Technische Universität Dortmund
Fakultät Statistik
Wintersemester 2022/2023

Fallstudien I

Projekt 2: Multiple Lineare Regression

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

Bericht von: Louisa Poggel

Mitglieder der Gruppe 1:
Caroline Baer
Daniel Sipek
Julia Keiter
Louisa Poggel

27.10.2022

Inhaltsverzeichnis

1 Einleitung

Dieses Projekt beschäftigt sich mit der Ermittlung der ortsüblichen Vergleichsmiete auf Basis des Mietspiegels der Stadt München aus dem Jahr 2015. Unter Berücksichtigung von gesetzlichen Vorgaben wird dazu eine repräsentative Zufallsstichprobe aus allen Mietobjekten Münchens gezogen. Ziel ist es ein multiples Regressionsmodell zu erstellen, dass die Nettomiete pro Monat möglichst gut anhand zahlreicher Regressoren bezüglich der Ausstattung und Wohnlage der Mietobjekte beschreibt. Dabei soll das Modell bei der Findung einer korrekten Miete für zukünftige Immobilien helfen und ein neues Mietobjekt richtig einordnen können.
Es wird sich herausstellen, dass die *Nettomiete pro Monat* gut durch einen linearen Zusammenhang ohne polynomiale Koeffizienten beschrieben wird. Dabei werden sich die Variablen *Wohnfläche*, *Ausstattung Küche*, *Wohnlage*, *Baujahr*, *Warmwasserversorgung*, *gefliestes Bad*, *Zentralheizung* und *Ausstattung Bad* als geeignete Regressoren herausstellen. Im folgendem Kapitel 2 wird zunächst die Problemstellung, inklusive aller betrachteten Variablen, genauer erläutert. Darauf folgt in Kapitel 3 eine Darstellung der Statistischen Methoden unterteilt in die Modellbildung und Variablenselktion (3.1), sowie die Modelldiagnose (3.2). Im Kapitel 4, der statistischen Auswertung, wird nach kurzer dekriptiver Auswertung der Variablen (4.1) das Modell gebildet (4.2) und anschließend diagnostiziert und verbessert (4.3). Zuletzt erfolgt das Unterkapitel 4.4 zur Interpretation der Koeffizienten des Parametervektors, bevor alle zentralen Ergebnisse in Kapitel 5 zusammengefasst werden.

2 Problemstellung

Um eine sachliche Entscheidung über die Festlegung der Miete für eine bestimmte Immobilie zu erleichtern spielt die Betrachtung der Vergleichsmiete eine entscheidende Rolle. Diese wird aus einem in zahlreichen Städten erstellten Mietspiegel gewonnen und steht dabei unter Beachtung gesetzlicher Definitionen. In diesen wird eine feste Grundgesamtheit aus betrachteten Mieten festgelegt aus der eine repäsentative Zufallsstichprobe gezogen werden soll. Diese schließt beispielsweise gestzlich festgelegte oder geförderte Mieten aus. Zudem resultiert aus dem Gesetzestext, dass die durchschnittliche Netto-

miete der Regressand ist. Dieser soll durch mehrere Regressoren erklärt werden, welche sowohl die Größe, Ausstattung, Beschaffenheit und Lage einschließlich der energetischen Ausstattung berücksichtigen soll.
In Rahmen dieses Projektes wird die Vergleichsmiete im Raum München, wie obig beschrieben, anhand von Daten eines Ausschnittes des Mietspiegels aus dem Jahr 2015, mithilfe eines multiplen linearen Regressionsmodell geschätzt. Dabei steht vor allem die Einordnung bzw. Prognose von neuen Beobachtungen im Vordergrund. Der vorliegende Datensatz *mietspiegel2015* besteht dabei aus 13 Variablen und 3065 Beobachtungen. Die detaillierten Beschreibungen der Variablen sind dabei in Tabelle 1 zu finden.
Der Datensatz beinhaltet zunächst die in diesem Projekt als Regressand verwendete metrisch, diskrete Variable *Nettomiete pro Monat*. Die ebenfalls als Regressand geeignete Variable *Nettomiete pro Monat und Quadratmeter* wird in diesem Projekt nicht betrachtet. Alle weiteren genannten Variablen gehören zu den Regressoren. Dazu gehört das metrisch, stetige Merkmal *Wohnfläche*, welches auf ganze Zahlen gerundet in Quadratmeter vorliegt. Das *Baujahr* der Immobilie und die *Anzahl an Zimmern* sind hingegen metrisch, diskrete Variablen. Alle weiteren Merkmale sind nominal und beschreiben sowohl die Ausstattung (bzgl. Energieversorgung, Inventar des Bads und der Küche) als auch die Wohnlage der Immobilie. Dabei ist ein Großteil der Variablen dichotom mit den Ausprägungen „0“ und „1“. Lediglich der *Bezirksname* besteht aus den 25 Bezirken der Stadt München. Für die in den Variablen *gute Wohnlage*, textitbeste Wohnlage zu findende Bewertung der Wohnlage wurde außerdem ein Gutachter hinzugezogen. Die Bezeichnungen „ja“ und „nein“ bei den Variablen *Warmwasserversorgung* und *Zentralheizung* geben jeweils an ob dies von Vermieter gestellt bzw. verfügbar ist. Durch eine vorherige Einteilung des *Baujahres* in Klassen, liegen durch die Klassenauflösung einige Werte reelwertig vor (beispielsweise 1957.5). Die Variable *Baujahr* an sich ist jedoch ein ganzzahliges, metrisches Merkmal. Die Qualität der vorliegenden Daten ist sehr gut, da in keiner der Variablen fehlende Werte vorliegen.

Variablenname (kurz)	**Ausprägung/Kodierung**		**Skalenniveau**
Nettomiete pro Monat (nm)	reellwertig in Euro (€)		metrisch, stetig
Nettomiete pro Monat und Quadratmeter (nmqm)	reellwertig in Euro (€)		metrisch, stetig
Wohnfläche (wfl)	ganzzahlig in Quadratmeter (m^2)		metrisch, stetig
Anzahl Zimmer (räume)	ganzzahlig als Anzahl		metrisch, diskret
Baujahr (bj)	reellwertig als Zeitpunkt (Jahr)		metrisch, diskret
gute Wohnlage (wohngut)	1 gute Lage	0 andere Lage	nominal, dichotom
beste Wohnlage (wohnbest)	1 beste Lage	0 andere Lage	nominal, dichotom
Warmwasserversorgung (ww)	1 nein	0 ja	nominal, dichotom
Zentralheizung (zh)	1 nein	0 ja	nominal, dichotom
gefliestes Bad (badkach)	1 nicht gefliest	0 gefliest	nominal, dichotom
Ausstattung Bad (badextra)	1 gehoben	0 nicht gehoben	nominal, dichotom
Ausstattung Küche (küche)	1 gehoben	0 nicht gehoben	nominal, dichotom
Bezirkname (bez)	Bezirkname in München		nominal

Tabelle 1: Variablen des Datensatzes

3 Statistische Methoden

Alle folgenden statistischen Methoden werden in der Version 4.1.1 der Software R durchgeführt (R Core Team (2021)). Dabei wird bei Ergebnissen, wenn nicht anderes angegeben, auf zwei Nachkommastellen gerundet.

3.1 Modellbildung und Variablenselektion

Elementar für dieses Projekt ist das klassiche allgemeine lineare Modell (Fahrmeir et al. (2009), S.62). Dieses besteht aus der Designmatrix $X \in \mathbb{R}^{n \times k}$, dem Parametervektor $\beta \in \mathbb{R}^k$ und den Zufallsvektoren $y, e \in \mathbb{R}^n$. Dabei ist y der Vektor der Beobachtungen und e ein unbeobachtbarer Vektor der Fehler. Die Dimensionen sind durch $n, k \in \mathbb{N}$ gegeben.

Nun ergibt sich die Modellgleichung als $y = X\beta + e$. Zudem soll gelten, dass der Erwartungswert und die Kovarianz von y existieren. Für den Erwartungswert gilt $\mathbb{E}(y) = X\beta$ und somit auch $\mathbb{E}(e) = 0$. Zudem sollen die Fehler e_i mit homoskedastischer Varianz normalverteilt sein, d.h. $e_i \sim N(0, \sigma^2)$. Außerdem ist eine Unkorreliertheit der Fehler untereinander erwünscht, sodass $Cov(e) = \sigma^2 I$ gilt. Hieraus ergibt sich für $i = 1, \ldots, n$ sowohl das in (1) definierte multiple lineare Regressionsmodell (Fahrmeir et al. (2009), S. 24) als auch die in (2) definierte polynomiale Regression (Fahrmeir et al. (2009), S. 153). Dabei werden Polynome vom Grad 1 bis $l \in \mathbb{N}$ angenommen.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e_i \tag{1}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_n x_i^l + e_i \tag{2}$$

Eine klassische Methode zur Bestimmung einer Schätzung des Parametervektos β ist die Kleinste Quadrate Schätzung (kurz: KQ-Schätzer). Dieser Schätzer ist definiert als die Lösung des folgenden Minimierungsproblemes, welches sich bei vollem Spaltenrang wie in (3) berechnen lässt (Fahrmeir et al. (2009), S. 90 bis 92).

$$\min_{\beta \in \mathbb{R}^k} \|y - X\beta\| \quad \text{d.h. falls } rg(X) = k \text{ ist } \hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

Falls schwache Multikollinearität vorliegt, sind mindenstens zwei Spalten der Designmatrix fast linear abhängig, was zu einer Erhöhung der Varianz des Schätzers $\hat{\beta}$ und somit zu einer Unzuverlässigkeit des KQ-Schätzer führt (Fahrmeir et al. (2009), S. 102). Wie Multikollinearität diagnostiziert werden kann wird näher in Kapitel 3.2 beschrieben. Um in diesem Fall eine aussagekräftige Schätzung zu erhalten wird auf eine alternative Schätzung in Form eines Shrinkage Schätzers zurückgegriffen. Dabei ist die Lasso (Least absolute shrinkage and selection operator) Schätzung, im Falle eines multiplen Regressionsmodell wie in (4) definiert (James et al. (2021) S. 241). Analog kann dieser auch für die polynmiale Regression verwendet werden.

$$\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\Big)^2 + v \sum_{j=1}^{k} |\beta_j| \quad v \in \mathbb{R},\ v > 0 \tag{4}$$

Der letzte Summand ist ein Strafterm, der die Koeffizienten von $\hat{\beta}$ schrumpfen lässt. Dabei wird das v durch eine Kreuzvalidierung unter Verwendung der Funktion `cv.glmnet()` aus dem Paket `glmnet` (Friedman et al. (2010)) passend gewählt.
Zur Variablenselektion werden zweiseitige t-Tests verwendet. Diese testen, wie in (5) be-

schrieben, ob ein geschätzter Koeffizient des Parametervektors $\hat{\beta}$ signifikant von Null verschieden ist. Dies geschieht mit dem wie in (6) definierten Test φ (Fahrmeir et al. (2009), S. 116). Dabei wird mit $n_e = n - p$ der Fehlerfreiheitsgrad mit der Anzahl von Regressoren $p \in \mathbb{N}$ bezeichnet.

$$H_0 : \hat{\beta}_i = 0 \quad vs. \quad H_1 : \hat{\beta}_i \neq 0 \quad T = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \quad \text{mit} \quad s_{\hat{\beta}_i} = \sqrt{\widehat{Var}(\hat{\beta}_i)} \tag{5}$$

$$\varphi := \begin{cases} 0 \quad \text{falls} \quad |T| \leq t_{n_e, 1-\frac{\alpha}{2}} \\ 1 \quad \text{falls} \quad |T| > t_{n_e, 1-\frac{\alpha}{2}} \end{cases} \tag{6}$$

Insbesondere wird der p-Wert als eine Überschreitungswahrscheinlichkeit des Tests φ betrachtet. Falls dieser kleiner als α ist, lässt sich die Nullphypothese unter Einhaltung des Signifikanzniveaus α ablehnen.
Bei Selektionsverfahren durch den p-Wert (James et al. (2021) S. 79) wird vorher ein cut-off Wert für α festgelegt. Bei der Rückwärtselimination werden, bei Start des vollen Modells, schrittweise alle Variablen eliminiert, die einen p-Wert kleiner als den cut-off Wert haben. Bei der Vorwärtsselektion wird dieser Prozess umgedreht und es werden schrittweise die Variablen mit dem kleinsten p-Wert aufgenommen, bis keiner der p-Werte unter dem cut-off Wert liegt. Bei einer schrittweisen Selektion werden beide Verfahren kombiniert.
Ein weiteres Selektionskriterium zur Bewertung der Anpassung des Modells ist das adjustierte Bestimmtheitsmaß R^2_{adj} (Fahrmeir et al. (2009), S. 160). Dieses setzt sich aus dem Bestimmtheitsmaß R^2 zusammen, welches den Anteil der durch das Modell erklärte Streuung an der Gesamtstreuung angibt (Fahrmeir et al. (2009), S. 99).

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \in [0, 1] \tag{7}$$

Werte nahe 1 sprechen für eine gute Anpassung des Modelles und Werte nahe Null für eine schlechte Anpassung. Da R^2 aber auch bei Hinzunahme von unwichtigen Regressoren steigt, ist das in (10) definierte adjustierte Bestimmtheitsmaß besser zur Variablenselektion geeignet.

$$\tilde{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2) \quad \leq 1 \tag{8}$$

Denn dieses bestraft die Hinzunahme von zusätzlichen Variablen mit einem Strafterm, der sich aus n der Anzahl Beobachtungen und $p \in \mathbb{N}$, die Anzahl der Regressoren, zusammensetzt. Somit kann dieses Maß auch negativ werden jedoch nicht größer als 1 werden. Das heißt, dass Werte nahe 1 weiterhin für eine gute Modellanpassung sprechen.

3.2 Modelldiagnositik

Zur Überprüfung der Modellannahmen werden die gwöhnlichen Residuen $\hat{e}_i = y_i - \hat{y}_i$ betrachtet, die in einigen Anwendungen mithilfe des i-ten Diagonalelemtes h_{ii} der Hat Matrix $H = X(X^TX)^{-1}X^T$ und dem mittleren quadratischen Fehler (MSE) wie in (11) standardisiert werden (Fahrmeir et al. (2009), S. 110).

$$e_i^* = \frac{\hat{e}_i}{\sqrt{MSE} \cdot \sqrt{1 - h_{ii}}} \qquad MSE = \frac{\hat{e}^T\hat{e}}{n_e} \tag{9}$$

Bei einem einfachen Residualplot werden die Residuen $\hat{e}$ gegen die angespassten Werte $\hat{y} = X\hat{\beta}$ abgetragen (James et al. (2021), S. 93 bis 94). Zudem wird ein Scale-plot mit analogem Prinzip des Residualplots unter Verwendung der Transformation der Residuen $\tilde{e_i} = \sqrt{|e_i^*|}$ genutzt. Diese Definition entpricht der in R implementierten Version aus der `plot.lm` Funktion unter Angabe des Arguments `which = 3` genutzt (R Core Team (2021)). Diese ist geeignet um Heteroskedastizität erkennen. Denn beim einfachen Residualplot kann der Effekt auftreten, dass mit größer werdendem $\hat{y}$ die Residuen größer werden, obwohl diese im Verhältniss genauso größ sind wie bei kleinen $\hat{y}$ Werten. Durch die Standardisierung wird dieser Effekt unterbunden.
Zur Überprüfung der Normalverteilungsannahme der Residuen wird ein Quantile-Quantile-Plot benutzt (Hartung et al. (2009), S.847). Bei diesem wird ebenfalls die in R implementierte Verion innerhalb `plot.lm` verwendet, in der die empirischen Quantile der standardisierten Residuen e^* gegen die theoretischen Quantile der Normalverteilung abgetragen werden(R Core Team (2021)). Falls ein Großteil der Punkte auf der Winkelhalbierenden liegt spricht dies für die Erfüllung der Normalverteilungsannahme.
Der Leverage score (deutsch: Hebel, Einfluss) ist ein Abstandsmaß, dass bezüglich einer unabhängigen Variable den Abstand einer Bobachtung i zu den übrigen Beobachtungen angibt (Fahrmeir et al. (2009), S. 177 bis 178). Das Maß ist für die i-te Beobachtung ist definiert als das, zwischen 0 und 1 liegende, i-te Diagnolalelement h_{ii} der Hatmatrix. Der durchschnittliche Leverage score beträgt p/n. Dabei werden Beobachtungen mit einem

Leverage score größer als $2 \cdot (p/n)$ als stark einflussreiche Beobachtungen (high-leverage points) bezeichnet. Diese Bobachtungen besitzen eine große Hebelwirkung bzw. Einfluss auf die Regressionsgerade.
Haben high leverage ponits zudem große Residuen, spricht man von einflussreichen Beobachtungen. Solche Beobachtungen können mithilfe der Cook's Distance bestimmt werden. Dazu berechnet man wie in (12) die Summe der quardierten Änderungen, wenn die i-te Beobachtung entfernt wird (Fahrmeir et al. (2009), S. 178).

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot \text{MSE}} \tag{10}$$

Hierbei ist eine Beobachtungen ab einem Wert von $D_i > 0.5$ auffällig und sollte ab dem Wert $D_i > 1$ auf jeden Fall näher untersucht werden .
Um zu ergründen ob Multikollinearität ein Problem darstellt, wird die Determinante von $X^T X$ berechnet. Denn diese ist im Falle schwacher Multikollinearität nahe an Null (Toutenburg (2003), S. 114). Zudem wird hier der Varianzinflationskoeffizien (VIF), definiert als $VIF_i = 1/1-R_i^2$ als Indikator für Multikollinearität genutzt (Fahrmeir et al. (2009), S. 170 bis 171). Dabei bezeichnet R_i^2 den multiplen Korrelationskoeffizient bei einer Regression wo x_i als abhängige Variable und alle weiteren Prädikatoren als unabhängige Variablen gesehen werde. Ein Wert des VIF größer als 10 spricht dabei für Multikollinearität. Die Umsetzung in R erfolgt über die Funktion `vif()` aus dem Paket `car` (Fox und Weisberg (2019)).

4 Statistische Auswertung

4.1 Deskriptive Beschreibung des Datensatzes

Der Tabelle 2 sind die deskriptiven Kennzahlen aller metrischen Variablen zu entnehmen. Interessant zu betrachten ist der zukünftige Regressand *Nettomiete pro Monat (nm)*, der eine rechtsschiefe, spitze Verteilung mit einer deutlich größeren Standardabweichung als die *Nettomiete pro Monat und Quadratmeter(nmqm)* aufweist. In der Abbildung 6, die im Anhang auf Seite 19 zu finden ist, lässt sich die Rechtsschiefe auch grafisch erkennen. Dabei wird zur besseren Sichtbarkeit der Verteilung die Beobachtung 1975 mit einer

Nettomiete pro Monat von 6000 Euro nicht dargestellt. Zur Berechnung der Schiefe und Wölbungsmaße wurde das moments (Komsta und Novomestky (2022)) verwendet. In Tabelle 3 sind die relativen Häufigkeiten der dichotomen Variablen, die alle zu den

	arithm. Mittel	Median	Standardabweichung	IQR	Schiefe	Wölbung
nm	763.06	700.00	338.16	360.46	2.59	25.47
nmqm	10.73	10.84	2.67	3.42	0.04	3.34
wfl	71.98	70.00	25.74	30.00	1.35	8.33
räume	2.70	3.00	0.98	1.00	0.46	3.60
bj	1964.21	1957.50	26.51	25.50	-0.18	2.31

Tabelle 2: Deskriptive Kennzahlen der metrischen Variablen

möglichen Regressoren gehören vorzufinden. Hier ist auffällig, dass bei den Merkmalen *wohnbest* und *ww* die Ausprägung „1“ nur sehr selten vorkommt. Eine vollständige Übersicht über die relativen Häufigkeiten der *Bezirke* ist im Anhang auf Seite 18 in Tabelle 5 zu finden. Die in Abbildung 1 dargestellten Korrelationen wurden mit dem Rangkorrela-

	wohngut	*wohnbest*	*ww*	*zh*	*badkach*	*badextra*	*kueche*
Ausprägung „0“	0.65	0.96	0.99	0.93	0.12	0.88	0.75
Ausprägung „1“	0.35	0.04	0.01	0.07	0.88	0.12	0.25

Tabelle 3: Relative Häufigkeiten der dichotomen Variablen (n = 3065)

tionskoeffizienten nach Spearman berechnet um einen Vergelich zwischen diskreten und metrischen Variablen möglich zu machen. Dabei weist die Variable *Wohnfläche* mit 0.75 die höchste Korrelation mit der abhängigen Variablen auf, gefolgt vom Merkmal *Anzahl Zimmer* mit einer moderaten Korrelation von etwa 0.61. Auch die *Nettomiete pro Monat und Quadratmeter* ist mit 0.44 leicht mit dem Regressand *Nettomiete pro Monat* korelliert. Denn beide Variablen beruhen auf der Nettomiete, wobei bei der *Nettomiete pro Monat und Quadratmeter* zusätzlich die Quadratmeter berücksichtigt werden. Aufgrund dieser Dopplung von Informationen wird die *Nettomiete pro Monat und Quadratmeter* als möglicher Regressor ausgeschlossen. Die höchste Korrelation besteht mit 0.86 zwischen den Variablen *Wohnfläche* und *Anzahl der Zimmer*. Dies könnte ein Hinweis auf Multikollinearität sein, der später überprüft wird. Ansonsten liegen kaum nennenswerte sehr leichte negative/positive Korrelationen zwischen den Variablen vor.

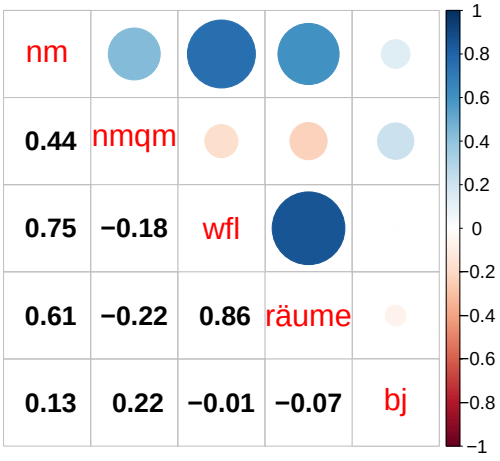

Abbildung 1: Korrelationen zwischen den metrischen Variablen

4.2 Modellbildung und Variablenselektion

Vor der Auswahl der Regressoren erfolgen einige Variablentransformationen. Die beiden Variablen *beste Wohnlage* und *gute Wohnlage* werden in einem neuen kategoriellen Merkmal *Wohnlage* mit den drei Ausprägungen „beste“, „gute“ oder „andere“ Lagekategorie eingeteilt.
Wie schon im vorherigen Kapitel erwähnt wird die Variable *Nettomiete pro Monat und Quadratmeter* als eine mögliche Einflussvariable ausgeschlossen. Die Verwendung der Variable *Wohnfläche* als erklärende Variable ist jedoch aufgrund der hohen Korrelation mit dem Regressand vielversprechend. Führt man, dadurch motiviert, explorativ eine einfache lineare Regression mit der *Wohnfläche* aus, erhält man schon ein recht hohes adjustiertes Bestimmtheitsmaß von $R^2_{adj} = 0.6133$ und einen p-Wert für β_1^{wfl}, der kleiner als $2 \cdot 10^{-16}$ ist. Das heißt die Nullhypothese, dass der Koeffizient β_1^{wfl} bezüglich der *Wohnfläche* Null ist, kann zum Niveau $\alpha = 0.001$ abgelehnt werden. Der in Abbildung 2 erkennbare Verlauf der Regressionsgerade stimmt grrößtenteil mit dem Verlauf der Punktewolke überein.

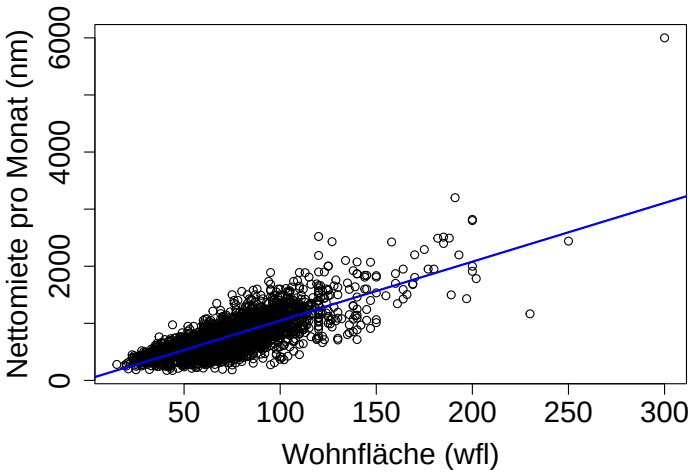

Abbildung 2: Einfache Regression durch die *Wohnfläche*

Aufgrund dem leicht gekrümmten Verlauf der Punktwolke könnte auch eine polynomialer Ansatz sinnvoll sein (vgl. Abbildung 7 auf Seite 19). Jedoch ergeben sich bei einer polynomialen Regression zweiten Grades, welche die *Nettomiete pro Monat* durch die *Wohnfläche* und die quadrierte *Wohnfläche* erklärt, Probleme mit der Multikollinearität (vgl. VIF von $wfl = 17.698541$, VIF von $wfl^2 = 11.256685$). Bei einer Vorwärtsselektion bzw. Rückwärtselimination mit einem cut-off Wert von 0.01 bleiben die beiden Terme der polynomialen Regression (und alle weitern Variablen bis auf den Bezirk) zwar erhalten. Wendet man aber aufgrund der Multikollinearität nach einer Kreuzvalidierung zur Bestimmung des Parameters $v = 0.4315911$ eine Lasso Regression an wird der quadratische Term mit 0.01975573 sehr klein geschätzt. Sogar der KQ-Schätzer schätzt den Einfluss mit 0.01956 nahe Null. Dies liegt auch daran, dass sich der Lasso-Schätzer aufgrund des kleinem v kaum von dem KQ-Schätzer unterscheidet. Somit wird dieser Ansatz im folgenden nicht weiter verfolgt.
Hingegen wird eine Vorwärtsselektion bzw. Rückwärtselimination ohne quadratische Terme auf alle möglichen Regressoren angewendet. Mit dem vorher festgelgten cut-off Wert für den p-Wert von $\alpha = 0.01$, werden bei Vorwärtsselektion der der Reihenfolge nach die Variablen *Wohnfläche*, *Ausstattung Küche*, *Wohnlage*, *Baujahr*, *Anzahl Zimmer*, *Warmwasserversorgung*, *gefliestes Bad*, *Zentralheizung* und *Ausstattung Bad* hinzugefügt. Zuletzt gilt zu entscheiden ob der *Bezirkname* aufgenommen werden soll. Selektion anhand weiteren Kriterien, wie die Minimierung des AICs oder die Maximierung des adjustieren

Bestimmtheitsmaßes weisen auf das volle Modell mit der Variable *Bezirk* hin. Da jedoch bei Betrachtung der Signifikanz dieses Merkmales nur die Dummy-Variable des Bezirkes „Ludwigvorstadt-Isarvorstadt“ den cut-off Wert von 0.01 einhält, wird auf die Hinzunahme des Merkmales *Bezirkname* zur Komplexitätsreduzierung des Modelles verzichtet.
Das nun resultierende Modell weist mit 0.6859 ein höheres adjustiertes Bestimmtheitsmaß als das einfache lineare Modell aus Abbildung 2 auf. Das heißt die Variablenhinzunahme trägt zur Erklärung der Variable *Nettomiete pro Monat* bei und hat das einfach Regressionsmodell verbessert. Bei der Rückwärtselimination erfolgt ebenfalls die Auswahl dieses Modelles, da alle p-Werte im vollen Modell unter $\alpha = 0.01$ liegen.
Noch näher zu betrachten ist die kritisch zu sehende Aufnahme der Variable *rooms*. Denn in die Abbildung 1 erkennbare hohe Korrelation mit der Variable *Wohnfläche* könnte zu Problemen führen. Zunächst wird untersucht, ob ein Problem durch Multikollinearität vorliegt. Die Determinante von X^TX liegt mit $1.672006 \cdot 10^{35}$ sehr weit weg von der Null. Auch der VIF von 3.558245 der Variable *Wohnfläche* und der Variable *Anzahl Zimmer* (3.513331) ist zwar etwas höher als bei den anderen Variablen, aber erst ein VIF von 5 bzw. 10 wird als ein kritsicher Wert für die Multikollinearität gesehen. Jedoch wird der Koeffizient der *Anzahl der Zimmer* mit einem fragwürdigen großen negativen Wert von -55.54 geschätzt. Das würde bedeuten, dass mehr Zimmer zu einer günstigeren Miete führen würde. Aufgrund dieses Ergebnisses, dass aus einer möglichen Wechselwirkung mit der *Wohnfläche* resultieren könnte, wird die Variable *Anzahl der Zimmer* entfernt. Das nun resultierende Modell hat ein nur geringfügig kleineres adjustiertes Bestimmtheitsmaß von 0.6786 und alle aufgenommenen Variablen haben einen p-Wert der unter dem cut-off Wert von 0.01 liegt. Jegliche Werte des VIF liegen unter 1.2 und die Determinante von X^TX liegt mit $2.000894 \cdot 10^32$ weit weg von der Null. Somit liegt keine Multikollinearität vor, die alternative Schätzmethoden fordern würde und der KQ-Schätzer kann für die Schätzung von $\hat{\beta}$ verwendet werden.

4.3 Modelldiagnostik

Im folgendem wird untersucht ob das Modell alle Modellannahmen erfüllt. Zunächst wird untersucht ob das Modell einflussreiche Beobachtungen (mit einem größem Hebelwert) enthält. In Abbildung 3 sind dazu die standardisierten Residuen gegen den Leverage abgetragen. Die orange vertikale Linie gibt den cut-off Wert von $2 \cdot {}^{k}/_{n}$ an. Es ist erkennbar, dass nur ein standardisiertes Residuum der Beobachtung 1975 eine

Cook's Distance größer als 0.5 und einen Leverage größer als $2 \cdot {}^{k}/_{n}$ hat. Diese Beobachtung ist also einflussreich und hat eine große Hebelwirkung. Auch die Beobachtungen 1263 und 231 sind auffällig, da diese einen hohen Leverage und eine tendenziell höhere Cook's Distance als die anderen Werte aufweisen. Generell ist zu erkennen, dass viele Residuen einen hohen Leverage haben. Dies könnte daran liegen, dass für große $\hat{y}$-Werte nur wenig Beobachtungen vorliegen, welche nun einen großen Einfluss auf die Regression ausüben.

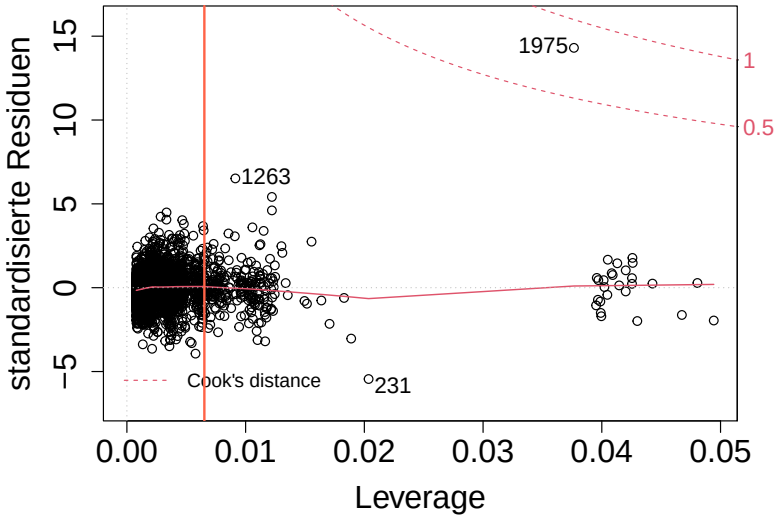

Abbildung 3: Leverage und Cook's Distance

Die obig genannten auffälligen Beobachtungen sind auch in der Abbildung 8 auf Seite 20 im Anhang erkennbar. Diese zeigt einen QQ-Plot, bei dem zu erkennen ist, dass im mittleren Teil ein Großteil der Punkte auf der Verbindungslinie zwischen dem ersten und dritten Quartil liegt, dessen Lage einer Winkelhalbierenden (in rot) gleicht. An den Rändern gibt es jedoch einige Punkte, die deutlich nach unten oder oben abweichen.
Somit werden die Beobachtungen 1975, 1263 und 231 probehalber entfernt, was neben dem Aspekt des Leverage zumindest zu einer Verbesserung der Ränder des Quantile-Quantile-Plot führt und somit der Erfüllung der Normalverteilungsannahme näher kommt. Zudem liegen die Residuen im Mittel bei einem Median von 2.11 vergleichsweise näher an Null als im vorherigen Modell mit einem Median von 3.34 und erfüllen somit auch besser die Modellannahme, dass der $\mathbb{E}(e) = 0$ ist. Dies ist auch gut in Abbildung 4 an der roten Linie erkennbar, die bis auf eine leichte anfängliche Schwankung bei Null liegt.

Ein Größteil der Residuen stammt aus dem Wertebereich zwischen 400 und 1000 der angepassten y-Werte und streut dort etwas weniger um die Null herum als in Wertebereich zwischen 1000 und 2500.

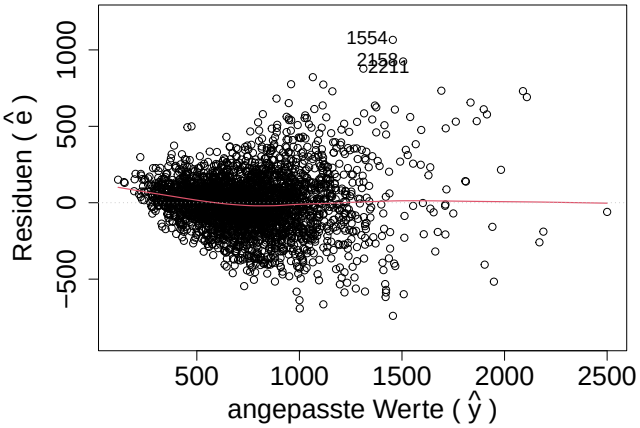

Abbildung 4: Residualplot ohne Beobachtungen 1975, 1263, 231

Um besser bewerten zu können ob die in Abbildung 4 zu beobachtende größer werdende Streuung der Residuen an einer Verletzung der Homoskedastizität liegt, wird in Abbildung 5 ein Scaleplot unter Verwendung von standardisierten Residuen betrachtet. Dort ist zu erkennen, dass die rote Linie zunächst konstant bleibt und dann von etwa 0.5 auf 1.8 ansteigt. Auch die Form der Punktewolke ähnelt annähernd einer Ellipse, dessen Achse der Ausrichtung der roten Linie entspricht. Beide Beobachtungen sprechen für Hetreoskedastizität, da die Residuen $\tilde{e}$ mit wachsenden $\hat{y}$ immer größer werden.

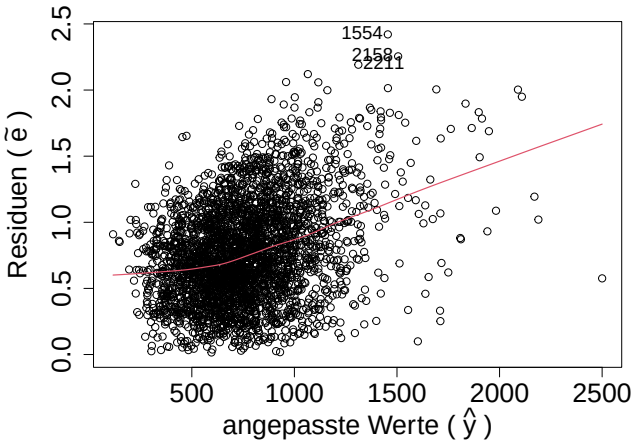

Abbildung 5: Scaleplot ohne Beobachtungen 1975, 1263, 231

Zusammenfassend lässt sich mit dem Modell also zumindest die Modellannahme, dass $\mathbb{E}(y) = X\beta$ ist durch die Herrausnahme der Beobachtungen 1975, 1263 und 231 verbessern. Wie am Qauntile-Quantile-Plot erkennbar ist, könnte man auch die Normalverteilungsannahme, unter Beobachtung von Abweichungen an den Rändern, akzeptieren oder zumindest als teilweise erfüllt sehen. Kritischer zu sehen ist das Vorliegen der Heteroskedastizität, die schon deutlich erkennbar ist. Da das Eliminieren der Beobachtungen 1975, 1263 und 231 die Erfüllung Modellannahmen tendenziell zu begünstigen scheint, ist somit auch ein Verlust des Bestimmtheitsmaßes in dritter Nachkommastelle von etwa 0.679 auf etwa 0.677 zu rechtfertigen. Das Entfernen verbessert auch aus inhaltlicher Sicht die Prognoseeigenschaft des Modelles, da untypische Beobachtungen nicht die Einordnung neuer Mietobjkete verzerren.

4.4 Interpretation der Koeffizienten des Parametervektors

Die Schätzungen des nun relsutierenden KQ-Schätzers (ohne Beobachtungen 1975, 1263 und 231) sind in Tabelle 4 zu findenden. Die zugehörigen p-Werte des Signifikanztests sind in Tabelle 6 auf Seite 18 im Anhang zu finden. Der Einfluss der (Ausprägungen) der Variablen *Wohnfläche*, „gehobene“ *Ausstattung der Küche*, „beste“ *Wohnlage*, „gute“ *Wohnlage*, *Baujahr*, „nicht vorhanden sein“ eines *gefliesten Bades* und „gehobene“ *Ausstattung des Bades* wird positiv auf die Miete eingeschätzt. Das heißt in Bezug auf

Intercept	-2196.526	*bj*	1.093
wfl	9.835	*ww* (nicht vorhanden)	-187.504
kueche (gehoben)	90.756	*badkach* (nicht vorhanden)	54.675
wohnlage (beste)	111.141	*zh* (nicht vorhanden)	-62.036
wohnlage (gute)	87.188	*badextra* (gehoben)	37.681

Tabelle 4: Koeffizienten des geschätzten Parametervektors $\hat{\beta}$

die numerischen Variablen *Wohnfläche* und *Baujahr*, dass wenn alle anderen Variablen konstant im Modell vorliegen und die jeweilige Variable um eine Einheit steigt, dass die *Nettomiete pro Monat* um etwa 9.835 bzw. 1.093 Euro steigt. Bei den dichotomen Variablen steigt die *Nettomiete pro Monat* um den jeweiligen Koeffizienten im Vergleich zur Referenzkategorie. Das heißt beispielsweise, dass die *Nettomiete pro Monat* in „bester“ *Wohnlage* um etwa 111.141 Euro teurer ist als in einer „anderen“ *Wohnlage*. Mit negativem Koeffizienten werden der Intercept und das „nicht vorhanden sein“ einer *Warmwasserversorgung* und *Zentralheizung* geschätzt. Dabei ist der Intercept nicht sinnvoll interpretierbar, da die *Wohnfläche* nicht Null werden kann. Die anderen beiden negativen Einflussvariablen sind analog zu den positiven, aber nun in Form einer Mietverringerung zu interpretieren.

5 Zusammenfassung

Zur Schätzung der Vergleichsmiete, anhand eines Ausschnittes des Münchener Mietspiegels aus dem Jahr 2015 wurde eine möglichst gutes multiples lineares Modell gesucht. Dieses soll die *Nettomiete pro Monat* mithilfe verschiedenster Regressoren, welche die Ausstattung, Energieversorgung und Wohnlage der Mietobjekte beschreiben, erklären.
Vor der Modellsuche wurde zunächst die Variable *Wohnfläche*, aufgrund der hohen Korrelation von 0.78 mit der Zielvariable, als vielversprechnder Regressor ausgemacht. Ansätze einer polynomialen Regression mit der *Wohnfläche* und weiteren zusätzlichen Regressoren erwiesen sich dabei als nicht sinnvoll.
Daher wurde stadtdessen der Ansatz eines multiplen linearen Regressionsmodelles verfolgt. Als Konsequnz einer Vorwärtsselektion wird dabei die Variable *Bezirk* eliminiert, sodass das Modell die Variablen *Wohnfläche*, *Ausstattung Küche*, *Wohnlage*, *Baujahr*, *Anzahl Zimmer*, *Warmwasserversorgung*, *gefliestes Bad*, *Zentralheizung* und *Ausstattung*

Bad enthält. Im weiteren Verlauf wurde die Variable *Anzahl der Zimmer* aufgrund logischer Überlegungen und möglichen Wechselwirkungen mit der *Wohnfläche* eliminiert. Im folgenden wurden dann die einlussreichen Beobachtungen 1975, 1263 und 231 entfernt um die Modellannahmen besser zu erfüllen und die Prognosefähigkeit des Modells zu verbessern. Eine detaillierte Darstellung der Koeffizienten von $\hat{\beta}$ ist in Tabelle 4 zu finden. Alle im Modell enthaltenen Variablen, bis auf das Fehlen von *Warmwasserversorgung* oder *Zentralheizung*, haben dabei einen positiven, mieterhöhenden Einfluss.
Leider stellte sich eine deutlich erkennbare Verletzung der Homskedastizität der Varianzen der Fehler heraus, was dazu führt dass die Schätzung zwar erwartungstreu aber nicht mehr effizient bleibt (Auer und Rottmann (2010), S.518 bis 520). Um dies zu verhindern könnte man eine gewichtete KQ-Schätung mit voher geschätzten Gewichten $\hat{w}_i = 1/\hat{\sigma_i^2}$ durchführen. Das heißt es gibt unter Umständen einen besseren Schätzer für $\hat{\beta}$. Zudem wird der Standardfehler verzerrt, was dazu geführt haben könnte, dass die Testentscheidungen im t-Test verfälscht wurden. In Anbetracht dieser Information könnten könnte es auch plausibel sein, diese Merkmale trotz hoher Signifikanz zu eliminieren. Denn die hohen negativen Schätzungen der mietverringernden Merkmale könnte an der nur sehr kleinen relativen Häufigkeit von Wohnungen ohne *Warmwasserversorgung* bzw. *Zentralheizung* liegen. Zudem könnte es sinnvoll sein Wechselwirkungsterme in das Modell zu integrieren (z.B. *Wohnfläche* und *Anzahl der Zimmer*). Auch ist es zunächst verwunderlich, dass ein *gefliestes Bad* mietverringernd wirkt. Auch hier könnte eine Wechselwirkung vorliegen. Denn *geflieste Bäder* sind häufiger in älteren Immobilien als in Neubauten zu finden, welche tendenziell eine höhere *Nettomiete pro Monat* haben. Jedoch könnte es auch sein, dass bei Mietobjekten tatsächlich Bäder ohne Fliesen bevorzugt werden. Dies könnte durch eine Plausibilitätsprüfung eines Fachmannes (z.B. Immobilinemarkler) näher ergründet werden.

Literatur

Auer, B. und H. Rottmann (2010). *Statistik und Ökonometrie für Wirtschaftswissenschaftler: eine anwendungsorientierte Einführung. 1. Auflage.*

Fahrmeir, L., T. Kneib und S. Lang (2009). *Regression: Modelle, Methoden und Anwendungen.* Statistik und ihre Anwendungen. 2. Auflage. Springer Berlin Heidelberg.

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

Friedman, J., T. Hastie und R. Tibshirani (2010). „Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33(1), S. 1–22. URL: `https://www.jstatsoft.org/v33/i01/`.

Hartung, J., B. Elpelt und K.-H. Klösener (2009). *Statistik Lehr- und Handbuch der angewandten Statistik. 15.Auflage.* Oldenbourg Verlag: München.

James, G., D. Witten, T. Hastie und R. Tibshirani (2021). *An Introduction to Statistical Learning: with Applications in R. Second Edition.* Springer.

Komsta, L. und F. 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 (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Toutenburg, H. (2003). *Lineare Modelle: Theorie und Anwendung. 2. Auflage.* Springer-Verlag Berlin Heidelberg.

Anhang

Bezirk	relative Häufigkeit
Allach-Untermenzing	0.01
Altstadt-Lehel	0.02
Au-Haidhausen	0.05
Aubing	0.02
Berg am Laim	0.03
Bogenhausen	0.05
Fledmoching-Hasenbergel	0.03
Hadern	0.03
Laim	0.03
Ludwigvorstadt-Isarvorstadt	0.05
Maxvorstadt	0.05
Milbersthofen-Am Hart	0.04
Moosach	0.03
Neuhausen-Nymphenburg	0.08
Obergiesing	0.05
Pasing-Obermenzing	0.04
Ramersdorf-Perlach	0.06
Schwabing-Freimann	0.05
Schwabing West	0.05
Schwanthalerhoehe	0.03
Sendling	0.04
Sendling-Westpark	0.04
Thalkirchen	0.06
Trudering-Riem	0.03
Untergiesing	0.04

Tabelle 5: Relative Häufigkeiten - Bezirke Münchens (n = 3065)

Intercept	$< 2 \cdot 10^{-16}$	*bj*	$9.27 \cdot 10^{-16}$
wfl	$< 2 \cdot 10^{-16}$	*ww* (nicht vorhanden)	$5.30 \cdot 10^{-07}$
kueche (gehoben)	$< 2 \cdot 10^{-16}$	*badkach* (nicht vorhanden)	$1.28 \cdot 10^{-07}$
wohnlage (beste)	$1.32 \cdot 10^{-09}$	*zh* (nicht vorhanden)	$1.23 \cdot 10^{-05}$
wohnlage (gute)	$< 2 \cdot 10^{-16}$	*badextra* (gehoben)	0.000432

Tabelle 6: Koeffizienten des finalen Modells - p-Werte des t-Tests

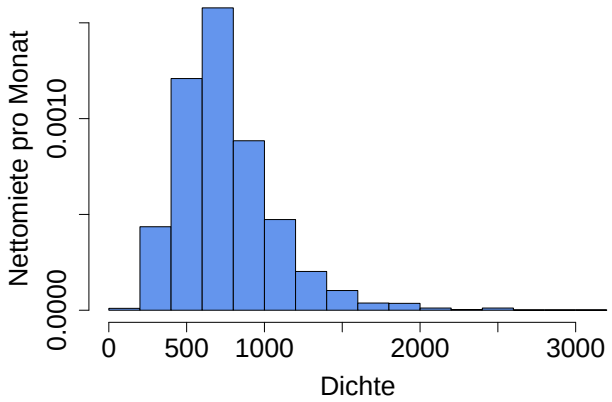

Abbildung 6: Verteilung der *Nettomiete pro Monat* ohne Beobachtung 1957

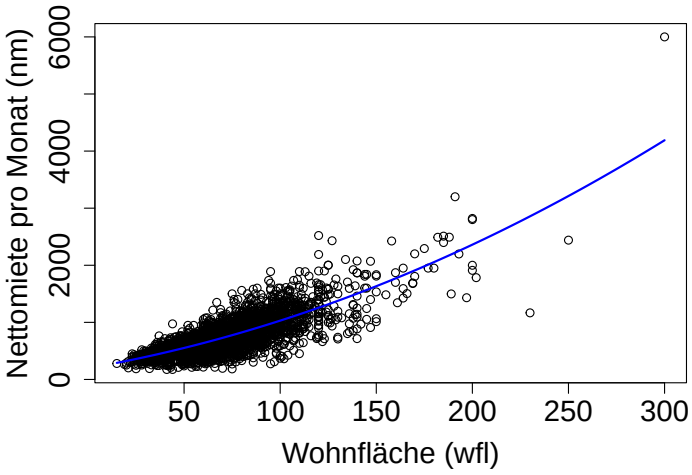

Abbildung 7: Polynomiale Regression zweiten Grades durch die *Wohnfläche*

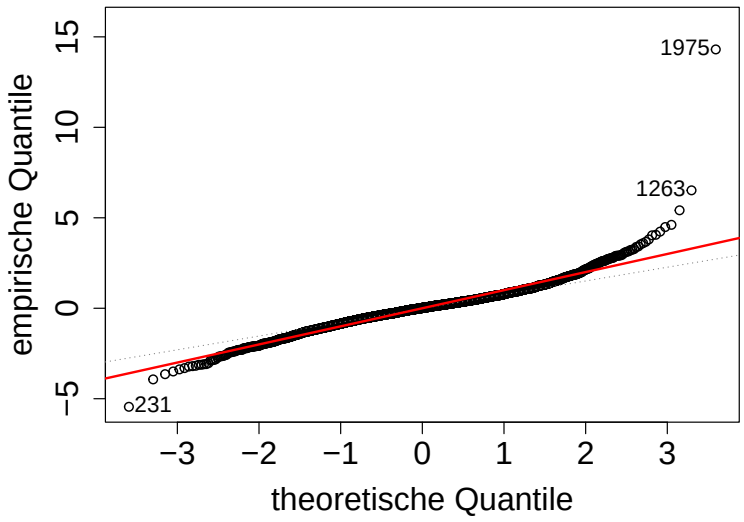

Abbildung 8: Quantile-Quantile-Plot