universität wien

MAGISTERARBEIT

Titel der Magisterarbeit

„Statistische Modellierung von Immobilienpreisen mithilfe eines additiven, nichtlinearen Modells"

Verfasser

Christian Pechhacker, Bakk.

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A 066 951
Studienrichtung lt. Studienblatt:	Magisterstudium Statistik
Betreuer:	o.Univ.Prof. Dr. Georg Ch. Pflug

Danksagung

Ich möchte mich herzlich bei meinem Betreuer o.Univ.Prof. Dr. Georg Ch. Pflug für die Betreuung und Unterstützung während der Verfassung dieser Arbeit bedanken.

Weiterer Dank gilt der Immobilien Rating GmbH sowie der Bank Austria UniCredit Group für die Bereitstellung der Daten, meinen Kollegen DI Ronald Weberndorfer, Dr. Wolfgang A. Brunauer sowie Univ.-Prof. Dr. Wolfgang Feilmayr.

Besonderer Dank gebührt meiner Familie für die sowohl mentale als auch finanzielle Unterstützung während meiner Studienzeit.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

1.1. Motivation

In den vergangenen Jahren kam der statistischen Analyse von Immobiliendaten immer größere Bedeutung zu. Dies liegt unter anderem an den Basel II Richtlinien, die seit dem 1. Januer 2007 alle Kredit- und Finanzdienstleistungsinstitute in der Europäischen Union betreffen. Demnach müssen Immobilien mindestens alle drei Jahre, in volatilen Märkten sogar noch häufiger, überprüft und der Marktwert gegebenfalls angepasst werden. Dies kann beispielsweise durch statistische Analysen geschehen, Banken steigern dadurch ihre Effizienz bei der Immobilienbewertung.

Angebot und Nachfrage bestimmen den Preis einer Immobilie. Eine generelle Schätzung erscheint allerdings schwierig, da Immobilien heterogene Güter darstellen und der Preis von unterschiedlichen Eigenschaften wie Lage, Größe oder Zustand mitbestimmt wird.
Stattdessen wird die Annahme getroffen, dass sich die Immobilie in diese verschiedenen Charakteristiken unterteilen lässt. Dafür eignen sich sogenannte hedonische Modelle, dabei wird ein Gut gedanklich in seine Merkmale aufgeteilt, mithilfe einer Regressionsanalyse wird der Einfluss dieser Qualitätseigenschaften auf den Preis ermittelt.

Dadurch besteht die Möglichkeit die preisrelevanten Charakteristiken von Immobilien sowie deren Einfluss zu ermitteln. Mithilfe eines hedonischen Modells kann sowohl der Wert einer bestimmten Eigenschaft, den Personen bereit sind für diese Eigenschaft zu zahlen, als auch der preismäßige Unterschied bestimmt werden, falls sich diese Charakterstik ändert.

Hedonische Preismodelle wurden zunächst in der Automobilindustie von Court[1] eingeführt und seit dem Artikel von Rosen[2] weitläufig angewandt. Eine gute Übersicht hedonischer Modelle im Kontext der Immobilienbewertung ist in der Arbeit von Brunauer[3] zu finden.
Bei der Analyse von Immobilienpreisen eignen sich klassische lineare Regressionsmodelle nur bedingt, da Variableneffekte teilweise hochgradig nichtlinear sein können. Goodman[4] empfahl anstatt der restriktiven linearen Form ein Box-Cox Modell, Anglin und Gencay[5] schlugen ein semi-parametrisches Modell vor. Dabei versuchten sie die abhängige

[1] Court, »Hedonic price indexes with automotive examples«.
[2] Rosen, »Hedonic Prices and Implicit Markets Product Differentiation in Pure Competition«.
[3] Brunauer, »Modeling House Prices using Multilevel Structured Additive Regression«.
[4] Goodman, »Hedonic price, price indices and housing markets«.
[5] Anglin und Gencay, »Semiparametric Estimation of a hedonic price function«.

Variable durch den gewohnten parametrischen Teil sowie durch einen nichtparametrisch geschätzten Teil zu erklären. Martins-Filho und Bin[6] verwendeten einen ähnlichen Ansatz basierend auf dem Backfitting-Algorithmus[7]. Brunauer[3] wendete mit „Mehrstufigen generalisierten strukturierten additiven Regressionsmodellen (STAR)" eine ziemlich flexible Methode im Bereich der semi-parametrischen Regression an[8] . Dabei können Regressionskoeffizienten nichtlinearer Terme auch einem anderen Regressionsmodell mit einem strukturierten, additiven Prädiktor gehorchen.

In dieser Arbeit wird ein - zur Klasse der semiparametrischen Regression gehörendes - additives Regressionsmodell geschätzt. Dadurch können für (kontinuierliche) erklärende Variablen auch nichtlineare Effekte modelliert werden. Allerdings besteht aufgrund der flexiblen Methode die Gefahr, dass unerklärbare Effekte mitmodelliert werden und sich diese entgegen unserer Erwartungen verhalten. Um diesem Problem zu entgehen, werden in dieser Arbeit gewisse Annahmen an die Effekte sowohl im parametrischen Teil als auch im nichtparametrischen Teil spezifiziert und mitmodelliert.

Zur Schätzung eines additiven Modells verwenden wir einen Datensatz bestehend aus 3152 über Österreich verteilten Einfamilienhäuser. Die Häuser werden in verschiedene Eigenschaften wie Baujahr oder Größe aufgeteilt. Diese Eigenschaften dienen bei einer Regressionsanalyse als unabhängige Variablen, die den (logarithmierten[9]) Preis erklären sollen. Das Modell kann folgendermaßen spezifiziert werden (eine detaillierte Beschreibung wird Kapitel 4 gegeben):

$$\mathbb{E}(ln(Preis_i)) = X_i^*\alpha + f_1(Baujahr_i) + \cdots + f_p(\text{Größe}_i)$$

Das Modell beschreibt logarithmierte Kaufpreise einserseits durch den üblichen linearen Part $X_i^*\alpha$. X_i^* gibt einen (Zeilen-)Vektor an, in dem sich (oftmals dummycodierte) Variablen beispielsweise zu Zustand oder Ausstattung befinden. Andererseits werden (logarithmierte) Preise durch (möglicherweise nichtlineare) Funktionen f_j $(j = 1, \ldots, p)$ beschrieben.

Erwartete Effekte können sowohl im linearen Part - zum Beispiel könnte man einen wertsteigernden Effekt eines Hauses in gutem Zustand erwarten - als auch in den nichtparametrisch geschätzten Funktionen f_j spezifiziert werden. So soll im Weiteren beispielswiese die Notation $f_\nearrow(\text{Größe}_i)$ auf eine nichtparametrische Funktion der Größe des Hauses hinweisen, für die ein monoton steigender Verlauf erwartet wird.

[6] Martins-Filho und Bin, »Estimation of hedonic price functions via additive nonparametric regression«.

[7] Nähere Details sind in Breiman und Friedman, »Estimating optimal transformations for multiple regression and correlations (with discussion)"« sowie Hastie und Tibshirani, *Generalized Additive Models* zu finden

[8] Nähere Details zur STAR- Modellen sind in Fahrmeir, *Regression. Modelle, Methoden und Anwendungen*, Brezger, Kneib und Lang, »Generalized structured additive regression based on Bayesian P-splines« sowie in Lang u. a., »Multilevel Structured Additive Regression« zu finden

[9] Nähere Details, warum logarithmierte Preise als abhängige Variable verwendet werden, werden in Kapitel 4 erläutert.

1.2. Gliederung

Der restliche Teil dieser Arbeit gliedert sich wie folgt: Kapitel 2 bietet einen Überblick über die verwendete Methodik, insbesondere wird das additive Modell inklusive Modellierung restringierter Terme beschrieben. Der verwendete Datensatz sowie die Modelle werden in Kapitel 3 und Kapitel 4 beschrieben. In Kapitel 5 werden die Ergebnisse präsentiert, eine Zusammenfassung befindet sich in Kapitel 6.

2. Konzept

2.1. Additive Modelle

Univariate Regressionsanalysen modellieren den Zusammenhang zwischen einer abhängigen Variablen (Regressand) und einer oder mehreren unahängigen Variablen (Regressoren). Das bekannte lineare Modell ist in vielen Situationen aufgrund der ziemlich restriktiven Annahme der Linearität ungeeignet, der Zusammenhang zwischen Regressand und Regressoren ist häufig nicht linear darstellbar. Deswegen kommt nicht- sowie semiparametrischer Regression immer größere Bedeutung zu.
Die flexible Modellierungsmöglichkeit hat allerdings den Nachteil, dass auch möglicherweise unerklärbare Effekte mitmodelliert werden. Um dieses Problem in den Griff zu bekommen wird in dieser Arbeit einerseits die Modellierung von Funktionen beschrieben, an deren Form gewisse Erwartungen vorliegen (Modellierung dieser „restringierten Funktionen" werden in Kapitel 2.2.3 und 2.2.4 beschrieben). Andererseits kann durch die Einführung von sogenannten Smoothing Parametern die Glätte der Funktionen kontrolliert werden, sodass unerklärbare Effekte nicht mitmodelliert werden.
Auch für den im linearen Modell üblichen parametrischen Teil sollen die Koeffizienten so gesetzt werden, dass sie mit unseren Erwartungen einhergehen. Aus technischen Gründen verläuft die Schätzprozedur nicht simultan. Zunächst wird das Modell gefittet, für das keine Restriktionen an den parametrischen Teil vorliegen (Kapitel 2.2 und Kapitel 2.3). Anschließend werden Koeffizienten nichtparametrisch modellierter Terme fest gehalten und das Modell mit dem restringierten parametrischen Teil reparametrisiert.

Generalisierte Additive Modelle

Generalisierte Additive Modelle (GAM), beschrieben in Hastie und Tibshirani[1], stellen eine flexible, in der Praxis häufig angewandte Applikation zum Modellieren von (möglicherweise) nichtlinearen Zusammenhängen einer abhängigen und mehreren Kovariaten dar. Der Erwartungswert von n unabhängigen Zufallsvariablen $Y_i, i = 1, \ldots, n$, stammend aus Verteilung der Exponentialfamilie, wird dabei über eine (bekannte) „Linkfunktion" g zu einem linearem Prädiktor η_i abgebildet:

$$g(\mu_i) = \eta_i \tag{2.1}$$

mit $\mu_i = \mathbb{E}(Y_i)$ und $Y_i \sim f_{\theta_i}(y_i)$. Die Dichte der Expontenialfamilienverteilungen kann durch

$$f_{\theta_i}(y_i) = exp(\frac{y_i\theta_i - b_i(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)) \tag{2.2}$$

[1] Hastie und Tibshirani, *Generalized Additive Models*.

dargestellt werden, wobei θ_i einen kanonischen Parameter bezeichnet. ϕ gibt einen Skalisierungsparamter gibt, der für alle Beobachtungen als konstant angenommen wird, a_i, b_i und c_i sind beliebige Funktionen.[2]

Der linerae Prädiktor η_i besteht dabei aus additven, möglicherweise nichtlinearen Effekten der Kovariablen $x_{1i}, x_{2i}, \ldots, x_{q_0,i}, z_{1i}, z_{2i}, \ldots, z_{pi}$, wobei die Kovariablen $x_{1i}, \ldots, x_{q_0,i}$ in einem Zeilenvektor $X_i^* = (x_{1i}, x_{2i}, \ldots, x_{q_0,i})$ zusammengefasst werden:

$$\eta_i = X_i^* \alpha + \sum_{j=1}^{p} f_j(z_{ji}) \tag{2.3}$$

Das generaliserte, additve Modell kann somit folgendermaßen formuliert werden:

$$g(\mu_i) = X_i^* \alpha + \sum_{j=1}^{p} f_j(z_{ji}). \tag{2.4}$$

Der Erwartungswert $\mu_i = \mathbb{E}(Y_i)$ einer Zufallsvariablen Y_i wird über eine Linkfunktion g zu einem linearen Prädiktor abgebildet. In diesem bezeichnet $X_i^* \alpha$ den - im linearen Modell üblichen - parametrischen Teil mit unbekanntem Parametervektor $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{q_0})^T$, während f_j eine nicht näher spezifizierte, möglicherweise komplexe, nichtlineare, glatte („smooth") Funktion darstellt.
Da unter anderem Normal-, Gamma-, Binomial-, Poisson- und Gleichverteilung zu der Klasse der Exponentialfamilienverteilung gehören, können mithilfe von generalisierten additiven Modellen eine Vielzahl von Zusammenhängen modelliert werden, auch für nicht kontinuierliche Regressanden.

Additive Modelle (AM)

Einen Spezialfall eines generalisierten additiven Modells stellt ein additives Modell dar, mit Identitäts -Linkfunktion $g(\mu_i) = \mu_i$ und unabhängigen, normalverteilten Zufallsvariablen Y_i $(i = 1, \ldots, n)$. Die Normalverteilung gehört zu der Familie der Exponentialverteilungen, da sich ihre Dichte

$$\begin{aligned} f_{\mu_i}(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2) \\ &= exp(-\frac{1}{2} log(2\pi\sigma^2)) exp(-\frac{y_i^2}{2\sigma^2} + \frac{y_i\mu_i}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2}) \\ &= exp(\frac{y_i\mu_i - \frac{\mu_i}{2}}{\sigma^2} + (-\frac{1}{2})(\frac{y_i^2}{\sigma^2} + log(2\pi\sigma^2))) \end{aligned}$$

in der Form der Gleichung (2.2) mit $\theta_i = \mu_i$, $\phi = \sigma^2$, $b_i(\theta_i) = \frac{\theta_i}{2}, a_i(\phi) = \phi$ und $c_i(y_i, \phi) = -\frac{1}{2}(\frac{y_i^2}{\sigma^2} + log(2\pi\sigma^2))$ darstellen lässt.

[2] Eine ausführliche Beschreibung generalisierter linearer Modelle ist in Nelder und Wedderburn, »Generalized Linear Models« sowie McCullagh und Nelder, *Generalized Linear Models* zu finden. Hastie und Tibshirani, *Generalized Additive Models* und Wood, *Generalized Additive Models. An Introduction with R* widmen sich generalisierten, additiven Modellen.

Im additiven Modell gilt dann folgender Zusammenhang:

$$\mu_i = \eta_i \tag{2.5}$$

mit $\mathbb{E}(Y_i) = \mu_i = \eta_i$ und $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
Im linearen Prädiktor

$$\eta_i = X_i^* \alpha + \sum_{j=1}^{p} f_j(z_{ji}). \tag{2.6}$$

bezeichnet $X_i^* \alpha$ wiederum den üblichen parametrischen Teil, während die $f_j's$ nicht näher spezifizierte, möglicherweise komplexe, nichtlineare, glatte Funktionen sind. Es ist leicht zu erkennen, dass das klassische lineare Modell einen Spezialfall eines additiven Modells repräsentiert, falls keine nichtlinearen Funktionen spezifiziert werden.
Die Kovariablen z_{ji} können dabei auch mehrdimensional sein, in dem Sinne, dass $f_j(z_{ji})$ auch eine höherdimensionale Funktion bezeichnen kann.
So kann der lineare Prädiktor beispielsweise folgende Form annehmen:

$$\eta_i = X_i^* \alpha + f_1(z_{i1}) + f_2(z_{i2}, z_{i3}) + z_{i4} \cdot f_3(z_{i5}, z_{i6})$$

$f_2 : \mathbb{R}^2 \to \mathbb{R}$ spezifiziert genauso wie f_3 eine zweidimensionale Funktion, wobei zur Funktion f_3 zusätzlich eine Kovariate multipliziert wird. Solche Modelle werden in der Literatur auch als „Varying-coefficient model"[3] bezeichnet.

Zusätzlich zur üblichen Schätzung des paramterischen Teils muss zur Schätzung des nichtparametrischen Teils eine flexible Methode benutzt werden, damit sowohl die Funktionen $f_j, j = 1, \ldots, p$ als auch die Stärke der Glättung geeignet repräsentiert werden können.
Wie diese nichtparametrischen Funktionen modelliert werden können, soll in folgendem Unterkapitel beschrieben werden.

[3] Hastie und Tibshirani, »Varying-coefficient models«.

2.2. Schätzung der nichtlinearen Funktionen

Eine Möglichkeit die unbekannte Funktion f_j der Kovariaten z_j zu schätzen besteht darin, die Kovariablen in höhere Potenzen zu transformieren und f_j durch ein Polynom $(q_j + 1)$-ter Ordnung zu schätzen:

$$f_j(z_j) = \sum_{l=0}^{q_j} z_j^l \gamma_{jl} \tag{2.7}$$

mit unbekannten Parametern $\gamma_{jl}, l = 1, \ldots, q_j$. Allerdings sind zur Modellierung komplexer Funktionen Polynome sehr hoher Ordnung notwendig. Damit erhöht sich aber auch die Anzahl der zu schätzenden Parameter, unerwartete Schwankungen können aufgrund der höheren Sensitivität und numerischen Instabilität auftreten.

Eine geeignetere Methode stellen sogenannte Splines dar[4]. Dabei wird der gesamte Wertebereich der Kovariablen z_j durch eine bestimmte Anzahl an Stützpunkten in verschiedene, disjunkte Teilbereiche getrennt. Diese Stützpunkte können gleichmäßig oder in einer anderem Form (beispielsweise quantilsweise) aufgeteilt sein, im weiteren Verlauf dieser Arbeit beschränken wir uns auf Funktionen mit gleichmäßig aufgeteilten Stützpunkten[5]. In jedem Teilbereich wird ein Polynom bestimmter Ordnung geschätzt. Dadurch ist aber nicht sicher gestellt, dass die Funktionswerte sowie Ableitungen an den Stützpunkten übereinstimmen, sodass die Funktion f_j im Allgemeinen weder glatt noch stetig ist. Die von DeBoor[6] eingeführten B - Splines repräsentieren eine spezielle Art von Splines, sodass die Funktion f_j stetig und glatt wird.

2.2.1. B - Splines

B - Splines beruhen auf der Idee von DeBoor. Dabei werden pro Funktion q_j Basisfunktionen $B_{jl}^{m_j}, l = 1, \ldots, q_j$ sowie $q_j + m_j + 2$ Stützpunkte $k_{j1} < k_{j2} < \cdots < k_{j,q_j+m_j+2}$, definiert, wobei $(m_j + 1)$ die Ordnung des B - Splines angibt.

Ein B - Spline der Ordnung $(m_j + 1)$, ausgewertet im Intervall $[k_{j,m_j+2}, k_{j,q_j+1}]$, kann nach DeBoor und Wood[7] wie folgt dargestellt werden:

$$f_j(z_j) = \sum_{l=1}^{q_j} B_{jl}^{m_j}(z_j)\gamma_{jl}, \tag{2.8}$$

wobei die Basisfunktionen rekursiv definiert sind:

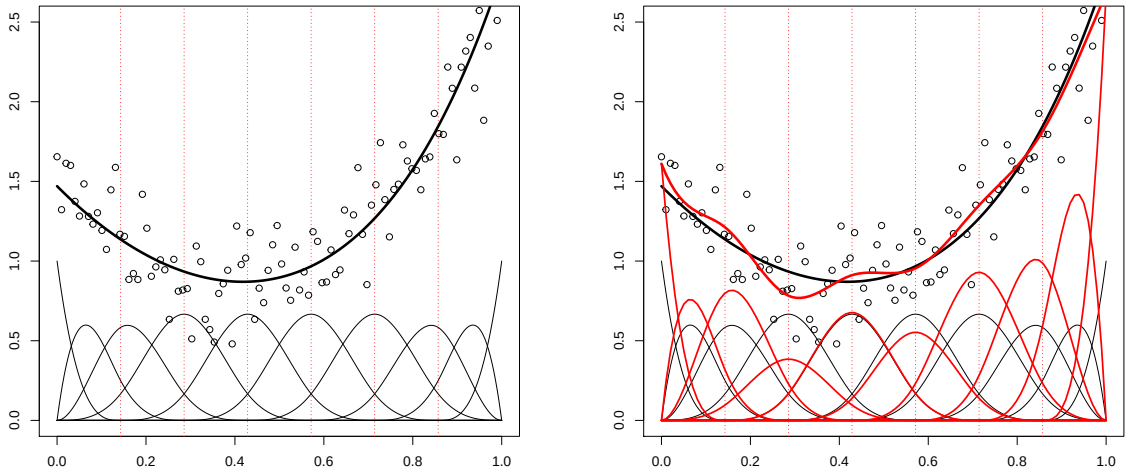
$$B_{jl}^{m_j}(z_j) = \frac{z_j - k_{jl}}{k_{j,l+m_j+1} - k_{jl}} B_{jl}^{m_j-1}(z_j) + \frac{k_{j,l+m_j+2} - z_j}{k_{j,l+m_j+2} - k_{j,l+1}} B_{j,l+1}^{m_j-1}(z_j) \tag{2.9}$$

[4] Erstmals eingeführt von Schoenberg, »Contributions to the problem of approximation of equidistant data by analytic functions«

[5] Die Interpretation von Penalties, die in Kapitel 2.3.1 beschrieben werden, ist für ungleichmäßig aufgeteilten Stützpunkten erschwert.

[6] DeBoor, *A Practical Guide to Splines*.

[7] Wood, *Generalized Additive Models. An Introduction with R*.



(a) (b)

Abbildung 2.1.: 100 simulierte Datenpunkte im Intervall [0,1] gemäß der Funktion $f(x) = x^5 + exp(x)^{-1} + 3(x-0.3)^2 + cos(x)/2 - 0.3$. 10 B-Spline Basisfunktionen 4. Ordnung sind in schwarz dargestellt (Panel (a)). Die mit den geschätzten Koeffizienten multiplizierten Basisfunktionen werden in Panel (b) rot markiert, Aufsummieren ergibt die geschätzte Funktion.

$$B_{jl}^{-1}(z_j) = \begin{cases} 1, & \text{falls } k_j \leq z_j \leq k_{j,l+1} \\ 0, & \text{sonst} \end{cases}, l = 1, \ldots, q_j$$

γ_{jl}, l=1, ..., q_j stellen dabei unbekannte, zu schätzende Koeffizienten dar.

In Abbildung 2.1(a) werden gemäß der Funktion

$$f(x) = x^5 + \frac{1}{exp(x)} + 3 \cdot (x - 0.3)^2 + \frac{cos(x)}{2} - 0.3 + \epsilon$$

mit $\epsilon \sim \mathcal{N}(0, 0.5)$ 100 Datenpunkte im Intervall [0,1] simuliert und zehn B- Spline Basisfunktionen vierter Ordnung („kubischer B- Spline") geschätzt..

Diese Basisfunktionen werden mit den Schätzern [8] der unbekannten Koeffizienten $\gamma_{jl}, l = 1, \ldots, q_j$ multpliziert (Abbildung 2.1(b)). Summieren dieser „aufgeblasenen" B-Spline Basisfunktionen ergibt die geschätzte Funktion (durchgezogene, rote Linie).

[8] Nähere Details zur Schätzung werden in Kapitel 2.3 erläutert.

B - Splines stellen eine äußerst flexible Methode dar, bei ausreichend hoher Anzahl an Basisfunktionen q_j kann jede Funktion exakt approximiert werden. Da die Basisfunktionen nur an $(m_j + 1)$ benachbarten Knoten positive Werte annehmen („local support property"), sind Rechenkosten sehr gering. Außerdem sind B- Splines bis zur m-ten Ableitung stetig, sodass die geschätzte Funktion stetig und so glatt wie gewünscht ist.

Allerdings stellt sich die Frage bezüglich der Anzahl der zu spezifizierenden Basisfunktionen q_j sowie den Grad m_j der verwendeten Polynome. Falls diese Parameter zu hoch gewählt werden, besteht Gefahr der Überanpassung („Overfitting"), zu wenige reichen nicht aus um die Funktion adäquat darzustellen. Diesem Problem widmen wir uns in Kapitel 2.3.1.

Im folgenden Unterkapitel soll gezeigt werden, wie sich Funktionen f_j, an deren Form keine speziellen Anforderung gestellt werden, mithilfe der B- Spline Basisfunktionen konventionell in Matrixnotation formulieren lassen. Fortan werden solche Funktionen als „unrestringiert" bezeichnet.

2.2.2. Modellierung unrestringierter Funktionen

In Matrixnotation kann die Funktion f_j aus Gleichung (2.8) geschrieben werden als

$$f_j = Z_j \gamma_j, \tag{2.10}$$

wobei $\gamma_j = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{j,q_j})^T$ den zu schätzenden Parametervektor und

$$Z_{ji} = (B_{j1}^{m_j}(z_{ji}), B_{j2}^{m_j}(z_{ji}), \ldots, B_{j,q_j}^{m_j}(z_{ji}))$$

die i-te Zeile der Modellmatrix Z_j bezeichnet. B_{jl} gibt dabei die l-te von q_j Basisfunktionen für die Funktion f_j an, die in Gleichung (2.9) definiert wurden.

Zum Zwecke der Konsistenz mit weiteren Kapiteln werden die Matrizen

$$\boldsymbol{\Sigma_j} = I_{q_j} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{2.11}$$

und

$$Z_j^* = Z_j = Z_j \boldsymbol{\Sigma_j}$$

definiert. Die Modellparameter γ_{jl} werden auf „Hilfsparameter" β_{jl} gesetzt:

$$\beta_{jl} = \gamma_{jl}, l = 1, \ldots, q_j.$$

Nach Einführung dieser Notation kann die Funktion f_j durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} \tag{2.12}$$

ausgedrückt werden, wobei $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{j,q_j})$ den unbekannten, zu schätzenden Parametervektor der Funktion f_j bezeichnet.

Auch wenn eine Spezifikation einer parametrischen Form einer Regressionsfunktion schwierig erscheint, liegen häufig gewisse Erwartungen an die Form der Funktion vor. Wie die unbekannten Koeffizienten γ_{jl} so eingeschränkt werden, dass eine Funktion die erwartete Form annimmt (fortan als „restringierte" Funktion bezeichnet) wird in den folgenden Kapitel 2.2.3 und Kapitel 2.2.4 beschrieben.

Hier wird gezeigt, dass sich auch restringierte Funktionen durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} \tag{2.13}$$

darstellen lassen.

2.2.3. Modellierung restringierter Funktionen

Modellierung isoton restringierter Funktionen

Die Funktion f_j aus Gleichung (2.8) soll nun so restringiert werden, dass sie monoton steigt. Analog zum Konzept von Pya[9], werden die unbekannten Modellparameter γ_{jl} so eingeschränkt, dass

$$f_j'(z_j) > 0.$$

Nach DeBoor[10] ist die erste Ableitung des B- Splines aus Gleichung (2.8) mit gleichmäßig aufgeteilten Stützpunkten

$$f_j'(z_j) = \frac{1}{h_j} \sum_{l=2}^{q_j} B_{jl}^{m-1}(z_j) \triangle^1 \gamma_{jl}, \tag{2.14}$$

wobei $\triangle^1 \gamma_{jl} = \gamma_{jl} - \gamma_{j,l-1}$ die erste Differenz der Modellparameter und $h_j = k_{j2} - k_{j1}$ die Distanz zwischen zwei benachbarten Knotenpunkten angibt.

Nachdem alle B-Spline Basisfunktionen per Definition nichtnegativ sind, ist zum Erreichen von $f_j'(z_j) > 0$ die Bedingung

$$\triangle^1 \gamma_{jl} > 0$$

hinreichend.

Demzufolge produziert eine steigende Sequenz der Modellparameter γ_{jl}, $l = 1, \ldots, q_j$ eine isotone Funktion f_j. Um dies zu bewerkstelligen, werden analog zu Pya die restringierten Modellkoeffizienten γ_{jl} zu unbekannten, unrestringierten „Hilfsparameter" β_{jk} umgeschrieben, sodass

[9] Pya, »Additive models with shape constraints«.
[10] DeBoor, *A Practical Guide to Splines*.

$$
\begin{aligned}
\gamma_{j1} &= \beta_{j1} \\
\gamma_{j2} &= \beta_{j1} + exp(\beta_{j2}) \\
\gamma_{j3} &= \beta_{j1} + exp(\beta_{j2}) + exp(\beta_{j3}) \\
&\vdots \\
\gamma_{j,q_j} &= \beta_{j1} + exp(\beta_{j2}) + exp(\beta_{j3}) + \cdots + \exp(\beta_{j,q_j})
\end{aligned}
$$

Nach Definition der $q_j \times q_j$ Matrix

$$
\boldsymbol{\Sigma_j} = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
1 & 1 & 0 & \cdots & 0 \\
1 & 1 & 1 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
1 & 1 & 1 & \cdots & 1
\end{pmatrix} \tag{2.15}
$$

sowie des Vektors

$$
\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j,1}, exp(\beta_{j,2}), exp(\beta_{j,3}), \ldots, exp(\beta_{j,q_j}))^T
$$

kann der unbekannte - aber mit einer steigenden Einträgen restringierte - Modellparametervektor der j-ten Funktion $\boldsymbol{\gamma}_j = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{j,q_j})^T$ durch

$$
\boldsymbol{\gamma}_j = \boldsymbol{\Sigma_j} \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}
$$

ausgedrückt werden, sodass die Funktion f_j aus Gleichung (2.8) folgendermaßen geschrieben wird:

$$
f_j = Z_j \boldsymbol{\Sigma_j} \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}.
$$

$Z_{ji} = (B_{j1}^{m_j}(z_{ji}), B_{j2}^{m_j}(z_{ji}), \ldots, B_{j,q_j}^{m_j}(z_{ji}))$ gibt wiederum die i-te Zeile der Modellmatrix Z_j an.

$$
\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j,1}, exp(\beta_{j,2}), exp(\beta_{j,3}), \ldots, exp(\beta_{j,q_j}))^T
$$

bezeichnet den Vektor der restringierten Modellparameter für die Funktion f_j, während

$$
\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{j,q_j})^T
$$

als Vektor der unrestringierten Modellparameter („Hilfsparameter") der Funktion f_j bezeichnet wird.

Nach Einführung der Notation

$$
Z_j^* = Z_j \boldsymbol{\Sigma_j}
$$

kann die Funktion f_j wie im Falle unrestringierter Funktionen (siehe Gleichung (2.13)) wiederum ausgedrückt werden durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$$

Die Symbol $f_{\nearrow}$ bezeichne fortan isoton restringierte Funktionen.

Modellierung antiton restringierter Funktionen

Monoton fallende Funktion erhält man durch die Restriktion $\triangle^1 \gamma_{jl} < 0$, sodass $f_j'(z_j) < 0$. Anstatt der Matrix $\boldsymbol{\Sigma_j}$ aus Gleichung (2.15) wird hierfür die Matrix

$$\boldsymbol{\Sigma_j} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 1 & -1 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & -1 & -1 & \cdots & -1 \end{pmatrix} \tag{2.16}$$

verwendet. Die Funktion f_j kann wiederum durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$$

mit $Z_j^* = Z_j \boldsymbol{\Sigma_j}$ ausgedrückt werden. Das Symbol $f_{\searrow}$ soll antiton restringierte Funktionen bezeichnen.

2.2.4. Zusätzliche Restriktionen an die Krümmung

Zusätzlich zur Monotonie kann manchmal auch eine Restriktion an die Krümmung der Funktion $f_j(z_j) = \sum_{l=1}^{q_j} B_{jl}^{m_j} \gamma_{jl}$ sinvoll sein.
Analog zu Pya sollen die unbekannten Modellparameter $\gamma_{jl}, l = 1, \ldots, q_j$ so gesetzt werden, dass

$f_j(z_j)'' \begin{cases} \geq 0, & \text{für konvex} \\ \leq 0, & \text{für konkav} \end{cases}$ restringierte Funktionen.

Nach DeBoor[11] ist die zweite Ableitung des B-Splines mit gleichmäßig aufgeteilten Stützpunkten

$$f_j''(z_j) = \frac{1}{h_j^2} \sum_{l=3}^{q_j} B_{jl}^{m_j - 1}(z_j) \triangle^2 \gamma_{jl}, \tag{2.17}$$

wobei $h_j = k_{j2} - k_{j1}$ wiederum die Differenz zwischen zwei benachbarten Stützpunkten angibt und $\triangle^2 \gamma_{jl} = \gamma_{jl} - 2\gamma_{j,l-1} + \gamma_{j,l-2}$ ein Maß der Krümmung für die Funktion f_j bietet. $\triangle^2 \gamma_{jl} = 0$ entspricht einem linearen Verlauf, $\triangle^2 \gamma_{jl} > 0$ einem konvexen und $\triangle^2 \gamma_{jl} < 0$ einem konkaven.

[11] DeBoor, *A Practical Guide to Splines*.

Demzufolge ist die Bedingung $\begin{cases} \triangle^2\gamma_{jl} \geq 0, & \text{für konvex} \\ \triangle^2\gamma_{jl} \leq 0, & \text{für konkav} \end{cases}$ restringierte Funktionen hinreichend ($l = 1, \ldots, q_j$), sodass sich nun vier verschiedene Kombinationsmöglichkeiten an Restriktionen für die Funktion f_j ergeben:

▷ isoton, konvex restringierte Funktionen $f_{\nearrow}$

▷ isoton, konkav restringierte Funktionen $f_{\curvearrowright}$

▷ antiton, konvex restringierte Funktionen $f_{\curvearrowleft}$

▷ antiton, konkav restringierte Funktionen $f_{\searrow}$

Wie in den vorangegangen Kapitel lassen sich diese Funktionen durch

$$f_j = Z_j \boldsymbol{\Sigma_j} \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} \tag{2.18}$$

darstellen, wobei $Z_{ji} = (B_{j1}^{m_j}(z_{ji}), B_{j2}^{m_j}(z_{ji}), \ldots, B_{j,q_j}^{m_j}(z_{ji}))$ wiederum die i-te Zeile der Modellmatrix Z_j und

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j1}, exp(\beta_{j2}), exp(\beta_{j3}) \ldots, exp(\beta_{j,q_j}))^T$$

den unbekannten, restringierten Parametervektor angibt.

$$\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{j,q_j})^T$$

bezeichnet abermals den Vektor der unrestringierten Modellparameter („Hilfsparameter").

Die $q_j \times q_j$ Matrix $\boldsymbol{\Sigma_j}$[12] hängt je nach Wahl der vier verschiedenen Restriktionen ab, wobei die l-te Zeile durch $\boldsymbol{\Sigma}_{\boldsymbol{j}[l,]}$ notiert werden soll:

1. für isotone, konvexe Smooths:

$$\boldsymbol{\Sigma_j} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 2 & 1 & 0 & \cdots & 0 \\ 1 & 3 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & q_j - 1 & q_j - 2 & q_j - 3 & \cdots & 1 \end{pmatrix} \tag{2.19}$$

Diese Matrix wurde so konstruiert, dass für den Ausdruck $\boldsymbol{\Sigma}_{\boldsymbol{j}[l,]} - 2 \cdot \boldsymbol{\Sigma}_{\boldsymbol{j}[l-1,]} + \boldsymbol{\Sigma}_{\boldsymbol{j}[l-2,]}$ für $l = 3, \ldots, q_j$ nur der Eintrag in der l-ten Spalte 1 beträgt, alle anderen Einträge haben den Wert 0. Nachdem ab dem dritten Element in $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ alle Einträge positiv

[12] Für nähere Details sei auf Pya, »Additive models with shape constraints« verwiesen.

sind, ist die Bedingung $\triangle^2\gamma_{jl} = \gamma_{jl} - 2\gamma_{j,l-1} + \gamma_{j,l-2} \geq 0$ für $l = 3, \ldots, q_j$ erfüllt, sodass eine isotone, konvexe Funktion gewährleistet ist.

Beispielweise ergibt $\boldsymbol{\Sigma_j}_{[3,]} - 2 \cdot \boldsymbol{\Sigma_j}_{[2,]} + \boldsymbol{\Sigma_j}_{[1,}$ den Zeilenvektor $\begin{pmatrix} 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$, der nach Multiplikation von $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ mit $exp(\beta_3)$ ein positives Ergebnis liefert.

2. für isotone, konkave Smooths:

$$\boldsymbol{\Sigma_j} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & \cdots & 2 & 2 & 1 \\ 1 & 3 & 3 & 3 & \cdots & 3 & 2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & q_j - 1 & q_j - 2 & q_j - 3 & \cdots & 3 & 2 & 1 \end{pmatrix} \tag{2.20}$$

Diese Matrix wurde so konstruiert, dass für den Ausdruck $\boldsymbol{\Sigma_j}_{[l,]} - 2 \cdot \boldsymbol{\Sigma_j}_{[l-1],} + \boldsymbol{\Sigma_j}_{[l-2,]}$ für $l = 3, \ldots, q_j$ nur der $(q_j + 3 - l)$-te Eintrag -1 beträgt, alle anderen Einträge sind 0, sodass nach Multiplikation mit $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ die Bedingung $\triangle^2\gamma_{jl} = \gamma_{jl} - 2\gamma_{j,l-1} + \gamma_{j,l-2} \leq 0$ für $l = 3, \ldots, q_j$ erfüllt ist.

3. für antiton, konvexe Smooths:

$$\boldsymbol{\Sigma_j} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 & \cdots & -1 & -1 & -1 \\ 1 & -2 & -2 & -2 & \cdots & -2 & -2 & -1 \\ 1 & -3 & -3 & -3 & \cdots & -3 & -2 & -1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & -(q_j - 1) & -(q_j - 2) & -(q_j - 3) & \cdots & -3 & -2 & -1 \end{pmatrix} \tag{2.21}$$

4. für antiton, konkave Smooths:

$$\boldsymbol{\Sigma_j} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & -1 & 0 & \cdots & 0 \\ 1 & -3 & -2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -(q_j - 1) & -(q_j - 2) & -(q_j - 3) & \cdots & -1 \end{pmatrix} \tag{2.22}$$

Nach Definition von $Z_j^* = Z_j \boldsymbol{\Sigma_j}$ werden auch diese Funktionen als

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$$

geschrieben.

2.2.5. Modellierung zweidimensionaler Funktionen

Dieses Unterkapitel beschäftigt sich mit der Modellierung von Funktionen $f_j(z_j^{(1)}, z_j^{(2)})$ mit einem zweidimensionalen Input-Vektor $(z_j^{(1)}, z_j^{(2)})$. Zunächst werden die einzelnen Funktionen wie gehabt ausgedrückt:

$$f_j(z_j^{(1)}) = \sum_{l=1}^{q_j^{(1)}} B_{jl}^{m_j}(z_j^{(1)})\gamma_{jl}^{(1)} = Z_j^{(1)}\gamma_j^{(1)}, f_j(z_j^{(2)}) = \sum_{k=1}^{q_j^{(2)}} B_{jk}^{m_j}(z_j^{(1)})\gamma_{jl}^{(2)} = Z_j^{(2)}\gamma_j^{(2)} \tag{2.23}$$

mit entsprechenden Design Matrizen Z_j und unbekannten Parametervektoren $\gamma_j^{(1)}$ und $\gamma_j^{(2)}$.

Um die zweidimensionale Funktion zu repräsentieren, können die Parameter γ_j^1 durch den B-Spline der zweiten Kovariate ausgedrückt werden:

$$f_j(z_j^{(1)}, z_j^{(2)}) = \sum_{l=1}^{q_j^{(1)}} \sum_{k=1}^{q_j^{(2)}} B_{jlk}^{m_j}(z_j^{(1)}, z_j^{(2)})\beta_{jlk} \tag{2.24}$$

mit

$$B_{jlk}^{m_j}(z_j^{(1)}, z_j^{(2)}) = B_{jl}^{m_j}(z_j^{(1)}) \cdot B_{jk}^{m_j}(z_j^{(2)})$$

und unbekannten Koeffizienten β_{jlk}, wobei $l = 1, \ldots, q_j^{(1)}$ und $k = 1, \ldots, q_j^{(2)}$ Laufindizes darstellen.

In Matrixnotation ergibt sich für die zweidimensionale Funktion:

$$f_j(z_j^{(1)}, z_j^{(2)}) = Z_j\beta_j$$

$$\beta_j = (\beta_{11}, \beta_{12}, \ldots, \beta_{1,q_j^{(2)}}, \beta_{21}, \ldots, \beta_{2,q_j^{(2)}}, \ldots, \beta_{q^{(1)}{}_j,1}, \ldots, \beta_{q_1,q_2})^T \tag{2.25}$$

bezeichnet den unbekannten Vektor der Hilfsparameter, die Design - Matrix Z_j lässt sich durch $Z_j = Z_j^{(1)} \otimes Z_j^{(2)}$ ausdrücken, wobei $\otimes$ das Kronecker-Produkt[13] bezeichnet.

Pya[14] beschrieb, wie sich auch zweidimensional restringierte Funktionen durch

$$f_j = Z_j\boldsymbol{\Sigma_j}\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} \tag{2.26}$$

[13]Nähere Details findet man in Steeb, *Kronecker Product of Matrices and Applications*.
[14]Pya, »Additive models with shape constraints«.

formulieren lassen. Die Gestalt von $\boldsymbol{\Sigma_j}$ und $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ hängt dabei je nach Art der Restriktion ab, hier sollen einige Beispiele gegeben werden[15]:

- ▷ Für zweidimensionale Funktionen ohne spezifizierte Restriktion wird $\boldsymbol{\Sigma_j} = I_{q_j^{(1)}} \otimes I_{q_j^{(2)}}$ und $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = \beta_j$ aus Gleichung (2.25) verwendet.
- ▷ Falls für die zweidimensionale Funktion sowohl für die erste als auch die zweite Kovariate ein monoton steigender Verlauf angenommen wird, wird die Matrix $\boldsymbol{\Sigma_j} = \boldsymbol{\Sigma_j^{(1)}} \otimes \boldsymbol{\Sigma_j^{(2)}}$ benutzt, wobei sowohl $\boldsymbol{\Sigma_j^{(1)}}$ als auch $\boldsymbol{\Sigma_j^{(2)}}$ die selbe Gestalt haben wie die Matrix aus Gleichung (2.15), nur mit passender Dimension.

  $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j11}, exp(\beta_{j12}), \ldots, exp(\beta_{j1,q_j^{(2)}}), exp(\beta_{j21}), \ldots, exp(\beta_{j,q_j^{(1)},1}), \ldots, exp(\beta_{j,q_j^{(1)},q_j^{(2)}}))^T$ bezeichnet den Vektor der restringierten Modellkoeffizienten. Die Symbologie $f_{\nearrow\nearrow}$ deutet zweidimensionale Smooths an, wobei der Einfluss beider Kovariaten als monoton steigend angenommen wird.
- ▷ Wird nur entlang der ersten Kovariate ein monoton steigender Verlauf erwartet, bestimmt sich $\boldsymbol{\Sigma_j} = \boldsymbol{\Sigma_j^{(1)}} \otimes I_{q_j^{(2)}}$ wobei für $\boldsymbol{\Sigma_j^{(1)}}$ die Matrix aus Gleichung (2.15) verwendet wird. Der Vektor der restringierten Modellkoeffizienten ist folgendermaßen festgelegt:

  $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j11}, \ldots, \beta_{j1,q_j^{(2)}}, exp(\beta_{j21}), \ldots, exp(\beta_{j2,q_j^{(2)}}), \ldots, exp(\beta_{j,q_j^{(1)},1}), \ldots, exp(\beta_{j,q_j^{(1)},q_j^{(2)}}))^T$

  Die Notation $f_{\nearrow\cdot}$ soll fortan diese Art der Modellierung beschreiben.

Nach Definition von $Z_j^* := Z_j \boldsymbol{\Sigma_j}$ kann die zweidimensionale Funktion f_j wiederum durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$$

beschrieben werden.

2.3. Schätzung der Parameter

In den vorangegangenen Unterkapitel wurde gezeigt, dass alle nichtparametrisch modellierten Funktionen $f_j, j = 1, \ldots, p$ durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} \tag{2.27}$$

ausgedrückt werden können, wobei sowohl $Z_j^* = Z_j \boldsymbol{\Sigma_j}$ als auch $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ jeweils von der spezifizierten Restriktion an die Funktion f_j abhängen:

[15] Auf eine detaillierte Beschreibung weiterer Möglichkeiten sei auf Pya, »Additive models with shape constraints« verwiesen

▷ Für $\boldsymbol{\Sigma_j}$ wird je nach Restriktion eine der Matrizen aus den Gleichungen 2.11, 2.15, 2.16, 2.19, 2.20, 2.21 oder 2.22 verwendet.

▷ Der unbekannte Parametervektor lässt sich durch

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = \begin{cases} (\beta_{j1}, \beta_{j2}, \beta_{j3} \ldots, \beta_{j,q_j})^T & \text{falls } f_j \text{ unrestringiert} \\ (\beta_{j1}, exp(\beta_{j2}), exp(\beta_{j3}) \ldots, exp(\beta_{j,q_j})^T & \text{falls } f_j \text{ restringiert} \end{cases}$$

ausdrücken.

Somit lässt sich das additve Modell

$$\mathbb{E}(Y_i) = X_i^* \alpha + \sum_{j=1}^{p} f_j(z_{ji})$$

mit unabhängig, normalverteilten Zielvariablen Y_i bei adäquater, restriktionsabhängiger Verwendung von $\boldsymbol{\Sigma_j}$ sowie $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ vereinfacht durch

$$\mathbb{E}(Y_i) = X_i^* \alpha + \sum_{j=1}^{p} Z_{ji} \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} \tag{2.28}$$

dargestellen. Dabei bezeichnen α sowie $\tilde{\boldsymbol{\beta}}_j, j = 1, \ldots, p$ die unbekannte, zu schätzende Modellkoeffizienten.

Allerdings ist das additve Modell üblicherweise nicht eindeutig identifizierbar, da ein konstanter Wert c bei eine Funktion addiert und bei einer anderen subtrahiert werden könnte, ohne das Ergebnis zu ändern. Um dieses Identifizierungproblem zu umgehen wird bei jeder unrestringierten Funktion eine „Zentrierungsnebenbedingung" ([16]) eingeführt, sodass

$$\sum_{i=1}^{n} f_j(z_{ji}) = 0$$

Für restringierte Funktionen f_j werden die Modellparameter $\gamma_{j1} = \beta_{j1}$ analog zu Pya auf den Wert 0 gesetzt, da diese dem Intercept der Funktion f_j entsprechen.

Der lineare Prädiktor aus Gleichung (2.3) lässt sich nun durch

$$\eta_i = X_i \tilde{\boldsymbol{\beta}} \tag{2.29}$$

ausdrücken, wobei

$$X_i = [X_i^* : Z_1^* : Z_2^* : \cdots : Z_p^*]$$

die i- te Zeile der Design- Matrix X darstellt und

$$\tilde{\boldsymbol{\beta}} = (\alpha^T, \tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T, \ldots, \tilde{\boldsymbol{\beta}}_p^T)^T$$

[16] Wood, *Generalized Additive Models. An Introduction with R.*

den gesamten Vektor der unbekannten Modellparameter angibt. Der Vektor

$$\beta = (\alpha^T, \beta_1^T, \beta_2^T, \ldots, \beta_p^T)^T$$

wird als Vektor der „Hilfsparameter" bezeichnet.

Das additive Modell aus 2.1 lässt sich somit folgendermaßen formulieren:

$$\mu_i = X_i \tilde{\boldsymbol{\beta}}$$

mit $\mu = \mathbb{E}(\boldsymbol{Y}), \boldsymbol{Y} \sim \mathcal{N}(\mu, \sigma^2)$.

Für n unabhängige Realisierungen y_i der Zufallsvariablen $\boldsymbol{Y_i}$ kann die Likelihood von β geschrieben werden als

$$L(\beta) = \prod_{i=1}^{n} f_{\mu_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2}$$

Logarithmieren ergibt die Log -Likelihood:

$$l(\beta) = \sum_{i=1}^{n} (-\frac{1}{2} log(2\pi) - log(\sigma) - \frac{1}{2\sigma^2}(y_i - \mu_i)^2) \tag{2.30}$$

$$= -\frac{n}{2} log(2\pi) - n \cdot log(\sigma) - \frac{1}{2}(y - \mu)^T (y - \mu), \tag{2.31}$$

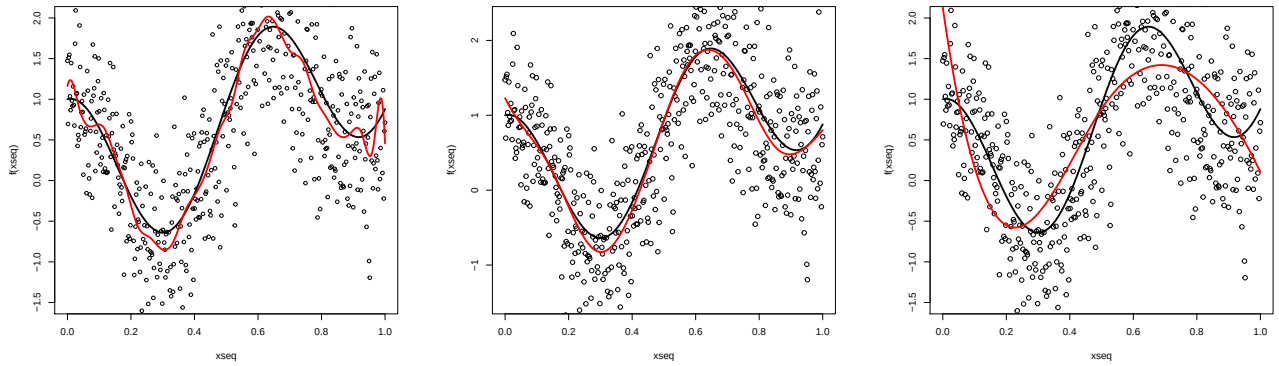
wobei die Abhängigkeit von β in $\mu = X\tilde{\boldsymbol{\beta}}$ enthalten ist.

Bevor die unbekannten Parameter β durch Nullsetzen der differenzierten Log-Likelihood geschätzt werden, soll auf die in Kapitel 2.2.1 kurz beschriebene Problematik näher eingegangen werden, wie viele Basisfunktionen q_j für die Funktion f_j angebracht sind und welcher Grad m_j für die Polynome in den durch die Stützpunkte getrennten Teilbereiche spezifiziert werden sollen.

2.3.1. P - Splines

Für die Schätzung der nichtparametrischen Funktion f_j ist die Anzahl der angegebenen Basisfunktionen q_j von deutlich größerer Bedeutung als der Grad des Polynoms m_j des B-Splines. In der Praxis hat sich die Verwendung kubischer Polynome (m_j=3) etabliert, da hierdurch die Funktion bis zur zweiten Ableitung stetig ist und sich somit für das freie Auge eine glatte Funktion ergibt.

Die Stärke der Glättung hängt vielmehr mit der Anzahl der Basisfunktion zusammen. Ist die Anzahl der Basisfunktionen q_j für die Funktion f_j zu hoch, besteht das Risiko, dass sich die geschätzte Funktion zu sehr den Daten anpasst und dadurch auch nicht



(a) 30 Basisfunktionen (b) 10 Basisfunktionen (c) 5 Basisfunktionen

Abbildung 2.2.: 500 Datenpunkte werden anhand der (schwarz markierten) Funktion $f(x) = cos(10x) + exp(x) - 1$ im Intervall [0,1] simuliert und mit unterschiedlicher Anzahl an Basisfunktionen gefittet (rot gekennzeichnet). Wegen der vielen Schwankungen erscheint die Spezifikation von 30 Basisfunktionen als zu hoch (Panel (a)). 10 Basisfunkktionen ergeben einen adäquaten Fit (Panel (b)), während die Funktion von 5 Basisfunktionen (Panel (c)) nicht mehr adäquat dargsetllt werden kann.

erklärbare Effekte mitmodelliert werden (häufig auch als „Überanpassung/ Overfitting" bezeichnet). Andererseits ist im Fall von zu wenig spezifizierten Basisfunktionen die geschätzte Funktion zu starr und zu wenig von den Daten erklärt („Unteranpassung/ Underfitting"). Abbildung 2.2 soll dieses Problem verdeutlichen, hier wird die Funktion

$$f(x) = cos(10x) + exp(x) - 1$$

mit 30, 10 und 5 Basisfunktionen modelliert (jeweils mit kubischen B- Splines, m_j=3). Während die Verwendung von 10 Basisfunktionen einen adäquaten Fit der Funktion ergibt, erscheint die Wahl von 30 Basisfunktionen deutlich zu hoch wie durch die vielen Schwankungen zu erkennen ist. Die Wahl von 5 Basisfunktionen wiederum erscheint als zu gering, um die Funktion adäquat zu beschreiben.

Einen Ausweg aus dieser Problematik bieten die unter dem Namen „P - Splines"[17] eingeführten penalisierten B - Splines an. Unter Verwendung einer fixen Anzahl an B-Spline Basisfunktionen werden die zu schätzenden Koeffizienten einerseits wie gehabt durch die zu fittenden Daten, anderseits aber auch durch einen zusätzlichen Penalisierungsterm bestimmt, der zu viele Schwankungen in der Funktion bestrafen soll.

Dieser Penalisierungsterm in der Funktion $f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ wird mit $\lambda_j P_j$ bezeichnet, wobei

[17] Eilers und Marx, »Flexible smoothing with B-splines and penalties«.

P_j ein Maß der Schwankungen in f_j angibt und $\lambda_j \in \mathbb{R}^+$ einen unbekannten, zu schätzenden „Smoothing Parameter" darstellt, der die Glättung der Funktion f_j kontrollieren soll. Die zu schätzende Funktion wird demnach durch

$$f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} - \frac{1}{2}\lambda_j P_j$$

dargestellt, wobei der Faktor $\frac{1}{2}$ wegen einer besseren Repräsentierung der Ableitungen inkludiert wurde.

Im Extremfall $\lambda_j = 0$ wird der Penalisierungsterm nicht berücksichtigt, hier wird die Funktion f_j durch die angegebe Anzahl an Basisfunktionen modelliert. Mit steigendem erhält der Penalisierungsterm mehr Bedeutung, die Funktion passt sich nicht mehr so genau an die Daten an, dafür wird der Verlauf glatter. Die Wahl der Glättung der Funktion f_j hängt demnach nur noch vom Smoothing Parameter λ ab, dessen Schätzung in Kapitel 2.3.3 beschrieben wird.

In folgenden Unterkapitel wird gezeigt, dass sich der Penalisierungsterm P_j des P - Splines in der Funktion f_j immer als

$$P_j = \beta_j^T \boldsymbol{S_j} \beta_j$$

schreiben lässt, wobei $\boldsymbol{S_j}$ eine $q_j \times q_j$ Penalisierungsmatrix bezeichnet.

Penalties für Funktionen ohne Restriktion

Für Funktionen f_j, die keiner bestimmten Restriktion unterliegen, bietet die Krümmung ein gutes Maß für die Schwankungen der Funktion f_j:

$$P_j = \sum_{l=1}^{q_j-2} (\beta_{j,l+2} - 2\beta_{j,l+1} + \beta_{jl})^2 \tag{2.32}$$

Niedrige Werte von P_j deuten eine geschmeidige, glatte Kurve ohne große Schwankungen an, da sich benachbarte Paramater nur wenig unterscheiden. $P_j = 0$ ergibt eine lineare Funktion.

In Matrixdarstellung kann der P - Spline geschrieben werden als $P_j = \beta_j^T \boldsymbol{S_j} \beta_j$,
wobei

$$\boldsymbol{S_j} = \begin{pmatrix} 0 & 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & -2 & 5 & -4 & 1 & 0 & 0 & \cdots \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & \cdots \\ 0 & 0 & 1 & -4 & 6 & -4 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

die Penalisierungsmatrix der Dimension $q_j \times q_j$ bezeichnet.

Penalties auf Funktionen mit Monotonierestriktion

Für iso-/ antiton restringierte Funktionen f_j wird analog zu Pya[18] der P- Spline Penalty P_j basierend auf den quadrierten Differenzen benachbarter Hilfsparameter definiert, beginnend mit dem 2. Hilfsparamter:

$$P_j = \sum_{l=2}^{q_j-1} (\beta_{j,l+1} - \beta_{j,l})^2 = \beta_j^T \boldsymbol{S_j} \beta_j,$$

mit der $q_j \times q_j$ Matrix

$$\boldsymbol{S_j} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ 0 & 0 & -1 & 2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Penalisierung basierend auf den Hilfsparametern β_j macht Sinn, nachdem nahe beieinander liegende Hilfsparameter β_j ähnliche Inkremente in den tatsächlichen Modellkoeffizienten γ_j erzeugen.

In Abbildung 2.3 werden im Intervall [0,2] 100 Datenpunkte gemäß der monoton steigenden (Panel (a)) / fallenden (Panel (b)) Funktion $f(x) = exp(x)$ / $f(x) = -exp(x)$ mit standardnormalverteilten Fehlern simuliert und die nichtparametrische Schätzung unter Verwendung gewöhnlicher B-Splines mit jener verglichen, die ein isoton restringierter P - Spline ergibt. Die restringiert modellierte Funktion passt sich der zugrunde liegenden Funktion wesentlich besser an und sieht deutlich glatter aus als der B - Spline.

Penalties auf Funktionen mit Monotonie- & Krümmungsrestriktion

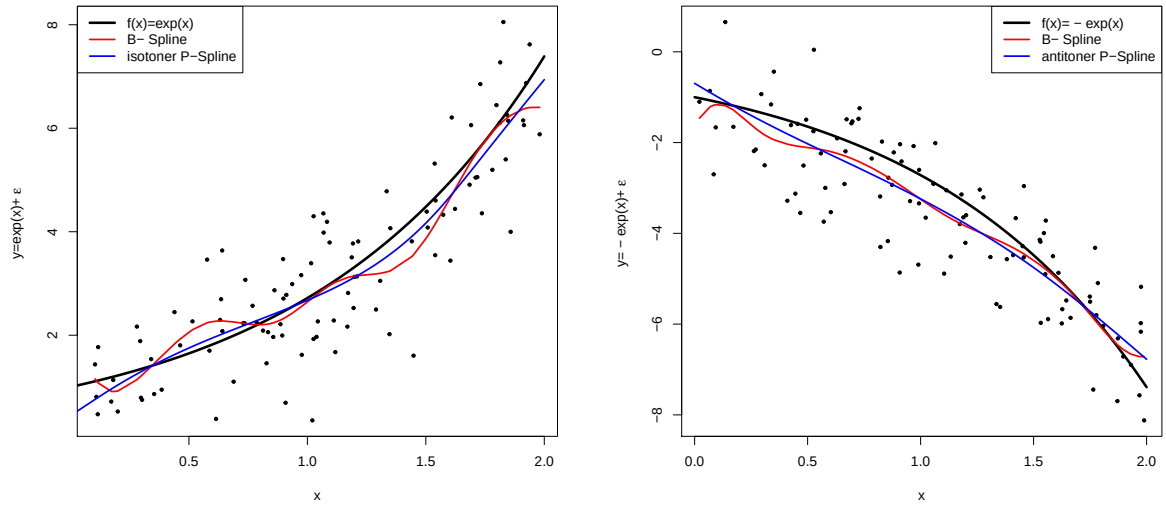
Für P- Splines, die sowohl Monotonie- als auch Krümmunsrestriktionen enthalten, werden wiederum quadratierte Differenzen benachbarter Hilfsparameter zur Penalisierung verwendet, diese startet allerdings erst ab dem dritten Hilfsparameter β_{j3}, da der zweite Hilfsparameter für die Steigung der gefitteten Kurve zuständig ist. Somit wird für gemischt- restringierte Modelle der Strafterm

$$P_j = \sum_{l=3}^{q_j-1} (\beta_{j,l+1} - \beta_{j,l})^2 = \beta_j^T \boldsymbol{S_j} \beta_j, \tag{2.33}$$

mit der $q_j \times q_j$ Penalisierungsmatrix

$$\boldsymbol{S_j} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -1 & 0 & \cdots \\ 0 & 0 & -1 & 2 & -1 & \cdots \\ 0 & 0 & 0 & -1 & 2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2.34}$$

[18] Pya, »Additive models with shape constraints«.



(a) isotone Funktion $f(x) = exp(x)$ (b) antitone Funktion $f(x) = -exp(x)$

Abbildung 2.3.: Vergleich zwischen dem Fit von B- Splines und monoton restringierter P-Splines basierend auf 100 simulierten Datenpunkten

verwendet. Pya[19] zeigte, dass eine solche Penalisierung zu einer quadratischen Funktion führt, falls nur $\lambda_j \to \infty$.

2.3.2. Penalisierte Log- Likelihood

Die Funktion f_j aus Gleichung (2.27) wird durch $f_j = Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} + \frac{1}{2}\lambda_j \beta_j^T \boldsymbol{S_j} \beta_j$ ersetzt, sodass für das additive Modell aus Gleichung (2.6) gilt:

$$\begin{aligned} \mu_i &= \eta_i \\ &= X_i^* + \sum_{j=1}^{p} (Z_j^* \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} - \frac{1}{2}\lambda_j \beta_j^T \boldsymbol{S_j} \beta_j) \end{aligned}$$

mit $\mathbb{E}(\boldsymbol{Y}) = \mu_i$ und $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Für Notationszwecke wird die gesamte $q \times q$ Penaltymatrix $\boldsymbol{S}$ definiert als Blockmatrix mit Einträgem der einzelnen Penaltymatrizen $\boldsymbol{S_j}$ in der Hauptdiagonale, wobei $q = q_0 + q_1 + q_2 + \cdots + q_p$ dabei die Gesamtanzahl aller im Modell verwendeten Parameter angibt. Die Smoothing Parameter λ_j, die die Stärke der Glättung der j-ten Funktion kontrollieren, werden ebenso wie der Varianzparameter σ^2 in die gesamte Penaltymatrix inkludiert/absorbiert:

[19] Pya, »Additive models with shape constraints«.

$$\boldsymbol{S} = \frac{1}{\sigma^2}\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_1\boldsymbol{S_1} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_2\boldsymbol{S_2} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_3\boldsymbol{S_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda_p\boldsymbol{S_p} \end{pmatrix}$$

Der lineare Prädiktor η_i lässt sich nun folgendermaßen darstellen:

$$\mu_i = X_i\tilde{\boldsymbol{\beta}} - \frac{1}{2}\beta^T\boldsymbol{S}\beta$$

Anstatt die Log-Likelihood aus Gleichung (2.30) zu maximieren wird die penalisierte Log-Likelihood $l_p(\boldsymbol{\beta})$, definiert durch

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta} \tag{2.35}$$

$$= -\frac{n}{2}log(2\pi) - nlog(\sigma) - \frac{1}{2}(y-\mu)^T(y-\mu) - \frac{1}{2}\beta^T\boldsymbol{S}\beta \tag{2.36}$$

in Abhängigkeit von β maximiert, wobei die Abhängigkeit von β auch in μ steckt. Es ist einfach zu sehen, dass Maximieren der penalisierten log-Likelihood äquivalent ist mit Minimierung des Terms

$$\frac{1}{2}(y-\mu)^T(y-\mu) + \frac{1}{2}\beta^T\boldsymbol{S}\beta \tag{2.37}$$

Bei als bekannt angenommenem Parametervektor $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)^T$ werden die „Hilfsparameter" β geschätzt. Bilden der ersten Ableitung nach β ergibt den sogenannten Score - Vektor

$$\boldsymbol{u}_p(\beta) = \frac{\partial l_p(\beta)}{\partial\beta} = \sum_{i=1}^{n}(y_i - \mu_i)\frac{\partial\mu_i}{\partial\beta} - \boldsymbol{S}\beta \tag{2.38}$$

$$= (y-\mu)\frac{\partial\mu}{\partial\beta} - \boldsymbol{S}\beta \tag{2.39}$$

Dabei hängt die partielle Ableitung $\frac{\partial\mu}{\partial\beta}$ von der Restriktion an die Funktion f_j ab:

▷ Für Funktionen ohne Restriktion gilt:

$$\frac{\partial\mu}{\partial\beta_{j,l}} = [X]_{j,l} \text{ für } l = 1, \ldots, q_j,$$

▷ Für Funktionen, die einer bestimmten Restriktion gehorchen, gilt:

$$\frac{\partial\mu}{\partial\beta_{j,1}} = [X]_{j,1} \text{ sowie } \frac{\partial\mu}{\partial\beta_{j,l}} = [X]_{j,l}exp(\beta_{j,l}) \text{ für } l = 2, \ldots, q_j$$

Nach Definition der Matrizen

$$\boldsymbol{C_j} = \begin{cases} diag(1, exp(\beta_{j2}), exp(\beta_{j3}), \ldots, exp(\beta_{j,q_j})) & \text{falls } f_j \text{ restringiert} \\ \boldsymbol{I}_{q_j} & \text{falls } f_j \text{ unrestringiert} \end{cases}$$

und

$$\boldsymbol{C} = diag(\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_p})$$

kann der Score Vektor $\boldsymbol{u}_p(\beta)$ in Matrixnotation ausgedrückt werden:

$$\boldsymbol{u}_p(\beta) = (X\boldsymbol{C})^T(y - \mu) - \boldsymbol{S}\beta$$

Nullsetzen des Score Vektors $\boldsymbol{u}_p(\beta)$ ergibt Schätzungen $\hat{\beta}$ für die Hilsparamter β. Daraus lassen sich nun Schätzer der restringierten Modellkoeffizienten $\tilde{\boldsymbol{\beta}}$ leicht ermitteln.

Diese (nichtlinearen) Gleichungen besitzen im Allgemeinen allerdings keine analytische Analysen, weshalb numerische Lösungen angewandt werden müssen. In Pya[20] und Wood[21] wird die Methode der „penalisierten, iterativen wiedergewichteten Kleinst- Quadrate /penalized iteratively re-weighted least squares (P- IRLS)" basierend auf dem Fisher-Scoring Algorithmus vorgestellt. Ist $\beta^{[k]}$ ein aktueller Schätzer für β, so bezeichnet

$$\beta^{[k+1]} = \beta^{[k]} + \mathcal{I}(\beta^{[k]})^{-1}\boldsymbol{u}_p(\beta^{[k]}),$$

den Schätzer der nächsten Iteration, wobei $\mathcal{I}(\beta) = -\mathbb{E}(H(\beta))$ Fisher's Informationsmatrix bezeichnet. $H(\beta)$, die Hesse-Matrix der penalisierten Log-Likelihood Funktion kann mithilfe der Gleichung (2.35) berechnet werden. Die Iterationsprozedur wird gestoppt, sobald sich die Ergebnisse stabilisieren.

Die Hesse Matrix bestimmt sich wie folgt:

$$H(\beta) = \frac{\partial^2 l_p(\beta)}{\partial\beta_j\beta_k} = -(X\boldsymbol{C})\frac{\partial\mu}{\partial\beta} + (X\boldsymbol{C'})^T(y - \mu) - \boldsymbol{S},$$

wobei $\boldsymbol{C'}$ die (elementweise) erste Ableitung der Matrix $\boldsymbol{C}$ nach β bezeichnet und demnach von der Restriktion abhängt:

$$\boldsymbol{C'_j} = \begin{cases} diag(0, exp(\beta_{j2}), exp(\beta_{j3}), \ldots, exp(\beta_{j,q_j})) & \text{falls } f_j \text{ restringiert} \\ \boldsymbol{0} & \text{falls } f_j \text{ unrestringiert} \end{cases}$$

sodass $\boldsymbol{C} = \text{diag}(\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_p})$.

[20] Pya, »Additive models with shape constraints«.
[21] Wood, *Generalized Additive Models. An Introduction with R*.

Nach Einsetzen des Ausdrucks für $\frac{\partial \mu}{\partial \beta}$ kann die Hesse-Matrix forgendermaßen bestimmt werden:

$$H(\beta) = -(X\boldsymbol{C})^T X\boldsymbol{C} + (X\boldsymbol{C}')^T - \boldsymbol{S}$$

Für einen gegebenen Schätzer $\beta^{[k]}$ kann der Schätzer der nächsten Iteration $\beta^{[k+1]}$ bestimmt werden durch

$$\beta^{[k+1]} = \beta^{[k]} + ((X\boldsymbol{C})^T X\boldsymbol{C} - (X\boldsymbol{C})^T + \boldsymbol{S})^{-1}((X\boldsymbol{C})^T(y-\mu) - \boldsymbol{S}\beta)$$

Für eine detailierte Beschreibung für generalisierte additive Modelle sei hier auf Wood[22] sowie Pya[23] verwiesen.

Modell ohne Restriktionen

Im Falle, dass keine Funktion $f_j, j = 1, \ldots, p$ restringiert wird, stimmt die Matrix $\boldsymbol{C}$ mit der Einheitsmatrix I_q überein, die Hilfsparamter β entsprechen den Modellparameter $\tilde{\beta}$. In diesem Fall ist eine analytische Lösung möglich und von folgender Gestalt:

$$\tilde{\boldsymbol{\beta}}^{\boldsymbol{unres}} = (X^T X + \boldsymbol{S})^{-1} X^T y$$

Hier ist der Einfluss der Penaltymatrix $\boldsymbol{S}$, die die Smoothing Parameter $\lambda_1, \ldots, \lambda_p$ enthält, deutlich zu erkennen. Nehmen alle Smoothing Parameter den Wert 0 an

$$(\lambda_j = 0 \forall j = 1, \ldots, p)$$

werden mögliche Schwankungen nicht berücksichtigt.

2.3.3. Effektive Freiheitsgrade und Selektion des Smoothing Parameters

Für die Schätzung der Hilfparameter β wurden bekannte Werte für $\lambda_1, \ldots, \lambda p$ angenommen. Wie diese geschätzt werden können, soll dieses Unterkapitel verdeutlichen. λ_j kontrolliert bekanntlich das Maß der Glättung der Funktion f_j. Im Extremfall $\lambda_j = 0$ ist die penalisierte Log-Likelihood ident mit der unpenalisierten Log-Likelihood, das Problem des Overfitting könnte auftreten. Auf der anderen Seite wird der Verlauf der Funktion f_j mit steigendem λ_j immer unflexibler und es sollten dadurch effektiv weniger Freiheitsgrade zur Schätzung verbraucht werden als im Extremfall $\lambda_j = 0$, wo die geschätzte Funktion höchstmögliche Komplexität besitzen kann.

Damit treten folgende Fragen auf:

- Wie soll der Wert der effektiv verbrauchten Freiheitsgrade ermittelt werden?
- Welche Werte soll der Smoothing- Parametervektor λ annehmen?

[22] Wood, *Generalized Additive Models. An Introduction with R.*
[23] Pya, »Additive models with shape constraints«.

Effektive Freiheitsgrade

Im gewöhnlichen linearen Modell stimmen die Freiheitsgrade mit der Anzahl der zu schätzenden Parameter überein. Ebenso würde ein unpenalisiertes additives Modell genau so viele Freiheitsgrade besitzen wie Modellparameter. Die Verwendung von P- Splines mit Penaltymatrix $\boldsymbol{S}$ reduziert allerdings die Anzahl der effektiv verbrauchten Freiheitsgrade.
Um diese zu bestimmen wird ein analoges Konzept zum üblichen linearen Modell gebildet: Im linearen Modell mit Designmatrix X der Dimension $n \times q$ hat die Hut - Matrix $\boldsymbol{H}$ folgende Gestalt:

$$\boldsymbol{H} = X(X^TX)^{-1}X^T,$$

sodass die gefitteten Werte $\hat{y}$ geschrieben werden können als

$$\hat{y} = \boldsymbol{H}y.$$

Die Anzahl der verbrauchten Freiheitsgrade im klassischen linearen Modell kann beispielsweise durch

$$\mathbf{spur}(\boldsymbol{H}) = \mathbf{spur}(X(X^TX)^{-1}X^T) = \mathbf{spur}(X^TX(X^TX)^{-1}) = \mathbf{spur}(I_q) = q$$

bestimmt werden. Analog hierzu wird im additiven Modell mit Penaltymatrix $\boldsymbol{S}$ die Anzahl der effektiven Freiheitsgrade τ definiert als

$$\tau = \mathbf{spur}(\boldsymbol{F}),$$

wobei die Einträge der Hut - Matrix $\boldsymbol{F}$ =((X$\boldsymbol{C}$)$^TX\boldsymbol{C} - (X\boldsymbol{C})^T + \boldsymbol{S})^{-1}(X\boldsymbol{C})^TX\boldsymbol{C}$ zum Zeitpunkt der Konvergenz des Algorithmus ausgewertet werden. Detaillierte Ausführungen sind in Wood und Pya zu finden.

Im Spezialfall, wo keine restringierte Funktion f_j vorliegt, ist $\boldsymbol{F}^{unres}$ durch

$$\boldsymbol{F}^{unres} = X(X^TX + \boldsymbol{S})^{-1}X^T$$

gegeben.

Schätzung von λ

Zur Schätzung des Smoothing Parameters λ wird das Kriterium der generalisierten Kreuzvalidierung (GCV)[24]

$$V_g = \frac{nD(\hat{\beta})}{(n-\tau)^2} \tag{2.40}$$

bezüglich λ minimiert. Die Abhängigkeit von λ ist dabei in τ enthalten.

$$D(\hat{\beta}) = 2(l_{sat} - l(\hat{\beta}))\sigma^2 \tag{2.41}$$

[24]vorgeschlagen von Hastie und Tibshirani, *Generalized Additive Models*

bezeichnet die Devianz des Modells. Die Notation l_{sat} bezeichnet dabei die Maximum - Likelihood des saturierten Modells mit einem Parameter pro Beobachtung, sodass $\hat{\mu} = y$. Die Devianz ist ein in Analogie zum linearen Modell verallgemeinert entwickeltes Konzept zur Beurteilung der Modellgüte und stimmt im additiven Modell aus Gleichung (2.1) mit der Quadratsumme der Residuen (RSS) im linearen Modell überein. Ein effizienter Algorithmus zur simultanen Schätzung der Smoothing Parameter $\lambda_j, j = 1, \ldots, p$ wird in Wood[25] und in Pya vorgestellt.

2.4. Restriktionen auf den parametrischen Part

Für das additive Modell liegt somit eine Schätzung $\hat{\beta}$ für den Vektor der „Hilfsparameter"
$\beta = (\alpha^T, \beta_1^T, \beta_2^T, \ldots, \beta_p^T)$ vor. Nun sollen auch Modellparameter des parametrischen Parts restringiert werden.
Hierzu bezeichne R eine $r \times q_0$ Restriktionsmatrix mit linear unabhängigen Reihen sowie r einen Vektor der Dimension r, sodass

$$C_{R,r} = \{\alpha \in \mathbb{R}^{q_0} : R\alpha \geq r\}. \tag{2.42}$$

Das Set $C_{R,r}$ nennt man einen polyedrischen Kegel.

Im gesamten Vektor der Modellkoeffizienten können diese Restriktionen durch Definition von

$$R^* = (R : \vec{0}),$$

mit $R^* \in \mathbb{R}^{r \times q}$ dargestellt werden, sodass

$$C_{R^*,r} = \{\beta \in \mathbb{R}^q : R^*\beta \geq r\} \tag{2.43}$$

Allerdings soll nur der parametrische Teil α restringiert werden, die Modellparameter der nichtlinearen Terme werden auf dem bereits geschätzten Wert festgehalten. Hierzu bezeichne $\beta^- = (\beta_1^T, \beta_2^T, \ldots, \beta_p^T)^T$ den Subvektor der unbekannten Modellkoeffizienten ohne α und $\hat{\beta}^- = (\hat{\beta}_1^T, \hat{\beta}_2^T, \ldots, \hat{\beta}_p^T)^T$ den Subvektor der geschätzten Hilfsparameter.

Nach Definition von

$$C^-_{R^*,r} = \{\beta \in \mathbb{R}^q : R^*\beta \geq r \quad , \beta^- = \hat{\beta}^-\} \tag{2.44}$$

kann das zusätzlich auf den parametrischen Part restringierte Minimierungsproblem aus Gleichung (2.37) als Minimierungsproblem mit Nebenbedingungen

$$min\{\beta \in C^-_{R^*,r} : (y - \mu)^T(y - \mu)\}, \tag{2.45}$$

formuliert werden, wobei die Abhängigkeit von β in $\mu = X\tilde{\beta}$ enthalten ist. Dieses Problem ist ein wohlbekanntes quadratisches Minimierungsproblem und kann mit geeigneter

[25] Wood, »Fast stable direct fitting and smoothness selection for generalized additive models«.

Software gelöst werden.

3. Datenbeschreibung

In dieser Arbeit werden Preise von Einfamilienhäusern in Österreich analysiert. Wie in Kapitel 1.1 beschrieben, wird in hedonischen Modellen ein Haus gedanklich in seine Eigenschaften wie Größe oder Zustand zerlegt, die Kaufpreise sollen mithilfe eines Regressionsmodells erklärt werden. Dieses Kapitel soll die verwendeten Daten vorstellen und näher beleuchten.

Die verwendeten Daten stammen aus folgenden Quellen:

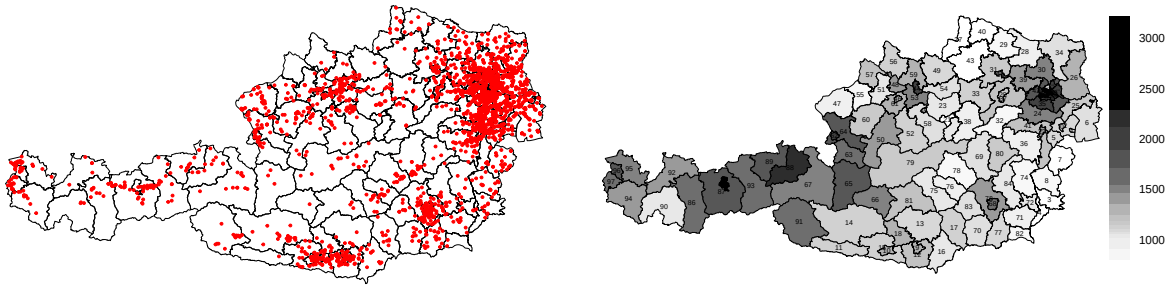
- ▷ Daten, die einem bestimmten Objekt zugeordnet werden können („Objektdaten") stammen vom Bewertungssystem der Bank Austria (BA). Möglicherweise wertrelevante Faktoren wie Zustand und Ausstattung werden von Gutachtern bestimmt sowie vorhandene Kaufpreise inklusive Kaufjahr erfasst.

- ▷ Lagebeschreibende Einflussfaktoren werden durch die Variablen Grundpreis, Abiturientenanteil sowie Lärm modelliert.
  Grundpreise stammen von der Zeitschrift GEWINN[1], die jährlich Grundstückspreise pro Gemeinde anhand von Direktabfragen bei den Gemeinden sowie statistischen Berechnungen des Fachbereichs Stadt- und Regionalforschung der TU Wien unter Leitung von Prof. Feilmayr erhält.
  Der Anteil der Abiturienten wurde auf Zählsprengelebene von der Technischen Universität Wien erhoben und soll erklären ob eine höhere Bildung die Lage und somit auch den Preis beeinflusst.
  Die Variable Lärm basiert auf 50x50 Meter genauen Rasterkacheln, sie beschreibt den Umgebungslärm, der von Autobahnen, Hauptstrassen, Nebenstrassen sowie Eisenbahnstrecken stammt. Nähere Details zur Modellierung der Lärmvariable sowie weiteren kleinräumigen - auf Rasterebene basierenden - Variablen wird in Weberndorfer[2]beschrieben.
  Zusätzlich liegen pro Objekt lagegenaue Geokoordinaten basierend auf dem kartesischen Koordinatensystem vor, sodass Distanzen zwischen Beobachtungspunkten ermittelt werden können.

Modelliert werden Objekte, die einen Preis mit dazugehörigem Datum aufweisen. Die Preisinformation kann sich dabei auf ein reelles Kaufgeschäft, einen Angebotspreis oder

[1] http://www.gewinn.com/immobilien/preisuebersichten/oesterreich/

[2] Weberndorfer, »Modellierung von wertrelevanten Mikrolageparametern fuer die automatisierte Immobilienbewertung«, 41ff.



(a) Verteilung der Datenpunkte (b) bezirksaggregierte Preise pro Quadratmeter

Abbildung 3.1.: Panel (a) zeigt die Verteilung der Beobachtungen über Österreich, während in Panel (b) bezirksaggregierte Preise pro Quadratmeter Nutzfläche zu erkennen sind.

einen geschätzten Preis eines qualifizierten Gutachters beziehen.

Zur Verringerung des Messfehlers wurden unplausible Beobachtungen von der Analyse ausgeschlossen, dazu gehören:

- ▷ Preise unter 30.000€ oder 1.000€/m^2 sowie Preise über 2.000.000€ oder 20.000€/m^2
- ▷ Grundstücksflächen über 1.500m^2
- ▷ Zur Nutzfläche gehörende Flächen (Erdgeschoß-, Obergeschoß-, Kellerfläche, ...) über 200 Quadratmeter
- ▷ Objekte, die vor 1900 beziehungsweise nach 2012 gebaut wurden
- ▷ Beobachtungen mit fehlender Geocodierung
- ▷ Objekte, die vor 2005 ge- / verkauft wurden, wurden auf das Jahr 2005 gesetzt.

Nach Ausschluss unplausibler Information verbleiben 3152 Einfamilienhäuser im Datensatz, diese werden in Abbildung 3.1(a) dargestellt. Dabei fällt die hohe Beobachtungsdichte in Niederösterreich und innerhalb der Städte Graz, Linz und Klagenfurt sowie deren urbanen Umgebung auf, während die Beobachtungsdichte besonders in Westösterreich noch recht gering ist.

Abbildung 3.1(b) zeigt bezirksaggregierte Preise pro Quadratmeter Nutzfläche. Die teuersten Bezirke befinden sich in Wien und Umgebung sowie in Westösterreich, hier sind besonders die Bezirke Innsbruck (Stadt), Kitzbühel und Salzburg (Stadt) zu erwähnen. Allerdings können daraus keine generellen Schlüsse gezogen werden, dass diese Gegenden teurer sind oder nur aufgrund anderer Effekte teurer erscheinen. Beispielsweise könnten sich in dem Datensatz viele neu gebaute Häuser in Westösterreich befinden und der erhöhte Preis auch aufgrund des Baujahreseffekt und nicht alleine von der Lage erklärbar sein.

3.1. Deskriptive Statistiken

	Kürzel	Mittelwert	Standardabw.	Minimum	Maximum
Preis	p	240.042	123.124	30.000	985.759
ln(Preis)	lnp	12.3	0.5	10.3	13.8

Tabelle 3.1.: Deskriptive Statistik der abhängigen Variable

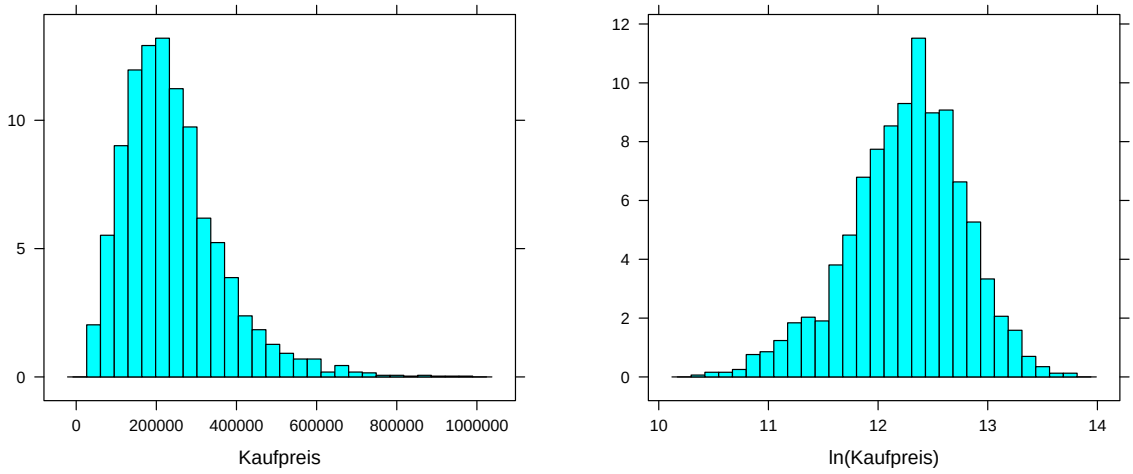
	Kürzel	Mittelwert	Standardabw.	Minimum	Maximum
Baujahr	bj	1982	26	1900	2012
Sanierungsdifferenz	san_diff	7	14	0	82
Grundstücksfläche	gst	629.6	289.6	80.0	1300.0
Dachgeschoss (Lagerzw.)	dg_lager	10.0	27.0	0.0	150.0
Dachgeschoss (Wohnzw.)	dg_wohn	27.2	33.6	0.0	164.4
Erdgeschossfläche	eg	81.1	26.4	35.0	165.0
Keller (Lagerzw.)	keller_lager	48.7	38.5	0.0	145.0
Keller (Wohnzw.)	keller_wohn	5.8	22.3	0.0	143.9
Obergeschossfläche	og	24.7	35.9	0.0	165.0

Tabelle 3.2.: Deskriptive Statistik der objektrelevanten, kontinuierlichen Einflussfaktoren

Tabelle 3.1 zeigt eine deskriptive Statistik der zu erklärenden Kaufpreise. Diese befinden sich zwischen 30.000 € und knapp 1 Million €. Aufgrund der Tatsache, dass Kaufpreise immer positive Werte annehmen und der deutlich rechtschiefen Verteilung werden die Preise logarithmiert. Histogramme zur Verteilung der Kaufpreise sowie logarithmierten Kaufpreise befinden sich in Abbildung 3.2. Die Preise sind deutlich rechtsschief verteilt, nach Log - Transformation passen sich die Daten annähernd einer Normalverteilung an.

Deskriptive Statistiken objektrelevanter Variablen sind in Tabelle 3.2 zu finden.

- ▷ Das Baujahr der Immobilie wurde ebenso erfasst wie das fiktive Baujahr, das sich aufgrund diverser Sanierungen vom Baujahr unterscheiden kann. Die Variable Sa-



(a) Histogramm der Preise

(b) Histogramm der logarithmierten Preise

Abbildung 3.2.: Histogramm der endogenen Variable

nierungsdifferenz gibt die Differenz der Zeitspanne zwischen dem fiktiven und dem tatsächlichen Baujahr an, sodass auch sanierte Einfamilienhäuser mithilfe des Einflusses dieser Variable mitmodelliert werden können.

▷ Dachgeschoß- sowie Kellerflächen werden unterschieden, ob sie hauptsächlich als Lagermöglichkeit dienen oder als Wohnmöglichkeit genutzt werden. Anhand der Tabelle 3.2 ist zu erkennen, dass Kellerflächen vorwiegend eine Lagermöglichkeit bieten, während Dachgeschosse oft auch als Wohnräume genutzt werden.

Tabelle 3.3 sowie Tabelle 3.4 geben weitere objektrelevante Eigenschaften sowie deren beobachtete Häufigkeit an. Während für die Variablen in Tabelle 3.3 mehrere Ausprägungsmöglichkeiten zur Wahl stehen, sind in Tabelle 3.4 nur die Ausprägungen ja (ist vorhanden) und nein (nicht vorhanden) möglich. Die Spalte „Eff" kennzeichnet die erwarteten Effekte. Referenzkategorien werden dabei mit „R" abgekürzt, das Symbol + deutet auf einen erwartet preissteigernden Effekt hin, - auf einen erwartet wertmindernden.

Nennenswerte Beispiele für objektrelevante Merkmale wären:

▷ allgemeine Eigenschaften des Hauses (Zustand, Ausrichtung der Wohnräume, Wandaufbau):
Zustände liegen in Schulnotensystem (Zustandsnoten 1 bis 5) vor und wurden durch qualifizierte Gutachter eingeschätzt. Auswirkungen vom Zustand zum Preis sollten direkt proportional sein, sodass ein guter Zustand einen preissteigernderen Einfluss besitzen sollte als ein mittelmäßiger Zustand.

Für süd- sowie westseitige Ausrichtung der Wohnräume werden im Vergleich mit nord- und südseitig ausgerichteten Wohnräumen positive Zusammenhänge erwartet.

- ▷ Eigenschaften des Badezimmers/ WCs (Größe, Zustand, Typ, Heizmöglichkeit):
  Eigenschaften des Badezimmers wurden ziemlich genau erfasst, da ein guter Zustand des Badezimmers auch gute Zustände der übrigen Räumlichkeiten vermuten lässt.

- ▷ Heizart (Zentralheizung/ Individualheizung, zusätzliche Fußbodenheizung, Wandheizung):
  Moderne Zentralheizungen sollten im Vergleich mit Individualheizungen einen positiven Einfluss auf der Preis besitzen. Für zusätzliche Heizmöglichkeiten einer Fußboden- sowie Wandheizung werden preissteigernde Effekte erwartet.

- ▷ Hochwertige Ausstattung (hochwertige Türen/Fenster, Alarmanlagen):
  Zur Klasse hochwertiger Ausstattungsmerkmale gehören beispielsweise Holz- oder Alufenster, Nurglastüren sowie die Existenz einer Alarmanlage oder Videoüberwachung. Vorhandensein dieser Eigenschaften sollte sich positiv auf den Preis auswirken.

- ▷ Außenanlagen (hochwertiger Gartenzaun, Schwimmbad)

- ▷ Zusatzflächen (Balkon, Terrasse, Garage)

- ▷ Art (Fertigteilhaus, Reihenhaus)

- ▷ Preisinformation

Tabelle 3.5 stellt deskriptive Statistiken zu lagerelevanten Variablen dar.

- ▷ Die Variable Lärm, in Dezibel gemessen, beschreibt den Umgebungslärm, der von stark befahrenen Strassen sowie Eisenbahnstrecken verursacht wird. Die Erhebung erfolgt auf 50x50 Meter genauen Rasterkacheln, nähere Details hierzu findet man in Weberndorfer[3]. Ein negativ korrelierter Zusammenhang zwischen Lärm und dem Preis eines Einfamilienhauses wird unterstellt. So sollte sich beispielsweise die Nähe zu einer Autobahn wertmindernd auswirken.

- ▷ Die Erhebung des Abiturientenanteils erfolgt zählsprengelgenau vom Institut Regionalforschung der Technischen Universität Wien unter Leitung von Prof. Feilmayr.

- ▷ Mittlere Grundstückspreise wurden der Zeitschrift GEWINN entnommen und liegen auf Gemeindebene vor. Aufgrund der rechtsschiefen Verteilung wurde diese Variable logarithmiert. Nach Annahme gehen mit höheren Grundstückspreisen höhere Preise für ein Einfamilienhaus einher.

- ▷ Zusätzlich wird die Makrolage eines Objektes von Gutachtern erfasst, Tabelle 3.6 enthält die möglichen Ausprägungen mit beobachteter Anzahl. Hier werden keine speziellen Effekte angenommen.

[3] Weberndorfer, »Modellierung von wertrelevanten Mikrolageparametern fuer die automatisierte Immobilienbewertung«, 41ff.

	Anzahl	Eff
nur Wanne	1219	R
nur Dusche	438	o
beides	1454	+
keine	41	-

(a) Badegelegenheiten

	Anzahl	Eff
< 4qm	341	-
4-10 qm	2308	R
> 10qm	503	+

(b) Größe des Badezimmers

	Anzahl	Eff
mittel	1835	R
sehr gut	194	+
schlecht	1123	-

(c) Zustand des Badezimmers

	Anzahl	Eff
sehr gut	1864	R
gut	130	-
mittel	941	-
schlecht	10	-
sehr schlecht	207	-

(d) Zustand des Hauses

	Anzahl	Eff
Zentralheizung	2904	R
Indiviudalheizung	248	-

(e) Heizung

	Anzahl	Eff
Süden	2469	R
Norden	559	-
Osten	53	-
Westen	71	o

(f) Orientierung der Wohnräume

	Anzahl	Eff
Ziegel oder Beton	2604	R
einfache Holzkonstruktion	516	o
hochwertige Holzkonstruktion	26	o
andere Bauweise	6	o

(g) Wandaufbau

	Anzahl	Eff
kein	1999	R
mittel	34	+
gross	1119	++

(h) Balkon

Tabelle 3.3.: Objektrelevante Eigenschaften mit mehreren Ausprägungsmöglichkeiten sowie beobachteter Anzahl und erwartetem Effekt. R bezeichnet die Referenzkategorie, + deutet auf einen positiv erwarteten Effekt hin, - auf einen negativ erwarteten. Effekte, an die vorab keine speziellen Forderungen gestellt werden, sind mit o gekennzeichnet.

	nein	ja	Eff
Bad verfliesst	424	2728	+
Fußbodenheizung im Bad	2338	814	+
nur Heizstrahler im Bad	3035	117	-
Fenster im WC	1279	1873	o
elektr. Garagentor	2395	757	+
eingeschränkte Zufahrt zur Garage	3028	124	-
Fußbodenheizung	2049	1103	+
Wandheizung	3099	53	+
Alarmanlage	2713	439	+
hochwertige Fentser/ Türen	2829	323	+
hochwertiger Gartenzaun	2782	370	+
Swimmingpool	2745	407	+
schlechte Raumaufteilung	2759	393	-
Terrasse vorhanden	2549	603	+
Garage vorhanden	2037	1115	+
Fertigteilhaus	987	2165	o
Reihenhaus	2621	531	o
Angebotspreis	2809	343	o
Bau von gemeinnützigem Bauträger	3125	27	o

(a) Existenz (möglicherweise) wertbeeinflussender Merkmale: Erwartete Effekte werden bei Vorhandensein einer Eigenschaft in Spalte Eff mit + (wertsteigernder Effekt), - (wertmindernder Effekt) und o (kein näher spezifizierter Effekt) angegeben

	Anzahl
2005	108
2006	27
2007	43
2008	221
2009	622
2010	788
2011	813
2012	525
2013	5

(b) Kaufjahr

Tabelle 3.4.: Weitere objektspezifischen Merkmale

	Kürzel	Mittelwert	Standardabw.	Minimum	Maximum
Lärm	laerm	63.47	11.09	4.77	95.06
ln(Abiturientenanteil)	ln_abi	2.90	0.46	0.88	4.11
ln(Grundpreis)	ln_grundp	4.85	0.72	2.84	7.01
X - Koordinate	xco	566777.85	109379.31	114994.00	681187.00
Y - Koordinate	yco	393116.74	58010.45	236919.00	512442.00

Tabelle 3.5.: Deskriptive Statistik der lagerelevanten Einflussfaktoren

	Anzahl
im Zentrum	633
am Stadt-/Ortsrand	2304
Villengegend	76
periphere Lage	139

Tabelle 3.6.: Makrolage

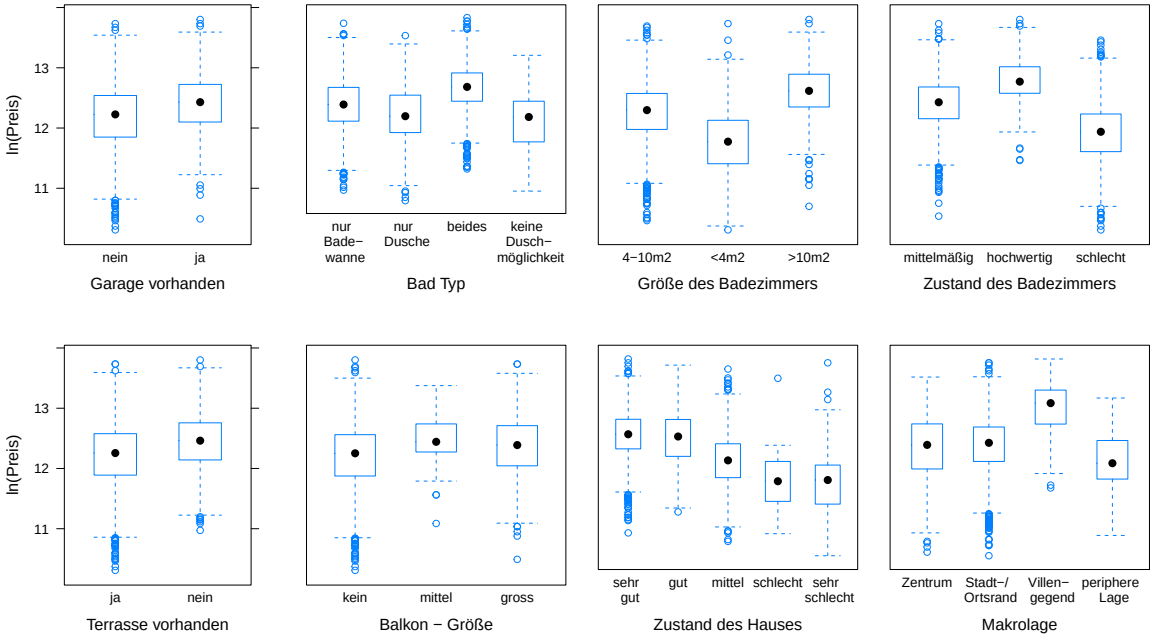


Abbildung 3.3.: Kastengrafiken ausgewählter objektrelevanter Eigenschaften

3.2. Grafische Beschreibung der Zusammenhänge

Dieses Kapitel soll aufgrund zwei- und mehrdimensionaler Grafiken die Daten näher erklären und Zusammenhänge mit den Kaufpreisen wiedergeben.

Abbildung 3.3 zeigt Kastengrafiken (Boxplots) zwischen logarithmierten Preisen und einigen (möglicherweise) wertbeeinflussenden Objekteigenschaften.

- ▷ Es scheint, als hätten Einfamilienhäuser, die nur eine Badewanne als Badegelegenheit haben einen minimal stärkeren Effekt auf den Preis als Objekte, bei denen nur eine Dusche vorhanden ist. Dienen sowohl Badewanne als auch Dusche zur Bademöglichkeit, dürfte sich dieser Effekt erhöhen. Ein negativer Zusammenhang im Fall keiner Duschgelegenheit und dem Preis ist anhand der Boxplots nicht zu erkennen, möglicherweise bedingt durch die geringe Ausprägungshäufigkeit (41).

- ▷ Kleine Badezimmer (Größe unter $4m^2$) dürften einen negativen Einfluss auf den Preis besitzen im Vergleich mit durchschnittlich großen Badezimmer (zwischen 4 und 10 Quadratmeter). Überdurchschnittlich große Badezimmer dürften erwartungsgemäß einen preissteigernden Effekt besitzen.

- ▷ Es macht den Anschein, dass Badezimmer in schlechtem Zustand Preise erwartungsgemäß verringern lässt, während eine Preissteigerung für hochwertig ausgestattete Badezimmer vermuten werden kann.

- ▷ Anhand der Boxplots sieht es aus, dass weder die Existenz einer Terrasse noch die eines Balkons wesentlich beeinflussende Merkmale auf den Preis eines Einfamilienhauses darstellen.

- ▷ Erwartungsgemäß dürften sich die Preise eines Objekts mit abnehmender Zustandsnote fallend auswirken. Dabei fällt ein deutlicher Sprung zwischen den ersten beiden Kategorien (sehr guter/guter Zustand) und den übrigen auf.

- ▷ Anhand des Boxplots für Makrolage ist von einer deutlichen Preissteigerung von Häusern in Villengegenden auszugehen, allerdings kann dieser Effekt möglicherweise auch durch dahinter stehenden Variablen erklärt werden. Ein in einer Villengegend angesiedeltes Objekt ist wahrscheinlich selbst eine Villa, der erhöhte Kaufpreis ist demnach nicht (nur) durch die Gegend, sondern beispielsweise auch durch die höhere Grundstücksfläche oder gehobenere Ausstattung erklärbar.

Ähnliche Schlussfolgerungen lassen sich auch für übrige Objekteigenschaften ziehen. Die Verwendung des Regressionsmodells in Kapitel 4 soll den Effekt des Einflusses einer bestimmten Eigenschaft wiedergeben.

Abbildung 3.4 zeigt Streudiagramme zwischen (logarithmierten) Kaufpreisen und mittleren (logarithmierten) Grundstückspreisen. Zudem wird das Alter der Immoblie in vier Kategorien aufgeteilt, die Graustufen der Punkte verdeutlichen die Größe des Erdgeschoßes. Dabei fällt der deutliche Unterschied zwischen Regionen mit niedrigen Grundstückspreisen und hohen Grundstückspreisen auf. Ebensfalls sind auch größere Preise für Objekte, die innerhalb der letzten 10 Jahre errichtet wurden und ein relativ großes Erdgeschoß besitzen, zu beobachten, auch wenn sie nicht in Regionen mit sehr hohen mittleren Grundstückspreisen stehen.

Abbildung 3.5 zeigt eine Streudiagrammmatrix zwischen dem (logarithmierten) Kaufpreis, dem Alter der Immobilie und den Flächengrößen des Grundstücks, des Erdgeschoßes sowie des Obergeschoßes. Punkte werden je nach Zustandsnote der Immobilie verschieden eingefärbt. Hier lässt sich ein deutlicher Zusammenhang zwischem dem Alter und dem Kaufpreis erkennen. Ein weiterer bemerkenswerter Punkt besteht darin, dass sich Objekte, die kein Obergeschoß besitzen, meist auch in einem sehr schlechten Zustand (Zustandsnote 5) befinden.

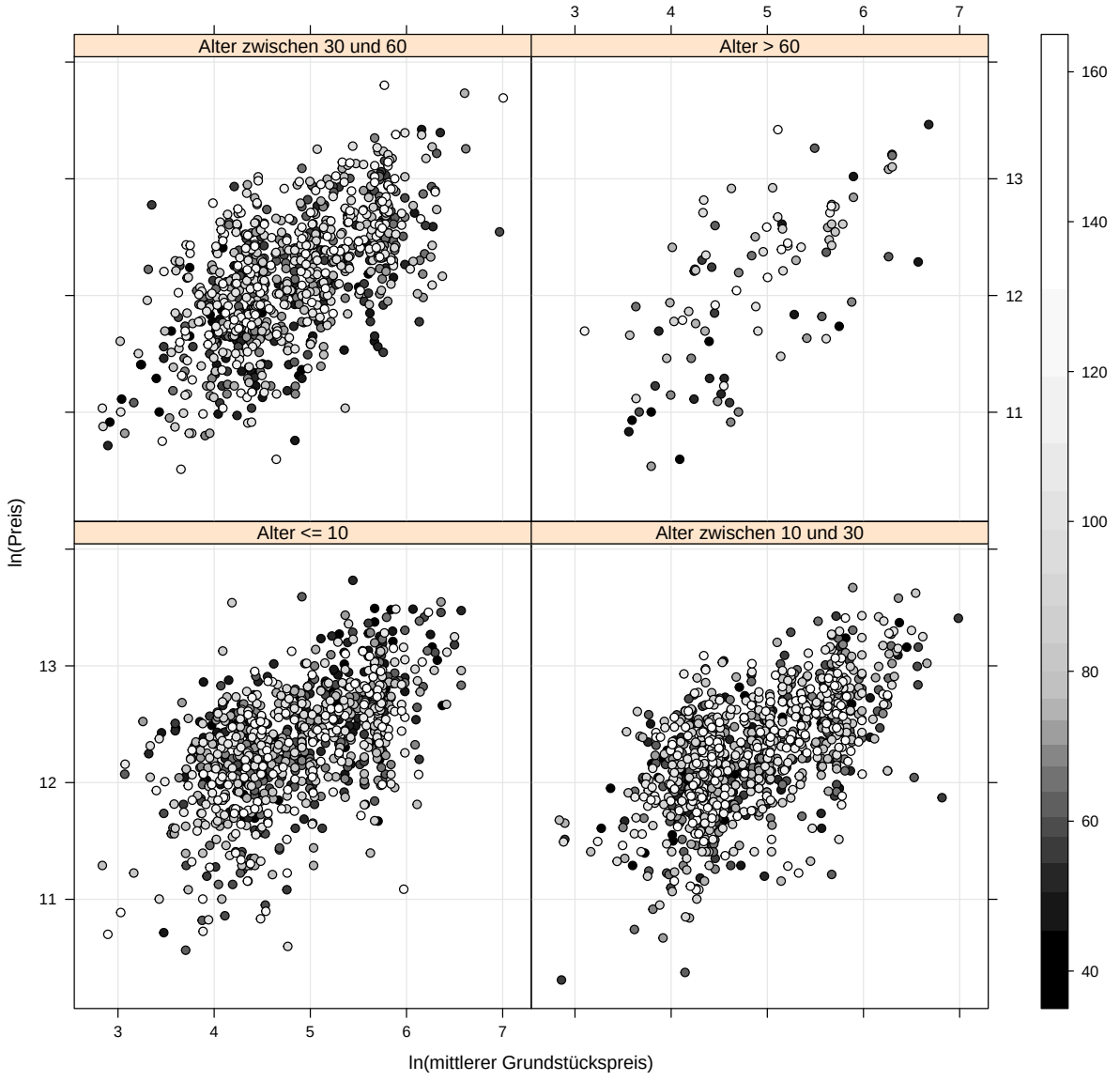


Abbildung 3.4.: Das Alter der Immobilie wird in vier Kategorien eingeteilt, wobei für jede Kategorie ein Streudiagramm zwischen (logarithmierten) mittleren Grundstückspreisen und den (logarithmierten) Kaufpreisen abgebildet wird. Die Graustufe ergibt sich aus der Größe des Erdgeschoßes.

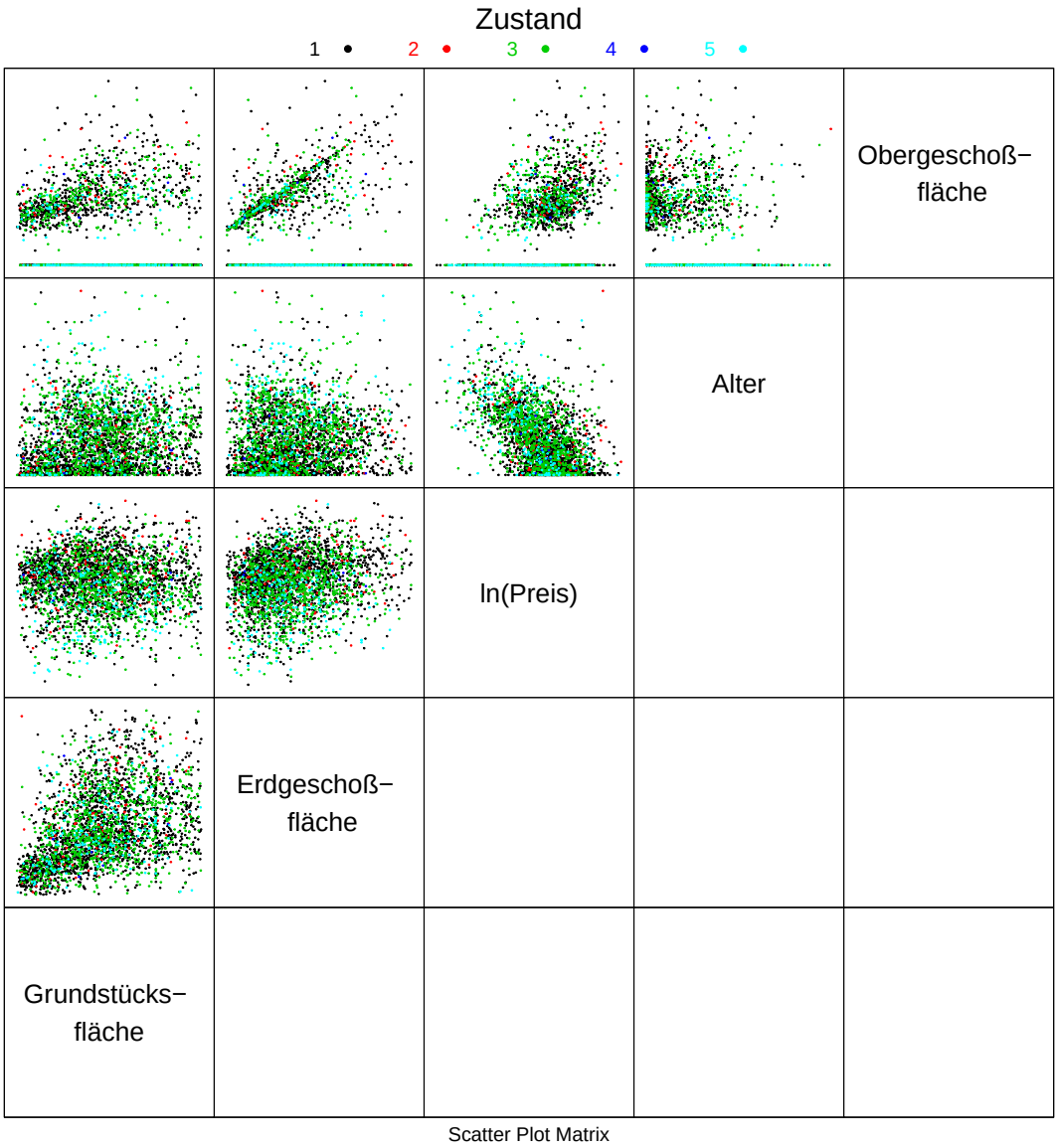


Abbildung 3.5.: Streudiagramme zwischen dem (logarithmierten) Preis, Flächengrößen von Grundstück, Erdgeschoß sowie Obergeschoß und dem Alter des Hauses. Die Farben ergeben sich aus den Zustandsnoten des Objektes.

4. Beschreibung der Modelle

4.1. Basismodell

Da ohne geeignete Transformation der Kaufpreise die Annahme der normalverteilten Zielgrößen im additiven Modell nicht zu rechtfertigen sind, werden die Kaufpreise logarithmiert, die resultierende Verteilung nähert sich der einer Normalverteilung an, siehe Abbildung 3.2.

Somit gilt für das Modell mit dem logarithmierten Preis (lnp) als abhängige Variable:

$$\mathbb{E}(lnp_i) = \eta_i$$

mit $lnp_i \sim \mathcal{N}(\eta_i, \sigma^2)$.

Die resultierende Verteilung der Kaufpreise ist somit lognormalverteilt, $p \sim lognormal(\eta_i, \sigma^2)$, es ergibt sich ein multiplikativer Zusammenhang zwischen Preisen und erhobenen Kovariaten. Der Erwartungswert der untransformierten Preise wird folgendermaßen ausgedrückt:

$$\mathbb{E}(p_i) = exp(\eta_i + \sigma^2/2) = exp(X_i^*\alpha) \cdot exp(f_1(z_{1i})) \ldots exp(f_p(z_{pi})) \cdot exp(\sigma^2/2)$$

Der lineare Prädiktor η_i ändert sich durch Erhöhung des Werts der Kovariate x_1^* um eine Einheit um den Faktor $exp(\alpha_1)$. Erhöhung der Kovariaten eines nichtlinearen Terms um eine Einheit, beispielsweise z_1 mit dazugehörigem Funktionswert $f_1(z_1)$ bewirkt eine Änderung um den Faktor $exp(f_1(z_1 + 1) - f_1(z_1))$.

Für das Basismodell wurde folgende Form unterstellt und mit den in Kapitel 2 vorgestellten Methoden gefittet:

$$\begin{aligned} \mathbb{E}(lnp_i) = X_i^*\alpha + f(bj_i) + f_{\nearrow}(san\_diff_i) + f_{\curvearrowright}(gst_i) + f_{\curvearrowright}(eg_i) + f_{\curvearrowright}(og_i) + \\ f_{\nearrow}(dg\_lager_i) + f_{\nearrow}(dg\_wohn_i) + f_{\nearrow}(keller\_lager_i) + f_{\nearrow}(keller\_wohn_i) + \\ f_{\searrow}(laerm_i) + f_{\nearrow.}(ln\_grundp_i, ln\_abi_i) + \\ kaufjahr_i \cdot f(xco_i, yco_i) \end{aligned} \tag{4.1}$$

Die Schätzung verläuft unter der in Kapitel 2.3 vorgestellen Logik. Pro nichtparametrisch modellierter Funktion werden f_j jeweils 9 B-Spline Basisfunktionen ($q_j = 9$) und Polynome vierter Ordnung ($m_j = 4$) spezifiziert.

- ▷ X_i^* enthält alle objektrelevanten Eigenschaften, die in Tabelle 3.4, 3.3 und 3.6 angegeben wurden. Referenzkategorien wurden in diesen Tabellen mit dem Kürzel „R" angegeben.

- ▷ Obwohl ein positiver Effekt auf den Preis für später errichtete Häuser vermuten werden kann, wird die nichtlineare Funktion des Baujahres ohne Restriktion geschätzt, da sich dieser Effekt in bereits bestehender Literatur recht volatil verhalten kann, siehe Brunauer[1]. So könnten vor dem ersten Weltkrieg gebaute Objekte einen höheren Effekt auf den Kaufpreis haben als beispielsweise Objekte, die während der Kriege errichtet wurden.

- ▷ Zur Modellierung der nichtlinearen Funktion Sanierung wurde ein monoton steigender Verlauf angenommen.

- ▷ Brunauer[1] zeigte, dass mit grösseren Flächen erwartungsgemäß steigende Preise einhergehen, dieser Effekt allerdings für zunehmend grössere Flächen abgeschwächt wird. So ist eine größere Preissteigerung zwischen 400 m^2 und 600 m^2 grossen Grundstücksflächen im Vergleich mit 800 m^2 und 1000 m^2 grossen zu erwarten. Deswegen wurde für die Modellierung von Grundstücks-, Erdgeschoß- und Obergeschoßfläche jeweils eine isotone, konkave Funktion angenommen.

- ▷ Obwohl für die zusätztliche Flächen des Dachgeschosses und Kellers, jeweils unterschieden nach Nutzungsart, ein ähnlicher Verlauf der Funktion wie für die oben beschriebene Flächen wahrscheinlich ist, werden hier lediglich isoton restringierte Funktionen angenommen.

- ▷ Die Variable Lärm wird mit einer antiton restringierten Funktion modelliert. Besonders hohe Lärmquellen wie unmittelbare Nähe zu Eisenbahnstrecken oder stark befahrenen Hauptstrasse sollten einen deutlich erkennbaren negativen Einfluss auf den Preis besitzen.

- ▷ Zur Modellierung der wesentlich lagebeschreibenden Kovariaten Abiturientanteil und Grundpreis (jeweils logarithmiert) wird eine zweidimensionale Funktion spezifiert, wobei für den logarithmierten Grundpreis ein positiver Effekt angenommen wird, die Richtung des Effekts für den logarithmierten Abiturientenanteil bleibt unspezifiziert.

- ▷ Immobilienpreise unterliegen einer starken zeitlich- räumlichen Heterogenität. So werden bespielweise in vielen Regionen Preisanstiege innerhalb einer gewissen Periode beobachtet, während im selben Zeitraum Immobilienpreise in anderen Regionen fallen. Um diesen zeitlich-räumlichen Effekt zu modellieren wird eine zweidimensionale Funktion mit den Kovariaten X- und Y-Koordinate in Interaktion mit dem Kaufjahr spezifiziert.

Basierend auf dem Basismodell wird eine Modellselektion durchgeführt, verschiedene Modelle anhand eines Gütekriteriums verglichen. Das Modell mit dem besten Gütekriterium wird als „finales Modell" bezeichnet.

[1] Brunauer, »Modeling House Prices using Multilevel Structured Additive Regression«.

4.2. Gewichtung der Flächen mittels Sliced Inverse Regression

In diesem Unterkapitel soll die Anzahl der Variablen reduziert werden, ohne dabei relevante Information zu verlieren. Die zur Nutzfläche gehörenden Variablen Erdgeschoß- und Obergeschoßfläche sowie Keller- und Dachgeschoßflächen (jeweils nach Nutzungsart unterschieden) sollen durch eine Variable „gewichtete Nutzfläche" ersetzt werden. Einfaches Aufsummieren dieser Flächen stellt allerdings keine geeignete Methode dar, da den Flächen vermutlich unterschiedliche Bedeutung zukommt. Beispielweise werden 10 Quadratmeter eines Kellers, der nur als Lagermöglichkeit dient, vermutlich weniger wert sein als 10 Quadratmeter Erdgeschoßfläche. Eine geeignetere Methode bieten Dimensionsreduktionsverfahren an.
Dabei soll ein q- dimensionaler Vektor unabhängiger Variablen x auf einen K- dimensionalen Teilraum projiziert werden, um die abhängige Variable y zu beschreiben:

$$y = h(\beta_1 x, \beta_2 x, \ldots, \beta_K x, \epsilon),$$

wobei β_j unbekannte Vektoren beschreibt, der systematische Fehler ϵ unabhängig von x ist und h eine beliebige, unbekannte Funktion darstellt. Die Linearkombination der β_j's wird als effektiv dimensionsreduzierende (EDR) Richtung bezeichnet, der durch die β_j's generierte lineare Teilraum als EDR-Raum. Nach Schätzen der EDR Richtungen können nichtparametrische Methoden aufgrund der geringeren Dimension erfolgreicher sein. In der von Li[2] „Sliced Inverse Regression" werden die EDR Richtungen geschätzt, indem man die Rollen von y und x vertauscht. Dadurch wird dem Dimensionsproblem ausgewichen. $\mathbb{E}(x|y)$ zeichnet mit variierendem y die sogenannte inverse Regressionskurve. Unter schwachen Annahmen kann gezeigt werden, dass die zentrierte inverse Regressionskurve $\mathbb{E}(x|y) - \mathbb{E}(x)$ im linearen Teilraum, aufgespannt durch $\beta_k \Sigma_{xx}$ enthalten ist. Σ_{xx} bezeichnet hier die Kovarianzmatrix von x. Daraus folgerte Li, dass die standardisierte Regressionskurve $\mathbb{E}(z|y)$ mit $z = \Sigma_{xx}^{-1/2}(x - \bar{x})$ im linearen Teilraum, aufgespannt durch die standardisierten EDR Richtungen $\zeta_i, \ldots, \zeta_K$ enthalten ist, sodass sich die Kovarianzmatrix von $\mathbb{E}(z|y)$ in orthogonale Richtungen der $\zeta_k, k = 1, \ldots, K$ zerlegen lässt. Durch Suche der K grössten Eigenwerte und Eigenvektoren von $\mathbb{E}(z|y)$ erhält man die standardisierten EDR Richtungen.
Li schlug folgenden Algorithmus zur Schätzung der EDR Richtungen vor:

1. Standardisiere x durch $\tilde{x_i} = \hat{\Sigma}_{xx}^{-1/2}(x_i - \bar{x})$, wobei $\hat{\Sigma}_{xx}$ und $\bar{x}$ die Stichprobenkovarianz und den Stichprobenmittelwert bezeichnet.

2. Der Wertebereich von y wird in h Intervalle $I_1, \ldots, I_h$ unterteilt. p_h bezeichne die relative Anzahl im Teilbereich I_{p_h}.

3. In jedem Teilbereich wird der Stichprobenmittelwert m_h geschätzt.

[2] Li, »Sliced inverse regression for dimension reduction«.

4. Führe eine gewichtete Hauptkomponentenanalyse aus: Forme die gewichtete Kovarianzmatrix $\hat{V} = \sum_{h=1}^{H} p_h m_h m_h^T$. Finde Eigenwerte und Eigenvektoren von $\hat{V}$.

5. Bezeichne die grössten K Eigenvektoren mit $\zeta_k, k = 1, \ldots, K$ und gib $\hat{\beta}_k = \zeta_k \hat{\Sigma}_{xx}^{-1/2}$ zurück.

Schritte 2 und 3 produzieren aufgrund der Aufteilung in Intervalle eine Schätzung der standardisierten Regressionskurve $\mathbb{E}(z|y)$, daher der Name „Sliced Inverse Regression".

Nachdem hier der Fokus nur auf eine Reduktion der Flächenvariablen gelegt wird, werden anhand des finalen Modells die partiellen Residuen der Fläche definiert:

$$\begin{aligned} \hat{lnp}_i^{flaeche} = lnp_i - X^*\hat{\alpha} - f(bj_i) - f_{\nearrow}(san\_diff_i) - f_{\searrow}(laerm_i) - \\ f_{\nearrow.}(ln\_grundp_i, ln\_abi_i) - kaufjahr_i \cdot f(xco_i, yco_i) \end{aligned} \tag{4.2}$$

Die partiellen Residuen $\hat{lnp}_i^{flaeche}$ sind dadurch bis auf den Einfluss der Flächenvariablen bereinigt und die EDR Richtungen der Flächenvariablen für Erd- und Obergeschoß sowie für Keller- und Dachgeschoße (jeweils unterschieden nach Nutzungsart) können geschätzt werden. Um interpretierbare EDR-Richtungen zu erhalten wird das Gewicht der Erdgeschoßfläche auf den Wert 1 gesetzt.

Durch Linearkombination mit den gefundenen Gewichten kann eine „gewichtete Nutzfläche" generiert und ein Modell geschätzt werden, dass anstatt den verschiedenen Flächenvariablen für Erdgerschoß, Keller, etc. nur noch die gewichtete Nutzfläche als nutzflächenbeschreibende Variable besitzt:

$$\begin{aligned} \mathbb{E}(lnp_i) = X_i^*\alpha + f(bj_i) + f_{\nearrow}(san\_diff_i) + f_{\nearrow}(gst_i) + \\ f_{\nearrow}(nutzflaeche_gewi) + f_{\searrow}(laerm_i) + \\ f_{\nearrow.}(ln\_grundp_i, ln\_abi_i) + kaufjahr_i \cdot f(xco_i, yco_i) \end{aligned} \tag{4.3}$$

4.3. Verwendete Software

Sämtliche Modelle wurden mit der Statistiksoftware R berechnet, Version 2.14. Zur Schätzung des additiven Modells wurde hauptsächlich das Paket von Pya[3] verwendet, Grömping[4] stellt ein Paket zur Lösung quadratischer Optimierungsprobleme mit Ungleichungsnebenbedingungen zur Verfügung, das teilweise auf dem Paket „quadprog" von Berwin A.[5] aufbaut. Diese Pakete führen zu den restringierten Schätzern des parametrischen Parts. Weisberg[6] beinhaltet einen Algorithmus zur Schätzung der EDR - Richtungen in „Sliced Inverse Regression".

Lattice Grafiken wurden mithilfe des Pakets von Sarkar[7] erzeugt, zur Darstellung räumlicher Daten wurden die Pakete von Pebesma und Bivand[8], Lewin-Koh u. a.[9] sowie Bivand, Pebesma und Gomez-Rubio[10] verwendet. Für die Darstellung der Spline-Funktionen in Kapitel 2 diente das Paket „fda" von Ramsay u. a.[11].

[3] Pya, *scam: Shape constrained additive models*.
[4] Grömping, »Inference With Linear Equality And Inequality Constraints Using R: The Package ic.infer«.
[5] Berwin A. *quadprog: Functions to solve Quadratic Programming Problems.*
[6] Weisberg, »Dimension Reduction Regression in R«.
[7] Sarkar, *Lattice: Multivariate Data Visualization with R*.
[8] Pebesma und Bivand, »Classes and methods for spatial data in R«.
[9] Lewin-Koh u. a., *maptools: Tools for reading and handling spatial objects*.
[10] Lewin-Koh u. a., *maptools: Tools for reading and handling spatial objects*.
[11] Ramsay u. a., *fda: Functional Data Analysis*.

5. Resultate

In diesem Kapitel werden die Resultate der in Kapitel 4 beschriebenen Modelle präsentiert. Zunächst werden die nichtparametrisch modellierten Funktionen sowie die Schätzung des linearen Parts für das Basismodell aus Gleichung (4.1) gezeigt. In Gleichung (4.3) wird das Basismodell einerseits mit jenem Modell verglichen, welches aus einer schrittweisen Modellselektion anhand eines Gütekriteriums hervorgeht, andererseits mit einem dimensionsreduzierten Modell, das nur noch die „gewichtete Nutzfläche" als flächenbeschreibende Variable beinhaltet.

5.1. Schätzung

5.1.1. Nichtparametrisch modellierte Kovariaten

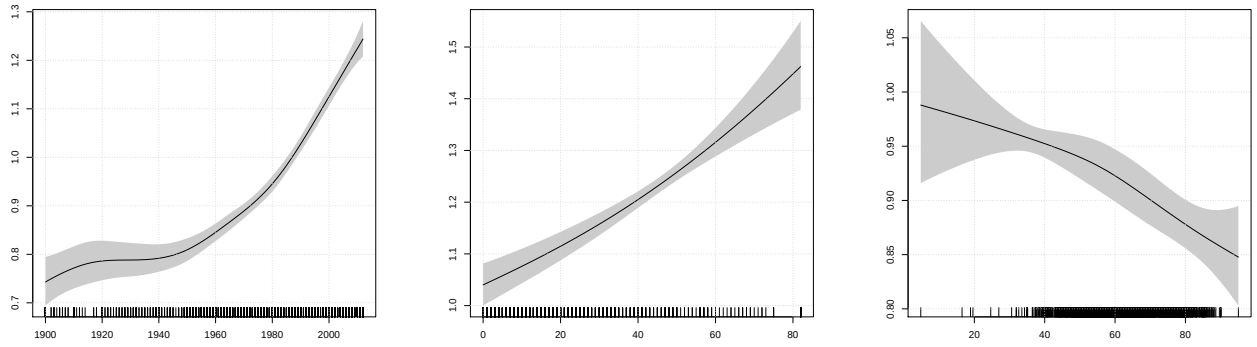
Nachdem die Stärke der Effekte in folgenden Abbildungen recht unterschiedlich ist, werden sie nicht auf der selben Skala gezeigt, aufgrund der einfacheren Interpretierbarkeit auf der exponentiellen Skala.

Abbildung 5.1 zeigt die Effekte der kontinuierlichen Kovariaten Baujahr, Sanierung und Lärm inklusive 95% punktweisen Konfidenzintervall [1]. Alle anderen Variablen werden dabei konstant auf der Referenzgruppe beziehungsweise dem mittlerem Effekt gehalten. Lärm wird hier als objektrelevante Eigenschaft betrachtet, da auch in kleinräumigen Gebieten unterschiedliche Lärmwerte möglich sind, so ist beispielsweise nicht von einem durchschnittlichen Lärm in einem Gebiet (z.B. Gemeinde) auszugehen.

Der Effekt des Baujahres (Panel (a)) ist ziemlich stark, er deckt den Bereich von knapp 70 Prozent bis 130 Prozent ab. So ist bei konstanter Haltung aller übriger Kovariablen ein Anfang des 20. Jahrhunderts errichtetes Haus ca. 70% von einem durchschnittlich erbautem Haus im Jahr 1982 wert, während neu errichtete Häuser 1.3 mal so viel wert sind wie durchschnittlich errichtete Objekte. Der positive Effekt ist besonders nach Ende des zweiten Weltkriegs zu erkennen, während sich davor errichtete Objekte auf ähnlichen Preisniveaus befinden.

Der Effekt der Sanierung verläuft annähernd linear, der Preisanstieg von Immobilien, die 20 Jahre nach dem Baujahr saniert wurden, beträgt sind knapp 10 Prozent. Objekte, die 80 Jahre nach dem Baujahr saniert wurden, sind um den Faktor 1.4 mehr wert,

[1] Auf eine Beschreibung zur Erzeugung der punktweisen Konfidenzintervalle sei auf Pya, »Additive models with shape constraints« verwiesen.



(a) Baujahr (b) Sanierung (c) Lärm

Abbildung 5.1.: Nichtlineare Plots der Effekte für Alter (Baujahr und Sanierung) sowie Lärm

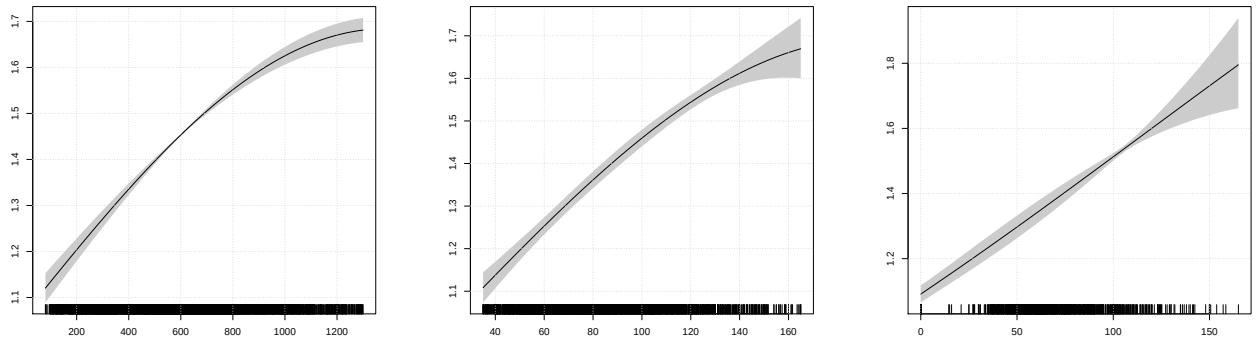
allerdings ist die Alterswertminderung in Betracht zu ziehen. Tabelle 5.1 gibt die relativen Wertveränderungen in Abhängigkeit von jeweils 5 Ausprägungen des Baujahres sowie der Sanierungsdifferenz an. Als Referenz dienen dabei unsanierte Häuser, die 1982 errichtet wurden. Unsanierte Objekte, die 1920 errichtet wurden verringen sich um den Faktor 0.818, während eine Sanierung um 60 Jahre, sodass 1980 als das fiktive Baujahr angesehen wird, den Wert des Hauses sogar leicht steigen lässt. Hingegen haben 1950 gebaute und 1990 sanierte Objekte den Faktor knapp kleiner 1. 1982 gebaute und 20 Jahre später sanierte Häuser haben den gleichen Effekt wie unsanierte Objekte mit Baujahr 1990.

Panel (c) beschreibt den Effekt des Umgebungslärms. Das relativ breite Konfidenzintervall für Lärmwerte unter 35 Dezibel entsteht aufgrund der kleinen Beobachtungszahl in diesem Intervall. Große Lärmquellen vermindern den Wert um etwas mehr als 15% im Vergleich zu Objekten, die fast keinem Lärm ausgesetzt sind, die relative Differenz zu durchschnittlich Lärm ausgesetzten Objekten (mit ca. 62 Dezibel) beträgt knapp 10 Prozent.

	0	20	40	60	80
1920	0.818	0.876	0.947	1.034	1.138
1950	0.842	0.902	0.975	1.065	.
1982	0.998	1.07	.	.	.
1990	1.069	1.145	.	.	.
2012	1.294	.	.	.	.

Tabelle 5.1.: Vergleich der Effekte für Baujahr (Zeile) und Sanierungsdifferenz (Spalte)

Abbildung 5.2 zeigt die geschätzten Effekte der Grundstücksfläche sowie jene des Erdgeschoß- und Obergeschoßes. Die Spanne des Effekt der Grundstücksfläche (Panel (a)) beträgt knapp 170 Prozent. So beträgt die Wertsteigerung für Grundstücksflächen



(a) Grundstücksfläche (b) Erdgeschossfläche (c) Obergeschossfläche

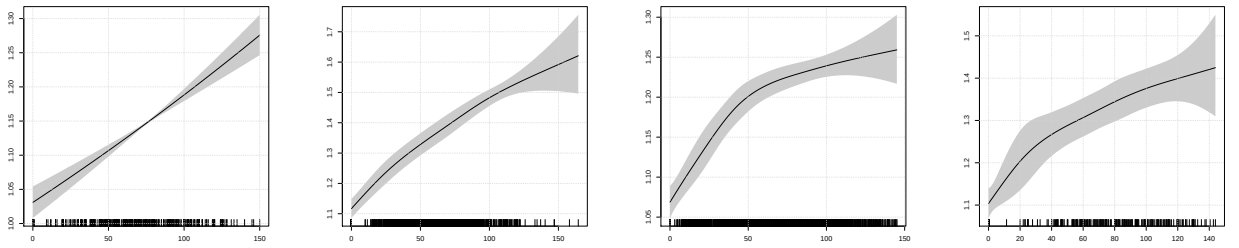
Abbildung 5.2.: Nichtlineare Plots der wichtigsten Flächenvariablen

über 1200 Quadratmeter 170 Prozent im Vergleich zu Objekten mit nur geringer Grundstücksfläche. Im Vergleich mit durchschnittlich großen Grundstücken (mit ca. 630 Quadratmeter) beträgt die Wertsteigerung ca. 30 Prozent. Nachdem für den nichtlinearen Effekt der Grundstücksfläche eine isoton und konkave Funktion angenommen wurde schwächt sich der Effekt der Preissteigerung mit zunehmenden Grundstücksflächen.
Dieser Effekt ist auch für Erdgeschoßflächen zu erkennen (Panel (b)). Objekte mit 60 Quadratmeter Erdgeschoßfläche sind um ca. 10 Prozent weniger wert als Objekte mit durchschnittlich großen Erdgeschoßflächen (ca. $80m^2$), während die Preissteigerung für ein Haus mit 100 Quadratmeter Erdgeschoßfläche im Vergleich zu durchschnittlich großen Erdgeschoßflächen knappe 5 Prozent beträgt.

Die Stärke des Effekts der Obergeschoßfläche (Panel (c)) ist noch ausgeprägter als jener der Erdgeschoßfläche. Einer der Gründe hierfür könnte sein, dass sich auch relativ viele Häuser ohne Obergeschoß im Datensatz befinden und somit der Wertebereich eine größere Spanne besitzt. Anhand der Abbildungen ist eine Preissteigerung von knapp 20% im Vergleich zwischen $50m^2$ und $100m^2$ Obergeschoßflächen abzulesen, jene für Erdgeschoßflächen beträgt 25%.

Abbildung 5.3 beschreibt die Effekte der weiteren Flächen. Der Verlauf der Funktionen ist ähnlich, einzig der Effekt der Dachgeschoßfläche für Lagerzwecke ist annähernd linear. Dafür unterscheiden sich die Stärken der Effekte. Während zum Lagerzweck genutzte Dachgeschoß- und Kellerflächen den Wert der Immobilie um maximal nur 30 Prozent steigern, ist der Verlauf und die Größe des Effekts von Dachgeschoßen, die zum Wohnen genutzt werden vergleichbar mit jenen der Erdgeschoßfläche. Auch zum Wohnzweck genutzte Kellerflächen besitzen ein höheres Niveau.

Die nichtparametrische Schätzung der zweidimensionalen Funktion mit den Kova-



(a) Dachgeschossfläche (Lagerzwecke)

(b) Dachgeschossfläche (Wohnzwecke)

(c) Keller (Lagerzwecke)

(d) Keller (Wohnzwecke)

Abbildung 5.3.: Nichtlineare Plots der weiteren Flächenvariablen

riaten ln_abi sowie ln_grundp wird in Abbildung 5.4 gezeigt. Der (logarithmierte) Grundpreis beeinflusst die Preise erwartungsgemäß deutlicher als der (logarithmierte) Akademikeranteil, allerdings ist eine deutliche Preissteigerung in Gebieten, wo sowohl Grundpreis als auch Akademikeranteil hohe Werte besitzen, erkennbar. Ebenfalls kann abgelesen werden, dass der Abiturientenanteil in Regionen mit hohen Grundpreisen ziemlich stark ist. Die Schätzung wird dabei nicht extrapoliert, sodass kein Effekt von in den Daten nicht vorhandenen Kombinationen von Abiturientenanteil und Grundpreis vorliegt. Beispielsweise ist keine Region zu finden, in denen der Abiturientenanteil sehr hoch, die Grundpreise wiederum sehr gering sind.

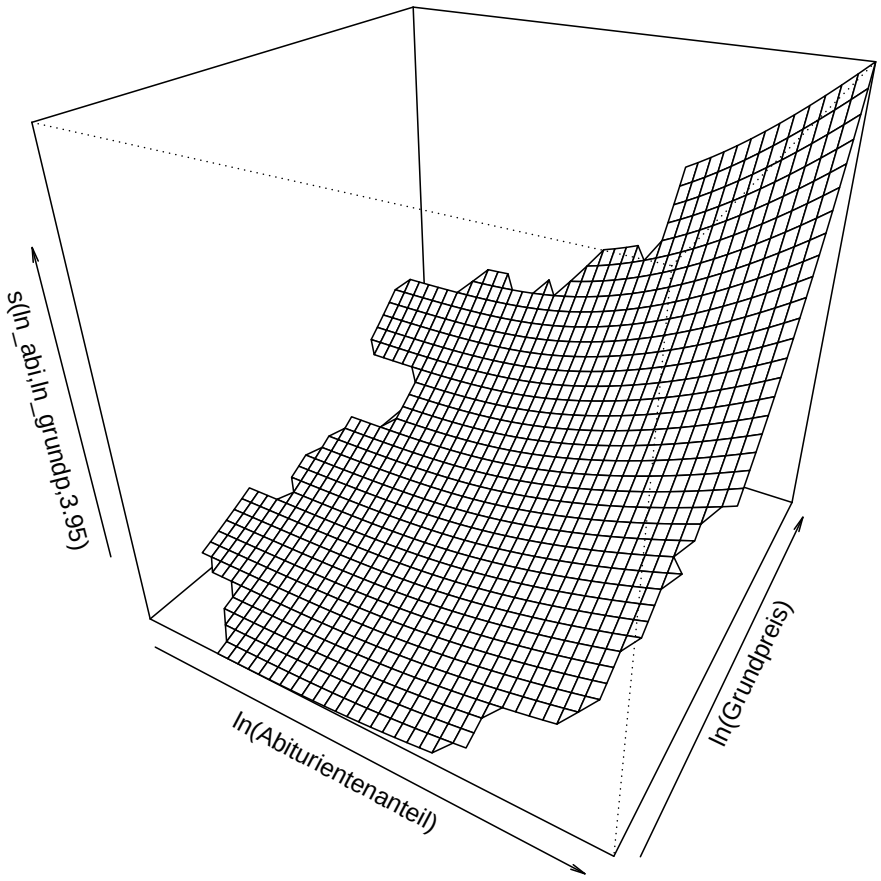


Abbildung 5.4.: Zweidimensionaler Plot der lagebeschreibenden Variablen. 3.95 gibt die Anzahl der effektiv verbrauchten Freiheitsgrade für diese zweidimensionale Funktion an.

5.1.2. Schätzung des parametrischen Teils

Tabelle 5.2 enthält die erwarteten Effekte sowie Schätzungen des üblichen parametrischen Parts. Unrestringierte Schätzer geben jene Schätzwerte an, falls der erwartete Effekt unberücksichtigt bleibt, restringierte Schätzer sind Lösungen des in Gleichung (2.45) vorgestellten quadratischen Optimierungsproblems.

Mit * markierte Kovariaten weisen auf Unterschiede auf dem 5% Signifikanzniveau hin, ** auf dem 1% Signifikanzniveau. Interpretationsmäßig verändert sich der Preis multiplikativ mit dem exponierten Parameter im Vergleich zur Referenzgruppe. So ändern sich Objekte am Stadt- / Ortsrand bei konstanter Haltung aller anderen Kovariaten um den Faktor $exp(-0.033) = 0.967$ im Vergleich zu Häuser im Zentrum, während Villengegenden den Wert der Immobilie um den Faktor $exp(0.081) = 1.084$ erhöhen.

Beinahe alle unrestringierten Schätzer verhalten sich gemäß der Erwartungen bis auf wenige Ausnahmen:

- ▷ Für die Existenz einer Garage wird ein (leicht) negativer Wert prognostiziert, welcher nicht unserer Erwartung entspricht.
- ▷ Einfamilienhäuser in sehr schlechtem Zustand werden um knapp 3% besser bewertet als Objekte, die nur einen schlechten Zustand aufweisen. Dieser unerwartete Effekt begründet sich möglicherweise in der selten beobachteten Ausprägung „schlechter Zustand" (nur 10 Objekte).
- ▷ Ein Preisanstieg von Immobilien mit ostseitig orientierten Wohnräumen ist entgegen unserer Vorstellung im Vergleich mit Immobilien mit südseitig gelagerten Wohnräumen zu erkennen.

Allerdings ist keiner jener Parameter, die unseren Erwartungen widersprechen, signifikant, weder auf dem 1% noch auf dem 5% Signifikanzniveau. Nachdem sich beinahe alle unrestringierten Schätzer erwartungsgemäß verhalten, sind nicht allzu große Unterschiede zwischen unrestringierten und restringierten Schätzungen zu erkennen.

Das wesentlich wertbeinflussende Merkmal stammt vom Zustand des Hauses, Objekte in sehr gutem Zustand sind in etwa um 30% mehr wert als Objekte in sehr schlechtem Zustand. Weitere wesentlich wertsteigernde Merkmale ist ein guter Zustand des Bades (+2.7%) inklusive Fußbodenheizung (+2.9%), die Existenz einer Terrasse (+3.8%) und eines Swimmingpools (+4.7%) sowie eines elektrischen Garagentores (+4.2%) oder Ausstattung mit hochwertigen Fenster oder Türen (+6.2%). Andererseits bestehen besonders wertmindernde Eigenschaften in der Nutzung eines kleines Badezimmers (-5.2%), die Existenz lediglich einer Individualheizung anstatt einer Zentralheizung (-8%) sowie eine schlechte Aufteilung der Räume (-5.1%).

	Bezeichnung	Eff	unrestringiert	restringiert
Makrolage	am Stadt- / Ortsrand **	o	-0,033	-0,033
	Villengegend **	o	0,08	0,081
	periphere Lage **	o	-0,06	-0,059
Grösse des Badezimmers	Grösse des Bads unter 4m² **	-	-0,054	-0,053
	Grösse des Bads über 4m²	+	0,001	0,001
Zustand des Badezimmers	sehr gut	+	0,026	0,027
	schlecht **	-	-0,037	-0,037
Typ des Badezimmers	nur Dusche **	o	-0,037	-0,037
	beides	+	0,016	0,017
	keine Duschgelegenheit	-	-0,034	-0,034
Zustand	gut - sehr gut *	-	-0,041	-0,036
	mittelmäßig- gut *	-	-0,052	-0,057
	schlecht - mittelmäßig **	-	-0,197	-0,17
	sehr schlecht - schlecht	-	0,027	0
Orientierung der Wohnräume	nach Norden *	-	-0,025	-0,026
	nach Osten	-	0,036	0
	nach Westen	o	0,034	0,029
Balkon	mittel	+	0,009	0,009
	groß	++	0,014	0,014
Heizung	Individualheizung **	-	-0,075	-0,075
Wandaufbau	einfache Holzkonstruktion	o	-0,01	-0,009
	hochwertige Holzkonstruktion	o	0,044	0,048
	andere Bauweise	o	-0,117	-0,107
weitere Ausstattungsmerkmale	Garage vorhanden	+	-0,016	0
	Terrasse vorhanden ***	+	0,038	0,038
	Fußbodenheizung im Bad **	+	0,028	0,029
	Bad verfließt	+	0,002	0,011
	nur Heizstrahler in Bad	-	-0,03	-0,03
	Fenster im WC	o	0,061	0,061
	elektrisches Garagentor **	+	0,042	0,035
	eingeschränkte Zufahrt	-	-0,019	-0,024
	Fußbodenheizung **	+	0,036	0,035
	Wandheizung	+	0,025	0,025
	Alarmanlage **	+	0,033	0,033
	hochwertige Fenster / Türen **	+	0,06	0,061
	hochwertiger Gartenzaun **	+	0,034	0,035
	Swimmingpool **	+	0,047	0,047
	schlechte Raumaufteilung **	-	-0,054	-0,053
	Fertigteilhaus	o	-0,001	-0,001
	Reihenhaus **	o	0,036	0,037
	Angebotspreis **	o	0,084	0,085
	Bau von gemeinnützigem Bauträger **	o	-0,123	-0,124

Tabelle 5.2.: Schätzung des parametrischen Parts inklusive erwartetem Effekt. Unrestringierte Schätzwerte werden mit restringierten verglichen. ** markierte Variablen deuten Signifikanz auf dem 1% Niveau an, * markierte auf dem 5% Signifikanzniveau

5.2. Vergleich der Modelle

Anhand des Basismodells (siehe Gleichung (4.1)) wurde eine schrittweise Modellselektionsprozedur durchgeführt, die in jedem Durchlauf jene Kovariate entfernt, die den höchsten P-Wert aufweist. Schlussendlich wurde jenes Modell als das finale spezifiziert, welches das Kriterium der generalisierten Kreuzvalidierung (GCV - Score, siehe Gleichung (2.40)) minimiert. Dabei stellte sich heraus, dass die Kovariaten „Wandaufbau", „Größe des Balkons" und die dummycodierten Kovariaten „Garage vorhanden", „Bad verfließt", „nur Heizstrahler im Bad", „eingeschränkte Zufahrt zur Garage", „Wandheizung", und „Fertigteilhaus" keinen quantifizierbaren Einfluss auf den Preis eines Einfamilienhauses besitzen und werden deshalb im finalen Modell nicht berücksichtigt.

	GCV	AIC	Devianz	EDF
Basismodell	0.0409	-1134.1	116.1	155.6
finales Modell	0.0406	-1153.6	116.1	145.9
flächengewichtetes Modell	0.0395	-1244.5	113.1	141.1

Tabelle 5.3.: Vergleich ausgewählter Modelle anhand verschiedenen Gütekriterien. EDF bezeichnet die effektiv verbrauchten Freiheitsgrade.

Tabelle 5.3 vergleicht relevante Kennzahlen für die Modellgüte. Dabei ist zu erkennen, dass die Devianz des finalen Modells jener des Basismodells gleicht, allerdings werden knapp 10 Freiheitsgrade weniger verbraucht, das AIC-Gütekriterium[2] sinkt um knapp 20 Einheiten. Die Spalte EDF bezeichnet τ, die Anzahl der effektiv verbrauchten Freiheitsgrade (siehe Gleichung (2.3.3)).

5.2.1. Flächengewichtung mittels Sliced Inverse Regression

	Gewicht
Erdgeschoß	1.00
Obergeschoß	1.12
Dachgeschoß (Wohnzweck)	1.02
Dachgeschoß (Lagerzweck)	0.43
Keller (Wohnzweck)	0.72
Keller (Lagerzweck)	0.46

Tabelle 5.4.: Hauptrichtung der Flächengewichtung. Weitere Richtungen sind nicht angegeben, da in folgenden Analysen hierfür kein quantifizierbarer Einfluss erkennbar war.

[2] Akaike, »A new look at the statistical model identification«.

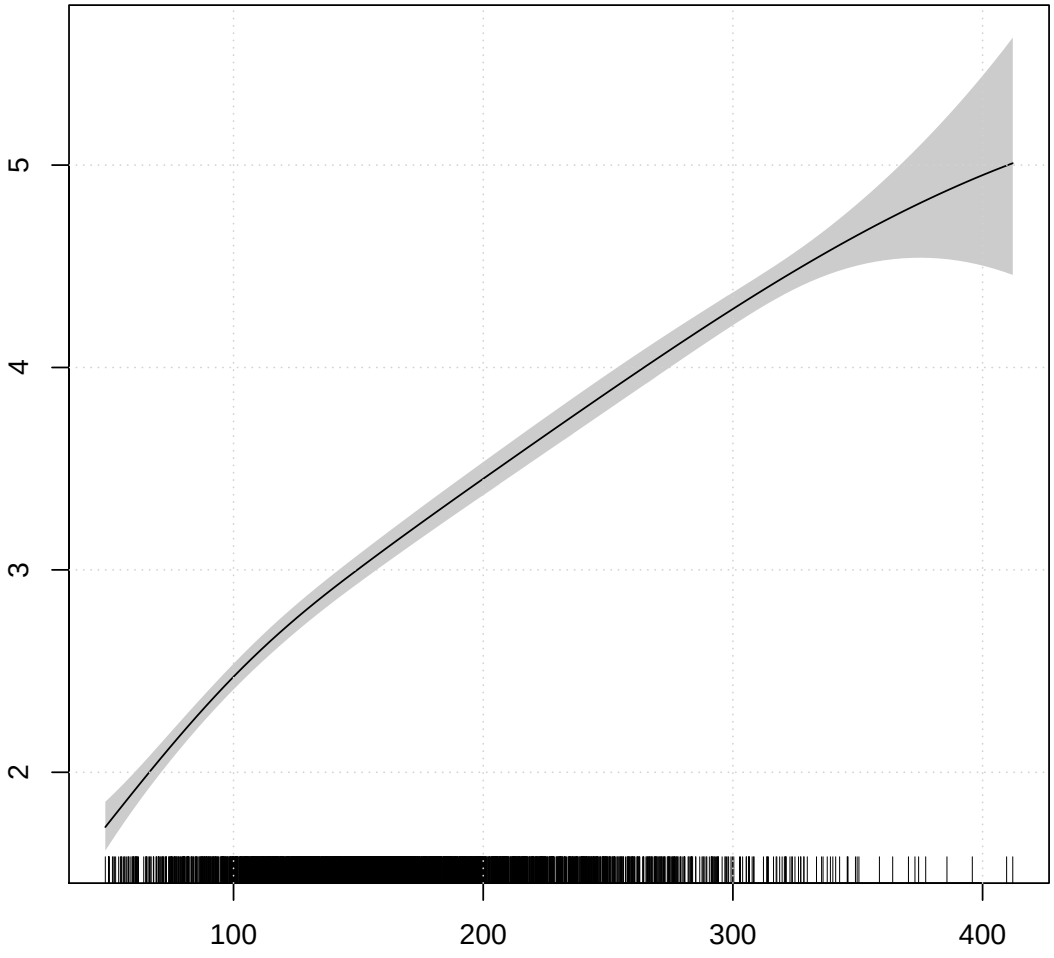


Abbildung 5.5.: Nichtlinearer Effekt der gewichteten Nutzfläche

	Mittelwert	Standardabweichung	Minimum	Maximum
gewichtete Nutzfläche	167.78	55.80	48.65	412.09

Tabelle 5.5.: Deskritpive Statistik zur gewichteten Nutzfläche

Tabelle 5.4 gibt die Hauptrichtung der in Kapitel 4.2 beschriebenen Flächengewichtung an. Dabei ist zu erkennen, dass Obergeschoße und zum Wohnen genutzte Dachgeschoße sogar leicht höhere Gewichte besitzen wie das Erdgeschoß. Demnach hat ein $112m^2$ großes Erdgeschoss dasselbe Gewicht wie ein 100m^2 großes Obergeschoß.
Der Effekt eines als Wohnraum genutzten Kellers beträgt 72% von jenem des Erdgeschoßes. Als Lagerflächen dienende Keller und Dachgeschoße besitzen nicht einmal 50% des Gewichts von der Erdgeschoßfläche. Beispielsweise erwirkt eine Erhöhung um $20m^2$ von nicht zum Wohnen benutzten Keller weniger als eine Erhöhung der Erdgeschoßfläche um $10m^2$.

Multiplizieren dieser geschätzten Gewichte mit den beobachteten Daten und anschließendem Addieren ergibt die „gewichtete Nutzfläche", eine deskriptive Statistik ist in Tabelle 5.5 zu finden. Zusätzliche Richtungen wurden zwar berechnet, in 5.4 allerdings nicht angegeben, da bereits die erste Richtung die Daten ausreichend gut erklärt, sodass Modelle mit mehr als einer Richtung schlechtere Gütekriterien aufweisen. Für das in Gleichung (4.3) spezifizierte Modell ist eine weitere Reduktion des Gütekriteriums der generalisierten Kreuzvalidierung zu erkennen, auch das AIC verringert sich markant (siehe Tabelle 5.3).

5.2.2. Identifizierung der wertbeeinflussendsten Merkmale

In diesem Unterkapitel sollen jene 5 Merkmale einer Immobilie angegeben werden, die ihren Wert am wesentlichsten beeinflusst. Dafür wurde wiederum eine schrittweise Modellselektion durchgeführt, das erste Modell inkludiert nur den Intercept. Danach wurde für jeden Term des Basismodells in Gleichung (4.1) einzeln ein Modell geschätzt und schlussendlich jenes genommen, welches das Kriterium der generalisierten Kreuzvalidierung (GCV; siehe Gleichung (2.40))[3] minimiert. Dieses Modell gibt somit den wertbeeinflussensten Term an.
Um ein zusätzlich wichtiges Merkmal zu identifizieren wurden alle verbleibenden Terme einzeln in das Modell aufgenommen und jenes Modell spezifiziert, welches das Kriterium der generalisierten Kreuzvalidierung minimiert. Diese Logik wurde schrittweise fortgesetzt, bis die fünf wesentlich wertbeinflussenden Kriterien gefunden wurden. Diese werden in Tabelle 5.6 inklusive verschiedener Gütekriterien dargestellt.

Der wichtigste Term zur Erklärung des Preises besteht in der zweidimensionalen Funktion mit den lageerklärenden Kovariaten „logarithmierter Grundpreis" sowie „logarith-

[3]Hastie und Tibshirani, *Generalized Additive Models*.

	erklärende Variable	GCV	AIC	Devianz	EDF
1	nur Intercept	0.2771	4886.5	870.2	1
2	.. + f(ln(Abiturientenanteil), ln(Grundpreis))	0.1773	3483	554.3	7.8
3	.. + f(Baujahr)	0.1243	2366.7	385.2	21.2
4	.. + f(Erdgeschossfläche)	0.095	1524.1	293.8	25.5
5	.. + Zustand	0.0814	1037	251.1	28.6
6	.. + f(Grundstücksfläche)	0.0714	625	219.9	30.8

Tabelle 5.6.: Einfluss der wichtigsten Merkmale basierend auf einer schrittweisen Modellselektion. Modellgüte wird anhand generalisierter Kreuzvalidierung (GCV), Akaika- Kriterium (AIC) sowie der Devianz gemessen. EDF bezeichnet die Anzahl der effektiv verwendeten Freiheitsgrade.

mierter Abiturientenanteil". Der Wert der generalisierten Kreuzvalidierung sinkt auf 0.178, ein nur die Konstante inkludierendes Modell hat den wesentlich höheren Score von 0.279. Zusätzlich zu diesem lagebeschreibenden Term wurde als zweit wichtigster Term der Effekt des Baujahres beobachtet, danach bietet die Fläche des Erdgeschoßes die wichtigste preisbeschreibende Kovariate. Viert wichtigster Term zur Erklärung des Preises besteht im Zustand des Hauses, gefolgt von der nichtparametrischen Schätzung der Grundstücksfläche. Unter Verwendung dieser fünf Terme beträgt der GCV- Score etwa 0.716, jener des finalen Modells liegt bei 0.406.

5.2.3. Österreich - Simulation

Um den Effekt der lagerelevanten Kovariaten zu verdeutlichen, werden Preise desselben Objekts über ganz Österreich simuliert. Dabei werden nur die Werte der lagebeschreibenden Kovariablen Grundpreis und Akademikeranteil je nach Lage verändert, als X- und Y-Koordinate dient der Zentroid jeder Gemeinde. 2010 wurde als Baujahr spezifiziert, die Größe des Erd- sowie des Obergeschoßes wurde jeweils $80m^2$ gesetzt, Keller sowie Dachgeschoß sind nicht vorhanden. Die Simulation wurde einmal mit 2008 als Kaufjahr durchgeführt, ein anderes mal mit 2012, um zeitlich-räumliche Wertveränderungen darzustellen. Übrige (objektrelevante) Kovariaten wurden auf Referenzwerte gesetzt. Abbildung 5.6 beinhaltet das Ergebnis der Simulation.

Ein ziemlich differenziertes Bild über Österreich ist zu erkennen. Deutliche Unterschiede gibt es zwischen West- und Ostösterreich. Die höchsten Preise sind in Wien zu finden, sofern es in den Bezirken überhaupt Einfamilienhäuser gibt. So erreicht im das simulierte Haus im 19. Wiener Gemeindebezirk Werte knapp über 900.000€. Unmittelbar an Wien grenzende Gemeinden besitzen ebenfalls ein relativ hohes Preisniveau, während das Objekt im restlichen Niederösterreich und Burgenland höchstens knapp 300.000€ wert ist. Oberösterreich, Kärnten sowie die Steiermark besitzen ein vergleichbares Preisniveau, hier ist das Objekt zwischen 150.000€ und 250.000€ wert, nur innerhalb und knapp au-

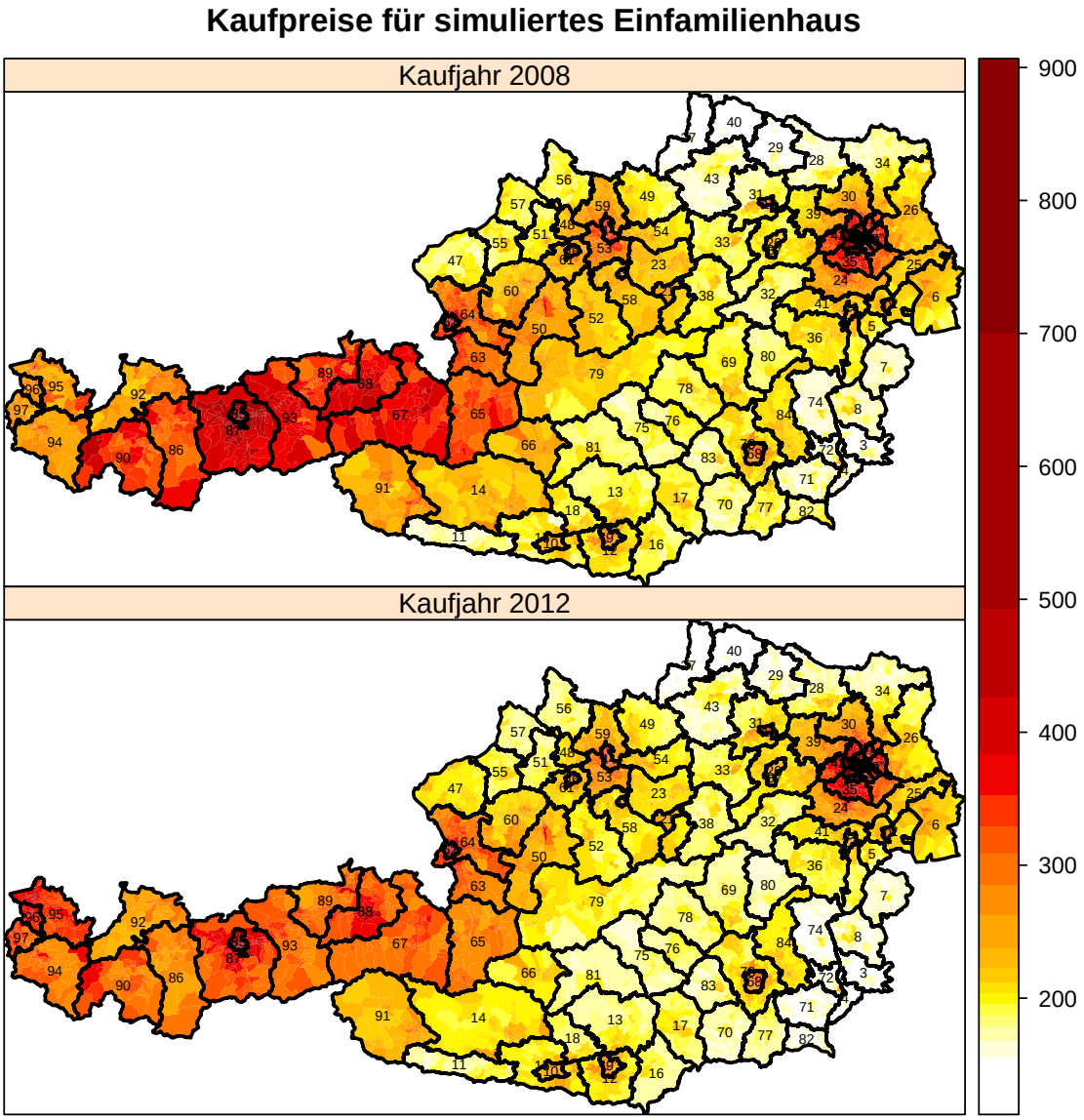


Abbildung 5.6.: Österreich -Simulation eines Hauses mit denselben Objekteigenschaften: Baujahr 2010, Erd- und Obergeschoßfläche: $80m^2$. Die Simulation wurde einmal mit 2008 als Kaufjahr durchgeführt, ein anderes mal mit 2012, um zeitlich-räumliche Veränderungen wiederzugeben. Angaben in Tausend Euro, eine Auflistung der Bezirke befindet sich im Anhang.

ßerhalb der Landeshauptstädte Linz, Klagenfurt und Graz steigt das Preisniveau an, so ist das Haus in bester Linzer Lage etwas mehr als 400.000€wert. Westösterreich besitzt allgemein ein deutlich höheres Preisniveau, hier kostet die Immobilie fast das doppelte wie ein im Burgenland angesiedeltes Objekt. Als besonders teure Gegenden sollen hier Salzburg, Kitzbühel, Innsbruck sowie Bregenz genannt werden, so ist die Immobilie in diesen Gebieten überall mindestens 350.000€ wert, Spitzenwerte liegen bei knapp über 500.000€.
Eine Wertminderung in Westösterreich von 2008 und 2012 fällt auf, hier muss allerdings die recht sparsame Beobachtungsdichte mitberücksichtigt werden.

5.3. Güte des Modells

In diesem Kapitel soll die Güte der Anpassung des finalen Modells beschrieben werden. Erhaltene Schätzungen für die Kaufpreise vergleichen wir mit den tatsächlichen und geben weitere Verbesserungsvorschläge.

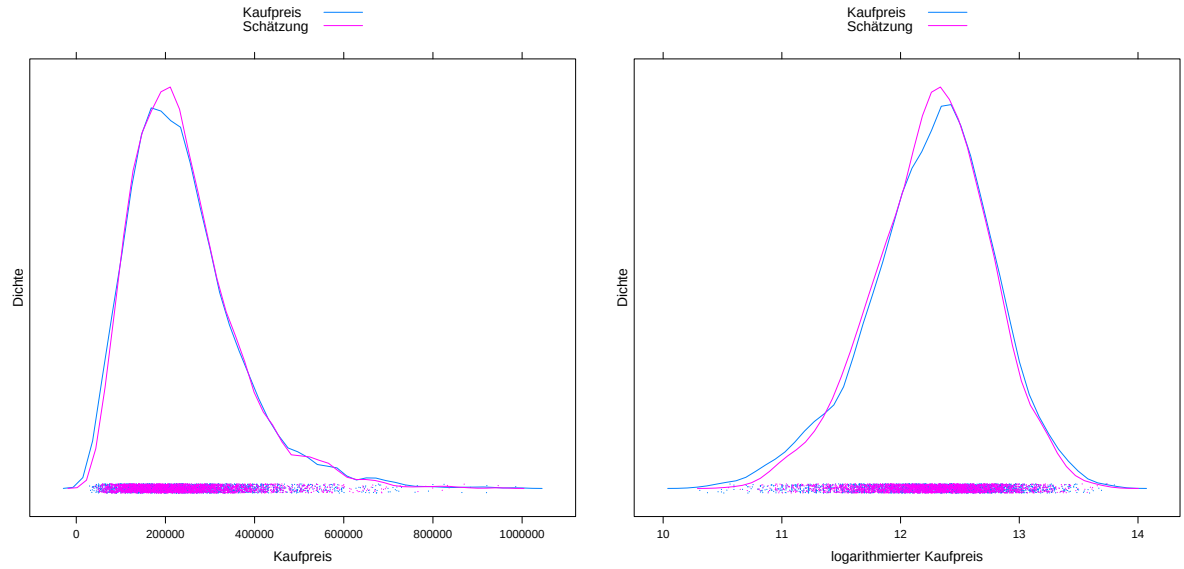
Tabelle 5.7 zeigt die Perzentile der tatsächlichen Kaufpreise sowie jene für die geschätzten Kaufpreise. Die dazugehörigen Kerndichteschätzer sind in Abbildung 5.7(a) zu finden, während Panel (b) Kerndichteschätzer für den Regressanden „logarithmierter Kaufpreis" zeigt. Hier ist zu erkennen, dass der Median bei knapp 220.000€ liegt, jeweils 50% der beobachteten Objekte sind über sowie über diesem Wert. Der Median für die Verteilung der geschätzten Kaufpreise ist annähernd ident. Allerdings steigt dieser Unterschied desto mehr man sich vom Median entfernt. So sind Schätzungen für unterdurchschnittlich teure Immobilien höher als tatsächliche Kaufpreise, für überdurchschnittlich teure Objekte ist ein umgekehrter Effekt zu beobachten. Beispielweise wird das Objekt mit dem niedrigsten Kaufpreis (30.000 €) um knapp 9.000€ zu hoch geschätzt, die Schätzung des teuersten Objektes (ca. 986.000€) liegt ungefähr 40.000€ darunter. Eine Möglichkeit extreme Quantile vorherzusagen bietet „Quantils-Regression"[4]. Ein bayesianischer Zugang wird in Yue und Rue[5] beschrieben.

Abbildung 5.8 zeigt einen auf Bundesland aggregierten Vergleich zwischen mittleren Kaufpreisen sowie den mittleren Schätzungen. Dabei fallen abermals die teuren Regionen Wien, Salzburg sowie Tirol auf, vorhergesagte Preise stimmen mit den tatsächlichen Kaufpreisen annähernd überein. Nur Niederösterreich und Vorarlberg werden erkennbar höher geschätzt, Wien und Tirol etwas unter den realisierten Kaufpreisen.

Abbildung 5.9 zeigt relevante Grafiken zur Modellanalyse.

- ▷ Panel (a) vergleicht den Regressanden „logarithmierter Kaufpreis" mit den dazugehörigen Schätzungen. Kaufpreise, die um mehr als 15% überschätzt werden, sind dabei in orange dargestellt, während rot Objekte angibt, deren Schätzungen um mindestens

[4]Koenker und Hallock, »Quantile Regression«.
[5]Yue und Rue, »Bayesian inference for structured additive quantile regression models«.



(a) Kaufpreis (b) logarithmierter Kaufpreise

Abbildung 5.7.: Kerndichteschätzer für die Verteilung des Kaufpreises (links) sowie des logarithmierten Kaufpreises (rechts). Kerndichteschätzer tatsächlicher Kaufpreise werden in blau dargestellt, die dazugehörige Schätzung in pink.

15% zu niedrig sind. Dabei fällt eine leichte Überschätzung besonders in dem Intervall zwischen 11 und 12 (also Kaufpreise zwischen 60.000€ und 160.000€) auf, Objekte werden deutlich seltener unterschätzt, am häufigsten tritt dieser Effekt in Regionen mit hohen Preisniveaus auf.

- ▷ In Panel (b) ist ein Histogramm der Residuen zu erkennen. Residuen sind annähernd normalverteilt.
- ▷ Die Residuen in Panel (c) sind ziemlich gleichmäßig um den Wert 0 verteilt, allerdings fällt die größere Spanne der negativen Residuen auf, wodurch Immobilien öfter und deutlicher überschätzt als unterschätzt werden. Andere nennenswerte Strukturen der Residuen sind anhand Panel (c) nicht zu erkennen.

Tabelle 5.8 enthält den Vergleich zwischen den tatsächlichen Kaufpreisen sowie deren Schätzungen. Die relative Differenz wird gemäß der Formel

$$rel\_diff = \frac{vorhergesagter Kaufpreis - Kaufpreis}{Kaufpreis}$$

ermittelt. Der Kaufpreis von 746 Immobilien wird einigermaßen richtig (absolute relative Differenz unter 5%) prognostiziert, weitere 1260 Objekte weisen höchstens eine relative Differenz von 15% auf. Nur 56 Objekte werden um mehr als 30% unterschätzt, allerdings 270 Objekte um mehr als 30% überschätzt. Für weitere 426 Immobilien beträgt die relative Differenz mehr als 15%.

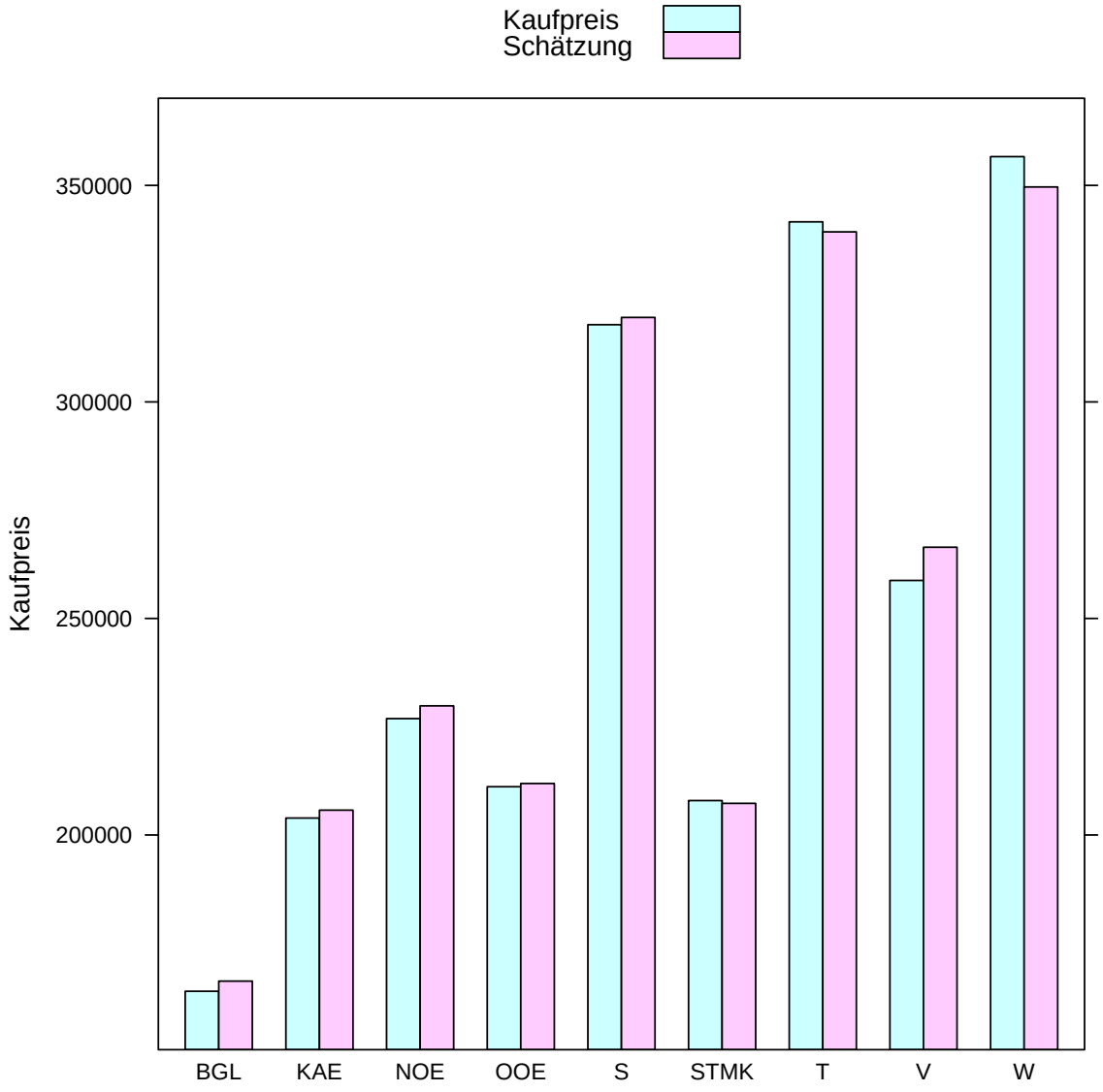


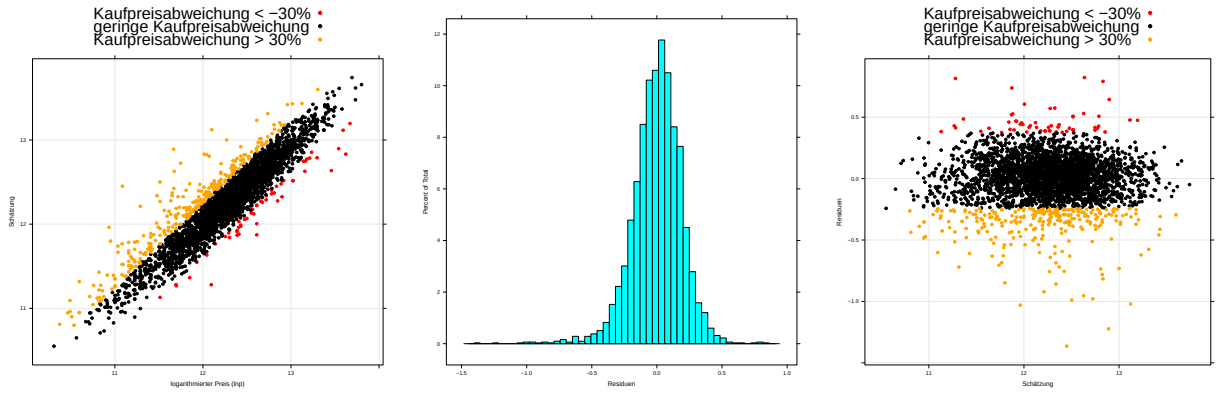
Abbildung 5.8.: Bundeslandaggregierter Vergleich der Kaufpreise mit der Schätzung

	Kaufpreis	vorhergesagter Kaufpreis
0%	30000	38959
10%	106527	112201
20%	140000	143464
30%	166744	171267
40%	192000	196502
50%	220082	221053
60%	249000	247082
70%	280068	280165
80%	323497	323603
90%	395999	390065
100%	985759	946453

Tabelle 5.7.: Vergleich der Dezile zwischen Kaufpreisen und Modellschätzung

	Anzahl
< -30%	56
-15% – -30%	394
-15% – -5%	675
perfekt	746
5% – 15%	585
15% – 30%	426
> 30%	270

Tabelle 5.8.: Relative Abweichung des geschätzten Kaufpreises zum tatsächlichen



(a) (b) (c)

Abbildung 5.9.: Modellanalyse: Vergleich des Regressanden mit der Schätzung (Panel (a)), Panel (b) zeigt ein Histogramm der Residuen, Panel (c) plottet den linearen Prädiktor gegen die Residuen . Extrem vom Kaufpreis abweichende Schätzungen werden farblich markiert.

6. Zusammenfassung

In dieser Arbeit wurden Immobilienpreise für Einfamilienhäuser in Österreich untersucht, als Datengrundlage dienten 3152 über ganz Österreich verteilte Objekte.
Zu diesem Zwecke wurden zunächst umfangreiche deskriptive sowie grafische Analysen getätigt, anschließend wurde versucht, Kaufpreise mithilfe eines additiven Modells vorherzusagen. Möglicherweise nichtlineare Effekte wurden anhand nichtparametrischer Methoden modelliert, teilweise unter einer speziellen Bedingung an die Monotonie und Krümmung der Funktion. Auch Koeffizienten des parametrischen Teils wurden so restringiert, dass sie mit unseren Erwartungen übereinstimmen.

Die wichtigsten erklärenden Merkmale der Immobilienpreise bestehen in der Lage des Objekts - in dieser Arbeit durch den (logarithmierten) Grundpreis und Abiturientenanteil beschrieben - dem Baujahr, der Fläche der Erdgeschoßes und des Grundstücks sowie dem Zustand des Objekts. Wien stellt generell die teuerste Umgebung dar, auch in Westösterreich angesiedelte Städte wie Salzburg, Innsbruck, Kitzbühel oder Bregenz besitzen ein hohes Preisniveau.

Einige Punkte für zukünftige Forschung sollen diskutiert werden:

- Kaufpreise werden bundesland-aggregiert relativ genau geschätzt, allerdings zeigte sich, dass die Schätzung einzelner Objekte teilweise recht danebenliegt. So werden Objekte mit hohem Kaufpreis tendenziell niedriger geschätzt, während Objekte mit geringem Kaufpreis oftmals zu hoch eingestuft wurden, weshalb eine Schätzung mithilfe Quantilsregression vorgeschlagen wird.
- Eine weitere Verbesserungsmöglichkeit besteht in der Verwendung von mehreren, kleinräumigen, lagebeschreibenden Variablen. Die Wahl von (gemeindeaggregierten) Grundstückspreisen erscheint aufgrund der Volatilität innerhalb einer Gemeinde als ungeeignet, stattdessen könnten Preise durch weitere Variablen auf kleinräumiger Rasterebene beschrieben werden[1]. Eine weitere Möglichkeit besteht darin, den Lageeffekt durch additive, gemischte Modelle (additive mixed models - AMM) oder durch räumlich-autoregressive Modelle (spatially autoregressive models - SAR) zu erklären.

[1] In Weberndorfer, »Modellierung von wertrelevanten Mikrolageparametern fuer die automatisierte Immobilienbewertung« wird die Erzeugung sowie Modellierung von wertrelevanten Mikrolageparametern für die Immobilienbewertung beschrieben.

- Mögliche Interaktionseffekte werden im Modell bisher nicht berücksichtigt und könnten in zukünftigen Forschungen betrachtet werden. Brunauer[2] schlug vor, anstatt des logarithmierten Kaufpreises den logarithmierten Kaufpreis pro Quadratmeter als Regressanden zu verwenden, was implizit einer Interaktion zwischen der Nutzfläche und den übrigen Eigenschaften einer Immobilie entspricht. Weiters sind zusätzlich Interaktionseffekte zwischen ähnlichen Charakteristiken denkbar, beispielsweise könnte die Existenz sowohl einer Fußbodenheizung als auch einer Wandheizung nicht mehr rein additiv (auf den logarithmierten Kaufpreis) wirken.

[2]Brunauer, »Modeling House Prices using Multilevel Structured Additive Regression«.

A. Anhang

	Bezirksname		Bezirksname
0	Eisenstadt(Stadt)	60	Vöcklabruck
1	Rust(Stadt)	61	Wels-Land
2	Eisenstadt-Umgebung	62	Salzburg(Stadt)
3	Güssing	63	Hallein
4	Jennersdorf	64	Salzburg-Umgebung
5	Mattersburg	65	Sankt Johann im Pongau
6	Neusiedl am See	66	Tamsweg
7	Oberpullendorf	67	Zell am See
8	Oberwart	68	Graz(Stadt)
9	Klagenfurt Stadt	69	Bruck an der Mur
10	Villach Stadt	70	Deutschlandsberg
11	Hermagor	71	Feldbach
12	Klagenfurt Land	72	Fürstenfeld
13	Sankt Veit an der Glan	73	Graz-Umgebung
14	Spittal an der Drau	74	Hartberg
15	Villach Land	75	Judenburg
16	Völkermarkt	76	Knittelfeld
17	Wolfsberg	77	Leibnitz
18	Feldkirchen	78	Leoben
19	Krems an der Donau(Stadt)	79	Liezen
20	Sankt Pölten(Stadt)	80	Mürzzuschlag
21	Waidhofen an der Ybbs(Stadt)	81	Murau
22	Wiener Neustadt(Stadt)	82	Radkersburg
23	Amstetten	83	Voitsberg
24	Baden	84	Weiz
25	Bruck an der Leitha	85	Innsbruck-Stadt
26	Gänserndorf	86	Imst
27	Gmünd	87	Innsbruck-Land
28	Hollabrunn	88	Kitzbühel
29	Horn	89	Kufstein
30	Korneuburg	90	Landeck
31	Krems(Land)	91	Lienz
32	Lilienfeld	92	Reutte
33	Melk	93	Schwaz
34	Mistelbach	94	Bludenz

35	Mödling
36	Neunkirchen
37	Sankt Pölten(Land)
38	Scheibbs
39	Tulln
40	Waidhofen an der Thaya
41	Wiener Neustadt(Land)
42	Wien-Umgebung
43	Zwettl
44	Linz(Stadt)
45	Steyr(Stadt)
46	Wels(Stadt)
47	Braunau am Inn
48	Eferding
49	Freistadt
50	Gmunden
51	Grieskirchen
52	Kirchdorf an der Krems
53	Linz-Land
54	Perg
55	Ried im Innkreis
56	Rohrbach
57	Schärding
58	Steyr-Land
59	Urfahr-Umgebung
95	Bregenz
96	Dornbirn
97	Feldkirch
98	Wien 1.,Innere Stadt
99	Wien 2.,Leopoldstadt
100	Wien 3.,Landstraße
101	Wien 4.,Wieden
102	Wien 5.,Margareten
103	Wien 6.,Mariahilf
104	Wien 7.,Neubau
105	Wien 8.,Josefstadt
106	Wien 9.,Alsergrund
107	Wien 10.,Favoriten
108	Wien 11.,Simmering
109	Wien 12.,Meidling
110	Wien 13.,Hietzing
111	Wien 14.,Penzing
112	Wien 15.,Rudolfsheim-Fünfhaus
113	Wien 16.,Ottakring
114	Wien 17.,Hernals
115	Wien 18.,Währing
116	Wien 19.,Döbling
117	Wien 20.,Brigittenau
118	Wien 21.,Floridsdorf
119	Wien 22.,Donaustadt
120	Wien 23.,Liesing

Literatur

Akaike, H. »A new look at the statistical model identification«. In: *Transactions on Automatic Control* 19 (1974), S. 716–723.

Anglin, P. M. und R. Gencay. »Semiparametric Estimation of a hedonic price function«. In: *Journal of A* 11 (1996), S. 633–648.

Berwin A. Turlach R, Andreas Weingessel. *quadprog: Functions to solve Quadratic Programming Problems.* R package version 1.5-4. 2011. URL: `http://CRAN.R-project.org/package=quadprog`.

Bivand, Roger S., Edzer J. Pebesma und Virgilio Gomez-Rubio. *Applied spatial data analysis with R.* Springer, NY, 2008. URL: `http://www.asdar-book.org/`.

Breiman, L. und J. H. Friedman. »Estimating optimal transformations for multiple regression and correlations (with discussion)"«. In: *Estimating optimal transformations for multiple regression and correlations (with discussion)".* 80 (1985), S. 580–619.

Brezger, A., T. Kneib und S. Lang. »Generalized structured additive regression based on Bayesian P-splines«. In: *Computational Statistics and Data Analysis* 50 (2006), S. 947–991.

Brunauer W., Lang S. & Umlauf N. »Modeling House Prices using Multilevel Structured Additive Regression«. Diss. Faculty of Economics und Statistics, University of Innsbruck, 2010.

Court, A. »Hedonic price indexes with automotive examples«. In: *The dynamics of automobile demand* (1939), pp. 99–117.

DeBoor, C. *A Practical Guide to Splines.* Cambridge University Press, 1978.

Eilers, P.H. und B.D. Marx. »Flexible smoothing with B-splines and penalties«. In: *Statistical Science* 11 (1996), S. 89 –121.

Fahrmeir L., Kneib T. & Lang S. *Regression. Modelle, Methoden und Anwendungen.* Springer, 2007.

Goodman, A. C. »Hedonic price, price indices and housing markets«. In: *Journal of Urban Economics* 5 (1978), S. 471–484.

Grömping, Ulrike. »Inference With Linear Equality And Inequality Constraints Using R: The Package ic.infer«. In: *Journal of Statistical Software* (2010). Forthcoming.

Hastie, T. und R. Tibshirani. *Generalized Additive Models.* Chapman & Hall, 1990.

– »Varying-coefficient models«. In: *Journal of the Royal Statistical Society* 55 (1993), S. 757 –796.

Koenker, R. und K. F. Hallock. »Quantile Regression«. In: *Journal of Economic Perspectives* 15 (2001), S. 143–156.

Kranewitter, H. *Liegenschaftsbewertung.* Manz'Sche Verlags- U. Universitbuchhandlung, 2010.

Lang, S. u. a. »Multilevel Structured Additive Regression«. In: *Statistics and Computing* (2012).

Lewin-Koh, Nicholas J. u. a. *maptools: Tools for reading and handling spatial objects*. R package version 0.8-14. 2012. URL: `http://CRAN.R-project.org/package=maptools`.

Li, K-C. »Sliced inverse regression for dimension reduction«. In: *Journal of the American Statistical Association* 86 (1991), S. 316 –342.

Martins-Filho, C. und O. Bin. »Estimation of hedonic price functions via additive nonparametric regression«. In: *Empirical Economnics* 30 (2005), S. 93–114.

McCullagh, P. und J. Nelder. *Generalized Linear Models*. Chapman und Hall, 1989.

Nelder, J. und R. Wedderburn. »Generalized Linear Models«. In: *Journal of the Royal Statistical Society* 135 (1972), S. 370–384.

Pebesma, Edzer J. und Roger S. Bivand. »Classes and methods for spatial data in R«. In: *R News* 5.2 (2005), S. 9–13. URL: `http://CRAN.R-project.org/doc/Rnews/`.

Pya, N. »Additive models with shape constraints«. Diss. University of Bath, 2010.

Pya, Natalya. *scam: Shape constrained additive models*. R package version 1.1-5. 2012. URL: `http://CRAN.R-project.org/package=scam`.

Ramsay, J. O. u. a. *fda: Functional Data Analysis*. R package version 2.2.7. 2011. URL: `http://CRAN.R-project.org/package=fda`.

Rosen, S. »Hedonic Prices and Implicit Markets Product Differentiation in Pure Competition«. In: *Journal of Political Economy* 82(1) (1974), S. 34–55.

Sarkar, Deepayan. *Lattice: Multivariate Data Visualization with R*. ISBN 978-0-387-75968-5. New York: Springer, 2008. URL: `http://lmdvr.r-forge.r-project.org`.

Schoenberg, I. J. »Contributions to the problem of approximation of equidistant data by analytic functions«. In: *Quart. Appl. Math.* 4 (1946), 45–99 und 112–141.

Steeb, W. H. *Kronecker Product of Matrices and Applications*. BI-Wiss.Verlag, 1991, S. 16.

W. A. Brunauer S. Lang, W. Feilmayr. »Hybrid multilevel STAR models for hedonic house prices«. In: *Jahrbuch fuer Regionalwissenschaft* (2013).

Weberndorfer, R. »Modellierung von wertrelevanten Mikrolageparametern fuer die automatisierte Immobilienbewertung«. 2013, Dissertation.

Weisberg, S. »Dimension Reduction Regression in R«. In: *Journal of Statistical Software,* 7 (2002).

Wood, S. N. »Modelling and smoothing parameter estimation with multiple quadratic penalties«. In: *Journal of the Royal Statistical Society (B)* 62.2 (2000), S. 413–428.

Wood, S. »Fast stable direct fitting and smoothness selection for generalized additive models«. In: *Journal of the Royal Statistical Society* 70 (2008), S. 495–518.

Wood, S.N. *Generalized Additive Models. An Introduction with R*. Chapman und Hall, 2006a.

Yue, Y. und H. Rue. »Bayesian inference for structured additive quantile regression models«. In: 2009.

Zusammenfassung

Statistische Analyse von Immobiliendaten erlangte in den vergangenen Jahren immer größer werdende Bedeutung. Hedonische Preismodelle zerlegen eine Immobilie gedanklich in verschiedene Eigenschaften wie Lage, Größe oder Zustand. Man versucht den Kaufpreis mithilfe von Regressionsmodellen zu schätzen.

Klassische lineare Regressionsmodelle eignen sich für die Analyse von Immobiliendaten allerdings nur bedingt, da Variableneffekte teilweise hochgradig nichtlinear sein können. Diese Arbeit beschäftigt sich mit der Modellierung von über Österreich verteilten Einfamilienhäuser anhand eines additiven Modells. Dabei werden Kaufpreise sowohl von der im linearen Modell verwendeten parametrischen Form als auch von nichtparametrisch modellierten Funktionen beschrieben, die eine äußerst flexible Schätzung erlauben.
Der Nachteil, der durch diese flexible Schätzung entsteht, besteht darin, dass auch möglicherweise unerklärliche Effekte mitmodelliert werden, da sich die nichtparametrisch geschätzten Funktionen zu sehr den Daten anpassen. Auch wenn die Spezifizierung einer bestimmten parametrischen Form schwierig erscheint, liegen trotzdem häufig gewisse Erwartungen an die Monotonie oder Krümmung der Funktion vor. Wie diese Annahmen mitmodelliert werden können soll ebenfalls in dieser Arbeit näher gebracht werden.

Einflüsse für den Effekt des Baujahres, verschiedener Flächengrößen wie beispielsweise Grundstücks- oder Erdgeschoßfläche werden ebenso präsentiert wie Effekte des Zustands und mehrerer Ausstattungsmerkmale. Geschätzte Kaufpreise werden mit den tatsächlichen verglichen, um die Güte des Modells zu beurteilen.
Zusätzlich werden jene fünf Eigenschaften angegeben, anhand deren der Wert einer Immobilie am besten widergespiegelt werden kann. Der Einfluss verschiedener Flächenvariablen (wie Erdgeschoß- oder Kellerflächen) soll weiters durch ein Dimensionsreduktionsverfahren auf den Effekt einer beschreibenden Variable reduziert werden.

Abstract

Statistical analysis of real estate data gained more and more attention during the last years. Hedonic pricing models divide a property into various characteristics such as location, size or condition. The purchase price is estimated with the aid of regression models.

Classical linear regression theory is not perfectly suitable for the analysis of real estate data because of the intense non linear behaviour of some variables. This paper addresses the modelling of single-family homes in Austria based on an additive model. Purchase prices are characterized by the parametric form known from the linear model as well as non-parametrically modelled functions, which allow extremely flexible estimation.
The disadvantage caused by this flexible method concerns the modelling of inexplicable effects, because the estimated function fits (even noisy) data too closely. Even if a specification of a certain parametric function appears difficult, there often exist some expectations of the monotonicity and curvature. This paper deals with the fact, how those expectations can be modelled.

Influences of the year of construction, the size of various areas, such as plot area or floor space are presented as well as effects of the condition and some equipment features. In order to evaluate model goodness, fitted prices are compared against the original ones. Additionally we highlight those five characteristics, which influence most the value of single-family homes. In order to reduce the number of variables, which describe the size of some areas (such as floor space or the size of the cellar), a dimension-reduction method is applied.

Curriculum Vitae

Persönliche Daten

Name:	Christian Pechhacker
Geburtstag:	09. August 1987
Geburtsort:	Baden
Staatsbürgerschaft:	Österreich

Schulbildung

1994-1998	Volksschule Gainfarn
1998-2006	Gymnasium Berndorf
26. Juni 2006	Matura
Okt. 2006 - Juni 2007	Zivildienst beim Roten Kreuz Bad Vöslau
September 2007	Beginn Bakkalaureatsstudium Statistik
19. Jänner 2011	Erhalt des Leistungsstipendiums nach dem Studienförderungsgesetz der Universität Wien
09. Februar 2011	Abschluss Bakkalaureatsstudium Statistik
März 2011	Beginn Magisterstudium Statistik
20. Jänner 2012	Erhalt des Leistungsstipendiums nach dem Studienförderungsgesetz der Universität Wien

Praktische Erfahrung

Jänner 2012 - Dezember 2012	Teilzeitbeschäftigter bei der Immobilien Rating GmbH (Wien)
seit Dezember 2012	Angestellter bei der Immobilien Rating GmbH (Wien)

Bad Vöslau, 12. Juni 2013