Technische Universität Dortmund
Fakultät Statistik
Wintersemester 2022/2023

Fallstudien I
Projekt 2

Multiple Lineare Regression

Dozent: Prof. Dr. Guido Knapp
M. Sc. Yassine Talleb

Caroline Baer

Louisa Poggel
Julia Keiter
Daniel Sipek

Gruppennummer: 1

17.11.2022

Inhaltsverzeichnis

1 Einleitung

Dieser Bericht behandelt die Schätzung der Nettomiete von Wohnungen in München anhand eines multiplen linearen Regressionsmodell auf Grundlage eines Teildatensatzes des Münchener Mietspiegels von 2015.
Dazu werden zunächst die gegebenen Daten und das Ziel der Untersuchung in Kapitel 2 näher erläutert, bevor in Kapitel 3 die Methoden zum Linearen Modell, der Modellselektion sowie der Modelldiagnostik vorgstellt werden. Diese werden dann in der statistischen Auswertung in Kapitel 4 nach einer kurzen deskriptiven Zusammenfassung der Daten angewendet. Abschließend erfolgt in Kapitel 4 eine Interpretation des erstellten Modells und in Kapitel 5 dann die Zusammenfassung der wichtigsten Ergebnisse.

2 Problemstellung

Der vorliegende Datensatz *mietspiegel2015* ist ein Ausschnitt des Münchener Mietspiegels von 2015. Dabei wurden Wohnungen zufällig und unabhängig voneinander aus dem Mietwohnungsbestand der Stadt München ausgewählt. Insgesamt wurden 3.219 Wohnungen ausgewählt und 3.131 ausgewertet. In dem hierbetrachteten Teildatensatz geht es um 3065 Wohnungen sowie 13 zugehörige Charakteristika. Darin enthalten sind die Nettomiete pro Monat in Euro (*Miete*) und die Nettomiete pro Monat pro Quadratmeter in Euro (*Quadrat-Miete*). Außerdem abgefragt wurden die Wohnfläche in Quadratmeter (*Fläche*), die Anzahl der *Zimmer* und das *Baujahr*. Des Weiteren erhoben wurde, ob die Warmwasserversorgung vom Vermieter gestellt wird (*Warmwasser*), eine Zentralheizung verfügbar ist (*Heizung*) und ob das Bad bis ungefähr zu Türhöhe an allen Wänden gefliest ist (*Fliesen*). Zusätzlich wurde die Ausstattung von Bad (*Badausstattung*) und Küche (*Küchenausstattung*) untersucht. Von einem Gutachter wurde darüberhinaus die Lage der Wohnung als *gute Lage*, *beste Lage* oder *andere Lagekategorie* eingeordnet. Der Name vom *Bezirk* in dem die Wohnung liegt wurde ebenfalls vermerkt.
Bei den Variablen *Miete*, *Quadrat-Miete*, *Fläche*, *Zimmer* und *Baujahr* handelt es sich um metrische Variablen und beim *Bezirk* um eine nominale Variable mit 25 Ausprägungen. Die Variablen *gute Lage*, *beste Lage*, *Warmwasser*, *Heizung*, *Fliesen*, *Badausstattung* und *Küchenausstattung* sind dichotom und nominal. Im hier untersuchten Teildatensatz gibt es keine fehlenden Werte, sodass zu allen 3065 Beobachtungen jeweils 13 Werte vorliegen.
Ziel des Berichts ist die Untersuchung eines Zusammenhangs zwischen der *Miete* als Regressand und den anderen Wohnungscharakteristika als Regressoren, wobei die *Quadrat-Miete* hierbei nicht einbezogen werden soll. Dazu wird ein multiples Regressionsmodell

erstellt, um Schätzungen der *Miete* anhand der abhängigen Variablen zu ermöglichen, und anschließend auf seine Anpassungsgüte hin untersucht.

3 Statistische Methoden

Die im Folgenden beschriebenen Methoden stammen, sofern nicht anders angegeben, aus Kapitel 3.1 und 3.6 von Fahrmeir et al. (2007). Dabei beschreibt n die Anzahl der Beobachtungen pro Variable, k die Anzahl der Regressoren, y den Regressand, $x_1, \ldots, x_k$ die Regressoren und x_{ji} die i-te Beobachtung des j-ten Regressors.

Die statistische Auswertung mit den hier aufgeführten Methoden wird mit der Software R Core Team (2022) Version 4.2.2 durchgeführt. Zusätzlich wichtige Pakete hierfür sind `moments` von Komsta und Novomestky (2022), `corrplot` von Wei und Simko (2021), `car` von Fox und Weisberg (2019)und `xtable` von Dahl et al. (2019).

3.1 Grundlagen Lineares Modell

Das allgemeine lineare Modell hat die Form $y = X\beta + e$, vgl. alternative Schreibweise in (1), wobei $X = (1, x_1, \ldots, x_k)$ die $n \times (k+1)$-dimensionale Designmatrix, $\beta = (\beta_0, \ldots, \beta_k)^\top$ der zu schätzende Koeffizientenvektor und $e = (e_1, \ldots, e_n)^\top$ der zufällige Fehlervektor ist.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i \quad \text{für} \quad i = 1, \ldots, n \tag{1}$$

Die Koeffizienten β_k geben an wie stark der Einfluss des Regressors x_k auf den Regressanden y ist (vgl. Toutenburg (2003), Kap. 4.1). Da der Fehlervektor e nicht beobachtbar ist, wird er mit den Residuen $\hat{e}_i = y_i - \hat{y}_i$ für $i = 1, \ldots, n$ abgeschätzt, wobei $\hat{y} = X\hat{\beta}$ die durchs Modell angepassten Werte und $\hat{\beta}$ die Schätzung des Koeffizientenvektors darstellt. Der Koeffizientenvektor $\hat{\beta}$ lässt sich durch Anwendung der Kleinsten-Quadrate-Methode, kurz KQ-Methode, schätzen. Dabei wird die Summe der quadrierten Abweichungen $\hat{\beta}_{KQ} := \sum_{i=1}^n (y_i - x_i^\top \beta)^2 = \sum_{i=1}^n e_i^2 = ||e||^2 = ||y - X\beta||^2$ betrachtet und derjenige Vektor β gesucht, der diese Summe minimiert (vgl. Fahrmeir et al. (2007), Kap. 3.2.1). Umformen von $||y - X\beta||^2$ liefert die Normalengleichungen $X^\top X\beta = X^\top y$, mit denen sich, falls die Inverse $\left(X^\top X\right)^{-1}$ existiert, $\hat{\beta}_{KQ} = \left(X^\top X\right)^{-1} X^\top y$ eindeutig bestimmen lässt. Die Inverse $\left(X^\top X\right)^{-1}$ existiert genau dann, wenn die Designmatrix X vollen Spaltenrang hat.

Dies gehört nach Fahrmeir et al. (2007) zu den folgenden Modellannahmen:

- $\mathbb{E}(e_i) = 0$, dh. $\mathbb{E}(y) = (X\beta)$
- Unabhängigkeit bzw. Unkorreliertheit der Fehler: $\mathbb{C}ov(e) = \sigma^2 \cdot I_n$
- Die Designmatrix X besitzt vollen Spaltenrang mit $\text{rg}(X) = k + 1$
- Die Fehler sind normalverteilt mit $e_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$
- Homoskedastische Varianzen σ^2 bei den Fehlern e_i $(i = 1, \ldots, n)$

In der Software `R` wird das lineare Modell mit `lm(`$y \sim x_1 + \cdots + x_k$`, data = datensatz)` erstellt und im Weiteren mit `lin.mod` abgekürzt.
Durch die Anwendung der Funktion `summary()` werden unter anderem die mittels KQ-Methode gechätzten Regressionskoeffizienten $\hat{\beta}_i$, der Standardfehler $\text{SE}(\hat{\beta})_i = \sqrt{\widehat{\mathbb{V}ar(\hat{\beta}_i)}}$, sowie die t-Statistik $t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$ für die Hypothesen $H_0 : \beta_i = 0$ gegen $H_1 : \beta_i \neq 0$ und der p-Wert ausgegeben. Dabei gibt der p-Wert das minimale Signifikanzniveau α an, zu dem die Hypothese $H_0 : \beta_i = 0$ verworfen werden kann. Für diese Auswertung wird das Niveau 1%, dh. $\alpha = 0.01$, festgelegt.
Des Weiteren werden das Bestimmtheitsmaß R^2 wie in Formel (2) und das adjustierte Bestimmtheitsmaß $\tilde{R}^2$ wie in (3) berechnet.

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \in [0, 1] \tag{2}$$

$$\tilde{R}^2 = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) \quad \leq 1 \tag{3}$$

Das Bestimmheitsmaß ist ein Maß zur Beurteilung der Anpassungsgüte des Modells (siehe Fahrmeir et al. (2007), Kap. 3.2.3), da R^2 umso näher an 1 ist, je kleiner die Residuenquadratsumme $\text{SSE} := \sum_{i=1}^n e_i^2$ ist. Dementsprechend liegt eine ideale Modellanpassung genau dann vor, wenn $R^2 = 1$ gilt, weil dies bedeutet, dass alle Residuen $\hat{e}_i = 0$ sind. Da das Bestimmheitsmaß den Nachteil hat, dass es automatisch größer wird, je mehr Regressoren ins Modell aufgenommen werden, beinhaltet das adjustierte Bestimmtheitsmaß $\tilde{R}^2$ den Strafterm $\frac{1}{n-k}$, der bei vergrößerter Anzahl an Variablen k, verhindert, dass $\tilde{R}^2$ automatisch größer wird. Das adjustierte Bestimmtheitsmaß kann auch negative Werte annehmen, ist aber ebenfalls nach oben durch 1 beschränkt und spricht für eine umso bessere Modellanpassung je näher $\tilde{R}^2$ an 1 liegt.

3.2 Modellselektion

Die Modellselektion findet anhand des Akaiken Informationskriteriums, kurz AIC, statt, welches wie in Formel (4) definiert ist.

$$AIC = n \cdot ln\left(\frac{\text{SSE}}{n}\right) + 2 \cdot (k+1) \tag{4}$$

Eine der in diesem Bericht verwendeten Methoden zur Modellselektion ist die in Kapitel 21.5.2 von Groß (2010) beschriebene Rückwärtselimination. Hierbei wird zu Beginn das vollständige Modell mit allen Variablen betrachtet. Dann wird einzeln je eine der Variablen entfernt und anhand des AIC's bewertet. Vom Modell das den kleinsten AIC aufweist werden erneut die Variablen einzeln entfernt und das jeweils entstehende Modell mittels AIC untersucht. Fortgeführt wird dieser Vorgang bis das Entfernen von Variablen keine Verringerung des AIC's mehr erbringt.
Ein ähnliches Vorgehen beinhaltet die Vorwärtsselektion (vgl. Groß (2010), Kap. 21.5.2) mit dem Unterschied, dass hier mit dem Modell ohne jegliche Variablen gestartet wird und die Variablen einzeln hinzugefügt werden bis der minimale AIC erreicht ist.
Die Kombination aus beiden Methoden ist die Schrittweise Regression, die wie die Rückwärtselimination mit dem vollen Modell startet und dann die Variablen einzeln entfernt, aber einen zusätzlichen Zwischenschritt hat. In diesem werden alle zuvor entfernten Variablen nochmal einzeln hinzugefügt, sodass eine zu Beginn bereits rausgenommene Variable in einem späteren Schritt nochmal eingefügt werden kann, wenn dies zu einem niedrigeren AIC führt.
In `R` sind alle drei Modellselektinsmethoden mit der Funktion `step()` implementiert und werden mit der Angabe `direction =` „backward“, „forward“ oder „both“ ausgewählt.

3.3 Modelldiagnostik

Zur Überprüfung der Modellannahme der normalverteilten Fehler wird der Quantile-Quantile-Plot, kurz Q-Q-Plot betrachtet. Dabei werden die empirischen Quantile der standardisierten Residuen gegen die theoretischen Quantile der Normalverteilung abgetragen. Die Verteilung entspricht umso mehr einer Normalverteilung je deutlicher der Großteil der Daten in der Mitte auf der Ursprungeraden mit Steigung= 1 liegt, denn dies heißt, dass die empirischen Quantile mit den theoretischen übereinstimmen (vgl. Hartung et al. (2009), Kap. XIV 1.9).
Die Residuen werden hierfür wie in Formel (5) angegeben standardisiert, da die Varianz der Residuen $\mathbb{V}ar(\hat{e}_i) = s^2(1 - h_{ii})$ ist, wobei s^2 die empirische Varianz bezeichnet und h_{ii}

das i-te Diagonalelement der Hat-Matrix $H = X(X^\top X)^{-1}X^\top$ (siehe Groß (2010), Kap. 20.5).

$$\tilde{e}_i = \frac{\hat{e}_i}{s \cdot \sqrt{1 - h_{ii}}} \tag{5}$$

Die Untersuchung auf homoskedastische Varianzen und $\mathbb{E}(e_i) = 0$ erfolgt mittels Residualplot. Bei diesem werden die standardisierten Residuen $\tilde{e}_i$ gegen die angepassten Werte $\hat{y}_i$ abgetragen (vgl. Fahrmeir et al. (2007), Kap. 3.4.3). Falls Homoskedastizität mit Erwartungswert Null vorliegt streuen die Residuen ohne erkennbares Muster um die Null herum. Bei Heteroskedastizität hingegen werden die Varianzen mit größer werdenden angepassten Werten entweder trichterförmig größer oder kleiner. Außerdem lässt sich am Residualplot die Unkorreliertheit der Fehler überprüfen, da die Residuen eine Sytematik aufzeigen, wenn diese Annahme verletzt ist.
Die Korrelation wird in dieser Auswertung mit dem Rang-Korrelationskoeffizienten nach Spearman ausgeführt, der in Kapitel I 9.4 von Hartung et al. (2009) wie in Formel (6) definiert wird. Dabei bezeichnet $R(x_{ji})$ den Rang der i-ten Beobachtung der Variable x_j nach aufsteigender Sortierung der Beobachtungen.

$$r^s_{j,t} = \frac{\sum_{i=1}^n R(x_{ji})R(x_{ti}) - n \cdot \overline{R(x_j)} \cdot \overline{R(x_t)}}{\sqrt{\left(\sum_{i=1}^n R(x_{ji})^2 - n \cdot \overline{R(x_j)^2}\right)\left(\sum_{i=1}^n R(x_{ti})^2 - n \cdot \overline{R(x_t)^2}\right)}} \tag{6}$$

In `R` lässt sich die Korrelation zwischen den Variablen mit dem Korrelationsplot mittels der Funktion `corrplot()` grafisch darstellen.
Eine hohe Korrelation zwischen den Regressoren kann zur Multikollinearität und somit zu einer instabilen Regressionsschätzung führen. Starke Multikollinearität liegt vor, wenn mindestens zwei Spalten der Designmatrix X linear abhängig sind, sodass X keinen vollen Spaltenrang besitzt (vgl. Toutenburg (2003), Kap. 4.5). Schwache Multikollinearität liegt vor, wenn keine exakte, aber annähernde lineare Abhängigkeit zwischen den Variablen vorliegt, was dazu führt das die Determinante von $X^\top X$ einen Wert nahe Null annimmt. Ebenfalls zur Überprüfung auf Multikollinearität anwendbar ist der Varianzinflationskoeffizient, kurz VIF, der nach Kapitel 4.5.3 von Toutenburg (2003) durch Formel (7) definiert wird. Dabei steht R_j^2 für den multiplen Korrelationskoeffizienten der Regression von x_j auf die anderen Variablen.

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{7}$$

Der VIF gibt an, um welchen Faktor die Varianz von $\hat{\beta}_j$ durch lineare Abhängigkeit vergrößert wird. Falls der $\text{VIF}_j > 10$ ist, liegt Multikollinearität vor.
Bei der Modelldiagnostik wird außerdem untersucht, ob es Beobachtungen mit besonders

großem Einfluss auf das Modell gibt. Dazu wird hier die von Fahrmeir et al. (2007) in Kapitel 3.6.4 beschriebene Cook's Distance betrachtet, wobei $\hat{y}_{(i)}$ für den angepassten Wert des Regressanden steht, wenn die i-te Beobachtung zuvor entfernt wurde (vgl. Formel (8)).

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^\top (\hat{y} - \hat{y}_{(i)})}{k \cdot \hat{\sigma}^2} \tag{8}$$

Dabei spricht ein Wert $D_i > 1$ für eine sehr einflussreiche und $D_i > 0.5$ für eine auffällige Beobachtung. Diese Beobachtungen werden auf Plausibilität geprüft und zum Vergleich aus dem Modell entfernt, um zu prüfen zu welchen Unterschieden in der Modellbildung dies führt.

4 Statistische Auswertung

4.1 Deskriptive Zusammenfassung der Daten

Die *Miete* liegt bei allen 3065 Beobachtungen zwischen 174.75€ und 6000.00€, und im Median bei 700.00€, aber beim arithmetischen Mittel bei 763.06€ im Monat (vgl. Tablle 5). Auch die Standardabweichung ist mit 338.16€ deutlich höher als der MAD mit 261.90€. Außerdem auffällig ist die stark leptokurtische Verteilung mit einem Wert von 25.47. Die *Quadrat-Miete* ist im Median 10.84€ und beim arithmetischen Mittel 10.73€. Insgesamt liegt sie zwischen 2.47€ und 22.13€ pro Monat. Die betrachteten Wohnungen weisen eine *Fläche* von 15.00m^2 bis 300.00m^2 auf und sind im arithmetischen Mittel 71.98m^2 groß. Die Anzahl der *Zimmer* geht von einem bis acht Zimmern und liegt im Median bei drei Zimmern pro Wohnung. Da das 1. Quartil bei zwei Zimmern und das 3. Quartil bei drei Zimmern liegt, hat die Hälfte aller untersuchten Wohnungen zwei oder drei Zimmer. Das *Baujahr* liegt zwischen 1918 und Mitte 2012 (2012.5), wobei die Hälfte aller Wohnungen zwischen Mitte 1957 (1957.5) und 1983 erbaut wurden.
Wie Tabelle 6 zu entnehmen ist, gibt es 110 Wohnungen in *bester Lage*, 1085 Wohnungen in *guter Lage* und dementsprechend 1870 Wohnungen die vom Gutachter in eine andere Lagekategorie eingeteilt wurden. Für 99.15% der Wohnungen wird die Versorgung mit *Warmwasser* vom Vermieter gestellt und bei 93.34% ist eine *Zentralheizung* verfügbar. Das Bad ist bei 12.40% mit *Fliesen* an allen Wänden bis zur Türhöhe versehen, und bei 11.78% liegt eine gehobene *Badausstattung* und bei 25.02% eine gehobene *Küchenausstattung* vor. Die meisten der betrachteten Wohnungen liegen im *Bezirk* Neuhausen-Nymphenburg mit 7.63% und Ramersdorf-Perlach mit 5.91%, während die wenigsten sich in Allach-

Untermenzing mit 0.82% und Altstadt-Lehel mit 1.53% befinden.

4.2 Modellbildung und Selektion

Eine allgemeine Betrachtung der *Miete* in Abhängigkeit der anderen Variablen zeigt, dass der Preis mit Zunahme der Größe der *Fläche* sowie der Anzahl an *Zimmern* steigt (vgl. Abbildung 1 und 2). In Bezug auf die anderen Variablen *Baujahr*, *Bezirk*, *gute Lage*, *beste Lage*, *Warmwasser*, *Heizung*, *Fliesen*, *Badausstattung* und *Küchenausstattung* liegen, weniger deutliche, aber ebenfalls lineare Zusammenhänge vor. Die Variable *Quadrat-Miete* wird im Weiteren nicht weiter betrachet.

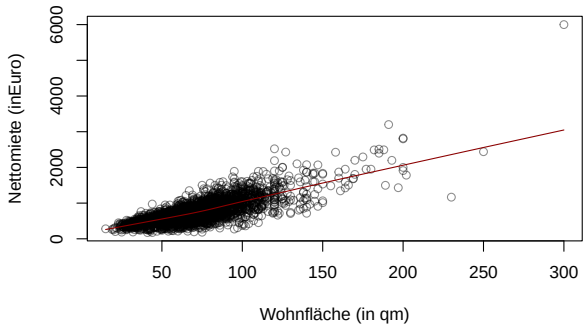

Abbildung 1: *Nettomiete* in Abhängigkeit der *Wohnfläche*

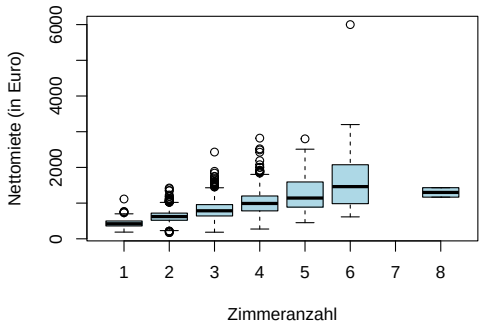

Abbildung 2: *Nettomiete* in Abhängigkeit der *Zimmeranzahl*

Zur besseren Interpretierbarkeit des multiplen Regressionsmodells wird die Variable *Baujahr* transformiert, indem von allen Werten das arithmetische Mittel (1964.21) abgezogen wird.

Zuerst betrachtet wird das Modell **lin-mod-gesamt**, welches mit der *Miete* als Regressanden und allen anderen Variablen als Regressoren erstellt wird. Dabei ergibt sich insgesamt ein p-Wert der kleiner als $2.2 \cdot 10^{-16}$ ist und das Bestimmtheitsmaß hat einen Wert von 0.7043, während das adjustierte Bestimmtheitsmaß den Wert 0.7010 hat. Dies spricht für eine gute Anpassung des Modells an die echten Daten. Die Residuen sind im Median bei 4.80 und streuen zwischen 2454.63 und -975.11 (siehe Tabelle 1). Die p-Werte der einzelnen Variablen zeigen, dass die Koeffizienten von *Fläche*, *Zimmer*, *Baujahr*, *gute Lage*, *beste Lage*, keine Versorgung mit *Warmwasser*, keine *Zentralheizung*, keine *Fliesen* und normaler *Küchenausstattung* zum Niveau 0.1%, dh. $\alpha = 0.001$, signifikant von Null verschieden sind. Der Koeffizient der normalen *Badausstattung* und der vom *Bezirk* Ludwigvorstadt-Isarvorstadt sind beide zum Niveau 1% signifikant verschieden von Null, während die restlichen Variablen das Niveau $\alpha = 0.01$ nicht einhalten können. Da nur

bei einem von insgesamt 25 Bezirken ein signifikanter Einfluss aufs Modell nachgewiesen werden kann, wird die Variable *Bezirk* im Folgenden aus dem Modell entfernt, das Modell neu angepasst und neu betrachtet.

Tabelle 1: Verteilung der Residuen im Modell mit allen Variablen

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-975.11	-95.77	4.80	93.97	2454.63

Das so entstehende Modell **lin-mod-ohne-bezirk** ohne die Variable *Bezirk* als Regressor hat ein geringfügig vermindertes Bestimmtheitsmaß mit 0.6869 und ebenso verringertes adjustiertes Bestimmtheitsmaß mit 0.6859. Die Residuen streuen bei diesem Modell im Vergleich zu lin-mod-gesamt bezüglich des Interquartilsabstands und der Spannweite zwar etwas mehr, aber dafür sind sie im Median mit 1.81 viel dichter an der Null (siehe Tabelle 2), was besser zur Annahme des Erwartungswerts der Fehler passt. Alle Variablen halten das Niveau $\alpha = 0.01$ ein und bis auf den Koeffizienten der normalen *Badaustattung* zusätzlich auch das Niveau 0.001 (vgl. Tabelle 10). Der p-Wert insgesamt ist weiterhin kleiner als $2.2 \cdot 10^{-16}$ und da die Anpassungsgüte mit der Entfernung des *Bezirks* nur minimal gesunken ist, bildet dieses Modell die Grundlage für die anschließende Variablenselektion.

Tabelle 2: Verteilung der Residuen im Modell lin-mod-ohne-bezirk

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-1022.92	-102.16	1.81	96.59	2491.59

Die Rückwärtselimination anhand des AIC's, welche mit der Software `R` wie in den Methoden (Kapitel 3) beschrieben, durchgeführt wird, ergibt, dass das Modell bezüglich des AIC's optimal ist, wenn keine weitere Variable aus dem Modell entfernt wird. Zu dem gleichen Ergebnis führt die Anwendung der Vorwärtsselektion und der Schrittweisen Regression, da der AIC beim Modell lin-mod-ohne-bezirk mit den Variablen *Fläche*, *Zimmer*, *gute Lage*, *beste Lage*, *Warmwasser*, *Heizung*, *Fliesen*, *Badausstattung* und *Küchenausstattung* am kleinsten ist.

4.3 Modelldiagnostik

Die Betrachtung der Korrelation in Abbildung 3 der metrischen Variablen verdeutlicht, dass die *Miete* in hoher Korrelation zu den Regressoren *Fläche* und *Zimmer* steht, jedoch sind diese beiden Einflussgrößen auch untereinander mit 0.8591 stark korreliert.

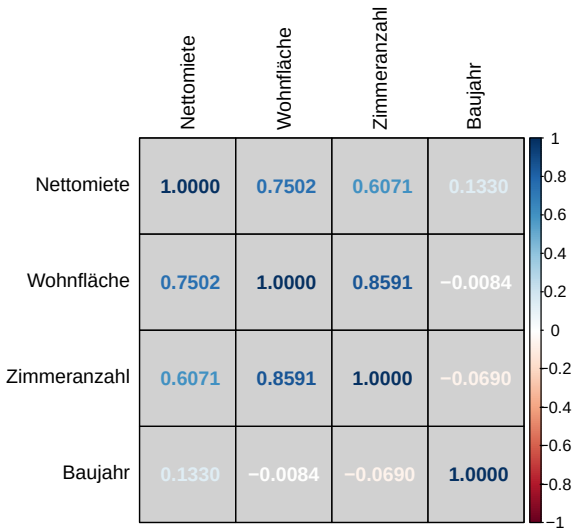

Abbildung 3: Korrelation der metrischen Variablen

Eine hohe Korrelation zwischen den Regressoren kann zu linearer Abhängigkeit der Variablen und somit zur Multikollinearität führen. Zur Überprüfung der Multikollinearität im Modell lin-mod-ohne-bezirk wird zunächst der Spaltenrang untersucht. Dieser ist mit $\text{rg}(X) = 11$ bei zehn Variablen plus Intercept voll und schließt somit eine starke Multikollinearität aus. Da die Determinante der transponierten Modellmatrix multipliziert mit der Modellmatrix mit $1.6720 \cdot 10^{35}$ stark von Null abweicht, lässt sich auch eine schwache Multikollinearität ausschließen. Bestätigt wird dies durch die Betrachtung des Varianzinflationskoeffizienten, welcher für die Variable *Fläche* mit 3.5582 zwar den höchsten Wert von allen Variablen hat, aber immer noch weit unter dem Grenzwert zehn liegt. Obwohl durch die Korrelation der Variablen *Fläche* und *Zimmer* keine Multikollinearität vorliegt, wird im Folgenden eine der beiden Variablen aus dem Modell genommen und dieses neu geprüft. Denn eine Betrachtung der Koeffizientenschätzungen in Tabelle 10 zeigt, dass aufgrund der starken Korrelation die Koeffizientenschätzung für *Zimmer* einen negativen Wert angenommen hat, was bezüglich der Interpretation unplausibel erscheint. Da die *Fläche* im Rahmen der Interpretierbarkeit aussagekräftiger ist, wird diese im Modell behalten und *Zimmer* entfernt.
Das Modell **lin-mod-ohne-bezirk-zimmer** hat mit einem Bestimmtheitsmaß von 0.6795 und einem adjustierten Bestimmtheitsmaß von 0.6786 eine etwas geringere Anpassungsgüte

als lin-mod-gesamt oder lin-mod-ohne-bezirk und auch die Residuen sind bei diesem Modell schlechter verteilt, da sie mehr streuen und im Median (3.34) mehr als lin-mod-ohne-bezirk (1.81) von der Null abweichen (vgl. Tabelle 3). Der Median liegt jedoch näher an der Null als beim Modell lin-mod-gesamt und auch die p-Werte aller Variablen sind kleiner als $\alpha = 0.01$.

Tabelle 3: Verteilung der Residuen im Modell lin-mod-ohne-bezirk-zimmer

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-1033.47	-100.95	3.34	95.54	2690.19

Die Untersuchung der Normalverteilungsannahme mittels Quantile-Quantile-Plot in Abbildung 4 zeigt, dass die standardisierten Residuen des Modells lin-mod-ohne-bezirk-zimmer näherungsweise normalverteilt sind. Es sind zwar auch schwere Ränder zu erkennen, aber der Großteil der standardisierten Residuen in der Mitte liegt auf der Ursprungsgeraden mit Steigung = 1, sodass mehrheitlich die theoretischen Quantile mit den empirischen Quantilen übereinstimmen.

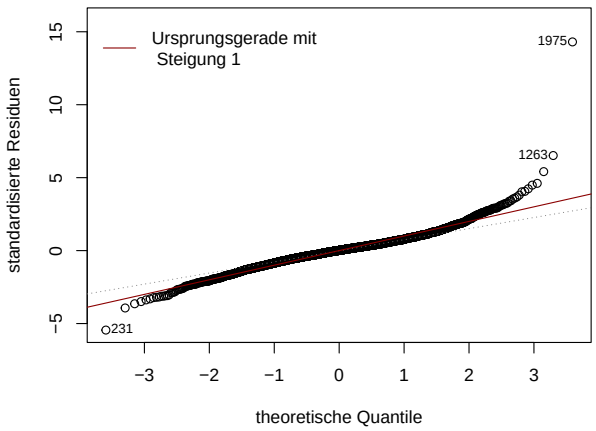

Abbildung 4: Überprüfung der Normalverteilung beim Modell lin-mod-ohne-bezirk-zimmer

Wie Abbildung 5 zu entnehmen ist, sind die drei Beobachtungen 231, 1263 und 1975 markiert, da sie bezüglich der Cook's Distance die höchsten Werte haben. Beobachtung 1975 ist mit einer Cook's Distance von 0.8005 jedoch die einzige die über dem Grenzwert von 0.5 liegt und somit als auffällige Beobachtung gilt.

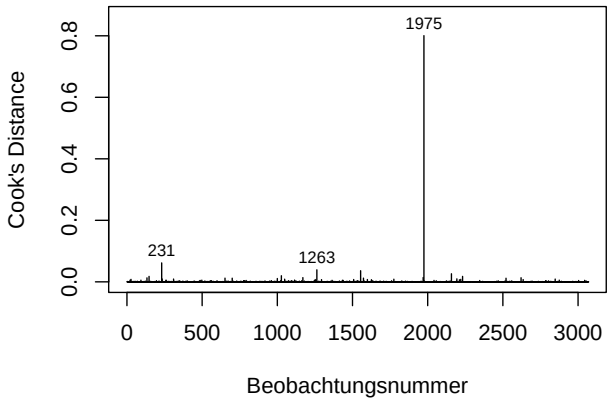

Abbildung 5: Auffälligkeiten bezüglich der Cook's Distance

Eine genauere Betrachtung dieser drei Beobachtungen in Bezug auf die *Fläche* und *Miete* der zugehörigen Wohnungen zeigt, dass alle drei in einem Bereich liegen, wo die Datenmenge zu Wohnungen mit vergleichbarer *Fläche* und *Miete* ziemlich gering ist und dementsprechend ihr Einfluss aufs Modell erhöht ist (siehe Abbildung 6). So erscheint die Wohnung zu Beobachtung 231 für die Größe der *Wohnfläche* recht günstig zu sein, liegt aber auch weder in *bester* noch *guter Lage* und besitzt eine normale *Bad-* und *Küchenausstattung* ohne verfügbare *Zentralheizung*. Bei Wohnung 1263 handelt es sich um eine vergleichsweise teure Wohnung in Bezug auf die *Fläche*, aber auch dies erscheint plausibel im Hinblick auf die *gute Lage*, die vom Vermieter gestellte Versorgung mit *Warmwasser* und die verfügbare *Zentralheizung*.

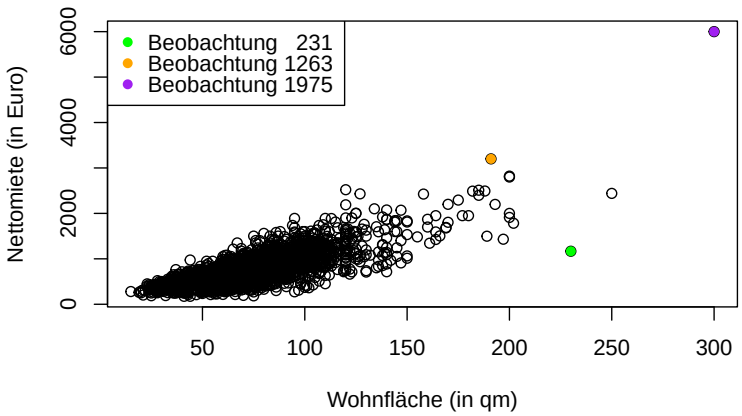

Abbildung 6: Markierungen der Beobachtungen mit erhöhter Cook's Distance

Auch die Beobachtung 1975 wirkt plausibel bei Betrachtung der *besten Lage*, gehobenen *Küchenausstattung* und besonders großen *Wohnfläche*. Dennoch wird, wie bei einer auffälligen Cook's Distance üblich, das Modell ohne diese Beobachtung angepasst und mit dem Modell mit Beobachtung 1975 verglichen. Daraus entsteht das Modell **lin-mod-ohne-bezirk-zimmer-1975**, welches eine deutliche Verbesserung der Residuenverteilung im Vergleich zu lin-mod-ohne-bezirk-zimmer aufweist. Wie Tabelle 4 zu entnehmen ist, weisen die Residuen dieses Modells eine viel geringere Spannweite als die zuvor behandelten Modelle auf und sind im Median mit 1.77 am nächsten an der Null. Die Anpassungsgüte ist zwar nun ein weiteres Mal minimal verringert worden ($R^2 = 0.6756$, $\tilde{R}^2 = 0.6746$), aber dafür sind nun die Koeffizienten aller Variablen zum Niveau 0.1% ($\alpha = 0.001$) signifikant von Null verschieden. Daher wird das Modell lin-mod-ohne-bezirk-zimmer-1975, welches die Variablen *Fläche*, *Baujahr*, *gute Lage*, *beste Lage*, *Warmwasser*, *Heizung*, *Fliesen*, *Badausstattung* und *Küchenausstattung* ohne Beobachtung 1975 enthält als finales Modell festgelegt.

Tabelle 4: Verteilung der Residuen im Modell lin-mod-ohne-bezirk-zimmer-1975

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-980.93	-99.73	1.77	93.58	1278.88

Eine erneute Überprüfung der Modellannahmen am finalen Modell lin-mod-ohne-bezirk-zimmer-1975 ergibt, dass die Residuen weiterhin näherungsweise normalverteilt sind und die Annahmen zum Erwartungswert und der Unkorreliertheit der Fehler ebenfalls erfüllt sind (vgl. Abbildung 7 und 8). Jedoch fällt in Abbildung 8 auf, dass die Residuen mit zunehmender Größe der angepassten Werte tendenziell breiter streuen und die Annahme der Homoskedastizität somit nicht vollständig erfüllt ist. Wie in Kapitel 4.4.2 von Rottmann und Auer (2010) erläutert, verändert die Heteroskedastizität der Varianzen nicht die Erwartungstreue der KQ-Schätzungen des Regressionskoeffizienten, führt aber dazu das die KQ-Schätzer nicht länger effizient sind, dh. minimale Varianz aufweisen.

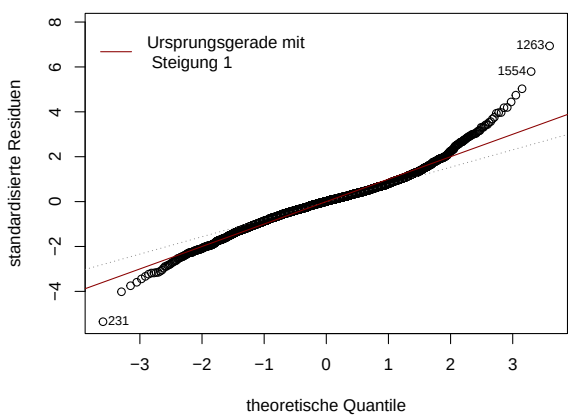

Abbildung 7: Q-Q-Plot der standardisierten Residuen für das finale Modell

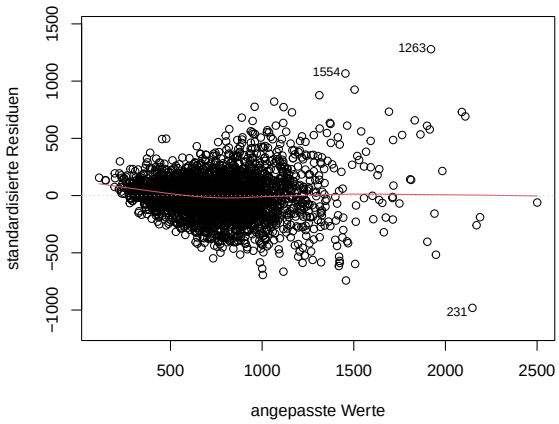

Abbildung 8: Residualplot für das finale Modell lin-mod-ohne-bezirk-zimmer-1975

Die Untersuchung der Multikollinearität für dieses finale Modell zeigt, dass sowohl schwache als auch starke Multikollinearität mittels vollem Spaltenrang und Determinante von $1.9256 \cdot 10^{32}$ ausgeschlossen werden kann. Belegt wird dies durch den Varianzinflationskoeffizienten, der nun für alle Variablen zwischen 1.0446 und 1.1745 und somit deutlich unter dem Grenzwert zehn liegt. Ebenso weist die Betrachtung der Cook's Distance auf keine auffälligen Beobachtungen mit besonders hohem Einfluss mehr hin, da die maximale Cook's Distance bei 0.0606 ist.

4.4 Interpretation des Modells

Wie Tabelle 12 zu entnehmen ist, besagt das finale Modell lin-mod-ohne-bezirk-zimmer-1975, dass die *Miete* pro hinzukommenden Quadratmeter *Fläche* und sonst gleichbleibenden Bedingungen im Mittel um circa 9.83€ pro Monat teurer wird. Aufgrund der Transformation des *Baujahrs* lässt sich sagen, dass die *Miete* pro Monat im Schnitt um je 1.08€ höher liegt für jedes Jahr, das seit dem durchschnittlichen Baujahr von 1964.21 vergangen ist. Eine *gute Lage* bewirkt einen Anstieg der *Miete* im Mittel um 88.82€ und eine *beste Lage* im Mittel um 111.63€ pro Monat, solange alle anderen Einflussfaktoren konstant bleiben. Die *Miete* fällt im Schnitt bei nicht gestellter Versorgung mit *Warmwasser* um 184.14€ und bei nicht verfügbarer *Zentralheizung* um 68.09€ ab. Ein nur teilweise mit *Fliesen* ausgestattetes Bad bewirkt im Mittel eine Preissteigerung von 54.63€. Die normale, im Vergleich zur gehobenen, *Badausstattung* senkt die *Miete* um 37.59€ und die normale Küchenausstattung um 90.56€ pro Monat.

5 Zusammenfassung

Die mittels multipler linearer Regression durchgeführte Untersuchung des Zusammenhangs zwischen der, im Teildatensatz des Münchener Mietspiegels 2015 erhobenen, *Nettomiete* pro Monat als Regressand und elf weiteren Wohnungscharakteristika als Regressoren hat gezeigt, dass die *Miete* einer Wohnung vor allem von der *Wohnfläche* und der Anzahl der *Zimmer* abhängt. Aufgrund der hohen Korrelation zwischen diesen beiden Einflussgrößen, wurde jedoch nur eine der beiden Variablen ins Modell mit aufgenommen, sodass dieses die Regressoren *Wohnfläche*, *Baujahr*, *gute Lage*, *beste Lage*, vom Vermieter gestellte Versorgung mit *Warmwasser*, verfügbare *Zentralheizung*, komplett vorhandene *Fliesung* im Bad, *Badausstattung* und *Küchenausstattung* beinhaltet.
Die ebenfalls im Datensatz enthaltene Variable *Bezirk* wurde, noch vor der Modellselektion anhand des AIC's, aus dem Modell entfernt, da sie das Niveau $\alpha = 0.01$ beim Signifkanztest nicht eingehalten hat. Des Weiteren wurde eine Beobachtung aus der Modellanpassung entfernt, da diese Beobachtung zu einer Wohnung gehört, zu deren Größe der *Wohnfläche* und höherpreisigen *Nettomiete* es keine vergleichbaren Beobachtungen im hier verwendeten Teildatensatz *mietspiegel2015* mit insgesamt 3065 Wohnungen gab und diese eine Beobachtung somit einen vergrößerten Einfluss auf die Modellbildung hatte.
Das erstellte multiple lineare Regressiosnmodell hat nach adjustiertem Bestimmtheitsmaß von 0.6746 eine gute Anpassungsgüte. Die Überprüfung der allgemeinen Modellannahmen wies darauf hin, dass die Annahme der homoskedastischen Varianzen der Fehler, im Gegensatz zu den anderen Annahmen, nicht beibehalten werden konnte. Daher sind die hier unter Verwendung der KQ-Methode geschätzten Koeffizienten zwar erwartungstreu, aber nicht mehr effizient, sodass es andere erwartungstreue Koeffizentenschätzungen geben kann, die eine geringere Varianz aufweisen.
Für eine weiterführende Untersuchung wäre von daher eine Regression mit gewichteten KQ-Schätzern in Betracht zu ziehen, da diese robust gegen Heteroskedastizität sind. Außerdem würde es sich anbieten ein Modell zu einem Datensatz zu erstellen, der mehr Beobachtungen zu teureren und größeren Wohnung beinhaltet, sodass die Schätzung der *Miete* von Wohnungen dieser Art exakter wäre.

Literaturverzeichnis

Literatur

Dahl, David B. et al. (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4. URL: https://CRAN.R-project.org/package=xtable.

Fahrmeir, Ludwig, Thomas Kneib und Stefan Lang (2007). *Regression. Modelle, Methoden und Anwendungen*. 1. Aufl. Springer Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-540-33933-5.

Fox, John und Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third. Sage: Thousand Oaks CA. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Groß, Jürgen (2010). *Grundlegende Statistik mit R. Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R*. 1. Aufl. Vieweg+Teubner Verlag Wiesbaden. ISBN: 978-3-8348-1039-7. DOI: https://doi.org/10.1007/978-3-8348-9677-3.

Hartung, Joachim, Bärbel Elpelt und Karl-Heinz Klösener (2009). *Statistik. Lehr- und Handbuch der angewandten Statistik*. 15. Aufl. Oldenbourg Verlag: München.

Komsta, Lukasz und Frederick Novomestky (2022). *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.14.1. URL: https://CRAN.R-project.org/package=moments.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rottmann, Horst und Benjamin Auer (2010). *Statistik und Ökonometrie für Wirtschaftswissenschaftler. Eine anwendungsorientierte Einführung*. 1. Aufl. Gabler Verlag Wiesbaden. ISBN: 978-3-8349-6372-7. DOI: https://doi.org/10.1007/978-3-8349-6372-7.

Toutenburg, Helge (2003). *Lineare Modelle. Theorie und Anwendungen*. 2. Aufl. Physica Heidelberg: Heidelberg. ISBN: 978-3-7908-1519-1. DOI: https://doi.org/10.1007/978-3-642-57348-4.

Wei, Taiyun und Viliam Simko (2021). *R package 'corrplot': Visualization of a Correlation Matrix*. (Version 0.92). URL: https://github.com/taiyun/corrplot.

Anhang

Tabelle 5: Deskriptive Kenngrößen der metrischen Variablen

	Nettom. (in €)	Nm. pro m^2 (in €)	Wohnfläche (in m^2)	Zimmer	Baujahr
arithm. Mittel	763.06	10.73	71.98	2.70	1964.21
Median	700.00	10.84	70.00	3.00	1957.50
Minimum	174.75	2.47	15.00	1.00	1918.00
Maximum	6000.00	22.13	300.00	8.00	2012.50
Spannweite	5825.25	19.66	285.00	7.00	94.50
1.Quartil	550.00	9.03	55.00	2.00	1957.50
3.Quartil	910.46	12.45	85.00	3.00	1983.00
IQR	360.46	3.42	30.00	1.00	25.50
Standardabw.	338.16	2.67	25.74	0.98	26.51
MAD	261.90	2.51	22.24	1.48	27.43
Schiefe	2.59	0.04	1.35	0.46	-0.18
Wölbung	25.47	3.34	8.33	3.60	2.31

Tabelle 6: Verteilungen der dichotomen Variablen

Variable	Ausprägung	absolute Häufigkeit	relative Häufigkeit
gute Lage	gute Lage	1085	0.3540
	andere Lagekategorie	1980	0.6460
beste Lage	beste Lage	110	0.0359
	andere Lagekategorie	2955	0.9641
Warmwasser	gestellt	3039	0.9915
	nicht gestellt	26	0.0085
Zentralheizung	verfügbar	2861	0.9334
	nicht verfügbar	204	0.0666
gefliestes Bad	gefliest	380	0.1240
	nicht gefliest	2685	0.8760
Badausstattung	gehoben	361	0.1178
	normal	2704	0.8822
Küchenausstattung	gehoben	767	0.2502
	normal	2298	0.7498

Tabelle 7: Verteilung der nominalen Variable Bezirk

Bezirk	absolute Hfgkeit	relative Hfgkeit
Allach-Untermenzing	25	0.0082
Altstadt-Lehel	47	0.0153
Au-Haidhausen	167	0.0545
Aubing-Lochhausen-Langwied	56	0.0183
Berg am Laim	101	0.0330
Bogenhausen	159	0.0519
Fledmoching-Hasenbergel	78	0.0254
Hadern	79	0.0258
Laim	100	0.0326
Ludwigsvorstadt-Isarvorstadt	154	0.0502
Maxvorstadt	168	0.0548
Milbersthofen-Am Hart	134	0.0437
Moosach	92	0.0300
Neuhausen-Nymphenburg	234	0.0763
Obergiesing	145	0.0473
Pasing-Obermenzing	118	0.0385
Ramersdorf-Perlach	181	0.0591
Schwabing-Freimann	140	0.0457
Schwabing West	165	0.0538
Schwanthalerhöhe	87	0.0284
Sendling	126	0.0411
Sendling-Westpark	122	0.0398
Thalkirchen	175	0.0571
Trudering-Riem	81	0.0264
Untergiesing	131	0.0427

Tabelle 8: Verteilung der Residuen in allen Modellen

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
lin-mod	-975.11	-95.77	4.80	93.97	2454.63
lin-mod-ohne-bez	-1022.92	-102.16	1.81	96.59	2491.59
lin-mod-ohne-bez-rooms	-1033.47	-100.95	3.34	95.54	2690.19
lin-mod-ohne-bez-rooms-1975	-980.93	-99.73	1.77	93.58	1278.88

Tabelle 9: summary(lin.mod)

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	65.5653	41.8440	1.57	0.1172
wfl	11.7193	0.2483	47.19	0.0000
rooms	-46.4250	6.4870	-7.16	0.0000
bj	1.6487	0.1473	11.20	0.0000
bezAltstadt-Lehel	0.7912	47.4605	0.02	0.9867
bezAu-Haidhausen	69.6583	40.1930	1.73	0.0832
bezAubing...	-53.2408	44.5461	-1.20	0.2321
bezBerg am Laim	-34.9110	41.3961	-0.84	0.3991
bezBogenhausen	1.5866	40.0529	0.04	0.9684
bezFledmoching-Hasenbergel	-81.9470	42.5829	-1.92	0.0544
bezHadern	-30.5425	42.5041	-0.72	0.4725
bezLaim	-27.0360	41.4876	-0.65	0.5147
bezLudwigvorstadt-Isarvorstadt	110.3494	40.4791	2.73	0.0064
bezMaxvorstadt	103.6501	40.4394	2.56	0.0104
bezMilbersthofen-Am Hart	4.1184	40.4493	0.10	0.9189
bezMoosach	-12.4847	41.7786	-0.30	0.7651
bezNeuhausen-Nymphenburg	46.2330	39.3093	1.18	0.2396
bezObergiesing	-16.3192	40.1664	-0.41	0.6846
bezPasing-Obermenzing	-0.0692	40.9292	-0.00	0.9987
bezRamersdorf-Perlach	-72.1614	39.5028	-1.83	0.0678
bezSchwabing-Freimann	68.7585	40.4486	1.70	0.0893
bezSchwabing West	40.6518	40.2934	1.01	0.3131
bezSchwanthalerhöhe	45.0293	42.2215	1.07	0.2863
bezSendling	30.3219	40.6416	0.75	0.4557
bezSendling-Westpark	-18.6270	40.6969	-0.46	0.6472
bezThalkirchen...	-23.3948	39.6960	-0.59	0.5557
bezTrudering-Riem	-34.9007	42.5945	-0.82	0.4126
bezUntergiesing	35.9715	40.5558	0.89	0.3752
wohngutGute Lage	44.9365	8.7791	5.12	0.0000
wohnbestBeste Lage	101.4078	20.1251	5.04	0.0000
ww0nein	-178.8789	37.8167	-4.73	0.0000
zh0nein	-78.6150	14.3448	-5.48	0.0000
badkach0nicht gefliest	52.3308	10.4720	5.00	0.0000
badextranormal	-32.8995	10.9525	-3.00	0.0027
kuechenormal	-85.5743	8.1683	-10.48	0.0000

Tabelle 10: summary(lin.mod_ohne_bez)

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	72.2735	19.9333	3.63	0.0003
wfl	11.9284	0.2509	47.54	0.0000
rooms	-55.5408	6.5566	-8.47	0.0000
bj	1.1074	0.1401	7.91	0.0000
wohngutGute Lage	82.4023	7.3415	11.22	0.0000
wohnbestBeste Lage	118.4039	18.9315	6.25	0.0000
ww0nein	-184.5388	38.6394	-4.78	0.0000
zh0nein	-67.2810	14.6376	-4.60	0.0000
badkach0nicht gefliest	52.3544	10.6974	4.89	0.0000
badextranormal	-30.6062	11.0553	-2.77	0.0057
kuechenormal	-85.3459	8.3193	-10.26	0.0000

Tabelle 11: summary(lin.mod_ohne_bez_rooms)

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	52.4284	20.0230	2.62	0.0089
wfl	10.1633	0.1414	71.89	0.0000
bj	1.1553	0.1416	8.16	0.0000
wohngutGute Lage	87.4565	7.4015	11.82	0.0000
wohnbestBeste Lage	131.3731	19.0868	6.88	0.0000
ww0nein	-187.5855	39.0826	-4.80	0.0000
zh0nein	-65.2646	14.8042	-4.41	0.0000
badkach0nicht gefliest	53.4245	10.8198	4.94	0.0000
badextranormal	-29.3194	11.1815	-2.62	0.0088
kuechenormal	-95.4618	8.3279	-11.46	0.0000

Tabelle 12: summary(lin.mod_ohne_bez_rooms_1975)

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	78.7248	19.4254	4.05	0.0001
wfl	9.8263	0.1385	70.96	0.0000
bj	1.0760	0.1369	7.86	0.0000
wohngutGute Lage	88.8216	7.1511	12.42	0.0000
wohnbestBeste Lage	111.6310	18.4877	6.04	0.0000
ww0nein	-184.1406	37.7580	-4.88	0.0000
zh0nein	-68.0905	14.3035	-4.76	0.0000
badkach0nicht gefliest	54.6290	10.4532	5.23	0.0000
badextranormal	-37.5936	10.8168	-3.48	0.0005
kuechenormal	-90.5553	8.0523	-11.25	0.0000