Ludwig-Maximilians-Universität

Institut für Statistik

Vergleich von multiplikativen und additiven Mietspiegeln

Bachelorarbeit

Andreas Singer

Betreuer: Prof. Dr. Göran Kauermann

München, 13. September 2015

Zusammenfassung

Die Spezifikation eines Regressionsmodells ist oft sehr komplex und zwingt den Statistiker viele Entscheidungen zu treffen. Auch bei der Erstellung eines Mietspiegels gilt es solche Entscheidungen zu treffen. Eine davon ist, ob ein additives oder ein multiplikatives Modell verwendet werden soll. In dieser Arbeit soll mit verschiedenen Techniken ein multiplikatives Mietspiegelmodell mit einem additiven Mietspiegelmodell verglichen werden. Zunächst werden die verwendeten Daten und Modelle erklärt werden. Zum Vergleich soll zum einen die Modellgüte zur Entscheidung herangezogen werden, zum anderen soll untersucht werden, welche Variante den Modellannahmen am besten entspricht. Ein besonderes Interesse birgt die Analyse der Vorhersagekraft für beide Modelle, da die Prädiktion die Hauptaufgabe eines Mietspiegelmodells darstellt. Des Weiteren wird mittels einer Simulation die Eignung der Modelle für Daten aus dem anderen Modell überprüft werden. Die Ausführung stützt sich dabei im Wesentlichen auf die Theorie des generalisierten additiven Modells von Wood (Wood, 2006). Die statistischen Analysen werden mit Hilfe des Programmpakets R anhand der Daten des Münchner Mietspiegels durchgeführt. Auch die verwendeten Modelle entsprechen dem Münchner Modell, bzw. lehnen sich daran an. Nach Betrachtung aller Ergebnisse wird ein leichter Vorteil des additiven Modells gegenüber dem multiplikativen Modell erkennbar sein.

Inhaltsverzeichnis

1 Einleitung

„Deutschland zieht die Mietpreisbremse“ (WirtschaftsBlatt, 2014).
Titel wie diese dominierten die Zeitungen am 6. März 2015, denn nach langjährigen Diskussionen verabschiedete der deutsche Bundestags am Vortag das „Mietrechtsnovellierungsgesetz“, das seit dem 1. Juni 2015 in Kraft ist. Ein wichtiger Punkt dieses Gesetzes, mit dem die inflationäre Steigerung der Mietpreise abgedämpft werden soll, ist die neu eingefügte über die Miethöhe bei Mietbeginn in Gebieten mit angespannten Wohnungsmärkten“, im Volksmund auch Mietpreisbremse genannt. Darin wird für jene Gebiete beschlossen, dass bei der Wiedervermietung einer Wohnung oder eines Hauses die Miete nicht mehr als 10 Prozent oberhalb der ortsüblichen Vergleichsmiete liegen darf.
So einfach dieses Gesetz auch klingen mag, bereitet dennoch ein Detail vielen Landesregierungen ernsthafte Probleme. Der kritische Punkt ist die Bestimmung jener ortsüblichen Vergleichsmiete, also dem Preis, der für eine Wohnung mit vergleichbaren Merkmalen durchschnittlich bezahlt wird. Voraussetzung zu deren Bestimmung ist die Erstellung eines qualifizierten Mietspiegels, bei dem die Nettomiete einer Wohnung meist durch ein Regressionsmodell beschrieben wird. Dazu werden die Merkmale von Wohnungen wie das Baujahr und die Wohnfläche sowie die Nettokaltmiete abgefragt, um auf diesen Daten ein Modell zu fitten. Anhand dieses Mietspiegels kann dann für jede Immobilie eine geschätzte Nettomiete berechnet werden. Jedoch muss diese Schätzung nicht immer korrekt sein. So wurde der Berliner Mietspiegel von 2013 kurz vor Inkrafttreten der Mietpreisbremse vor Gericht für nichtig erklärt, da er teilweise nicht nach anerkannten Wissenschaftlichen Methoden erstellt wurde. Auch andere Städte hatten in der Vergangenheit mit Klagen von Vermietern und Mietern zu kämpfen, so auch die Stadt München im Jahre 2011 und auch in Zukunft wird es wohl einige Mietspiegel geben, an deren Wahrheitsgehalt gezweifelt werden wird. Diese Beispiele zeigen, wie wichtig eine korrekte Methode bei der Erstellung eines Mietspiegels ist. Heute existiert eine Menge an unterschiedlichen Ansätzen. In der teuersten Stadt Deutschlands, München, wird ein Modell, das die vergleichbare Nettomiete durch den additiven Einfluss verschiedener Merkmale bestimmt, verwendet. Alternativ dazu wäre ein multiplikativer Ansatz möglich.
Diese Arbeit hat das Ziel anhand des Münchner Mietspiegels die Annahme multiplikativen Einflusses mit der Annahme eines additiven Einflusses zu vergleichen. Dazu werden zuerst die in dieser Arbeit verwendeten Daten aus dem Münchner Mietspiegel von 2015 erklärt. Anschließend soll auf das hier verwendete generalisierte additive Modell eingegangen werden. Dabei soll der Fokus zunächst auf den theoretischen Grundlagen der Modelle liegen und anschließend die Unterschiede der multiplikativen und additiven Modellierung aufgezeigt werden. Dabei sollen auch die Ergebnisse der Modellschätzung verglichen werden. Im darauf folgenden Teil sollen im Rahmen einer Modellvalidierung die Modellannahmen überprüft und ein erster Vergleich der beiden Modelle auf Grund von Gütekriterien angestellt werden. Ebenfalls soll ein Augenmerk auf die Vorhersagekraft der beiden Modelle gelegt werden. Dazu soll eine Kreuzvalidierung nähere Erkenntnisse bringen. Im letzten Teil wird durch eine Simulation der Umgang der Modelle mit Daten aus dem anderen Modell überprüft werden.

2 Mietspiegeldaten

Um ein Mietspiegelmodell zu fitten zu können werden zunächst Informationen über die Wohnungen benötigt. Diese Daten bilden die Grundlage für den Mietspiegel. Damit das Modell auch als Mietspiegel verwendet werden kann, müssen die erhobenen Daten repräsentativ für den gesamten relevanten Wohnungsmarkt der Stadt sein. Außerdem sollten auch alle Merkmale, die möglicherweise einen Einfluss haben, erhoben werden. Die in dieser Arbeit verwendeten Daten aus dem Münchner Mietspiegel erfüllen diese Voraussetzungen. Hier soll zunächst ein Überblick über diese Daten und deren Erhebung gegeben werden.

2.1 Datenerhebung

Im ersten Schritt der Datenerhebung wurde eine Zufallsstichprobe von Telefonnummern gezogen. Mit Hilfe der gezogenen Telefonnummern wurden in mehreren Schritten Interviews geführt. Unter anderem sollte dabei auch die Eignung der Wohnungen für den Mietspiegel untersucht werden. Im Bürgerlichen Gesetzbuch hat der Gesetzgeber dazu unter Paragraph 558c/d klare Vorgaben gegeben, welche Wohnungen zur Berechnung eines Mietspiegels herangezogen werden dürfen. Zudem gibt es „Hinweise zur Erstellung von Mietspiegeln“ (bun, 2002), die die Auswahl der Wohnungen regeln. Im verwendeten Datensatz finden sich nur die Wohnungen, die zwischen dem 01.01.2010 und Januar 2014 neu vermietet wurden oder der Mietpreis im selben Zeitraum geändert wurde oder auch beides. Außerdem wurde ausgeschlossen:

- preisgebundener Wohnraum.
- Wohnungen ohne Mietverhältnisse.
- Gewerblich genutzte Wohnungen.
- Durch Eigentümer selbst genutzter Wohnraum.
- Wohnraum zum vorübergehenden Gebrauch. Fälle mit einer vereinbarten Vertragsdauer von bis zu einem Jahr wurden nicht berücksichtigt.
- Vom Vermieter möblierter Wohnraum.
- Private Untermietverhältnisse.
- Studenten- und Jugendwohnheime.
- Wohnraum in Anstalten, Heimen oder Wohnheimen, bei denen die Mietzahlung auch Serviceleistungen abdeckt (z.B. Verpflegung oder Betreuung).
- Einzelzimmer.
- Einfamilienhäuser, Doppelhaushälften, Reihenhäuser.
- Penthouse-Wohnungen.

- Wohnungen, deren Küche, Bad und Toilette von zwei oder mehr Hauptmieterparteien gemeinsam genutzt werden.

- Wohnungen im Untergeschoss.

Des Weiteren gab es noch andere Ausschlusskriterien, deren Aufführung hier den Rahmen sprengen würde. Dabei wurden beispielsweise Wohnungen aussortiert, deren Merkmale nur sehr selten im Datensatz vorkamen, sodass keine verlässliche Schätzung dafür möglich wäre. Eine vollständige Zusammenfassung findet sich in der Dokumentation des Mietspiegels (Windmann u. Kauermann, 2015). Im abschließenden Teil der Datenerhebung fand das Hauptinterview statt. Hierbei bearbeiteten die Interviewer zusammen mit den ausgewählten Mietern einen Fragebogen. Zusätzlich wurden in einem Vermieterfragebogen auch die Vermieter befragt. Die daraus hervorgehenden Daten wurden auf Übereinstimmung überprüft und gegebenenfalls aussortiert. Das Ergebnis dieses Vorgehens ist der finale Datensatz, bestehend aus 3065 Wohnungen, die eine repräsentative Stichprobe aller rund 200000 Münchner Wohnungen bilden.
Genauere Informationen finden sich im ersten Kapitel der Dokumentation des Mietspiegels für München 2015 (vgl. Rösch u. a., 2015)

2.2 Variablenbeschreibung

Nicht alle Variablen, die im Zuge der Befragung erhoben wurden, wurden auch im Mietspiegelmodell berücksichtigt. Einige Variablen wurden nicht berücksichtigt, da sie sich im Münchner Mietspiegelmodell als nicht signifikant erwiesen. Andere wurden erst nach einer Modifizierung oder einer Kombination mit einer zweiten Variablen verwendet. Alle Variablen, die sich im endgültigen Mietspiegelmodell und in den Analysen dieser Arbeit finden, werden nun genannt und erklärt. Details zu den Variablen können in der Dokumentation des Mietspiegels für München 2015 (vgl. Rösch u. a., 2015) nachgelesen werden.

Nettomiete pro Quadratmeter (nmqm)

Jeder Mietspiegel hat zum Ziel eine mittlere Vergleichsmiete für jede Wohnung anzugeben. Das bedeutet, dass man für eine Wohnung mit bekannten Merkmalen aber einem unbekannten Mietpreis die Nettomiete bestimmen will. Nach dieser Aufgabenstellung eignet sich die Nettomiete als Zielgröße der Regression. Zur Bestimmung der Nettomiete wurden von den angegebenen Zahlungen der Mieter für die Wohnung zunächst Zuschläge für Gartenbenutzung oder ähnlichem abgezogen. Waren diese nicht bekannt, wurden Durchschnittswerte verwendet. Anschließend wurden, falls vorhanden, anfallende Beträge für Mietpreisermäßigung oder Mietminderung addiert. Im letzten Schritt wurden die Betriebskosten vom Mietpreis abgezogen. Der so ermittelte Betrag wird von nun an Nettomiete genannt. Im Münchner Mietspiegelmodell wird allerdings nicht die Nettomiete, sondern die Nettomiete pro Quadratmeter verwendet. Dieser Wert wird durch die Division der Nettomiete durch die Wohnfläche für jede Wohnung berechnet.

Wohnfläche (wfl.gekappt)

Diese Variable beschreibt die Wohnfläche der Wohnung in Quadratmetern, wobei die Fläche bei 20qm und 160qm gekappt wurde. Das bedeutet, dass Wohnungen mit weniger als 20qm den

Wert 20 und Wohnungen mit mehr als 160qm den Wert 160 erhielten. Dies war notwendig, da sich in diesen extremen Bereichen nur wenige Beobachtungen finden ließen und sich gleichzeitig deren Quadratmeterpreise stark unterschieden, so dass deren Aufnahme zu einer großen Schätzungenauigkeit geführt hätte. Das kappen der Variable ist damit eine Alternative zum direkten Ausschließen der Beobachtungen. Zu beachten ist jedoch, dass bei sehr kleinen oder sehr großen Wohnungen der Mietspiegel deshalb keine genaue Schätzung, sondern lediglich eine Orientierungshilfe bietet.

Baujahr (bjahr.katmean.imp)

Das Baujahr wird durch eine kategoriale Variable beschrieben und besteht aus bedingten Mittelwerten jeder Kategorie. Diese bedingten Mittelwerte wurden durch eine semiparametrische Regression mit mehreren Einflussgrößen aus den Mietspiegeldaten bestimmt. Durch dieses Verfahren war es möglich, Unstimmigkeiten zwischen Mieter- und Vermieterfragebögen, sowie fehlende Werte zu behandeln. Die hier verwendete Variable gibt das (vermutete) Baujahr in verschiedenen Jahreszahlkategorien an.

Wohnlagen (Faktor.ML.WL)

Die Wohnlage, in der sich eine Wohnung befindet, setzt sich im Münchner Mietspiegel aus zwei Teilen zusammen. Der eine Teil ist die Wohnlage, die durch Expertenmeinungen in einem Gutachterausschuss definiert wurde, der andere die Makrolagen. Die Wohnlagen des Gutachterausschusses orientieren sich an Merkmalen, wie der Infrastruktur und dem Bestand an Grünflächen. Dabei wurden die Wohnungen in eine der vier Kategorien der Wohnlage von einfacher bis bester Wohnlage eingeteilt. Die Makrolagen wurden für jeden Münchner Stadtbezirk anhand der schrittweisen Änderung der Kategorie bestimmt. Es wurden den Bezirken die Kategorie zugeteilt, die die höchste Modellgüte anhand des AICs (siehe Abschnitt 5.2.1) erzeugten. Die endgültige Variable entstand anschließend durch die Kombination der beiden Merkmale, wobei die Interaktion bereits enthalten ist.

Faktor.ML.WL	Wohnlage
0.2	Durchschnittliche Lage
0.3	Gute Lage
0.4	Beste Lage
1.2	Zentrale Durchschnittliche Lage
1.3	Zentrale Gute/Beste Lage

Tabelle 1: Kodierung der Variable $Faktor.ML.WL$

Gebäudetypen (Gebaeudetyp)

Das Gebäude, in dem sich die Wohnung befindet, wurde einer der Kategorien „Hochhaus“, „Wohnblock“, „Stadthaus“ oder „andere“ zugeordnet. Ausschlaggebend für die Einteilung waren vor allem Bauart und Grünflächen.

Haustyp (Haustyp.15)

Ähnlich wie der Gebäudetyp wurde auch der Haustyp einer Kategorie zugeordnet. Dabei wurde in „einfacher Nachkriegsbau“, „einfacher Altbau a“, „einfacher Altbau b“ und „andere“ unterschieden. Die Zuordnung erfolgte anhand des Baujahres und der Deckenhöhe.

Warmwasserversorgung (WW.Bereitung)
Diese Variable gibt an, ob die Wohnung eine vom Vermieter gestellte Warmwasserversorgung besitzt oder nicht. Eine „Unvollständige Warmwasserversorgung“ besteht bei Warmwasserversorgung in Bad oder Küche, aber nicht in beiden Räumen.
Heizung (Heizung.15)
Bei der Betrachtung der Beheizung werden nur die vier größten Wohnräumen, die Küche und das Bad betrachtet. Wird nur ein Teil dieser Räume beheizt, ist die Beheizung unvollständig.
Sanitärbereich (Sanitaerbereich)
Hier spielt die Anzahl und die Ausstattung der Badezimmer eine Rolle. Mögliche Ausprägungen sind das Fehlen eines Badezimmers („keine.Badausstattung“), ein „zweites Bad“ oder eine „besondere Badausstattung“, wie ein zweites Waschbecken oder eine Badewanne mit separater Dusche. Letzteres ist nicht möglich, falls es ein weiteres Badezimmer gibt.
Modernisierungsmaßnahmen (Modernisierung)
Diese Variable nimmt den Wert 0 an, falls seit dem Jahr 2005 weder die Heizung, noch die Warmwasserversorgung oder ein Badezimmer modernisiert wurden. Andernfalls wird die Variable mit 1 kodiert.
Küche (Kueche.Ausstattung) und (Gaskochfeld.15)
Eine gute Küchenausstattung liegt vor, wenn vom Vermieter eine Küche bereitgestellt wird, die mindestens über eine Kühlmöglichkeit, ein Kochfeld, eine Spülmöglichkeit und Einbauschränke verfügt. Die Variable Gaskochfeld.15 wird mit 1 kodiert, falls ein Gaskochfeld, aber keine gute Küchenausstattung vorhanden ist. Andrenfalls ist sie 0.
Fußboden (Boden.einfach), (Boden.einfach.teilweise) und (GuterBoden.neu)
Hat die Wohnung in keinem Wohnraum einen hochwertigen Boden wie Parkett, Teppich oder Fliesen hat die Variable Boden.einfach den Wert 1, ansonsten den Wert 0. Ebenso verhält es sich mit der Variable Boden.einfach.teilweise. Sie hat den Wert 1, wenn nur ein Teil der Wohnräume einen einfachen Boden hat und der Rest einen hochwertigen Boden. Die Variable GuterBoden.neu zeigt mit dem Wert 1 an, dass der Boden in allen Wohnräumen hochwertig oder neu ist.
Terrasse (Terrasse)
Hat die Wohnung eine Terrasse oder etwas vergleichbares in Süd- oder Westausrichtung, die mindestens 5qm groß ist, so hat die Variable den Wert 1.
Gegensprechanlage (Keine.Gegensprech)
Diese Variable ist mit 1 kodiert, falls in der Wohnung keine Gegensprechanlage oder Videogegensprechanlage vorhanden ist.
Fußbodenheizung (Fussboden.Heizung)
Existiert in der Wohnung eine Fußbodenheizung hat diese Variable den Wert 1.
elektrische Rolläden (Rolllaeden.elektrisch)
Diese Variable gibt analog zur vorhergehenden Variable die Existenz von elektrischen Rollläden an.

Dachgeschosswohnung (Hohe.DG)
Mit dem Wert 1 wird in dieser Variable eine Wohnung im Dachgeschoss, deren größter Raum eine Raumhöhe von mindestens 270cm misst, kodiert. Dachschrägen werden dabei nicht berücksichtigt.

3 Das Mietspiegel Modell

Für die Erstellung eines Mietspiegels ist es in Deutschland heutzutage üblich ein Regressionsmodell zu erstellen. Das Ziel dabei ist es, die Nettomiete durch möglichst alle Eigenschaften der Wohnungen, die einen Einfluss darauf haben, zu erklären. In der Praxis existieren sehr viele verschiedene Ansätze für ein solches Modell. Ein Modell, das sich bewährt hat und große Anerkennung erfährt, ist das Münchner Mietspiegelmodell (vgl. Windmann u. Kauermann, 2015). Alle Analysen dieser Arbeit werden auf dem Münchner Mietspiegel aufbauen und an den Münchner Daten durchgeführt werden. Zunächst soll hier jedoch das Münchner Modell und deren theoretische Grundlagen dargestellt werden.

3.1 Grundidee der Semiparametrischen Regression

Die ersten Regressionsmietspiegel wurden mit der Hilfe von einfachen linearen Modellen erstellt (S. 15 Gonzalez, 2002). Diese Technik erwies sich jedoch schnell als zu ungenau und nicht geeignet für die Verwendung als Mietspiegelmodell. In vielen Städten entwickelten sich daraufhin verschiedene Herangehensweisen; eine der bekanntesten ist das additiv-multiplikative Modell des Regensburger Mietspiegels. Der Münchner Mietspiegel 2015 verwendet hingegen ein generalisiertes additives Modell. Dieses Modell ist die Erweiterung eines generalisierten linearen Modells durch einen nicht-parametrischen Teil. Das ermöglicht die zusätzliche Modellierung nicht-linearer Einflüsse und ist damit deutlich flexibler als ein gewöhnliches lineares Modell. So kann der Einfluss von Variablen, denen eine besondere Wichtigkeit zugeordnet wird, sehr viel besser geschätzt werden. Beispielsweise ist es so möglich den Einfluss einer Variablen, der bei kleinen Werten der Variablen mit ihr sinkt und bei großen Werten steigt, mit ins Modell aufzunehmen.

3.2 Glättungssplines

Um eine solche detaillierte Modellierung bewerkstelligen zu können, ist es notwendig ein geeignetes Werkzeug dafür einzuführen. Im generalisierten additiven Modell werden dazu Splines verwendet. Stellt man sich Punkte in einem Streudiagramm vor, so kann man einen Glättungsspline als ein biegsames Lineal betrachten, dass sich den Punkten anpasst. Formal bedeutet das: Untersucht man den Einfluss einer metrischen Variablen z auf eine andere metrische, abhängige Variable y, nimmt man an, dass sich dieser Einfluss als

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

darstellen lässt. Allerdings gibt es eine Reihe von unterschiedlichen Ansätzen diese Funktion $f(z)$ zu berechnen. Hier werden die beiden Glättungssplines dargestellt, die im Mietspiegelmodell Verwendung finden werden.

3.2.1 P-Splines

Eine Möglichkeit, einen nicht-linearen Einfluss einer Kovariablen auf die Zielgröße zu modellieren, sind penalisierte Splines, kurz P-Splines (S. 293-314 Fahrmeir u. a., 2009) und (S. 152-154 Wood, 2006). Dabei wird ein Polynom-Spline durch einen Strafterm so modifiziert, dass eine zu raue Schätzung bestraft wird. Im Paket mgcv basieren diese P-Splines auf B-Splines. Die Idee von Polynom-Splines ist es, die Datenpunkte in einem Streudiagramm möglichst genau und flexibel durch eine glatte Funktion zu beschreiben. Zunächst wird der Wertebereich der Einflussgröße in m-1 Intervalle zerlegt, die durch m Knoten verbunden sind. Auf jedem Intervall wird ein Polynom vom Grad l geschätzt. Da bei der Zusammensetzung der einzelnen Polynome der Intervalle an den Knoten Sprünge und Kanten entstehen können, müssen weitere Glattheitsanforderungen gestellt werden. Bei Verwendung von B-Splines wird dieses Problem durch die Bildung von Polynomfunktionen vom Grad l über l+2 Knoten, die ausschließlich in diesem Intervall positiv definiert sind, gelöst. Die daraus resultierenden m+l-1 Basisfunktionen werden nun zusammengesetzt, sodass sich die Funktion

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

ergibt, wobei γ der zu schätzende Koeffizientenvektor zu den Basisfunktionen $B(z)$ ist. Somit ist eine stetige Funktion auf dem Wertebereich der Kovariablen definiert, die in jedem Knoten zweimal stetig differenzierbar ist. Die Glattheit der Funktion hängt allerdings entscheidend von der Anzahl der Knoten ab. Je mehr Knoten verwendet werden, desto rauer wird meist die Schätzung für die Funktion, weil sie zu stark auf die Daten angepasst wurde. In diesem Fall spricht man von Overfitting. Hat man jedoch zu wenige Knoten gewählt ist die Funktion nicht flexibel genug und kann detaillierte Verläufe nicht mehr korrekt darstellen. Um das Problem der Knotenwahl zu lösen, wird bei P-Splines ein B-Spline mit einem Strafterm der Form $\lambda \int (f''(z))^2 \, \mathrm{d}z$ ergänzt. Dadurch wird eine zu raue Schätzung bestraft. λ ist dabei ein Glättungsparameter. Je größer der Glättungsparameter gewählt wird, desto unflexibler wird die Schätzung. Im verwendeten Modell wird λ automatisch gewählt.

3.2.2 Thin-Plate Regression Splines

Das Paket mgcv für R verwendet in der Defaulteinstellung jedoch keine P-Splines, sondern die sogenannten Thin-Plate Regression Splines (S. 154-160 Wood, 2006) und (S. 27-29 Hasti u. Tibshirani, 1990). Mit deren Hilfe ist es möglich auch mehrdimensionale Splines zu schätzen. Das bedeutet, dass beispielsweise zwei Variablen gleichzeitig in die Schätzung eines Splines eingehen und damit auch eine glatte Interaktion mitgeschätzt werden kann. Ein weiterer Vorteil ist die automatische Wahl der Glättungsparameter und der Knoten. Simon Wood bezeichnet die Thin-Plate Regression Splines als die optimale Glättungsvariante. Durch ihre Multidimensionalität sind die Definition und die Erklärung dieser Splines sehr umfangreich und kompliziert. Da im Folgenden dieses Verfahren jedoch nur für eine Variable verwendet werden soll, wird die Erklärung auf den einfachen Fall begrenzt. Die Grundlage für den TPRS (nach Hasti u. Tibshirani, 1990) bilden kubische Glättungssplines. Ein kubischer Glättungsspline löst folgendes Problem: Es soll die Funktion f(x), die zweimal stetig differenzierbar ist, gefunden werden, die die penalisierte Residuenquadratsumme

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 \,\mathrm{d}t$$

minimiert. Dabei ist λ ein fester Glättungsparameter und $a \leq x_1 \leq ... \leq x_n \leq b$. Der erste Teil dieser Formel stellt sicher, dass sich die Funktion gut genug an die Daten anpasst, der zweite Teil bestraft ein Overfitting. Durch die Erweiterung dieses Problems auf den mehrdimensionalen Raum nennt man Thin-Plate Splines. Diese sind jedoch relativ aufwändig zu schätzen, weshalb stattdessen eine Approximation durch Thin-Plate Regression Splines verwendet wird. Auf das genaue Verfahren der Approximation wird hier nicht weiter eingegangen, kann aber in (S. 157ff Wood, 2006) nachgelesen werden. Die Vorteile eines solchen Splines sind, dass er effizient berechnet werden kann, auf eine Bestimmung der Knoten verzichtet und durch wenige Parameter definiert ist.

3.3 Generalisiertes Additives Modell

Im Münchner Mietspiegel wird für die Erklärung der Miete durch die Kovariablen ein generalisiertes additives Modell verwendet (vgl. Wood, 2006) und (vgl. Hasti u. Tibshirani, 1990). Dieses Modell erweitert das lineare Modell, bei dem der Einfluss jeder Kovariablen linear modelliert wird, um einen nichtparametrischen Teil:

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + \eta_i^{lin} + \epsilon_i = f_1(x_{1i}) + f_2(x_{2i}) + \beta_1 \cdot x_{3i} + \beta_2 \cdot x_{4i} + \epsilon_i = \eta_i^{add} + \epsilon_i$$

Dieser additive Teil ermöglicht das Modellieren einer nicht-linearen, glatten, flexiblen Funktion f. Jede Funktion wird mittels eines Streudiagramm-Glätters, wie den P-Splines oder Thin-Plate Regression Splines geschätzt. Dabei ist es auch erlaubt unterschiedliche Kovariablen im selben Modell durch verschiedene Smoother zu behandeln. Verwendet man ein Modell dieser Form tritt jedoch ein Identifikationsproblem $f_1(x_1) + f_2(x_2) = f_1(x_1) + c + f_2(x_2) - c$ mit $c \neq 0$ auf (S. 400 Fahrmeir u. a., 2009). Das bedeutet, dass die Höhe der einzelnen Funktionen willkürlich verändert werden kann. Die gebräuchlichste, und auch hier verwendete, Lösung dieses Problems ist die Zentrierung dieser Funktionen:

$$\sum_{i=1}^{n} f_1(x_{1i}) = \sum_{i=1}^{n} f_2(x_{2i}) = 0$$

Für die Verteilung der Zielgröße Y nimmt man eine aus der Exponentialfamilie ableitbare Verteilung an, im Fall des Mietspiegels eine Normalverteilung. Zusätzlich wird der Erwartungswert der abhängigen Variablen und der Prädiktor mit einer Responsefunktion $h(\eta) = \mu$ und einer Linkfunktion $g(\mu) = \eta$ verknüpft. Insbesondere seien hier die in den folgenden Teilen verwendeten Linkfunktionen der Identität $g(\mu) = \mu$ und der Log-Link $g(\mu) = log(\mu)$ erwähnt. Dadurch ergibt sich ein Modell in der folgenden Form:

$$g(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + X_i \cdot \Theta$$

(S. 121 Wood, 2006) μ_i entspricht dabei dem Erwartungswert der Zielgröße y_i mit einer Verteilung aus der Exponentialfamilie. X entspricht der Designmatrix des Modells und $\Theta = \beta_1 + ... + \beta_k$ dem zugehörigen Parametervektor. Die Funktionen f_1 und f_2 sind die Funktionen, die sich aus der Schätzung der Glättungssplines mit den Kovariablen x_1 und x_2 ergeben. Zusätzlich ist es auch möglich Interaktionen zwischen zwei Kovariablen zu betrachten. Da sich diese Interaktionen im Folgenden jedoch unberücksichtigt bleiben wird an dieser Stelle nicht weiter darauf eingegangen. Bei der Schätzung dieses Modells werden neben den Koeffizienten Θ aus dem linearen Teil des Prädiktors auch die Glättungsparameter der beiden Funktionen f_1 und f_2 geschätzt. Dazu wird das generalisierte Kreuzvalidierungskriterium (S. 49 Hasti u. Tibshirani, 1990)

$$GCV = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{y}_i}{1 - \frac{1}{n} tr(H)} \right\}^2$$

mit $tr(H)$ als der Spur der Glättungsmatrix verwendet. Das GCV ist ein Optimalitätskriterium für die Glättungsparameterwahl mit deren Hilfe ein geeigneter Kompromiss zwischen Verzerrung und Varianz des Glättungsverfahrens gefunden werden kann. Die Minimierung des GCVs ergibt dabei die optimalen Glättungsparameter. Die Schätzung des Modells geschieht anschließend durch P-IRLS, Penalisierte Iterativ-neugewichtete kleinste Quadrate (S. 169f,137ff Wood, 2006). Dieser Algorithmus löst das Problem des Maximierens der Likelihood, indem iterativ die Summe der penalisierten gewichteten kleinsten Quadrate

$$\sum_{i=1}^{n} w_i \left(y_i - \eta_i^{add} \right)^2$$

mit den Gewichten w_i minimiert werden. Dazu werden in jedem Schritt neue Daten und die zugehörigen Gewichte auf Grundlage des im vorhergehenden Schritt geschätzten Modells erstellt. Aus diesen Pseudodaten und -gewichten werden durch Minimierung der penalisierten KQ-Methode die neuen Koeffizienten geschätzt auf deren Basis wiederum im nächsten Schritt die Pseudodaten erstellt werden können. Dies geschieht so lange, bis der Algorithmus konvergiert, also die Koeffizienten gleich bleiben. Die Koeffizienten des letzten Schrittes ergeben schließlich die Schätzung der Parameter im fertig gefitteten Modell.

3.4 Zweistufige KQ-Schätzung

Bei einem solchen Modell ist es möglich, dass die Varianz der Störterme nicht homogen ist. Das bedeutet, dass die Varianz von den Werten anderer Variablen abhängig sein kann und beispielsweise mit dem Anstieg einer Kovariable auch steigt. Eine Möglichkeit, dieses Problem zu lösen, ist die zweistufige KQ-Schätzung (S. 50 Windmann u. Kauermann, 2015). Dabei wird zunächst ein ungewichtetes generalisiertes additives Modell nach Abschnitt 3.3 auf den Daten gefittet. Nun interessiert man sich, ob die Kovariablen einen Einfluss auf die Varianz der Residuen haben. Auf Grundlage der vorhergesagten Werte aus dem Modell $\hat{y}_i$ und der wahren Werte y_i, werden die quadrierten Residuen $r_i^2 = (y_i - \hat{y}_i)^2$ errechnet. Diese bilden die neue Zielvariable für den zweiten Schritt der Schätzung. Dabei wird beim Münchner Mietspiegel ein generalisiertes additives Gamma-Modell verwendet. Bei diesem Modell handelt es sich um ein generalisiertes additives Modell, wobei eine Gammaverteilung als Verteilung der Zielgröße angenommen wird. Als Prädiktor wird erneut der exakt gleiche Prädiktor wie im ersten Modell gewählt. Die aus der Schätzung hervorgehenden Koeffizienten und Funktionen geben den geschätzten Einfluss der Kovariablen auf die Residuen an. Später sollen jedoch nur die Kovariablen verwendet werden, deren Einfluss sich als signifikant erwiesen hat. Aus diesem Modell werden so die Gewichte $w_i = 1/E(r_i^2)$ als Kehrwert der geschätzten quadratischen Residuen für jede Beobachtung gewonnen. Diese Gewichte werden anschließend für die Schätzung im finalen Mietspiegelmodell verwendet.

3.5 Additives Mietspiegelmodell

Das additive Mietspiegel Modell dieser Arbeit entspricht exakt dem Münchner Mietspiegelmodell von 2015. Dabei wird die Nettomiete pro Quadratmeter (*nmqm*) durch die Kovariablen aus Abschnitt 2.2 erklärt, wobei die Wohnfläche und das Baujahr jeweils nichtparametrisch geschätzt werden. Für die Wohnfläche wird dabei ein Thin-Plate Regression Spline mit automatischer Wahl der Glättungsparameter verwendet. Bei der Modellierung des Baujahres kommt ein P-Spline-Ansatz zum Einsatz. Alle anderen Kovariablen werden linear berücksichtigt. Für die Zielgröße *nmqm* wird eine Normalverteilung angenommen und die Identität als Linkfunktion benutzt. Die Gewichte werden durch eine zweistufige KQ-Schätzung analog zu Abschnitt 3.4 KQ-Schätzung bestimmt. Dabei ergab sich ein signifikanter Einfluss der Variablen $wfl.gekappt$, $bjahr.katmean.imp$ und $Faktor.ML.WL$ auf die Varianz der Residuen. Diese drei Variablen wurden deshalb zur Berechnung der Gewichte verwendet. Es ergibt sich also das additive Mietspiegelmodell

$$
\begin{aligned}
\mathbb{E}(nmqm) =& f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add} \\
=& f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \beta_1 \cdot Faktor.ML.WL \\
&+ \beta_2 \cdot Gebaeudetyp + \beta_3 \cdot Haustyp.15 + \beta_4 \cdot WW.Bereitung + \beta_5 \cdot Heizung.15 \\
&+ \beta_6 \cdot Sanitaerbereich + \beta_7 \cdot Modernisierung + \beta_8 \cdot Kueche.Ausstattung \\
&+ \beta_9 \cdot Gaskochfeld.15 + \beta_{10} \cdot Boden.einfach + \beta_{11} \cdot Boden.einfach.teilweise \\
&+ \beta_{12} \cdot GuterBoden.neu + \beta_{13} \cdot Hohe.DG + \beta_{14} \cdot Keine.Gegensprech \\
&+ \beta_{15} \cdot Fussboden.Heizung + \beta_{16} \cdot Rolllaeden.elektrisch + \beta_{17} \cdot Terrasse
\end{aligned}
$$

Es ist offensichtlich, dass durch die Aufsummierung aller Einflüsse jede Kovariable additiv auf die Nettomiete pro Quadratmeter wirkt. Interpretiert werden kann das dadurch, dass die erwartete Nettomiete pro Quadratmeter für eine Wohnung mit Terrasse β_{17} Euro teurer ist, als für eine Wohnung ohne Terrasse, wobei die übrigen Kovariablen gleich bleiben. Die Miete erhöht sich folglich additiv. Damit ist auch der Einfluss jeder Kovariablen auf die Nettomiete additiv:

$$
\begin{aligned}
\mathbb{E}(nm) =& wfl.gekappt \cdot \Big(f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add} \Big) \\
=& wfl.gekappt \cdot \Big(f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \beta_1 \cdot Faktor.ML.WL \\
&+ \beta_2 \cdot Gebaeudetyp + \ldots + \beta_{17} \cdot Terrasse \Big)
\end{aligned}
$$

3.6 Multiplikatives Mietspiegelmodell

Das multiplikative Mietspiegelmodell, das in dieser Arbeit verwendet wird, basiert auf dem additiven Modell aus dem vorhergehenden Abschnitt. Um den Einfluss der Variablen multiplikativ darzustellen, wird das Modell an einer Stelle verändert. Nämlich wird nun nicht mehr die Identität als Linkfunktion verwendet, sondern der Log-Link. Die Linkfunktion lautet somit $g(\mu) = log(\mu)$ und die zugehörige Responsefunktion $h(\eta) = exp(\eta)$. Bei der Berechnung der Gewichte wurden die Variablen $Boden.einfach$ und $Boden.einfach.teilweise$ und $wfl.gekappt$ verwendet, da sich einzig ihre Einflüsse auf die Varianz der quadrierten Residuen als signifikant erwiesen. Das multiplikative Modell lässt sich somit als

$$
g\left(\mathbb{E}(nmqm)\right) = f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add}
$$

mit $g(\mu) = log(\mu)$ darstellen. Man kann den Erwartungswert von nmqm für die Wohnung i folgendermaßen berechnen:

$$
\mathbb{E}(nmqm) = exp\left(f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add} \right)
$$

Stellt man die Formel um, ergibt sich:

$$\begin{aligned}\mathbb{E}(nmqm) =& exp\left(f_1(wfl.gekappt)\right) \cdot exp\left(f_2(bjahr.katmean.imp)\right.\\ & \cdot exp\left(\beta_1 \cdot Faktor.ML.WL\right) \cdot exp\left(\beta_2 \cdot Gebaeudetyp\right) \cdot \ldots \cdot exp\left(\beta_{17} \cdot Terrasse\right)\end{aligned}$$

Es ist ersichtlich, dass die Parameter aus dem Modell nun multiplikativ zu interpretieren sind. Die erwartete Nettomiete pro Quadratmeter für eine Wohnung mit Terrasse ist $exp(\beta_{17})$ mal teurer, als die einer Wohnung ohne Terrasse, wobei die übrigen Kovariablen gleich bleiben. Allerdings ist zu beachten, dass jeder Parameter nur exponiert interpretiert werden kann. Der Erwartungswert für die Nettomiete ergibt sich analog zum additiven Modell durch die Multiplikation mit der Wohnfläche als

$$\mathbb{E}(nm) = wfl.gekappt \cdot exp\left(f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add}\right).$$

4 Ergebnisse der Modelle

Nachdem die beiden Mietspiegelmodelle im vorhergehenden Kapitel erklärt wurden, sollen nun deren Ergebnisse verglichen werden. Dazu wurden die Modelle mit Hilfe des Programmpaketes R und dem darin enthaltenen Paket mgcv geschätzt.

Vergleicht man die Schätzungen für die Parameter der beiden Modelle kann man erkennen, dass die Richtung des Einflusses in beiden Modellen bei allen Variablen, deren Einfluss linear geschätzt wurde, gleich ist. Das heißt, dass man beispielsweise für eine Wohnung mit Terrasse in beiden Modellen einen höheren Mietpreis erwartet, als für eine Wohnung mit den gleichen Eigenschaften aber ohne Terrasse. Man kann also keine großen Unterschiede zwischen den geschätzten Koeffizienten der beiden Modelle erkennen. Auch beim Betrag der geschätzten Parameter der beiden Modelle fallen keine bedeutenden Unterschiede ins Auge. So ist die Relation der Beträge für alle Koeffizienten in etwa gleich. Beträgt der Koeffizient im additiven Modell ungefähr 0.5, so ist der vergleichbare Koeffizient des multiplikativen Modells ungefähr 1.04. Für einen größeren Koeffizienten des additiven Modells ist auch der des multiplikativen Modells größer. Es scheint also, als wären die geschätzten Parameter der beiden Modelle zueinander proportional. Bei der Betrachtung der p-Werte fallen keine nennenswerten Unterschiede auf. Die p-Werte aller Koeffizientenschätzungen der beiden Modelle sind kleiner als 0.005 und somit signifikant zum Niveau 0.01.

Der Einfluss der nichtparametrisch geschätzten Wohnfläche kann, wie der Name bereits sagt, nicht auf den Wert eines Parameters reduziert werden. Allerdings ist es möglich, die geschätzte Funktion $f(wfl.gekappt)$ zu visualisieren. Diese Funktion sehen wir in Abbildung 1 für beide Modelle abgebildet. Die Interpretation der y-Werte der beiden Funktionen ist auf Grund der Zentrierung der Glättungsfunktionen nur schwer möglich. Im additiven Modell bedeutet $f(wfl.gekappt) = 1$, dass der erwartete Quadratmeterpreis der Wohnung im Vergleich zu einer Wohnung mit $f(wfl.gekappt) = 0$ und ansonsten gleichen Eigenschaften um einen Euro höher ist. Im multiplikativen Modell lässt sich $f(wfl.gekappt) = 0.1$ so interpretieren, dass der erwar-

Variable	Additives Modell	Multiplikatives Modell
Intercept	10.43	10.47
Faktor.ML.WL:0.3	0.61	1.06
Faktor.ML.WL:0.4	1.44	1.13
Faktor.ML.WL:1.2	0.73	1.08
Faktor.ML.WL:1.3	1.53	1.14
Gebaeudetyp:Hochhaus	-0.51	0.96
Gebaeudetyp:Stadthaus	0.77	1.06
Gebaeudetyp:Wohnblock	-0.60	0.94
Haustyp.15:EinfacherAltbau-a	-2.18	0.81
Haustyp.15:EinfacherAltbau-b	-1.14	0.90
Haustyp.15:EinfacherNachkriegsbau	-0.50	0.96
WW.Bereitung:fehlend	-1.63	0.80
WW.Bereitung:unvollstaendig	-0.64	0.95
Heizung.15:unvollstaendig	-0.73	0.94
Sanitaerbereich:Besondere.Badausstattung	0.47	1.04
Sanitaerbereich:Keine.Badausstattung	-1.17	0.87
Sanitaerbereich:Zweites.Bad	0.85	1.07
Modernisierung	1.32	1.11
Kueche.Ausstattung:Gute	0.81	1.07
Gaskochfeld.15	-0.84	0.90
Boden.einfach	-1.88	0.81
Boden.einfach.teilweise	-1.23	0.88
GuterBoden.neu	0.53	1.04
Terrasse	0.46	1.04
Keine.Gegensprech	-0.50	0.95
Fussboden.Heizung	0.42	1.04
Rolllaeden.elektrisch	0.68	1.05
Hohe.DG	1.12	1.14

Tabelle 2: Geschätzte Parameter für alle linear geschätzten Variablen
In der ersten Spalte finden sich die Variablen und deren Ausprägungen wieder. Nicht aufgeführt sind deren Referenzkategorien, die zur Interpretation der geschätzten Parameter notwendig sind. Die hier dargestellten Parameter des multiplikativen Modells wurden bereits exponiert, die des additiven sind unverändert.

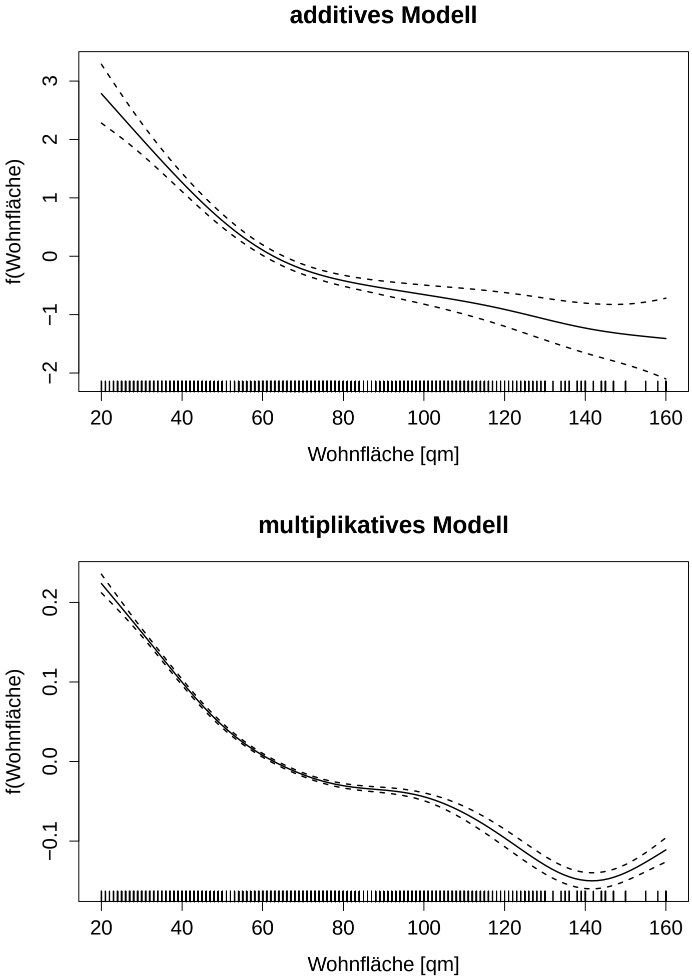

Abbildung 1: Geschätzter Einfluss der Wohnfläche ($f_2(wfl.gekappt)$) auf die Nettomiete pro Quadratmeter

tete Quadratmeterpreis der Wohnung, im Vergleich zu einer Wohnung mit $f(wfl.gekappt) = 0$ und ansonsten gleichen Eigenschaften, um $0.1 = 10\%$ teurer ist. Hierbei ist immer die Wohnfläche, bei der die Glättungsfunktion den Wert 0 annimmt, als Referenz zu betrachten. Diese Wohnfläche ist nicht die durchschnittliche Wohnfläche. Interessanter als die Werte der Funktion ist deren Verlauf. Man erkennt, dass beide Funktionen für Werte unter 140 eine fallende Tendenz aufweisen. Beide Funktionen sind bis etwa 65 Quadratmeter im positiven Bereich. Übereinstimmend fallen beide Funktionen bis zum Wert 70 qm stärker als für größere Werte. Für große Wohnungen ab 100 qm unterscheiden sich die Funktionen. Während das additive Modell einen eher linearen Verlauf schätzt, weist das multiplikative Modell ein Minimum bei ca. 140 qm auf. Allerdings sollten die Unterschiede in diesem Bereich nicht überbewertet werden, da hierbei deutlich weniger Wohnungen zur Schätzung zur Verfügung standen, was man auch am Konfidenzband des additiven Modells sehen kann. Wichtig ist jedoch, dass keine nennenswerten Unterschiede im Bereich zwischen 20 und 100 qm, in dem etwa 90 Prozent der Wohnungen liegen, ersichtlich sind.

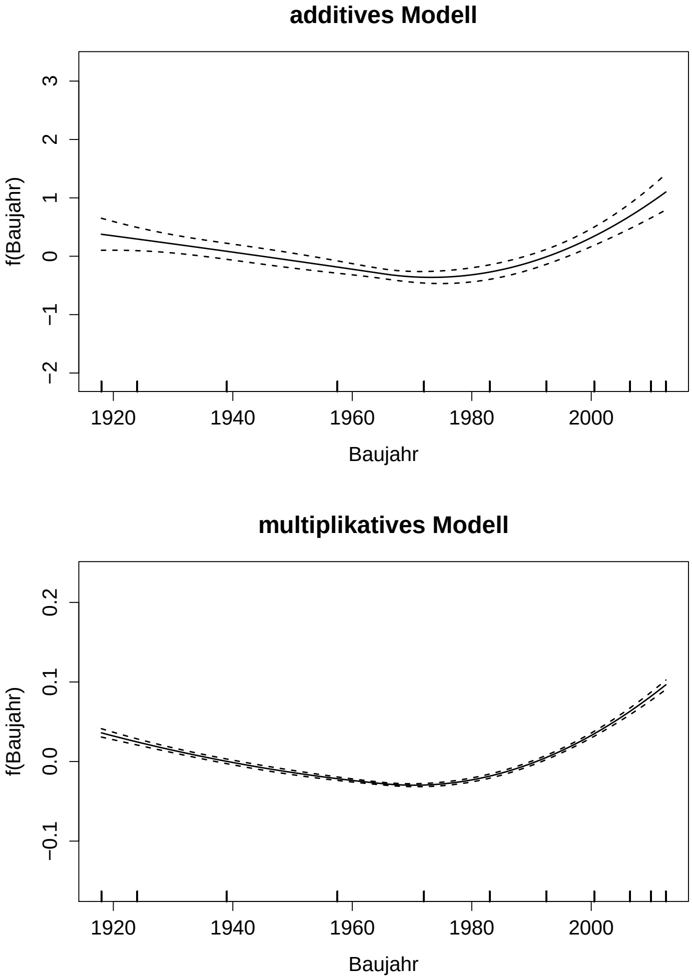

Abbildung 2: Geschätzter Einfluss des Baujahres ($f_2(bjahr.katmean.imp)$) auf die Nettomiete pro Quadratmeter

Auch das Baujahr wurde nichtparametrisch geschätzt und ist in Abbildung 2 dargestellt. Die beiden Funktionen unterscheiden sich optisch kaum. Beide weisen ein Miniumum bei ca. 1970 auf, wobei die Funktion nach 1970 stärker steigt als sie vor 1970 fällt. Das bedeutet, dass der durchschnittliche Quadratmeterpreis für eine neue Wohnung höher ist als für eine Wohnung mit gleichen Eigenschaften, die um 1970 entstanden ist. Betrachtet man eine Wohnung, die früher als 1970 errichtet wurde, erhöht sich der erwartete Quadratmeterpreis wieder im Vergleich zu einer vergleichbaren Wohnung von 1970, erreicht aber durchschnittlich nicht die Höhe des Preises einer sehr neuen Wohnung.
Abschließend kann man sagen, dass der Einfluss aller Variablen in beiden Modellen sehr ähnlich geschätzt wurde. Es ist also davon auszugehen, dass jede Variable sowohl im additiven Mietspiegelmodell als auch im multiplikativen Modell einen ähnlichen Einfluss hat. Daher lässt sich beim reinen Betrachten der Ergebnisse noch kein Urteil fällen, welches Modell zu bevorzugen ist.

5 Modellvalidierung

Auch wenn die Modelle sich auf den ersten Blick hinsichtlich ihrer Schätzwerte nicht unterscheiden, ist es möglich, dass sich ein Modell als deutlich besser erweist als das andere. Einerseits können die Modelle Annahmen, die bei der Modellierung getroffen wurden, verletzen, andererseits kann sich die Modellgüte unterscheiden. In diesem Kapitel soll nun zuerst eine Modelldiagnose der beiden Modelle durchgeführt und deren Ergebnis miteinander verglichen werden. Anschließend werden ausgewählte Gütemaße für beide Modelle berechnet und so die Güte der Modelle bestimmt.

5.1 Modelldiagnose

Eine Modelldiagnose dient dazu, die Richtigkeit der Modellierung zu überprüfen. Dieses Verfahren ist bei der Erstellung eines Modells unerlässlich. Soll ein Modell zuverlässige Ergebnisse liefern, muss die Modelldiagnose zeigen, dass die Modellannahmen zufriedenstellend eingehalten wurden. Die wichtigste Größen im Rahmen der Modelldiagnose sind die Residuen.

5.1.1 Residuen im GAM

Das Modell hat zum Ziel möglichst viel der Streuung der Zielvariable durch alle Kovariablen zu erklären. Jedoch kann meist nicht die gesamte Streuung dadurch erklärt werden. Die Teile, die nicht durch das Modell erklärt werden, sind die Residuen. Sie sind sehr wichtig um Modellannahmen zu überprüfen. Ähnlich wie im linearen Modell sind die Residuen des generalisierten additiven Modells als $\hat{\epsilon}_i = y_i - \hat{\mu}_i$ definiert. Bei generalisierten Modellen, wie dem generalisierten additiven Modells, kann es jedoch, je nachdem welche Verteilung für die Zielgröße angenommen wurde, zu Problemen bei der Diagnose des Modells kommen. So sollten die Residuen im Poissonmodell beispielsweise proportional zu den Werten der Zielgröße steigen. Dabei ist allerdings nur schwer zu überprüfen in welchem Verhältnis die Residuen steigen. Aus diesem Grund werden für generalisierte Modelle üblicherweise nur speziell standardisierte Residuen verwendet. Die üblichen Residuen für generalisierte Modelle sind die Pearson-Residuen und die Devianz-Residuen (S. 73 Wood, 2006). Die Pearson-Residuen entstehen durch die Standardisierung der einfachen Residuen:

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Die Pearson-Residuen sind damit eine Möglichkeit das Modell zu überprüfen jedoch weisen sie in der Praxis nicht immer die gewünschten Eigenschaften auf. Oftmals besser geeignet sind die Devianz-Residuen:

$$\hat{\epsilon}_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i}.$$

Dabei ist $D = \sum_{i=1}^n d_i$ die Devianz und damit d_i der Teil der Devianz, der der i-ten Beobachtung zugeordnet wird. Die Devianz kann wie die Residuenquadratsumme des linearen Modells gesehen werden und die Devianz-Residuen als Komponenten der Devianz bilden damit das Analogon zu den Residuen im linearen Modell. Vorteilhaft ist zudem, dass die Devianz-Residuen eines

korrekt gefitteten Modells standardnormalverteilt sind, also $\epsilon_i^d \sim N(0, 1)$.
Allgemein gelten für beide Arten von Residuen ähnliche Annahmen, deren Gültigkeit im gefitteten Modell durch Residualplots überprüft werden können:

1. Der Mittelwert der Residuen ist 0.

2. Der Mittelwert der Residuen sollte keinen Trend in Abhängigkeit von y oder einer beliebigen Kovariable aufweisen.

3. Die Varianz der Residuen sollte keinen Trend in Abhängigkeit von y oder einer beliebigen Kovariable aufweisen.

4. Die Residuen sollten annähernd einer Standardnormalverteilung folgen. Das bedeutet insbesondere, dass sie symmetrisch verteilt sein sollten.

Um diese Annahmen für die Residuen zu überprüfen werden üblicherweise Diagnoseplots verwendet. Besonders interessant sind dabei die Annahmen 2, 3 und 4.

5.1.2 Normal-Quantil-Plots

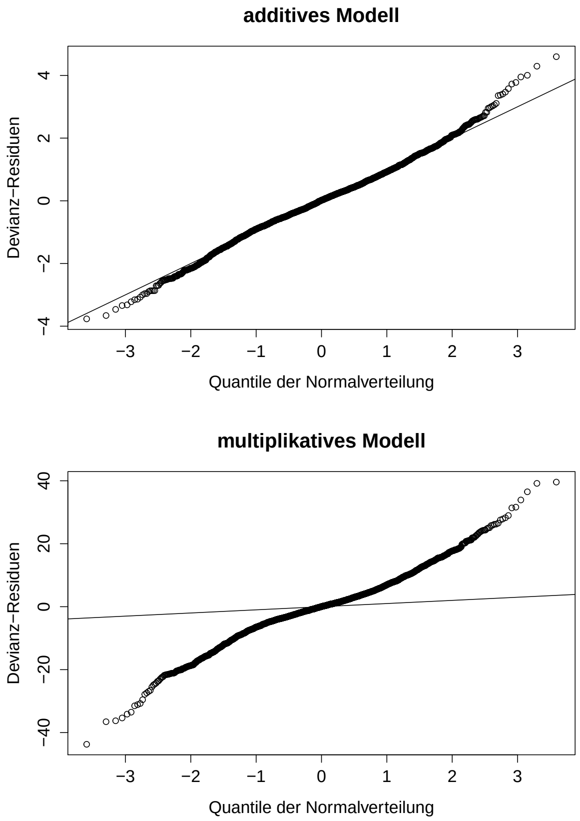

Abbildung 3: Normal-Quantil-Plots der Devianz-Residuen für das additive und das multiplikative Modell

Mit einem Normal-Quantil-Plot kann eine beliebige Verteilung mit einer Standardnormalverteilung verglichen werden. Hier soll die Verteilung der Residuen betrachtet werden. Befinden sich die eingezeichneten Punkte auf der Diagonale, so folgen die Residuen der Standardnormalverteilung und haben damit auch den Mittelwert 0.
Man kann in Abbildung 3 erkennen, dass die Residuen des additiven Modells zwar symmetrisch verteilt sind, jedoch etwas breitere Enden haben. Da die Abweichung jedoch nicht stark ist, ist kein schwererer Verstoß gegen die Modellannahmen zu erkennen. Die Devianz-Residuen des multiplikativen Modells jedoch weichen deutlich von der Standardnormalverteilung ab, was gegen eine gute Modellierung spricht. Um die Symmetrie zu untersuchen, wurden die Residuen daher in Abbildung 4 nochmals standardisiert verwendet.

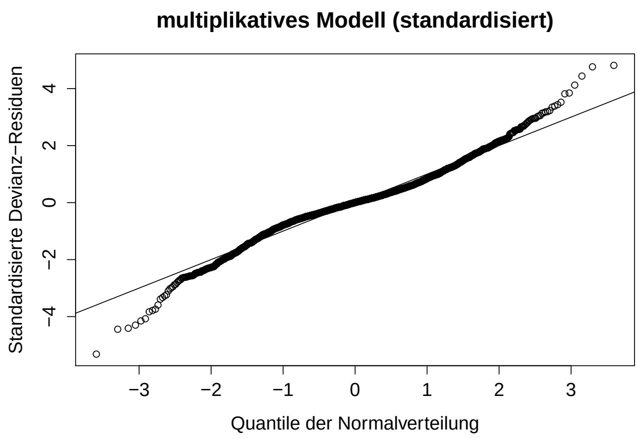

Abbildung 4: Normal-Quantil-Plot der standardisierten Davianz-Residuen des multiplikativen Modells

Es lässt sich erkennen, dass die Verteilung der Residuen des multiplikativen Modells annähernd symmetrisch ist und wieder deutlich breitere Enden hat. Allerdings bilden die abgebildeten Punkte eine schwankende Linie, die vor allem im negativen Teil starken Änderungen unterworfen ist. Man kann hier zwar erneut keine deutlich fehlerhafte Annahme feststellen, jedoch ist offensichtlich, dass die Residuen des additiven Modells der geforderten Verteilung eher entsprechen als die des multiplikativen Modells.

5.1.3 Residuenplots

Zur Überprüfung der Annahmen 2 und 3 werden Residuenplots verwendet. Dazu werden die Residuen gegen die gefitteten, also die vorausgesagten, Werte $\hat{y}_i$ aufgetragen. Die Annahmen sind dann für die Zielgröße erfüllt, falls die Residuen homoskedastisch um die 0 streuen. Dies ist an einer strukturlosen Punktwolke, deren Mittelpunkt auf Höhe der 0 ist, erkennbar.

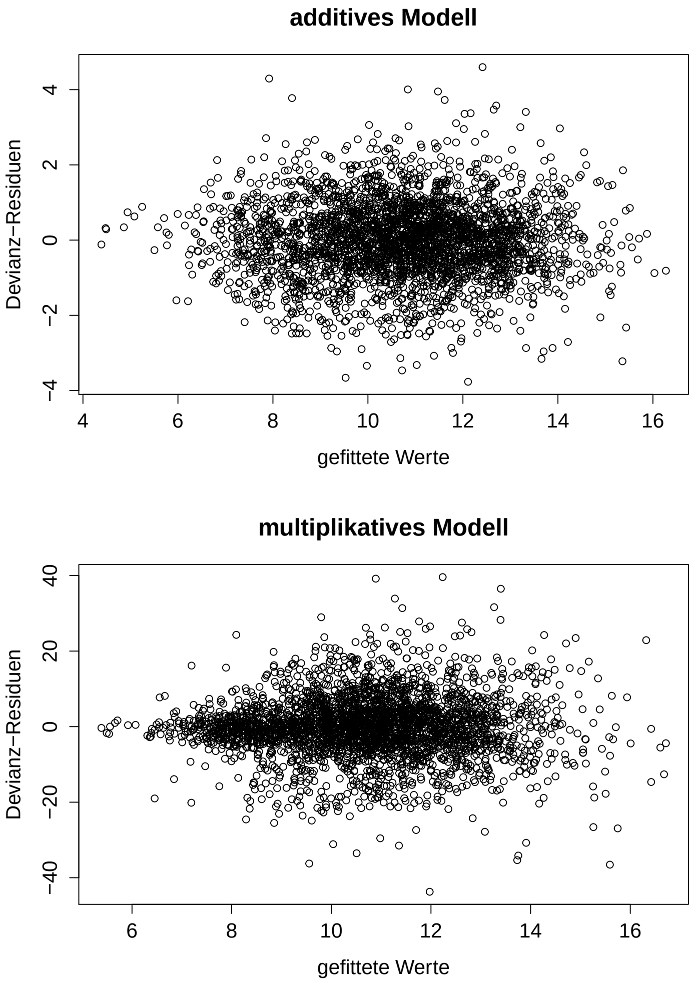

Abbildung 5: Residuenplots des additiven und multiplikativen Modells

Wenn man den Residuenplot des additiven Modells betrachtet, erkennt man eine gleichmäßige Punktwolke ohne strukturelle Besonderheiten. Die Residuen streuen gleichmäßig um die 0 und geben weder Anlass eine Varianzinhomogenität noch eine Abhängigkeit des Mittelwertes zu attestieren. Man kann also durchaus davon ausgehen, dass die beiden Annahmen für das additive Modell erfüllt sind. Der Plot des multiplikativen Modells hingegen lässt eine leichte Abhängigkeit der Varianz des vorhergesagten Wertes der Zielgröße erahnen. So scheint die Varianz der Residuen mit dem geschätzten Quadratmeterpreis zu steigen. Besonders deutlich wird das an den seitlichen Rändern der Grafik. Auf der linken Seite streuen die Werte sehr nahe um die 0, während auf der rechten Seite die Residuen deutlich weiter streuen. Damit kann ein Verstoß

gegen die 3. Annahme festgestellt werden. Der Mittelwert scheint jedoch nicht von der Zielgröße abzuhängen.
In diesem Kapitel wurde der Einfluss der Kovariablen auf die Varianz nicht überprüft, da die Gewichtung durch die zweistufige KQ-Schätzung bereits einer Varianzinhomogenität vorbeugt. Dadurch ist sichergestellt, dass die Kovariablen keinen Einfluss mehr auf die Varianz der Residuen nehmen.
Die Modelldiagnose führt zu dem Ergebnis, dass das additive Modell die Modellannahmen besser einhält als das multiplikative Modell. Während das additive Modell keinen Grund zum Anzweifeln der Annahmen liefert, zeigt das multiplikative Modell sowohl bei der Annahme der Varianzhomogenität als auch bei der Verteilung der Residuen Anzeichen einer fehlerhaften Modellierung. Bei der bloßen Betrachtung der Modelldiagnose ist also das additive Modell zu bevorzugen.

5.2 Modellvergleich anhand von Gütekriterien

Nachdem gerade die Annahmen der beiden Modelle überprüft wurden, soll nun auch die Güte der beiden Modelle untersucht werden. Um die Güte zweier Modelle zu vergleichen, bieten sich verschiedene Maße an, mit denen ein Schluss auf die Aussagekraft eines Modells gezogen werden kann. Das wohl bekannteste Gütekriterium dabei ist das Bestimmtheitsmaß R^2, anhand dessen man erkennen kann, wie viel Prozent der Streuung innerhalb der Zielgröße durch das Modell erklärt wird. In generalisierten Modellen kann das Bestimmtheitsmaß jedoch nur in Spezialfällen verwendet werden. Allerdings existieren auch Gütemaße, mit deren Hilfe auch im generalisierten additiven Modell ein Modellvergleich durchgeführt werden kann.

5.2.1 AIC

Eine Möglichkeit, die Güte zweier generalisierter additiver Modelle zu vergleichen, ist der Vergleich des Akaike-Informations-Kriteriums beider Modelle, kurz AIC (S. 477 Fahrmeir u. a., 2009) und (S. 111-113 Wood, 2006). Das Akaike-Informations-Kriterium ist für ein Modell mit dem Parametervektor $\Theta = (\Theta_1, ..., \Theta_p)$ durch die Formel

$$AIC = 2 \cdot (-l(\hat{\Theta})) + 2p$$

definiert. Der erste Teil der Formel besteht aus der Log-Likelihood $l(\hat{\Theta})$. Je besser das Modell an die Daten angepasst ist, desto größer wird die Log-Likelihood. Allerdings erhöht sich die Log-Likelihood meist bei dem Hinzufügen eines weiteren Parameters in das Modell. Um eine Überanpassung zu verhindern, besitzt das AIC mit $2p$ noch einen Strafterm, der eine zu große Menge an Parametern bestraft. In einem semiparametrischen Modell wie dem GAM wird der zweite Teil der Formel $2p$ üblicherweise durch die „'effektive' Dimension“(S. 477 Fahrmeir u. a., 2009) $df = sp(S)$ ersetzt, wobei S die Glättungsmatrix des Modells ist. Da die Log-Likelihood, die die Anpassung des Modells beschreibt, negativ in das AIC eingeht, ist ein Modell mit niedrigerem AIC dem mit höheren AIC vorzuziehen. Das Akaike-Informations-Kriterium eignet sich

damit besonders zur Variablenselektion, ist aber auch sehr gut zum Modellvergleich geeignet. Dafür spielt die Linkfunktion keine Rolle, allerdings verliert das AIC bei Modellen mit unterschiedlichen Zielgrößen oder Verteilungsfamilien seine Vergleichbarkeit.
Vergleicht man das AIC des additiven Mietspiegelmodells mit dem des multiplikativen Mietspiegelmodells, lässt sich ein deutlicher Unterschied erkennen. Das additive Modell weist ein AIC von 12730.46 auf, wobei das multiplikative Modell einen Wert von 13764.44 ergibt. Nachdem das AIC des multiplikativen Modells mehr als 1000 größer ist als das des additiven Modells, muss man davon ausgehen, dass ein Modell, das den Einfluss additiv schätzt, einem multiplikativen Modell überlegen ist.

5.2.2 Proportional deviance explained

Die Devianz ist ein Maß für die Anpassungsgüte in Likelihood-basierten Modellen wie dem generalisierten additiven Modell, das ähnlich wie die Residuenquadratsumme im gewöhnlichen linearen Modell interpretiert werden kann. Formal lässt sich die Devianz (S. 70 Wood, 2006) als

$$D = 2[l(\hat{\beta}_{max}) - l(\hat{\beta})]\Phi$$

darstellen, wobei $l(\hat{\beta}_{max})$ der maximalen Likelihood des saturierten Modells, also dem Modell, das ebenso viele Parameter schätzt wie es Daten gibt, und $l(\hat{\beta})$ der maximalen Likelihood des geschätzten Modells entspricht. Ein Modell mit einer niedrigeren Devianz erklärt die Daten damit besser als ein Modell mit hoher Devianz. Dies gilt jedoch nicht, falls für die Modelle unterschiedliche Linkfunktionen verwendet werden. Da dies beim Vergleich des multiplikativen und additiven Mietspiegelmodells der Fall ist, kann hier die Devianz nicht direkt verglichen werden. Eine Möglichkeit die Devianz zweier Modelle mit unterschiedlichen Linkfunktionen vergleichbar zu machen ist der proportional of deviance explained, der Anteil der Devianz, der durch das Modell erklärt wird (S. 84 Wood, 2006).

$$PDE = (D_0 - D_R)/D_0$$

Die PDE besteht aus der Null-Devianz D_0 und der Residual-Devianz D_R. Die Null-Devianz ist die Devianz eines Intercept-Modells, also mit nur einem konstanten Teil, und die Residual-Devianz ist die Devianz des betrachteten Modells. Damit lässt sich der PDE als der Anteil der Devianz, also der Streuung der Daten, betrachten, die durch das Modell erklärt werden. Damit bildet der proportional of deviance explained eine Generalisierung des Bestimmtheitsmaßes aus dem linearen Modell.
Aus der Berechnung der beiden Modelle in dieser Arbeit ergibt sich der proportional of deviance explained

$$PDE = (5771.967 - 3066.085)/5771.967 = 0.469$$

für das additive Modell und

$$PDE = (338864.8 - 203719.4)/338864.8 = 0.399$$

für das multiplikative Modell. Diese Werte sagen aus, dass das additive Modell 46.9% der Devianz erklären kann, wohingegen das multiplikative Modell mit 39.9% einen deutlich geringeren Anteil erklärt.
Es zeigt sich also, dass das additive Modell in Bezug auf das AIC und die PDE deutlich besser abschneidet als das multiplikative Modell. Damit beweist sich das additive Modell sowohl bei der Diagnose als auch beim Vergleich durch die Gütekriterien als das geeignetere Mietspiegelmodell.

6 Kreuzvalidierung

Im vorhergehenden Kapitel wurden die beiden Modelle bereits nach verschiedenen Gütekriterien und Maßen verglichen. Dabei wurden die beiden Modelle jeweils auf dem kompletten Datensatz der Erhebung im Rahmen des Münchner Mietspiegels von 2015 erstellt. Dieser Datensatz ist zwar repräsentativ für den Münchner Wohnungsmarkt, ist aber dennoch nur ein kleiner Teil der Grundgesamtheit (vgl. Rösch u. a., 2015). Orientiert man sich ausschließlich am fertig gefitteten Modell und einem Teildatensatz, kann es zu Overfitting kommen. Das bedeutet, dass sich das Modell so gut an die Daten anpasst, dass auch zufällige Abweichungen vom Mittelwert des Modells erklärt werden. Ist das der Fall, verliert das Modell seine genaue Vorhersagekraft für die Grundgesamtheit und lässt sich nur noch auf die Stichprobe korrekt anwenden. Doch genau dieser Punkt ist bei Mietspiegelmodellen entscheidend, schließlich ist es das Ziel den Mietpreis einer Wohnung vorherzusagen, selbst wenn deren Daten nicht zur Modellierung verwendet wurden. Dieses Problem lässt sich mit einer Kreuzvalidierung untersuchen.
Eine Kreuzvalidierung ist ein Verfahren, mit dem es möglich ist die Prognosefähigkeit eines Modells zu testen. Ebenso kann damit ein Rückschluss auf die Güte eines Modells gezogen werden. Die Grundidee der Kreuzvalidierung ist es den vorhandenen Datensatz in zwei Teile zu teilen, einen Testdatensatz oder Trainingsdatensatz und einen Validierungsdatensatz oder Prognosedatensatz. Der Trainingsdatensatz bildet dabei die Grundlage für die Modellschätzung während der Validierungsdatensatz zur Bestimmung des Prognosefehlers verwendet wird. Durch die Aufteilung in mehrere Teile wird die Situation, bei der ein Teil der Grundgesamtheit erhoben wurde und ein Teil unbekannt bleibt aber trotzdem zur Vorhersage verwendet werden soll, imitiert (S. 161 Fahrmeir u. a., 2009). In der Praxis besteht eine Fülle an Varianten der Kreuzvalidierung, die auf diesem Grundprinzip beruhen. Besonders prominente Beispiele hierfür sind die k-fache Kreuzvalidierung und das Leave-One-Out-Verfahren, ein Spezialfall der k-fachen Kreuzvalidierung.

6.1 Die k-fache Kreuzvalidierung

Die k-fache Kreuzvalidierung ist die wohl meist benutzte Kreuzvalidierungsmethode. Sie funktioniert nach folgendem Vorgehen (S. 161-162 Fahrmeir u. a., 2009) und (S. 409-413 Kuhlmann, 2007):

1. Der vollständige Datensatz mit n Beobachtungen wird in k Teildatensätze zerlegt. Dabei soll jeder Teildatensatz 1,...,K ungefähr die gleiche Größe haben. Verwendet man eine

stratifizierte k-fache Kreuzvalidierung, muss dabei darauf geachtet werden, dass die Teildatensätze jeweils die gleiche Verteilung haben. Das ermöglicht die Verringerung der Varianz des Prognosefehlers. Außerdem kann es unter Umständen passieren, dass im Validierungsdatensatz Kombinationen der Kovariablen vorkommen, die im Trainingsdatensatz nicht vorhanden sind. Im Folgenden wird trotzdem auf eine Stratifizierung verzichtet, da der Datensatz mit 3065 Beobachtungen groß genug ist, um diese Anforderungen auch ohne Stratifizierung einzuhalten.

2. Als Validierungsdatensatz wird zunächst der erste Teildatensatz bestimmt. Alle anderen k-1 Datensätze bilden die Trainingsdaten. Nun wird das zu überprüfende Modell auf dem Trainingsdatensatz gefittet. Anschließend wird das daraus resultierende Modell zur Prädiktion der Zielgröße für die Validierungsdaten verwendet. Die daraus prognostizierten Werte und die tatsächlichen Werte der Validierungsdaten werden dazu verwendet an dieser Stelle einen geeigneten Prognosefehler (siehe Abschnitt 6.3) zu berechnen.

3. Nacheinander wird jetzt der 2. Schritt für jeden Teildatensatz wiederholt. Das bedeutet, zunächst wird statt dem ersten der zweite Teildatensatz als Validierungsdatensatz ausgewählt, anschließend der dritte, solange bis alle k Teildatensätze genau einmal die Validierungsstichprobe gebildet haben. In jedem Schritt soll dabei der Prognosefehler bestimmt werden.

4. Jetzt wird der mittlere Prognosefehler über alle k Durchgänge von Schritt 2 hinweg bestimmt. Der mittlere Prognosefehler ergibt sich aus dem arithmetischen Mittel der berechneten Fehler in jedem Schritt der Kreuzvalidierung. Das Modell mit dem kleinsten mittleren Prognosefehler ist am Ende auszuwählen.

Bei der Leave-One-Out-Kreuzvalidierung werden die Teildatensätze in Schritt 1 so bestimmt, dass jede der n Beobachtungen einen eigenen Teildatensatz bildet. Das bedeutet, dass für jede einzelne Beobachtung ein Modell mit Hilfe aller anderen Beobachtungen des Datensatzes geschätzt und damit der Prognosefehler berechnet wird.

6.2 Wahl des Verfahrens

Grundsätzlich ist für den Vergleich des additiven Mietspiegelmodells und des multiplikativen Mietspiegelmodells eine k-fache Kreuzvalidierung mit jedem beliebigen $k\epsilon 2, ..., n$, also auch einer Leave-One-Out-Kreuzvalidierung möglich. Gängige Werte für k sind 5, 10, 20 und n, also eine Leave-One-Out-Kreuzvalidierung. Der Vorteil des Leave-One-Out-Verfahrens ist, dass der Prognosefehler nicht von der Wahl der Teildatensätze in Schritt 1 abhängt. Außerdem wird für die Schätzung des Modells in jedem Schritt die maximale Anzahl an Beobachtungen verwendet. Damit ist sichergestellt, das auch die nichtparametrischen Teile des Modells noch sehr gut geschätzt werden können. Jedoch bedeutet das auch, dass die Validierungsdatensätze jeweils sehr klein sind. Außerdem ist das Verfahren sehr rechenintensiv. Breiman und Spector (S. 306 Breiman, 1992) schließen zudem aus einer Simulationsstudie, dass schon eine 10-fache Kreuzvalidierung dem Leave-One-Out-Verfahren hinsichtlich der Eignung zur Modellwahl und -bewertung überlegen sein kann. Mit über 3000 Beobachtungen ist der Datensatz auch groß genug, um auch mit kleineren k gute Ergebnisse zu erzielen. Da das Modell mit 19 Einflussgrößen, von denen 2 nichtparametrisch geschätzt werden, sehr komplex ist, sollte der Trainingsdatensatz dennoch nicht zu klein werden. Die Entscheidung fällt damit auf $k = 20$. Hierbei dienen jeweils 95% der Daten, also ungefähr 2912 Beobachtungen als Trainingsdaten und 5% (153 Beobachtungen) als Validierungsstichprobe. Da die Ergebnisse dieses Verfahrens jedoch von der Aufteilung der Datensätze abhängt, wird die 20-fache Kreuzvalidierung 10 mal mit unterschiedlicher Datensatzaufteilung durchgeführt. Durch dieses Verfahren sinkt der Einfluss der Wahl der Datensätze und wird damit zuverlässiger.

6.3 Prognosefehler

Von entscheidender Bedeutung ist auch das gewählte Prognosemaß zur Bestimmung des Prognosefehlers. Auch dafür lassen sich in der Literatur viele verschiedene Beispiele finden, von denen alle ihre Stärken und Schwächen haben. Bisher konnte sich noch kein Maß als das „Beste“ erweisen. Deshalb empfiehlt Barrot (S. 429 Barrot, 2007) die Verwendung mehrerer Maße. Im Rahmen der folgenden Kreuzvalidierung sollen vier Prognosemaße verwendet werden. Der Mean Absolute Percentage Error, kurz MAPE (S. 419 Barrot, 2007), ergibt sich durch

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{x_i}\right| \cdot 100\%,$$

wobei n die Anzahl der Prognosen, $\hat{y}_i$ die i-te Prognose und y_i der wahre Wert der Variablen ist. Dieses Maß gibt die durchschnittliche absolute prozentuale Abweichung der Prognose $\hat{y}$ zum wahren Wert y an. $MAPE = 0$ bedeutet also, dass die Prognose perfekt ist, $MAPE = 1$ hingegen bedeutet, dass die Prognose durchschnittlich um 1% vom wahren Wert abweicht. Dabei spielt es keine Rolle, ob die Prognose kleiner oder größer als der wahre Wert ist. Der Median Absolute Percentage Error, auch MdAPE (S. 428 Barrot, 2007), hat die Form

$$MdAPE = med\left(\left|\frac{\hat{y}_i - y_i}{x_i}\right|\right) \cdot 100\%$$

und unterscheidet sich vom MAPE dadurch, dass nicht der Mittelwert, sondern der Median der Abweichungen interessant ist. Das bewirkt eine Unempfindlichkeit gegenüber Ausreißern, aber auch eine schlechtere Interpretierbarkeit. So spricht $MdAPE = 0$ nicht wie beim MAPE für eine perfekte Vorhersage sondern lediglich dafür, dass mindestens die Hälfte der Vorhersagen nicht vom wahren Wert abweichen. $MdAPE = 1$ heißt analog dazu, dass mindestens die Hälfte aller vorhergesagten Werte um maximal 1% vom wahren Wert abweichen. Ein einfacheres Maß für den Prognosefehler ist der Mean Error (ME) (S. 418 Barrot, 2007)

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i).$$

Dieser mittlere Fehler gibt an um wie viel die Prognose durchschnittlich von der Wahrheit abweicht. Im Gegensatz zu den ersten beiden Prognosemaßen wird dieser Fehler in der Einheit der betrachteten Variable gemessen und kann sowohl negativ als auch positiv sein. Ein Wert von $ME = 0$ bedeutet, dass sich die Über- und Unterschätzung der Prognose aufheben und nicht notwendiger Weise, dass die Vorhersage perfekt ist. Mit Hilfe dieses Maßes lässt sich auch erkennen, ob die Abweichungen im Mittel 0 sind oder ob tendenziel über- oder unterschätzt wird. Das bekannteste und am weitesten verbreitete Maß für die Prognosegüte ist der RMSE (S. 420 Barrot, 2007)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}.$$

Der Root Mean Squared Error ist bei einer perfekten Vorhersage 0. Je schlechter die Vorhersage, desto größer wird der RMSE.
Allgemein gilt für alle 4 Maße, dass das Modell, deren Wert näher an der 0 liegt, zu bevorzugen ist.

6.4 Durchführung für den Datensatz

Die in Abschnitt 6.1 beschriebene k-fache Kreuzvalidierung wurde für die beiden Modelle in R 10 mal durchgeführt. Die Mietspiegelmodelle entsprechen dabei genau denen, die auch für die bisherigen Analysen verwendet wurden. Auch die Gewichtung der Modelle durch die zweistufige KQ-Schätzung wurde beibehalten. Anders als bei der Erstellung der Mietspiegelmodelle im Abschnitt 3.5 und 3.6 wurde hier der Einfluss der Kovariablen auf die Residuen nicht mehr überprüft. In jedem Schritt der Kreuzvalidierung wurden also die Gewichte auf die gleiche Art und Weise geschätzt. Die Ergebnisse dieser Schätzung der Gewichte unterscheiden sich jedoch leicht von denen im Originalmodell, da nicht alle Daten zu deren Schätzung verwendet wurden. Da auf eine Variablenselektion im Rahmen der Regression auf den Residuen verzichtet wurde, kann es vorkommen, dass die Gewichte auch mit Variablen, die in diesem Schritt nicht signifikant auf die Residuen wirken, geschätzt werden. Dies ist jedoch unproblematisch, da die Schätzung dadurch nicht verfälscht wird. Die Kreuzvalidierung wurde nach dem oben beschriebenen Vorgehen als10 mal wiederholte 20-fache Kreuzvalidierung durchgeführt. Die Prognosefehler wurden auf Basis der Nettomieten berechnet und nicht wie in den bisherigen Analysen auf der Nettomiete pro Quadratmeter. Schließlich soll mit dem Mietspiegel in der Praxis die Nettokaltmiete und nicht der Quadratmeterpreis geschätzt werden. Die Nettomiete nm ergibt sich dabei aus dem Datensatz durch die Multiplikation der Variablen $wfl.gekappt$ und $nmqm$. Analog dazu ergibt sich die vorhergesagte Nettomiete durch das Produkt von $wfl.gekappt$ und der geschätzten Nettomiete pro Quadratmeter. Die mittleren Prognosefehler ergeben sich durch das arithmetische Mittel der mittleren Prognosefehlern aus jeder der 10 Durchführungen der 20-fachen Kreuzvalidierung.

6.5 Ergebnisse der Kreuzvalidierung

Die Kreuzvalidierung wurde mit Hilfe des Programmpaketes R durchgeführt. Daraus ergaben sich folgende Werte:

Prognosemaß	Additives Modell	Multiplikatives Modell
ME	0.25150	1.75387
MAPE	15.9069	15.9903
MdAPE	10.8770	10.9282
RMSE	152.809	154.220

Tabelle 3: Mittlere Prognosefehler des additiven und multiplikativen Modells in der 20-fachen Kreuzvalidierung

Die Ergebnisse zeigen, dass das additive Mietspiegelmodell von allen 4 Prognosemaßen als das bessere Modell angezeigt wird. Allerdings sind die Unterschiede nur sehr gering. Der mittlere Fehler ME zeigt an, dass die Nettomiete einer Wohnung im Rahmen der Kreuzvalidierung durch das multiplikative Modell um 1.75 Euro teurer als die wahre Miete geschätzt wurde. Das additive Modell überschätzte die Nettomiete durchschnittlich nur um 25 Cent. Betrachtet man diese 1.75 Euro in Relation zur durchschnittlichen Nettomiete von 760 Euro, erscheint der Fehler relativ gering. Dieser Eindruck bestätigt sich mit der mittleren absoluten prozentualen Abweichung MAPE. Daraus geht hervor, dass im additiven Modell die vorhergesagte Miete durchschnittlich um 15.91 Prozent von der wahren Miete abweicht. Der MAPE im multiplikativen Modell ist mit 15.99 Prozent nur um etwa 0.08 Prozentpunkte größer. Auch das Ausreißer-unempfindlichere Maß MdAPE weist das multiplikative Modell als das schlechtere aus. Die Hälfte der geschätzten Nettomieten weichen beim additiven Modell durchschnittlich um 10.88 Prozent von der wahren Miete ab, wohingegen dem multiplikativen Modell eine Abweichung um 10.93 Prozent bescheinigt wird. Der am häufigsten verwendete Prognosefehler ist das für Ausreißer anfällige Maß RMSE. Dieses Maß weist das additive Modell erneut als das Bessere aus. Bei der Analyse der Prognosekraft lässt sich also kein deutlicher Unterschied zwischen den beiden Modellen feststellen, allerdings zeigt sich eine leichte Tendenz zu Gunsten des additiven Modells.

7 Simulation

Bisher wurden die beiden Modelle auf ihre Eignung für die bestehenden Daten überprüft. Diese Daten werden allerdings für jeden Mietspiegel neu erhoben. Da ist es theoretisch möglich, dass sich auch die Struktur in den Daten so ändert, dass das multiplikative Modell etwas besser passen könnte. Zwar ist davon nicht auszugehen, aber dennoch ist es interessant, wie sich die Modelle verhalten, wenn sich die Struktur der Daten an das andere Modell anpasst. Gibt es ein Modell, das sich besser an die veränderte Struktur anpassen kann? Um diese Frage zu beantworten ist es notwendig die vorhandenen Daten zu verändern. Das Ziel ist es, einen Datensatz zu erhalten, bei dem der Einfluss der Kovariablen auf den Quadratmeterpreis exakt multiplikativ ist. In einem zweiten Datensatz sollen die Kovariablen exakt additiv wirken. Diese gewünschten

Eigenschaften erhalten wir, wenn wir die beiden Modelle auf den Münchner Mietspiegeldaten fitten und anschließend die Variable nmqm jeweils durch die geschätzten Werte ersetzen. Es ergeben sich dabei zwei neue Datensätze. Der additive Datensatz hat die Struktur:

$$\begin{aligned} nmqm_a =& f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \beta_1 \cdot Faktor.ML.WL \\ &+ \beta_2 \cdot Gebaeudetyp + \ldots + \beta_{17} \cdot Terrasse, \end{aligned}$$

wobei $nmqm_a = \widehat{nmqm}_{add}$ mit $\widehat{nmqm}_{add}$ den, durch das additive Modell vorhergesagten, Werten. f_1, f_2 und β_1 bis β_{17} entsprechen den in Kapitel 4 für das additive Modell geschätzten Funktionen und Parametern. Analog dazu ist die Struktur des zweiten Datensatzes gegeben durch

$$\begin{aligned} nmqm_m =& exp\left(f_1(wfl.gekappt)\right) \cdot exp\left(f_2(bjahr.katmean.imp)\right) \\ &\cdot exp\left(\beta_1 \cdot Faktor.ML.WL\right) \cdot exp\left(\beta_2 \cdot Gebaeudetyp\right) \cdot \ldots \cdot exp\left(\beta_{17} \cdot Terrasse\right), \end{aligned}$$

wobei $nmqm_m$ den vorhergesagten Werten durch das multiplikative Modell entspricht, also $nmqm_m = \widehat{nmqm}_{mul}$. f_1, f_2 und β_1 bis β_{17} entsprechen den in Kapitel 4 für das multiplikative Modell geschätzten Funktionen und Parametern. Diese beiden Datensätze bilden also extreme Beispiele für einen Wohnungsmarkt, bei dem die Nettomiete pro Quadratmeter durch die Eigenschaften additiv bzw. multiplikativ beeinflusst wird und für Wohnungen mit gleichen Eigenschaften nicht variiert. Nun werden diese Datensätze dazu verwendet, um die Modelle am jeweils anderen Datensatz zu fitten. Dabei soll der additive Datensatz zur Modellierung des multiplikativen Modells verwendet werden. Das additive Modell soll auf den multiplikativen Daten angepasst werden. Wir betrachten nun also das additive Modell auf den multiplikativen Daten

$$\mathbb{E}(nmqm_m) = f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add}$$

und das multiplikative Modell auf den additiven Daten

$$log\left(\mathbb{E}(nmqm_a)\right) = f_1(wfl.gekappt) + f_2(bjahr.katmean.imp) + \eta^{add}$$

In den Kapiteln 5 wurden bereits einige Möglichkeiten zum Vergleich zweier Modelle aufgezeigt. Das AIC kann jedoch nicht mehr zum Vergleich herangezogen werden, da die beiden Modelle nicht mehr die gleiche Zielgröße haben. Der proportional of deviance explained kann jedoch angewendet werden. Es ergeben sich die Werte $PDE_{add} = 0.9929$ für das additive Modell und $PDE_{mul} = 0.9928$ für das Multiplikative. Beide PDEs unterscheiden sich nur unwesentlich und zeigen an, dass beide Modelle etwa 99,3% der Devianz erklären. Nach Betrachtung dieses Maßes ergeben sich beide Modelle gleichermaßen genau bei der Modellierung der anderen Struktur. Auch die Prognosemaße aus Abschnitt 6.3 eignen sich zum Vergleich der beiden Modelle. Die Tabelle 4 enthält die berechneten Werte der Maße. Die Berechnung geschah analog zur Kreuzvalidierung auf Basis der Nettomiete und nicht der Nettomiete pro Quadratmeter. Im Gegensatz zur Kreuzvalidierung wird der Prognosefehler nicht als der mittlere Prognosefehler, sondern

Prognosemaß	Additives Modell	Multiplikatives Modell
ME	0.00000	-0.33870
MAPE	0.00968	0.01038
MdAPE	0.00544	0.00592
RMSE	9.95508	10.4808

Tabelle 4: Prognosefehler des additiven und multiplikativen Modells für die simulierten Daten

gleichzeitig aus allen Beobachtungen des Datensatzes bestimmt. Wir sehen hier, dass das additive Modell, das auf den multiplikativen Daten geschätzt wurde, für jedes Maß den niedrigeren Wert annimmt. Die Unterschiede sind allerdings absolut gesehen gering. Das liegt unter anderem daran, dass die verwendeten Daten keine zufälligen Fehler, die bei den Originaldaten teilweise sehr groß waren, mehr beinhalten. Dadurch verringert sich auch die Varianz der Daten und damit auch die Werte der Prognosemaße. Trotzdem reichen die beobachteten Unterschiede der Werte nicht aus, um das additive Modell mit Bestimmtheit als das besser geeignete für die jeweils andere Datenstruktur zu erkennen. Wie bei der Kreuzvalidierung lässt sich lediglich ein sehr leichter Vorteil für das additive Modell erkennen.

8 Zusammenfassender Vergleich

Um die Frage, ob ein additives oder ein multiplikatives Modell besser zur Erstellung eines Mietspiegels geeignet ist zu beantworten, wurden in dieser Arbeit eine Reihe von Analysen durchgeführt. Dabei wurden verschiedene Gütemaße der beiden Modelle verglichen, eine Modelldiagnose durchgeführt, die Vorhersagekraft mittels einer Kreuzvalidierung überprüft und zuletzt die Eignung für die Datenstruktur des anderen Modells getestet. In allen Bereichen schnitt das additive Modell besser ab. Teilweise waren die Ergebnisse jedoch sehr knapp, sodass kein eindeutiger Schluss gezogen werden kann. So ergaben sich nur sehr leichte Unterschiede beim Vergleich der Vorhersagekraft und der Eignung des additiven Modells für multiplikative Daten und anders herum. Deutlicher waren die Unterschiede beim Vergleich der Gütemaße. Hier konnte aufgezeigt werden, dass die Modellgüte für ein additives Modell höher ist als für ein multiplikatives. Auch bei der Diagnose der beiden Modelle erwies sich das additive Modell als geeigneter. Insgesamt scheint also die Annahme eines additiven Einflusses der Kovariablen der multiplikativen Annahme überlegen zu sein. Allerdings sind alle Ergebnisse in dieser Arbeit mit dem Hintergrund zu sehen, dass alle Untersuchungen auf den Daten und Modellen des Münchner Mietspiegels aufbauen. Das bedeutet, dass die Daten bereits so verändert wurden, dass beispielsweise extreme Ausreißer ausgeschlossen wurden. Außerdem wird im multiplikativen Modell derselbe Prädiktor wie im additiven Modell verwendet. Dieser Prädiktor enthält jedoch nur die Variablen, die bei der Variablenselektion bei der Erstellung des Münchner Mietspiegels von 2015 ausgewählt wurden. Die Frage, ob bei einer Variablenselektion am multiplikativen Modell andere Variablen zur Erklärung der Nettomiete ausgewählt worden wären, kann hier nicht beantwortet werden. Wäre das allerdings der Fall, wären andere Ergebnisse in manchen Teilen dieser Arbeit durchaus möglich. Trotzalledem kann man, vor allem aufgrund der Kapitel 5 und 6, ein additives Mietspiegelmodell als die bessere Variante der beiden untersuchten Modelle betrachten. Von einem multiplikativen Mietspiegelmodell im Sinne von 3.6 ist zudem abzuraten.

Literatur

[bun 2002] *Hinweise zur Erstellung von Mietspiegeln.* Berlin : Bundesministerium für Verkehr, Bau- und Wohnungswesen, 2002

[Barrot 2007] Barrot, C.: Prognosegütemaße. In: *Methodik der empirischen Forschung.* Gabler-Verlag, 2007, Kapitel 28

[Breiman 1992] Breiman, L. und P. S.: *Submodel Selection and Evaluation in Regression: The X-Random Case.* International Statistical Review,, 1992

[Fahrmeir u. a. 2009] Fahrmeir, L. ; Kneib, T. ; Lang, S.: *Regression - Modelle, Methoden und Anwendungen.* Berlin/Heidelberg : Springer-Verlag, 2009

[Gonzalez 2002] Gonzalez, M. J. R.: *Erstellung von Mietspiegeln: Ein Semiparametrisches Regressionsmodell zur Schätzung der ortsüblichen Vergleichsmiete*, Universität Freiburg in der Schweiz, Dissertation, 2002

[Hasti u. Tibshirani 1990] Hasti, T. ; Tibshirani, R.: *Generalized Additive Models.* Boca Raton, FL : Chapman & Hall/CRC, 1990

[Kuhlmann 2007] Kuhlmann, J.: Ausgewählte Verfahren der Holdout- und Kreuzvalidierung. In: *Methodik der empirischen Forschung.* Gabler-Verlag, 2007, Kapitel 27

[Rösch u. a. 2015] Rösch, B. ; Sauer, A. ; Langfritz, M.: Statistische Analyse der Nettomiete. In: *Mietspiegel für München ©2015. Statistik, Dokumentation und Analysen.* Landeshauptstadt München, Sozialrefarat Amt - für Wohnen und Migration., 2015, Kapitel 1

[Windmann u. Kauermann 2015] Windmann, M. ; Kauermann, G.: Erhebung der Daten und Berechnung der Nettomieten. In: *Mietspiegel für München ©2015. Statistik, Dokumentation und Analysen.* Landeshauptstadt München, Sozialrefarat Amt - für Wohnen und Migration., 2015, Kapitel 2

[WirtschaftsBlatt 2014] WirtschaftsBlatt: Deutschland zieht die Mietpreisbremse. In: *http://wirtschaftsblatt.at/home/nachrichten/europa/3878873/Deutschland-zieht-die-Mietpreisbremse* (2014), Oktober

[Wood 2006] Wood, S.: *Generalized Additive Models - An Introduction with R.* Boca Raton, FL : Chapman & Hall/CRC, 2006

Abbildungsverzeichnis

Tabellenverzeichnis

Erklärung der Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Bachelor-Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum	Unterschrift