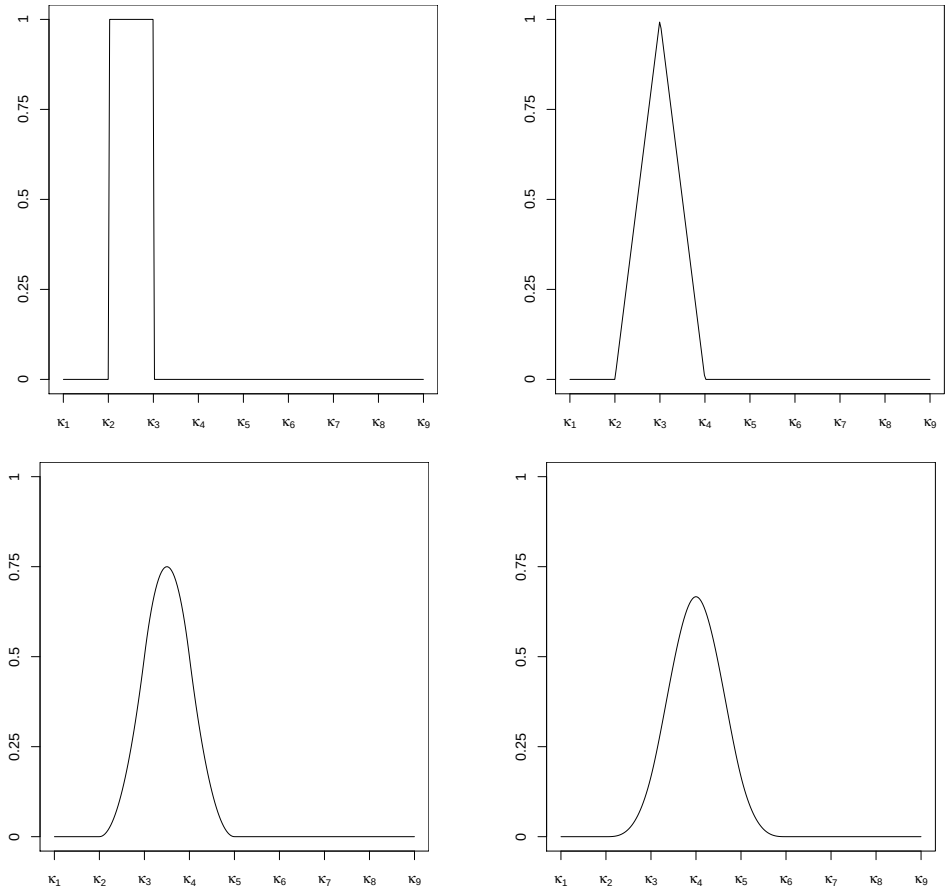


Figure 42: *Single B-spline basis functions for degrees* $l = 0, 1, 2, 3$ *and equidistant knots.*

Another spline basis: B-splines

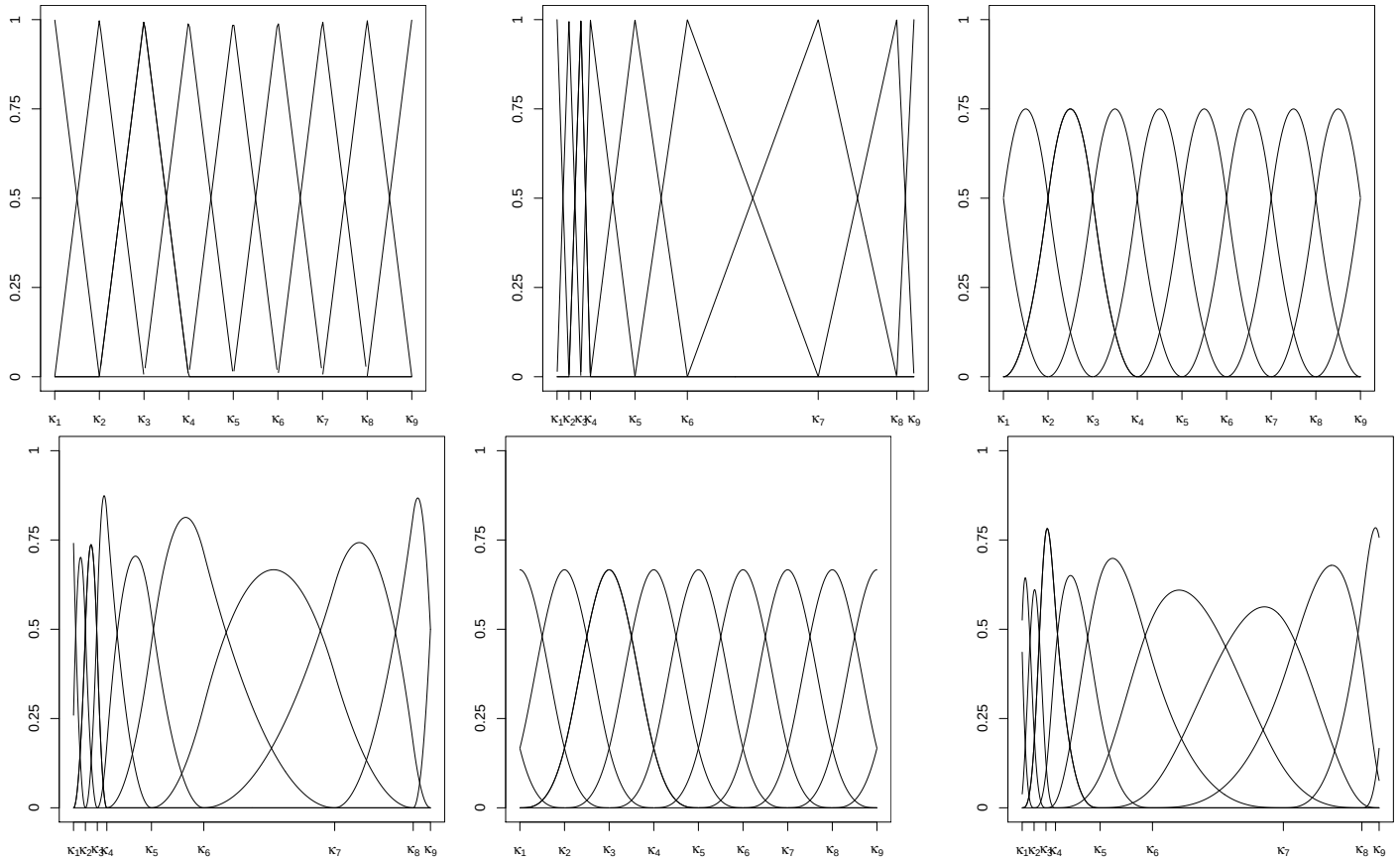


Figure 43: *B-spline bases of degree* $l = 1, 2, 3$ *with equidistant knots and unevenly distributed knots.*

Another spline basis: B-splines

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

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

- From Figure 42, we can immediately deduce the definition for B-splines of order $l = 0$ 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.

Another spline basis: B-splines

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

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

- To use the recursive definition of B-splines for the calculation of the basis functions, we need $2l$ *outer knots* outside of the domain $[a, b]$ in addition to the *interior knots* $\kappa_1, \ldots, \kappa_m$.

- This leads to the expanded knots sequence $\kappa_{1-l}, \kappa_{1-l+1}, \ldots, \kappa_{m+l-1}, \kappa_{m+l}$.

Another spline basis: B-splines

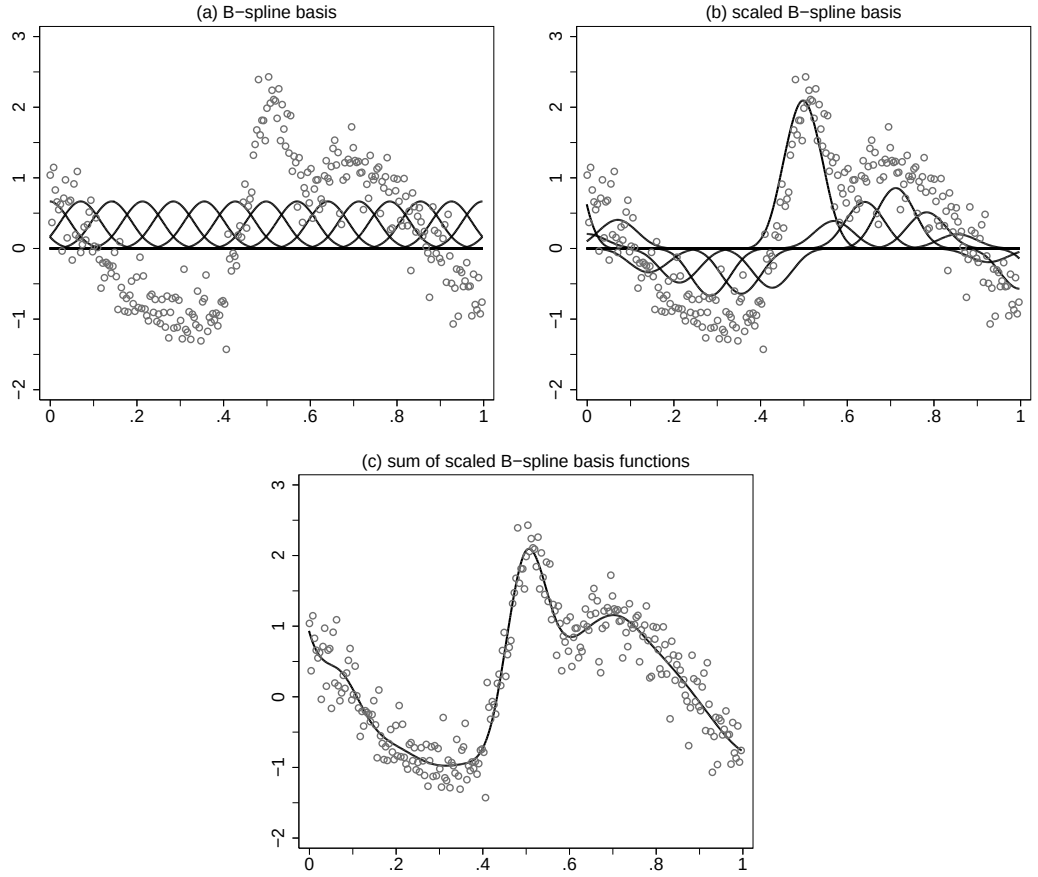


Figure 44: *Schematic representation of a nonparametric fit with cubic B-splines.*

Penalized-splines

- We first consider P-splines based on the TP basis, i.e.

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

- This basis consists of two different parts: the first $l + 1$ basis functions, which describe a global polynomial in z, and the truncated powers describing deviations from this polynomial.

- Estimated functions will be rough when there is a lot of variability in the second part of the model.

- Therefore, regularized function estimates may be defined by introducing a penalty for the coefficients of the corresponding basis functions.

Penalized-splines

- We can create such a penalty, e.g., using the sum of squared coefficients

$$\sum_{j=l+2}^{d} \gamma_j^2,$$

  so that large coefficients associated with the truncated powers are penalized.

- Instead of the usual residual sum of squares we now minimize the penalized residual sum of squares

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

Penalized-splines

- The penalty is constructed to discourage estimated functions that are too rough, thereby preventing overfitting to the data.

- The *smoothing parameter* $\lambda \geq 0$ controls the influence of the penalty.

- As $\lambda \to 0$, the effect of the penalty disappears so that the penalized residual sum of squares approximately corresponds to the standard residual sum of squares and we have an estimate for $\boldsymbol{\gamma}$ that is close to the least squares estimate.

- As $\lambda \to \infty$, the estimation criterion is dominated by the penalty such that $\hat{\gamma}_j = 0$ for $j = l+2, \ldots, d$, and the estimate for $f(z)$ approaches a polynomial of degree l.

Penalized-splines

- The main advantage of the penalization is that the smoothness of the estimated fit is no longer controlled by the number and the position of knots, but rather by one single real-valued parameter, namely the smoothing parameter.

- Since we use a large number of knots, the exact positioning of these knots is of minor importance.

- Figure 45 shows that penalized function estimates depend on the number of knots only very moderately (when an optimal value is chosen for the smoothing parameter).

- For a very small number of knots, the estimate may not be flexible enough. However, once a sufficiently large number of knots is used, the estimated function only slightly changes when the number of knots is increased further, provided that the smoothing parameter is adapted appropriately.

Penalized-splines

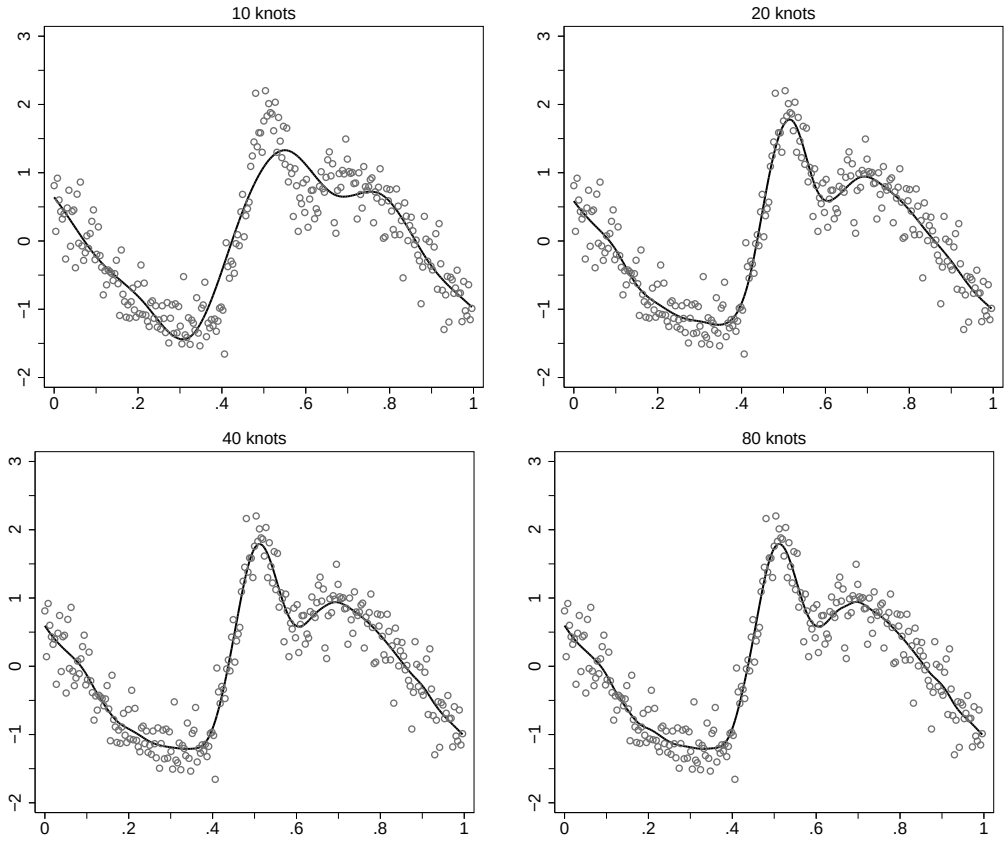


Figure 45: *Influence of the number of knots on estimated P-splines.*

Penalized-splines based on B-spline basis

- To characterize the smoothness of any type of function, the use of (squared) derivatives is appropriate, since these represent measures for the variability of a function.

- Penalties based on the second derivative such as

$$\lambda \int (f''(z))^2 dz, \tag{16}$$

  are particularly attractive since they measure the curvature of a function.

- This type of penalty forms the basis of the implementation of penalized splines in the R package `mgcv`

Penalized-splines based on B-spline basis

- A simple approximation to the derivative is available for B-splines, based on differences of neighboring coefficients.

- For instance, the first derivative of a B-spline can be written as a function of the first differences of the corresponding coefficient vector.

- Analogous to first order differences, we can use differences of a higher order r if we aim at a smooth function in terms of rth order derivatives.

Penalized-splines based on B-spline basis

- This leads to 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

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

Penalized-splines based on B-spline basis

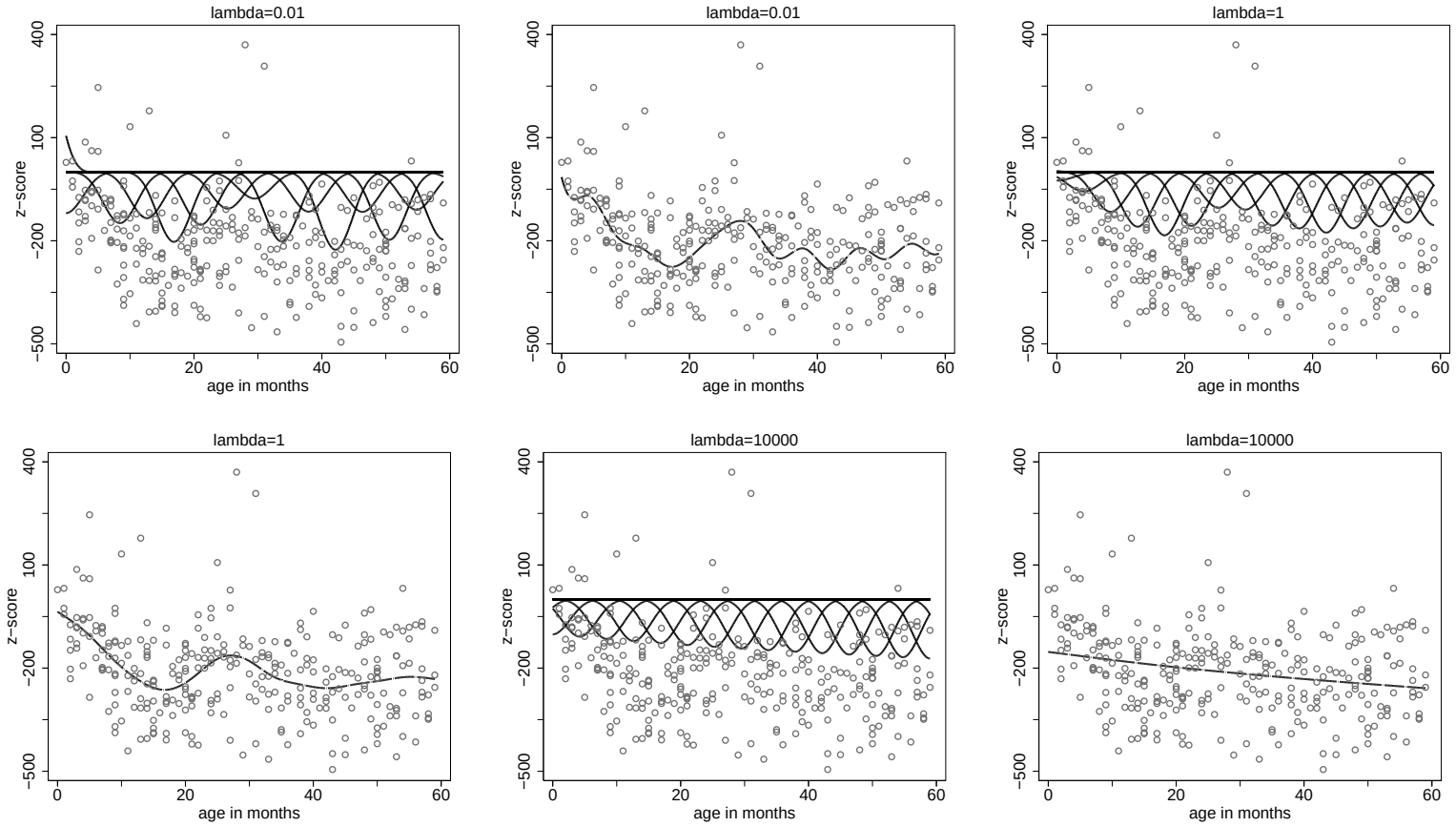


Figure 46: *Malnutrition in Tanzania: impact of the smoothing parameter on estimated P-splines with second order difference penalty.*

Penalized-splines based on B-spline basis

- For a large value of the smoothing parameter ($\lambda \to \infty$), the function estimate for $f(z)$ is close to linear in case of second order differences.

- More generally, as $\lambda \to \infty$, the fit approaches a polynomial of degree $r-1$ with r-th order differences, provided that the degree of the spline is at least as large as the order of the differences, i.e. $l \geq r$.

Penalized Least Squares Estimation

- For the derivation of the penalized least squares estimate (PLS estimate), it is advantageous to write the penalty in matrix notation.

- For the TP basis, this is simply achieved via

$$\lambda \sum_{j=l+2}^{d} \gamma_j^2 = \lambda \boldsymbol{\gamma}' \mathbf{K} \boldsymbol{\gamma}$$

with the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ and the *penalty matrix*

$$\mathbf{K} = \text{diag}(\underbrace{0, \ldots, 0}_{(\mathbf{l+1})}, \underbrace{1, \ldots, 1}_{(\mathbf{m-2})}).$$

Penalized Least Squares Estimation

- For the B-spline penalty based on differences, we start by writing the vector of first differences using the difference matrix

$$\mathbf{D_1} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

of dimension $(d-1) \times d$, yielding

$$\mathbf{D_1}\boldsymbol{\gamma} = \begin{pmatrix} \gamma_2 - \gamma_1 \\ \vdots \\ \gamma_d - \gamma_{d-1} \end{pmatrix}.$$

Penalized Least Squares Estimation

- Higher differences can be expressed recursively with the help of difference matrices

$$\mathbf{D_r} = \mathbf{D_1 D_{r-1}}.$$

- For example with $r = 2$ we obtain, the $(d-2) \times d$-difference matrix

$$\mathbf{D_2} = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

Penalized Least Squares Estimation

- This yields the penalty

$$\lambda \sum_{j=r+1}^{d} (\Delta^r \gamma_j)^2 = \lambda \boldsymbol{\gamma}' \mathbf{D}_\mathbf{r}' \mathbf{D}_\mathbf{r} \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}' \mathbf{K}_\mathbf{r} \boldsymbol{\gamma},$$

with first and second order difference penalty matrices

$$\mathbf{K_1} = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 1 \end{pmatrix}, \qquad \mathbf{K_2} = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix},$$

respectively.

Penalized Least Squares Estimation

- In general, we can write the penalized residual sum of squares as

$$\begin{aligned} \text{PLS}(\lambda) &= (\mathbf{y}-\mathbf{Z}\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\mathbf{K}\boldsymbol{\gamma} \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Z}\boldsymbol{\gamma} - \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{y} + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\gamma} + \lambda\boldsymbol{\gamma}'\mathbf{K}\boldsymbol{\gamma} \\ &= \mathbf{y}'\mathbf{y} - \mathbf{2}\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{y} + \boldsymbol{\gamma}'(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{K})\boldsymbol{\gamma} \end{aligned}$$

  using an appropriate penalty matrix $\mathbf{K}$.

- Taking the derivative with respect to $\boldsymbol{\gamma}$ and setting the derivative equal to zero yields the system of equations

$$-2\mathbf{Z}'\mathbf{y} + \mathbf{2}(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{K})\boldsymbol{\gamma} = \mathbf{0},$$

  and, therefore, the penalized least squares estimate

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{K})^{-1}\mathbf{Z}'\mathbf{y}. \tag{17}$$

Penalized Least Squares Estimation

- Consequently, the vector of estimated function evaluations $\hat{\mathbf{f}} = (\hat{\mathbf{f}}(\mathbf{z_1}), \ldots, \hat{\mathbf{f}}(\mathbf{z_n}))'$ is calculated as

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

- Note that the form of the penalized least squares estimate is very similar to the unpenalized least squares estimate, as it only differs by the additional term $\lambda\mathbf{K}$.

- This additional term represents the influence of the penalty within the penalized least squares criterion.

- For $\lambda = 0$, we again obtain the solution of the unpenalized optimization problem.

Example: Malnutrition in Tanzania

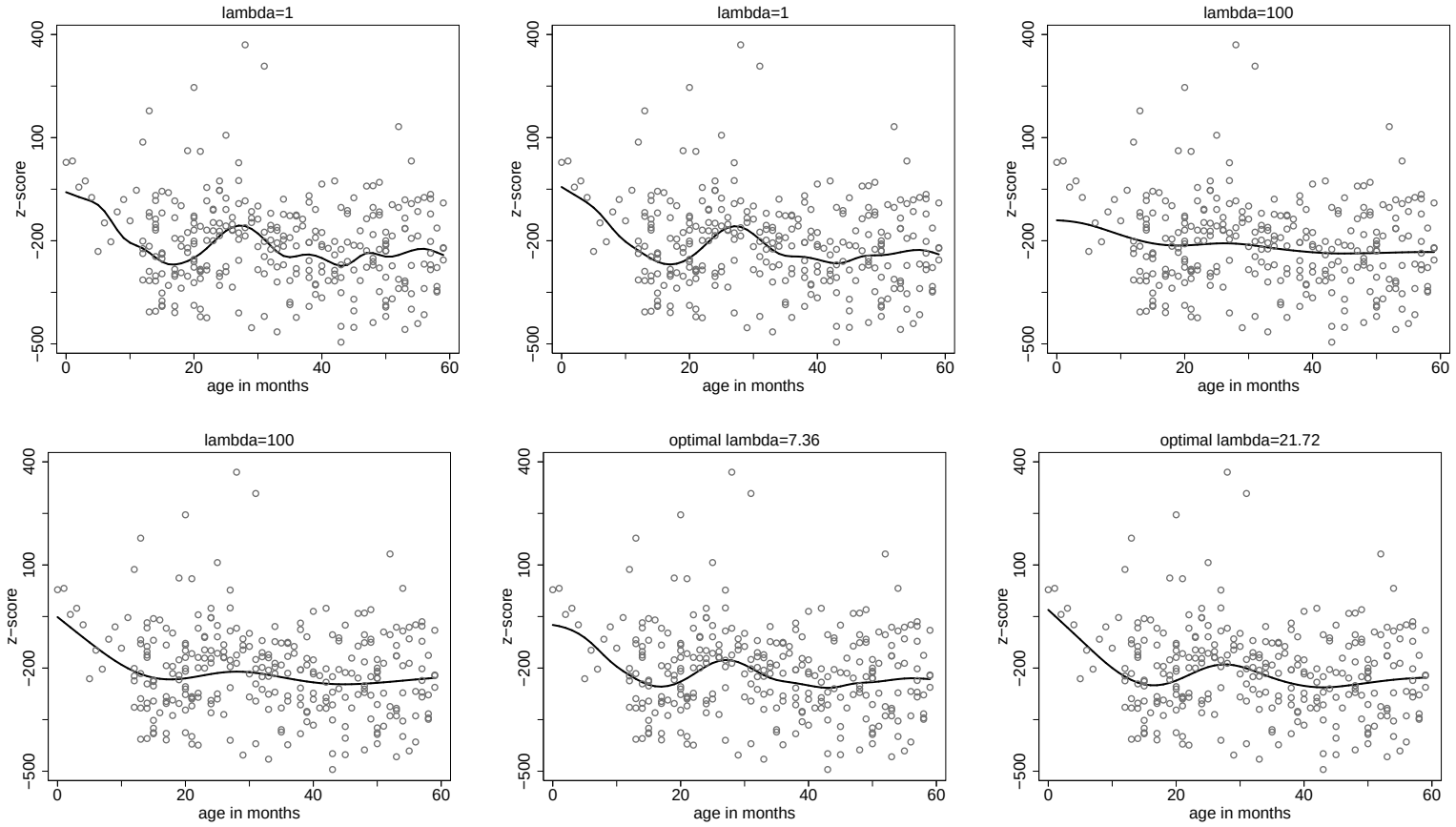


Figure 47: *Malnutrition in Tanzania: P-spline estimates based on first and second order differences and different smoothing parameters.*

Other scatterplot smoothers

- Smoothing splines
- Local polynomial smoothers

Advanced Regression Analysis

Additive models

Standard Additive Regression Model

- **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.

Standard Additive Regression Model

- The functions f_j are approximated using basis functions of the form

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

- Regardless of the chosen basis, we obtain the additive model

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

  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.

Munich Rent Index – Additive Model

- We specify the additive model

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

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

- Figure 48 shows the estimated nonlinear effects of living area and year of construction.

Munich Rent Index – Additive Model

- In the left panels the partial residuals are additionally included.
- Their calculation is completely analogous to partial residuals in linear models.
- For example, the partial residuals with regard to the apartment's living area are defined by

$$\hat{\varepsilon}_{area,i} = \hat{\varepsilon}_i + \hat{f}_1(area_i).$$

Munich Rent Index – Additive Model

- All in all, the estimated nonlinear effects are very similar to those found in the parametric model.

- Nevertheless, the fit to the data is better with the additive model used here. If we rather use the transformation $1/area$ for the living area and a cubic polynomial for modeling the effect of the year of construction, we obtain an AIC of 7380.3. With the nonparametric model, we obtain a lower AIC of 7362.6 such that the nonparametric model is preferable.

Munich Rent Index – Additive Model

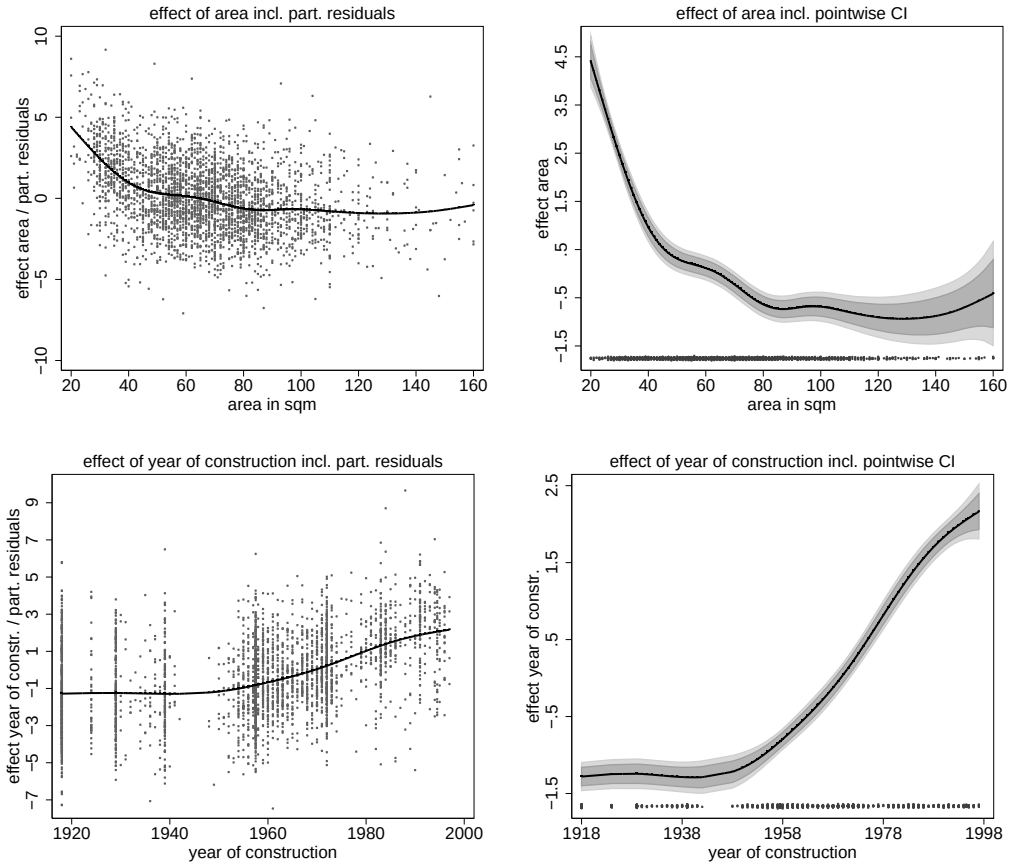


Figure 48: *Munich rent index: estimated nonlinear effects of area and year of construction.*

Penalized Least Squares Estimation

- Extending the penalized least squares criterion of the previous chapter, we minimize the criterion

$$\mathrm{LS}_{pen}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \beta) = \mathrm{LS}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \beta) + \sum_{j=1}^{q} \lambda_j \boldsymbol{\gamma}_j' \mathbf{K}_j \boldsymbol{\gamma}_j \tag{18}$$

  with respect to $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \beta$.

- In this setting, the least squares criterion is defined as

$$\mathrm{LS}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \beta) = (\mathbf{y} - \mathbf{Z}_1\boldsymbol{\gamma}_1 - \ldots - \mathbf{Z}_q\boldsymbol{\gamma}_q - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{Z}_1\boldsymbol{\gamma}_1 - \ldots - \mathbf{Z}_q\boldsymbol{\gamma}_q - \mathbf{X}\beta).$$

- The smoothing parameter λ_j governs the trade-off between fidelity to the data and smoothness controlled by the penalty $\boldsymbol{\gamma}_j' \mathbf{K}_j \boldsymbol{\gamma}_j$.

Backfitting Algorithm

- The idea of backfitting is to obtain estimates $\mathbf{f_1}, \ldots, \mathbf{f_q}, \hat{\beta}$ iteratively by the smoothing of partial residuals.

- Neglecting the errors ε, the following approximately holds:

$$\mathbf{f_j} \approx \mathbf{y} - \mathbf{f_1} - \ldots - \mathbf{f_{j-1}} - \mathbf{f_{j+1}} - \ldots - \mathbf{f_q} - \mathbf{X}\beta.$$

- In the case that we already have estimates for $\hat{\beta}$ and $\mathbf{f_l}$, $l \neq j$, we can view $\mathbf{y} - \sum_{\mathbf{l \neq j}} \mathbf{f_l} - \mathbf{X}\hat{\beta}$ as a (partial) residual vector (without $\mathbf{f_j}$).

- In order to estimate $\mathbf{f_j}$, we apply the scatterplot smoother $\mathbf{S_j}$ associated with $\mathbf{f_j}$ to this residual vector. We obtain

$$\mathbf{f_j} = \mathbf{S_j}(\mathbf{y} - \sum_{\mathbf{l \neq j}} \mathbf{f_l} - \mathbf{X}\hat{\beta}).$$

Backfitting Algorithm

- All estimators that are based on basis functions approaches have the form

$$\mathbf{f_j} = \mathbf{Z_j}(\mathbf{Z_j'Z_j} + \lambda_\mathbf{j}\mathbf{K_j})^{-1}\mathbf{Z_j'}(\mathbf{y} - \sum_{\mathbf{l}\neq\mathbf{j}} \mathbf{f_l} - \mathbf{X}\hat{\beta}).$$

- Given starting values for $\hat{\beta}, \mathbf{f_1}, \ldots, \mathbf{f_q}$, this concept is applied iteratively, and results in the *backfitting algorithm*.

Backfitting Algorithm

1. Initialization: set, for example, $\mathbf{f_1} = \ldots = \mathbf{f_q} = \mathbf{0}, \hat{\beta} = \mathbf{0}$.

2. Obtain updated estimates $\mathbf{f_j}, \mathbf{j} = \mathbf{1}, \ldots, \mathbf{q}$, through

$$\mathbf{f_j} = \mathbf{S_j}(\mathbf{y} - \sum_{\mathbf{l} \neq \mathbf{j}} \mathbf{f_l} - \mathbf{X}\hat{\beta}).$$

3. Obtain updated estimates $\hat{\beta}$, through

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{y} - \mathbf{f_1} - \ldots - \mathbf{f_q}).$$

4. Iterate the steps 2–3 until the estimated functions do not deviate more than a small given increment in two subsequent iterations.

Choice of smoothing parameters

- The complexity of a linear model is provided by the number of regression coefficients in the model.

- The number of parameters can for example be recovered by the trace of the prediction matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ since $\text{trace}(\mathbf{H}) = \mathbf{p}$.

- Since the prediction matrix of the linear model corresponds to the smoothing matrix $\mathbf{S}$ of a linear smoother, the *equivalent degrees of freedom* or *effective number of parameters* are analogously defined as

$$\text{df}(\mathbf{S}) = \text{trace}(\mathbf{S}). \tag{19}$$

- In fact, we reproduce the number of parameters in the model for basis function approaches without penalization, while in the case of penalization, the penalty effectively reduces the number of parameters.

Choice of smoothing parameters

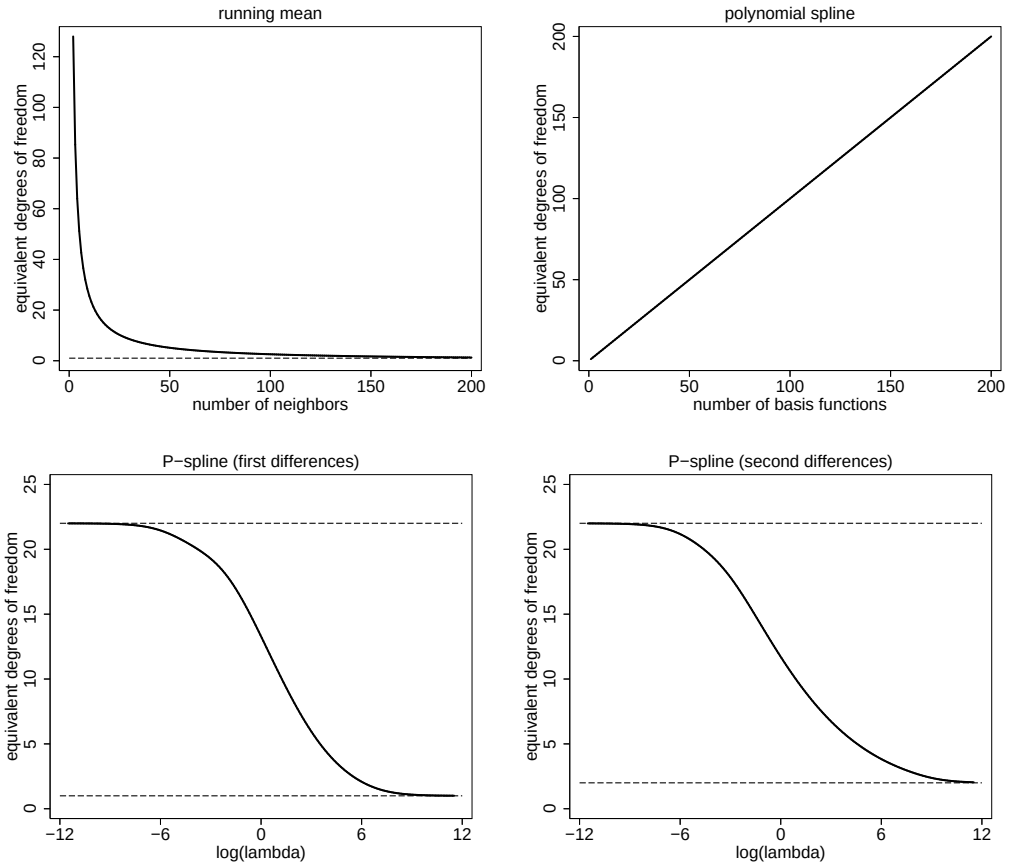


Figure 49: *Equivalent degrees of freedom as a function of the smoothing parameter. The dashed lines indicate limiting values of the equivalent degrees of freedom.*

Choice of smoothing parameters

- Figure 49 shows the connection between the original smoothing parameter and the resulting effective degrees of freedom for a selection of smoothing approaches and the simulated data example.

- For the running mean, a larger neighborhood leads to a smoother estimate and, hence, to a smaller effective number of parameters. As a limiting case, we obtain only one parameter, when the neighborhood contains all observations.

- For polynomial splines, the effective number of parameters corresponds exactly to the number of basis functions, so that a linear relationship results.

Choice of smoothing parameters

- When considering penalized splines, the effective number of parameters decreases with an increasing smoothing parameter (note that a logarithmic scale for the smoothing parameter is used to facilitate presentation).

- For $\lambda = 0$, we default back to the unpenalized case, so that the number of parameters corresponds to the number of basis functions used.

- For a very large smoothing parameter, the effective number of parameters apparently approaches a limiting value, which depends on the order of differences used when constructing the penalty.

Choice of smoothing parameters

- The AIC criterion is given by

$$\mathsf{AIC} = n \log(\hat{\sigma}^2) + 2(\mathsf{tr}(\mathbf{H}) + \mathbf{1}),$$

  where $\hat{\sigma}^2 = \sum (y_i - \hat{\eta}_i)^2 / n$ is the ML estimator of the error variance.

- For the generalized cross validation score GCV, we obtain

$$\mathsf{GCV} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{\eta}_i}{1 - \mathsf{tr}(\mathbf{H})/\mathbf{n}} \right)^2 .$$

Choice of smoothing parameters

- If we use the iterative backfitting algorithm to estimate the regression coefficients, we do not obtain the full smoother matrix and (exact) computation of the criteria is, thus, not possible. In practice, the trace of the full smoother matrix is often approximated via the sum of the traces of individual smoother matrices. Provided that the correlation among individual terms is not too strong, this approximation works well in practice.

- Until approximately the year 2000, a simple but not very effective grid search was widely used to minimize AIC or GCV, and thus problems with more than, say four or five smoothing parameters, were barely manageable.

- Currently, much more efficient (albeit technically complex) algorithms for the optimization of model choice criteria are available.

- An implementation can be found in the R package `mgcv`.

Advanced Regression Analysis

Quantile regression

Basic Idea

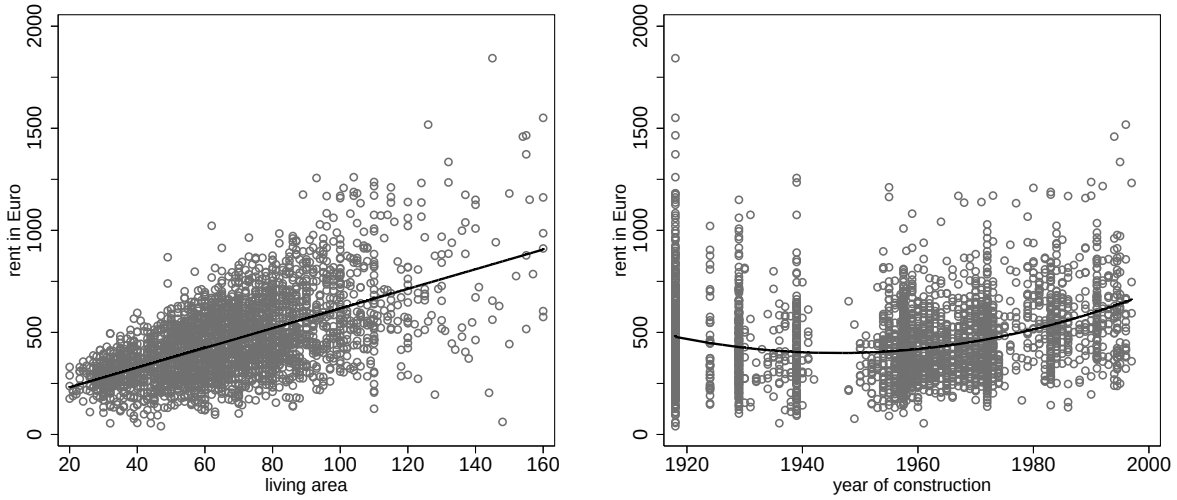


Figure 50: *Munich rent index: scatter plots of rents in Euro versus living area (left panel) and year of construction (right panel) together with a linear (left panel) and a quadratic (right panel) least squares fit.*

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

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

Linear quantile regression

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 $\hat{\beta}_\tau$.
- 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: $y = \mathbf{x}'\beta + \exp(\mathbf{x}'\alpha)\varepsilon$, $\varepsilon \sim \mathrm{N}(0, 1)$.

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

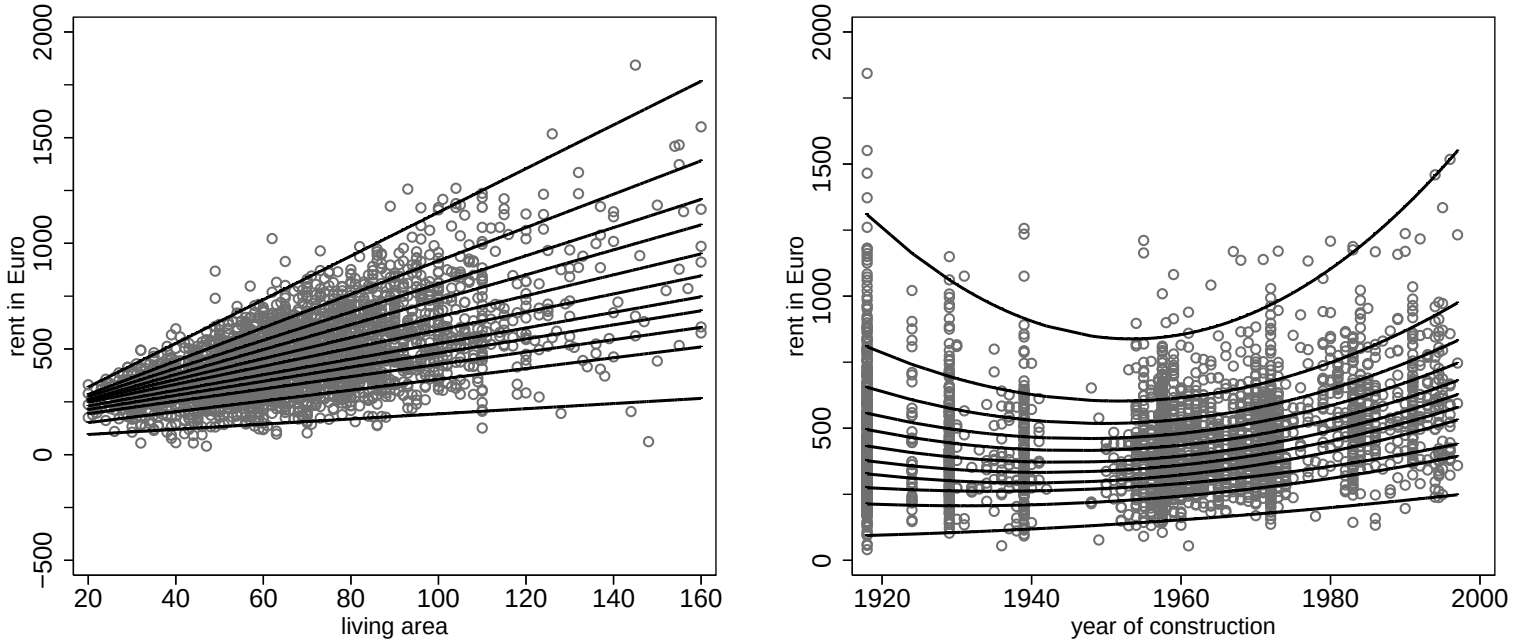


Figure 51: *Munich rent index: scatter plots of the rents in Euro versus living area (left column) and year of construction (right panel) together with linear/quadratic quantile regression fits for 11 quantiles.*

Example: Munich rent index

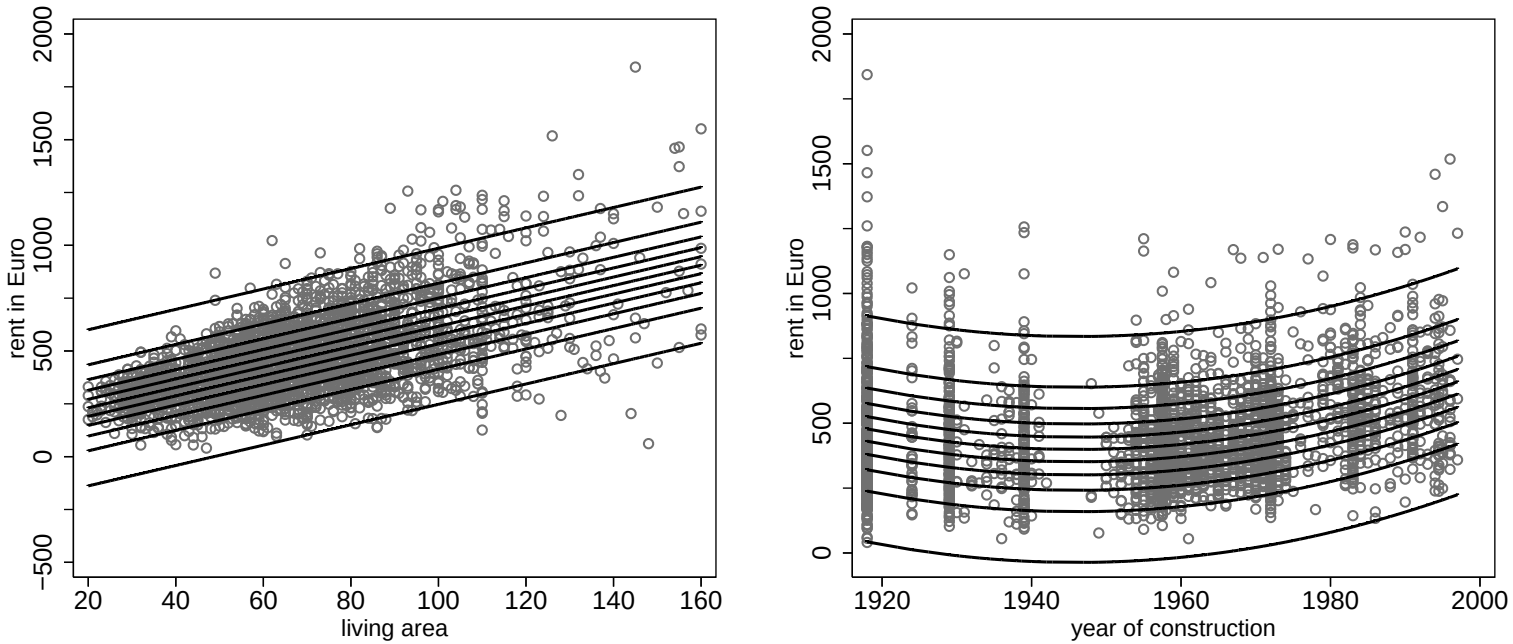


Figure 52: *Munich rent index: scatter plots of the rents in Euro versus living area (left column) and year of construction (right panel) together with quantiles determined from a homoscedastic linear model.*

Example: Munich rent index

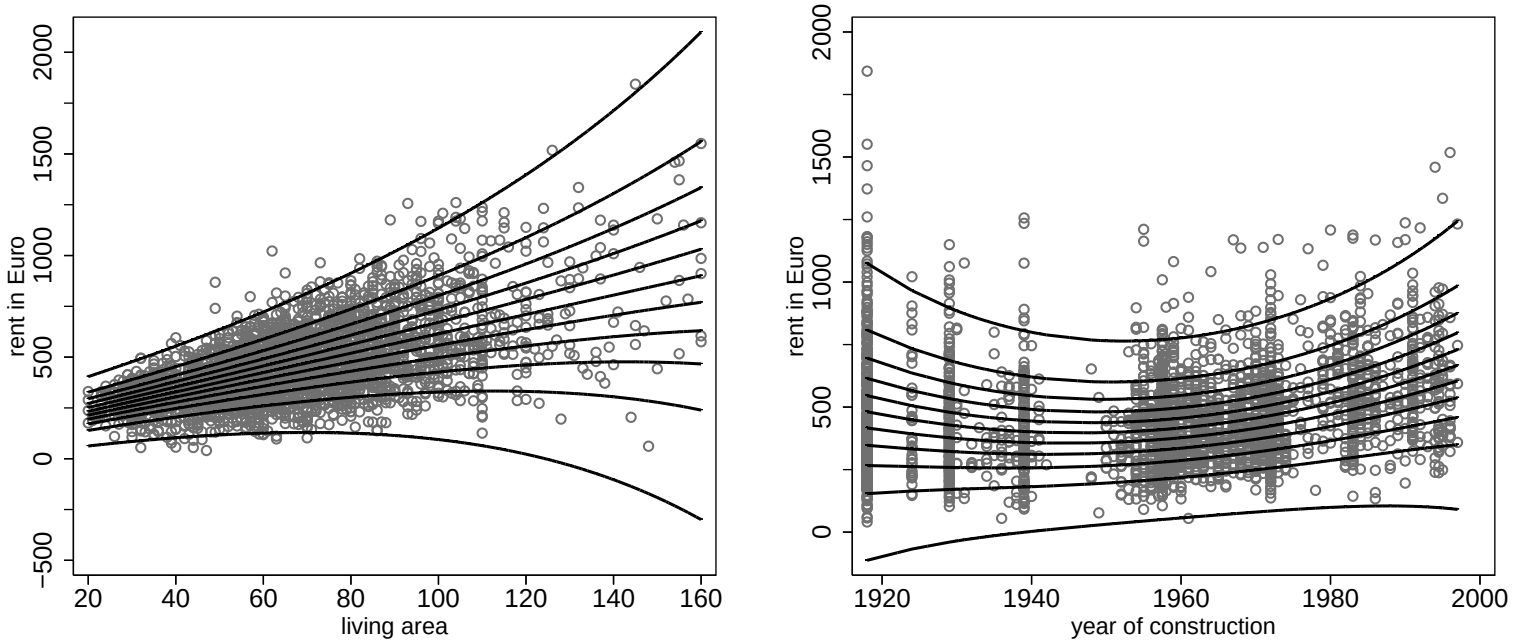


Figure 53: *Munich rent index: scatter plots of the rents in Euro versus living area (left column) and year of construction (right panel) together with quantiles determined from a heteroscedastic linear model.*

Example: Munich rent index

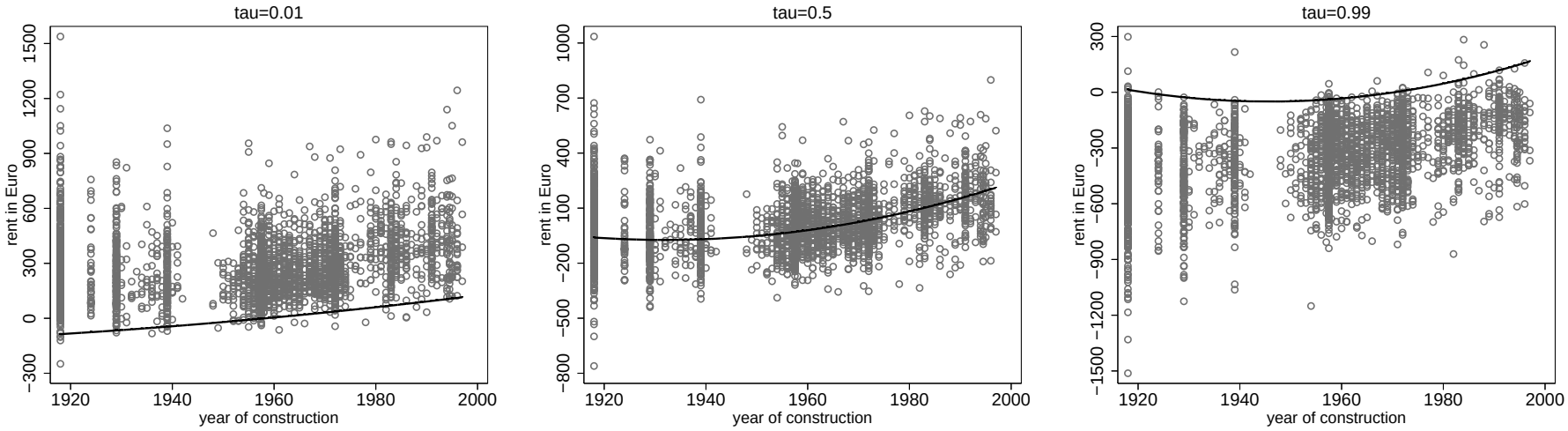


Figure 54: *Munich rent index: estimated effects of year of construction together with partial residuals for different quantiles.*

Example: Munich rent index

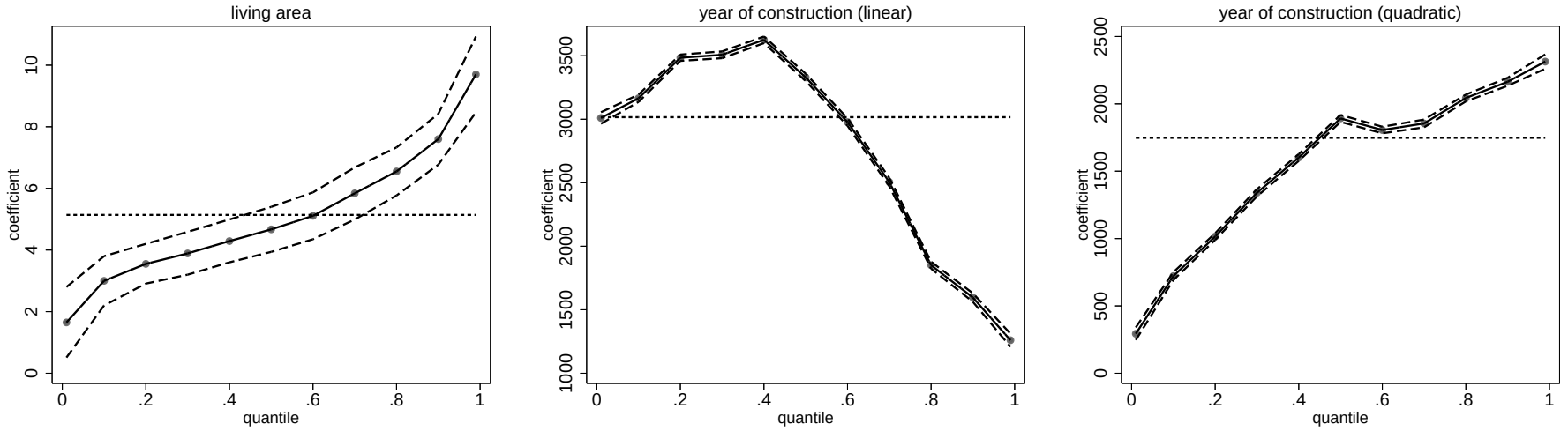


Figure 55: *Munich rent index: paths of estimated coefficients (solid line) together with 95% confidence intervals (dashed lines) obtained from inverting a rank-based test for various quantiles τ. The horizontal dotted line corresponds to the least squares estimate.*

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

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

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.

Bayesian quantile regression

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

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

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:

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

Additive quantile regression

- 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

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

Additive quantile regression

- $V(f'_\tau)$ 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.

Additive quantile regression

- 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 $\mathbf{V_j}\boldsymbol{\gamma_j}$ are now given by

$$\boldsymbol{\gamma}_j \,|\, \cdot \;\sim \mathrm{N}(\mathbf{m_j}, \boldsymbol{\Sigma_j})$$

with expectation and covariance matrix

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

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

Additive quantile regression

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