

Technische Universität Dortmund

Fakultät Statistik

Wintersemester 2022/23

Fallstudien I: Projekt 2

# **Deskriptive Analyse der Demografie einer klinischen Studie**

Dozenten:

Prof. Dr. Guido Knapp

Yassine Talleb, M. Sc.

Verfasserin:

Julia Keiter

Gruppe 1:

Caroline Baer

Julia Keiter

Louisa Poggel

Daniel Sipek

17.11.2022

# Inhaltsverzeichnis

|          |                                 |           |
|----------|---------------------------------|-----------|
| <b>1</b> | <b>Einleitung</b>               | <b>1</b>  |
| <b>2</b> | <b>Problemstellung</b>          | <b>1</b>  |
| 2.1      | Datenmaterial . . . . .         | 1         |
| 2.2      | Ziele des Projekts . . . . .    | 2         |
| <b>3</b> | <b>Statistische Methoden</b>    | <b>3</b>  |
| 3.1      | Modell und Hypothesen . . . . . | 3         |
| 3.2      | Testverfahren . . . . .         | 5         |
| 3.3      | Modellauswahl . . . . .         | 6         |
| 3.4      | Einflussanalyse . . . . .       | 8         |
| 3.5      | Modelldiagnostik . . . . .      | 8         |
| <b>4</b> | <b>Statistische Auswertung</b>  | <b>10</b> |
| <b>5</b> | <b>Zusammenfassung</b>          | <b>14</b> |
|          | <b>Literatur</b>                | <b>15</b> |
|          | <b>Anhang</b>                   | <b>16</b> |

# 1 Einleitung

München gehört zu den Paradebeispielen, wenn in den Medien von der (Un-)bezahlbarkeit der Mieten in deutschen Großstädten berichtet wird. Die Angebotsmieten in Euro pro Quadratmeter für Wohnungen in München sind seit 2012 bis zum zweiten Quartal 2022 um mehr als 63 % gestiegen (Statista Research Department, 2022). Welche Kriterien bei der Bildung der Mietpreise eine Rolle spielen, wird im sogenannten Mietspiegel von vielen deutschen Städten abgebildet. Der Mietspiegel soll als sachliche Entscheidungshilfe dienen, indem eine ortsübliche Vergleichsmiete, die nach BGB §558 insbesondere von

”den üblichen Entgelten, die in der Gemeinde oder einer vergleichbaren Gemeinde für Wohnraum vergleichbarer Art, Größe, Ausstattung, Beschaffenheit und Lage einschließlich der energetischen Ausstattung und Beschaffenheit in den letzten vier Jahren vereinbart oder, von Erhöhungen nach §560 abgesehen, geändert worden sind”,

als Orientierung gegeben wird. Inwiefern die Nettomieten von Variablen wie Art, Größe, Lage oder weiteren Zustandsvariablen abhängig sind, soll in diesem Projekt statistisch untersucht werden.

Im folgenden Kapitel 2 wird ein Ausschnitt des Datensatzes zum Münchener Mietspiegel aus dem Jahr 2015 beschrieben. Mithilfe von in Kapitel 3 vorgestellten statistischen Methoden wird der Datensatz in Kapitel 4 hinsichtlich der Variablenverteilungen kurz deskriptiv beschrieben. Desweiteren wird die Variable Nettokaltmiete von 3065 zufällig ausgewählten Wohnungen aus 25 münchener Stadtbezirken in Abhängigkeit zu den übrigen Variablen des Datensatzes gesetzt, die den Zustand der jeweiligen Wohnung beschreiben. Die Auswertung und Interpretation der Ergebnisse ermöglicht schließlich in Kapitel 5 die Zusammenfassung und eine kurze Diskussion der Ergebnisse.

## 2 Problemstellung

### 2.1 Datenmaterial

Der zugrunde liegende Datensatz *mietspiegel2015.csv* wurde im Auftrag der Landeshauptstadt München von TNS Deutschland erhoben (Sozialreferat der Landeshauptstadt München, 2015). Die Daten zu  $n=3065$  Wohnungen stellen in 13 Variablen die Grundlage dar, eine Datenanalyse zur Erstellung des Mietspiegels für München für das Jahr 2015 durchzuführen. Die Daten hat die Stadt München nicht mehr online gestellt. Die Dokumentation zum Datensatz ist dort allerdings noch verfügbar (Sozialreferat der Landeshauptstadt München, 2015).

Die Daten wurden Mietspiegelinterview erhoben. Die Erhebung erfolgte zum einen Teil in Form eines persönlich-mündlichen Interviews der Mieter und zum anderen Teil in Form

von schriftlichen Fragebögen, die an die Vermieter verschickt wurden. Die schriftlichen Fragebögen konnten analog oder digital ausgefüllt werden. Die Teilnehmenden wurden entweder persönlich oder durch ein Telefoninterview befragt.

In der nominal skalierten Variable *Bezirk* sind die 25 Münchener Stadtbezirke aufgeführt, in denen sich die Wohnungen jeweils befinden. Die übrigen zwölf interessierenden Variablen sind mit den jeweiligen Skalenniveaus und Ausprägungen in Tabelle 1 aufgeführt.

Tabelle 1: erhobene Variablen mit Messniveau und Ausprägungen

| Variable                                    | Skalenniveau      | Ausprägungen                       |
|---------------------------------------------|-------------------|------------------------------------|
| Nettomiete (pro Monat in €)                 | metrisch, diskret | {174.75, ..., 6000.00}             |
| Nettomiete (pro Monat in € und $m^2$ )      | metrisch, diskret | {2.47, ..., 22.13}                 |
| Wohnfläche (in $m^2$ )                      | metrisch, diskret | {15, ..., 300}                     |
| Anzahl der Zimmer                           | metrisch, diskret | {1, ..., 8}                        |
| Baujahr                                     | metrisch, diskret | {1918.0, ..., 2012.5}              |
| Bezirkname                                  | nominal           | {Hadern, Laim,...}                 |
| Gute Lage                                   | nominal, dichotom | {Gute Lage, andere Lagekategorie}  |
| Beste Lage                                  | nominal, dichotom | {Beste Lage, andere Lagekategorie} |
| Warmwasserversorgung vom Vermieter gestellt | nominal, dichotom | {0,1}                              |
| Zentralheizung verfügbar                    | nominal, dichotom | {0,1}                              |
| Gefliestes / Gekacheltes Bad                | nominal, dichotom | {0,1}                              |
| Ausstattung des Bades                       | nominal, dichotom | {0,1}                              |
| Ausstattung der Küche                       | nominal, dichotom | {0,1}                              |

Das Baujahr (ursprünglich kategorielle, an Klassenmitten orientiert). Die Ausprägungen 0 und 1 in den nominal, dichotomen Variablen sind wie folgt zu interpretieren: Bei der Variable Gute Lage steht 0 für eine andere Lagekategorie und 1 für eine gute Lage. Bei der Variable Beste Lage steht 0 für eine andere Lagekategorie und 1 für eine beste Lage. Bei der Variable Warmwasserversorgung vom Vermieter gestellt, in der Tabelle und im Folgenden mit Warmwasser abgekürzt, steht 0 für ja und 1 für nein. Bei der Variable Zentralheizung verfügbar, in Tabelle 1 und im Folgenden mit Heizung abgekürzt, steht 0 für ja und 1 für nein. Bei der Variable Gefliestes / Gekacheltes Bad, in der Tabelle und im Folgenden mit Fliese abgekürzt, steht 0 ja und 1 nein. Bei der Variable Ausstattung des Bades, in der Tabelle und im Folgenden mit Bad abgekürzt, steht 0 für normal und 1 für gehoben. Bei der Variable Ausstattung der Küche, in der Tabelle und im Folgenden mit Küche abgekürzt, steht 0 für normal und 1 für gehoben. Die Datenqualität ist aufgrund keinerlei fehlender Werte im gesamten Datensatz als sehr gut zu bewerten.

## 2.2 Ziele des Projekts

Ziel dieses Projekts ist, ein geeignetes multiples Regressionsmodell zur Schätzung der *Nettomiete* als Regressanden zu konstruieren. Da die Variable *Nettomiete in  $m^2$*  als Quotient der Variablen *Nettomiete* und *Wohnfläche* ein alternativer Regressand und keine unabhängige Variable ist, wird die *Nettomiete in  $m^2$*  in der Modellbildung nicht

berücksichtigt.

Welche Variablen einen signifikanten Einfluss zur Modellierung des Sachverhalts haben, und damit Regressoren genannt werden, wird durch geeignete Test- und Variablenselektionsverfahren geprüft. Nachdem untersucht wird, welche Modell- und insbesondere Fehlerannahmen das finale Modell erfüllt, wird abschließend erläutert, inwiefern die Regressoren die Entstehung *Nettomiete* im konstruierten Modell beeinflussen.

## 3 Statistische Methoden

### 3.1 Modell und Hypothesen

Das multiple lineare Regressionsmodell ist eine statistische Methode, um einen linearen funktionalen Zusammenhang zwischen einer abhängigen Variable  $Y$ , dem **Regressanden**, und  $k > 1$  unabhängigen Einflussvariablen  $X_j$  mit  $j = 1, \dots, k$ , den **Regressoren**, möglichst genau zu beschreiben. Es wird die folgende **multiple lineare Modellgleichung** angenommen:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i \quad , i = 1, \dots, n \quad (1)$$

Dabei bezeichne  $y_i$  eine Realisation des Regressanden  $Y$  und  $x_{ji}$  die Realisation des  $j$ -ten Regressors  $X_j$  für die  $i$ -te Beobachtung.  $\epsilon_i$  bezeichnet das so genannte **Residuum** der  $i$ -ten Beobachtung, eine Fehlervariable, die nicht beobachtet werden kann und die die Abweichung zwischen dem beobachteten und dem durch die Modellgleichung geschätzten Wert von  $Y$  angibt. Für die Residuen  $\epsilon_i$  werden die **Modellannahmen** getroffen, dass sie unabhängig identisch normalverteilt sind mit Erwartungswert 0 und konstanter Varianz  $\sigma_\epsilon^2$  (Homoskedastizität), kurz

$$\epsilon_i \sim N(0, \sigma_\epsilon^2) \quad \text{und} \quad \text{Cov}(\epsilon_i, \epsilon_j) = 0 \text{ für } i \neq j \quad (2)$$

(Holling, 2015).  $\beta_0, \dots, \beta_k$  sind unbekannte **Regressionskoeffizienten**, die möglichst gut geschätzt werden. Für sie sollen folgende **Hypothesen** getestet werden (Fahrmeir, 2016):

$$H_0^j : \beta_j = 0 \quad \text{vs.} \quad H_1^j : \beta_j \neq 0 \quad j = 0, \dots, k \quad (3)$$

In Worten:

$H_0^j$ : Ein signifikanter Erklärungswert des Regressors  $X_j$  für den Regressanden  $Y$  kann nicht nachgewiesen werden,  $X_j$  kann aus dem Regressionsansatz entfernt werden

vs.

$H_1^j$ : Der Regressor  $X_j$  hat einen signifikanten Erklärungswert für den Regressanden  $Y$  und muss im Regressionsansatz enthalten sein

Die statistische Auswertung in Kapitel 4 wird mit der statistischen Software **R** (R Core Team, 2022, Version 4.2.1) durchgeführt. In **R** wird das multiple lineare Regressionsmodell mit der Funktion `lm()` erstellt.

Liegen jeweils  $n$  Beobachtungswerte der Variablen  $Y$  und  $X_1, \dots, X_k$  vor, so ist das lineare Regressionsmodell in Matrixschreibweise formuliert als:

$$y_i = X\beta_i + \epsilon_i \quad (4)$$

mit

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{R}^k \quad \text{und} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

Die Realisationen der Regressoren werden in der **Designmatrix**  $X$  zusammengefasst. Diese Matrix besteht aus  $n$  Zeilen und  $k + 1$  Spalten, wobei in der ersten Spalte die Konstante 1 für den Achsenabschnitt der Regressionsgeraden steht und die Realisationen für  $X_1, \dots, X_k$  in den Spalten 2, ...,  $k + 1$  folgen (Holling, 2015). Die Normalverteilungsannahme der Residuen aus Formel 2 überträgt sich auf den unbeobachtbaren, unabhängig verteilten **Residuenvektor**  $\epsilon$ . Daraus folgt die Normalverteilung des beobachtbaren **Zufallsvektors**  $y \sim N(\mu, \sigma_\epsilon^2 I_n)$  mit  $\mu = X\beta$ .

Kategorielle Regressoren mit  $c$  geordneten oder ungeordneten Kategorien werden durch einen Vektor von  $c - 1$  so genannten **Dummy-Variablen**  $x^{(1)}, \dots, x^{(c-1)}$  kodiert:

$$x^{(p)} = \begin{cases} 1, & \text{Kategorie } p \text{ liegt vor,} \\ 0 & \text{sonst,} \end{cases} \quad p = 1, \dots, c - 1 \quad (5)$$

Falls die Referenzkategorie  $c$  beobachtet wird, so haben alle Dummy-Variablen den Wert 0 (Fahrmeir, 2007).

Die Regressionskoeffizienten  $\beta_0, \dots, \beta_k$  und die Fehlervarianz  $\sigma_\epsilon^2$  werden aus den Daten geschätzt. Der **Koeffizientenvektor**  $\beta$  wird so bestimmt, dass die Quadratsumme der Residuen  $(Y - X\beta)^T(Y - X\beta)$  bei fester Beobachtungsmatrix  $(y \ x_{i,1} \ \dots \ x_{i,k}), i = 1, \dots, n$ , minimiert wird (**Kleinste-Quadrate-Methode**) (Toutenburg, 2013, S. 90f.). Um eine eindeutige Lösung des Minimierungsproblems zu erhalten, muss zum einen  $k + 1 < n$  gelten, es dürfen also nicht mehr Parameter als Beobachtungen vorliegen und zum anderen darf keine Multikollinearität vorliegen, die Variablen dürfen also nicht voneinander abhängen. Wenn die Voraussetzungen erfüllt sind, ist

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (6)$$

der **Kleinste-Quadrate-Schätzer**(KQ-Schätzer) und die eindeutige Lösung des Mini-

mierungsproblems, wobei  $X^T$  eine Spiegelung der Matrix  $X$  an der Hauptdiagonalen und  $(X^T X)^{-1}$  die Invertierung des Matrixproduktes  $(X^T X)$  meint. Der Vektor der mit dem durch das multiple lineare Regressionsmodell geschätzten (gefitteten) Werten heißt dann  $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ . In R werden die Regressionskoeffizienten und die gefitteten Werte durch die `summary()` Funktion ausgegeben, in die das durch `lm()` erzeugte Lineare-Modell-Objekt (im Folgenden als `lm.objekt` abgekürzt) eingegeben wird.

Einen unverzerrten, effizienten und konsistenten (besten erwartungstreuen?) Schätzer für die **Fehlervarianz**  $\sigma_\epsilon^2$  stellt

$$\hat{\sigma}_\epsilon^2 = \frac{SSR}{n - (k - 1)} \quad (7)$$

dar (Holling, 2015, S. 327).  $SSR = \sum_{i=1}^n \hat{\epsilon}_i^2$  ist die Quadratsumme der Residuen  $\hat{\epsilon}_i = y_i - \hat{y}_i$ . Die Gesamtstreuung  $SSG = \sum_{i=1}^n (y_i - \bar{y})^2$  ist die Summe der erklärten Streuung  $SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  und der Reststreuung  $SSR$  (Holling, 2015).

Das **Bestimmtheitsmaß** gibt an, wie gut das multiple Regressionsmodell an die Daten angepasst ist, indem der Quotient aus erklärter Streuung und Gesamtstreuung gebildet wird (Fahrmeir, 2016, S. 456):

$$R^2 = \frac{SSE}{SSG} = 1 - \frac{SSR}{SSG} \quad (8)$$

Der Wertebereich des Bestimmtheitsmaß lautet  $0 \leq R^2 \leq 1$ . Bei einer perfekten Anpassung des Modells an die Werte ist  $R^2 = 1$  weil es in diesem Fall keine Reststreuung  $SSR$  gäbe. Wäre  $R^2 = 0$  würde die erklärte Streuung  $SSE$  bei 0 liegen, das Modell würde die Daten in keinsten Weise erklären (Fahrmeir, 2016).  $R^2$  ist kein unverzerrter Schätzer: Es wird umso größer, je größer die Zahl der unabhängigen Variablen im Modell ist. Eine an dieses Problem angepasste Version von  $R^2$  ist das **adjustierte Bestimmtheitsmaß**  $R_{adj}^2$  (Fahrmeir, 2007, S. 161)

$$R_{adj}^2 = 1 - \frac{n - 1}{n - k} (1 - R^2) \quad (9)$$

In R wird sowohl  $R^2$  als auch  $R_{adj}^2$  in der Ausgabe der `summary()` Funktion angegeben.

## 3.2 Testverfahren

Der **Signifikanztest** für einen bestimmten Parameter prüft die in Formel 3 getroffenen Hypothesen. Unter der Annahme der Gültigkeit der Nullhypothese  $H_0^j : \beta_j = 0$  und der Normalverteilungsannahme der Schätzer (Holling, 2015, S. 329) ist die Teststatistik des **t-Tests**

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \quad (10)$$

t-verteilt mit  $(n - k - 1)$  Freiheitsgraden, das heißt es gilt  $t_j \sim t_{n-k-1}$ .

$$\hat{\sigma}(\hat{\beta}_j) = \frac{\hat{\sigma}_\epsilon^2}{(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_{\cdot j})^2} \quad (11)$$

ist der Standardfehler des Schätzers  $\hat{\beta}_j$  für den Regressor  $X_j$  und  $R_j^2$  ist das multiple Bestimmtheitsmaß für die abhängige Variable  $X_j$  mit den übrigen Prädiktoren als unabhängige Variablen (Toutenburg, 2013, S. 119).

Die Nullhypothese wird zum Signifikanzniveau  $\alpha$  abgelehnt, falls  $|t_j| > t_{n-k-1, 1-\alpha/2}$ . Dabei bezeichnet  $t_{n-k-1, 1-\alpha/2}$  das  $(1 - \alpha/2)$ -Quantil der t-Verteilung mit  $(n - k - 1)$  Freiheitsgraden (Fahrmeir, 2016, S. 446).

In R werden die Werte der Signifikanztests, die für die einzelnen Regressoren durchgeführt werden, durch die Funktion `summary()` ausgegeben. Neben dem t-Wert wird der **p-Wert** mit der `summary()` Funktion aufgelistet. Der p-Wert gibt Aufschluss darüber, ob ein Regressor als erklärende Variable für den Regressanden signifikant ist. Er ist als die Wahrscheinlichkeit definiert, einen Teststatistikwert zu erhalten, der unter der Annahme der Nullhypothese den Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert annimmt (Fahrmeir, 2016, S. 387). So ist die Teststatistik dann zu verwerfen, wenn der p-Wert kleiner als das Signifikanzniveau  $\alpha$  ist.

### 3.3 Modellauswahl

Um ein Modell zu finden, das den Regressanden am "besten" beschreibt, werden verschiedene Methoden zur Modellauswahl verwendet. Eine beste Beschreibung ist dann erreicht, wenn ein Kompromiss gefunden wird zwischen möglichst guter Datenanpassung und zu großer Modellkomplexität durch zu viele Regressoren.

Für die Auswahlmethoden müssen folgende Kriterien erfüllt sein: Erstens muss die Modellauswahl aus den Daten für jedes angepasste Modell schätzbar sein und zweitens auf dem Maximum-Likelihood oder Bayesian Ansatz oder auch auf beiden Ansätzen beruhen (Burnham, 2004, S. 262).

Der gängige Ansatz für den Vergleich und die Auswahl statistischer Modelle ist der Signifikanztest (vgl. 3.2. Signifikanztests reagieren jedoch empfindlich auf recht kleine Abweichungen von der Nullhypothese, so dass in sehr großen Datensätzen alle einigermaßen sparsamen Modelle verworfen werden (Fahrmeir, 2007, S. 161) und es zu einer Überanpassung (overfitting) kommen kann. Das **adjustierte Bestimmtheitsmaß** (10) ist auch für Modelselektionen dem Bestimmtheitsmaß vorzuziehen und wird als Selektionskriterium insofern verwendet, dass das Modell mit dem größten  $R_{adj}^2$  gewählt wird. Jedoch ist auch  $R_{adj}^2$  anfällig, zu steigen, wenn Variablen mit t-Werten größer als eins in das Modell aufgenommen werden. In dem Fall würden wir Variablen mit einem p-Wert von ungefähr 0.3 zusätzlich aufnehmen, die vom Signifikanztest jedoch abgelehnt werden würden (Fahrmeir, 2007, S. 161).

Das **Akaike Informationskriterium** (AIC) ist ein Maß für die Qualität der Anpassung,



die auf dem Maximum-Likelihood Ansatz beruht. Es "bestraft" hohe Komplexität also eine zu hohe Zahl an Regressoren. Mit der logarithmierten Likelihood des  $k$ -dimensionalen Parametervektor  $\theta = (\theta_1, \dots, \theta_k)^T$  ist das AIC gegeben durch

$$AIC = -2l(\hat{\theta}) + 2k \quad (12)$$

Der Term  $2k$  bestraft ein überparametrisiertes Modell, da das Modell mit dem kleinsten AIC-Wert bei der Wahl zwischen verschiedenen bevorzugt wird (Fahrmeir, 2007, S. 477). Eine Alternative zum AIC ist das **Bayesian Information Criterion** (BIC), gegeben durch

$$BIC = -2l(\hat{\theta}) + \log(n)k \quad (13)$$

(Fahrmeir, 2007, S. 489) dar. Der Unterschied zum AIC liegt darin, dass beim BIC die logarithmierte Stichprobengröße  $n$  mit der Regressorenanzahl multipliziert wird und somit den Strafterm bildet. Das BIC bestraft Modelle mit vielen Parametern stärker als das AIC, sodass mit dem BIC Modelle mit geringerer Komplexität selektiert werden als mit dem AIC.

In R wird die Funktion `ols_step_all_possible(lm.objekt)` aus dem Paket `olsrr` (Hebbali, 2020) verwendet. Das Argument dieser Funktion ist das volle lineare Regressions Modell als `lm.objekt`. Für alle möglichen Modelle (Regressor-Anzahl von  $1, \dots, k$  und alle möglichen Regressor-Kombinationen) berechnet diese Funktion neben  $R^2_{adj}$ ,  $AIC$  und  $BIC$  acht weitere Selektionskriterien, die in diesem Projekt nicht vorgestellt werden.

Testbasierte Modellauswahlmethoden sind die schrittweisen Selektionsverfahren.

Bei der **Rückwärtselimination** werden mit dem vollen Modell, welches alle Regressoren einschließt, beginnend iterative Signifikanztests mit den Hypothesen aus 3 durchgeführt und nach jeder Iteration die Einflussgröße mit dem größten p-Wert aus dem Modell herausgenommen. Dies wird so lange durchgeführt, bis alle Regressoren im Modell einen p-Wert unter dem so genannten cut-off Wert (kann größer oder kleiner als Signifikanzniveau  $\alpha$  sein) haben (Hedderich und Sachs, 2015, S. 779).

Die **Vorwärtsselektion** dreht das Prinzip der iterativen Signifikanztests um indem hier mit dem *Nullmodell*, also nur mit der Konstanten  $\beta_0$  gestartet wird. Iterativ werden dann Regressoren in das Modell mit aufgenommen, die einen signifikanten Einfluss auf den Regressanden, das heißt einen p-Wert, der unter dem cut-off Wert liegt, haben. Dieses Verfahren wird so lange durchgeführt bis keine Einflussgröße mehr in das Modell aufgenommen werden können (Hedderich und Sachs, 2015, S. 780). In R Fox u. a., 2022

### 3.4 Einflussanalyse

Ziel einer Einflussanalyse ist es, zu untersuchen, ob es einzelne Beobachtungen im Datensatz gibt, die einen großen Einfluss auf die Schätzergebnisse im linearen Regressionsmodell haben. Eine Beobachtung mit großem Einfluss wird durch die **Hebelwirkung** (engl. **leverage**) erfasst, weswegen diese Beobachtung auch Hebelwert genannt wird (Hedderich und Sachs, 2015, S. 773). Die Hebelwirkung einer Beobachtung ist beschrieben durch  $h_{ii}$ , dem  $i$ -ten Diagonalelement der Projektionsmatrix  $H$ .  $h_{ii}$  wird auch Leverage Score genannt.

$$h_{ii} = [H]_{ii} \in \left[\frac{1}{n}, 1\right] \quad \text{mit} \quad H = X(X^T X)^{-1} X^T \in \mathbb{R}^{n \times k} \quad (14)$$

Da  $H$  idempotent ( $H^2 = H$ ) und symmetrisch ( $H^T = H$ ) ist (Fahrmeir, 2007, S. 93), gilt  $\text{Rang}(H) = k$  und die durchschnittliche leverage aller Beobachtungen ist  $\frac{k}{n}$ . Eine Beobachtung mit einer Hebelwirkung größer als  $2\frac{k}{n}$  wird als einflussreiche Beobachtung (high leverage) bezeichnet (Fahrmeir, 2007, S. 178). Eine "mittlere Leverage" wird durch  $\frac{k+1}{n}$  bestimmt (Hedderich und Sachs, 2015, S. 773). In R können die Hebelwirkungswerte der einzelnen Beobachtungen mit der Funktion `hatvalues(lm.objekt)` aufgerufen werden. Ein weiteres Maß für die Beurteilung des Einflusses einer Beobachtung  $x_i$  auf die Kleinst-Quadrate Regressionsanalyse ist die **Cooks-Distanz**.

$$\text{Cook's } D_i = \frac{(\hat{y}_{(i)} - \hat{y})^T (\hat{y}_{(i)} - \hat{y})}{k \cdot \hat{\sigma}_\epsilon^2} \quad (15)$$

Es wird der euklidische Abstand zwischen der Schätzung  $\hat{y}$  und  $\hat{y}_{(i)}$  (eine Schätzung von der Zielvariable  $Y$ , die auf allen Beobachtungen bis auf der  $i$ -ten beruht) mit der geschätzten Fehlervarianz (8) gewichtet. Beobachtungen mit  $D_i$  größer als 0.5 gelten als auffällig und Beobachtungen mit  $D_i$  größer als eins bedürfen in jedem Fall einer Überprüfung auf ihre Validität (Fahrmeir, 2007, S. 178). In R können die Cook's  $D_i$  Werte der einzelnen Beobachtungen mit der Funktion `cooks.distance(lm.objekt)` aufgerufen werden.

### 3.5 Modelldiagnostik

"Kein Modell ist korrekt, jedoch sollten die hinter einem Modell stehenden Annahmen zumindest approximativ erfüllt sein, um Fehlschlüsse zu vermeiden." (Fahrmeir, 2007, S. 167)

Die Überprüfung bzw. die Erfüllung der in Kapitel 3.1 vorgestellten Modellannahmen ist Bestandteil des Modellbildung und ausschlaggebend für die Bewertung der Güte (die Verlässlichkeit der abgeleiteten Schätzwerte  $\hat{y}_i$ ) des linearen Regressionsmodells. In R werden mit der Funktion `plot(lm.objekt)` sechs verschiedene Grafiken angezeigt, die die

Modellannahmen überprüfen. Vier dieser Grafiken werden im Folgenden vorgestellt.

Die Homoskedastizität der Residuen wird grafisch in **Residuenplot** (`plot(lm.objekt, which=1)`) überprüft. In Residuenplots werden Residuen  $\epsilon_i$  in einem Streudiagramm gegen die geschätzten Werte  $\hat{y}_i$  dargestellt. Bei Homoskedastizität sollten die Residuen regellos und mit konstanter Variabilität um Null streuen (vgl. Fahrmeir, 2007) (vgl. Abbildung 2, siehe Anhang S. 16). Wie Residuenplots bei Vorliegen einer Verletzung der Homoskedastizitätsverletzung aussehen, ist in Abbildung 3 (siehe Anhang S. 16) exemplarisch dargestellt. Neben der Homoskedastizität, kann in Residuenplots auch die anderen Modellannahmen geprüft werden. Beispiele von in Residuenplots dargestellten Daten, die in anderer Weise von der Modellannahme  $\epsilon_i \stackrel{u.i.v.}{\sim} N(0, \sigma_\epsilon^2)$  abweichen, sind in Abbildung 4 (siehe Anhang S. 16) dargestellt.

Neben Residualplots werden **Quantile-Quantile (QQ-)Plots** (`plot(lm.objekt, which=2)`), zur grafischen Überprüfung der Normalverteilungsannahme der Residuen verwendet. Es werden die standardisierten Residuen

$$\tilde{\epsilon}_i := \frac{\hat{\epsilon}_i}{\hat{\sigma}_\epsilon^2 \sqrt{1 - h_{ii}}} \quad i = 1, \dots, n \quad (16)$$

(Groß, 2010, S. 201) auf der y-Achse gegen die theoretischen Quantile der Normalverteilung auf der x-Achse geplottet (vgl. Kohn und Öztürk, 2016, S. 97). Zusätzlich wird in das Streudiagramm eine Winkelhalbierende als Referenzlinie eingezeichnet. Dies dient dem Vergleich der Ähnlichkeit zwischen der Verteilung der Stichprobe und der Normalverteilung. Liegt eine perfekte Normalverteilung der standardisierten Residuen vor, so liegen die Punkte entlang der Winkelhalbierenden (vgl. ??, siehe Anhang S. 17). Je ausgeprägter die Abweichung der Residuenverteilung von der Normalverteilung ist, desto mehr streuen die Punkte um die Referenzlinie im Streudiagramm (Hedderich und Sachs, 2015, S. 454). Dabei lassen bestimmte Streuungsmuster auf bestimmte Verteilungscharakteristiken rückschließen (vgl. Abbildung 6, siehe Anhang S. 17)

Auch die in Kapitel 3.4 vorgestellte Einflussanalyse wird mit `plot(lm.objekt)` dargestellt.

`plot(lm.objekt, which=4)` zeigt in einem **Cook's Distance Histogramm** die Werte  $D_i$  jeder Beobachtung in der Stichprobe. Die größten  $D_i$  Werte der Beobachtungen werden mit Angabe der jeweiligen Beobachtungsnummer hervorgehoben (Quelle??).

Ob einzelne Beobachtungen der Stichprobe durch extreme Werte und das Fehlen benachbarter Beobachtungen eine hohe Leverage haben, also dazu führen, dass das angepasste Regressionsmodell nah an diesen high-leverage Beobachtungen liegt kann im **Residual-Leverage plot** untersucht werden. Die mit `plot(lm.objekt, which=5)` erzeugte Grafik zeigt ein Streudiagramm, in dem standardisierte Residuen auf der y-Achse gegen die Leverage Scores der einzelnen Beobachtungen  $h_{ii}$  abgebildet sind. Außerdem sind die Cook's distance Schwellenwerte  $D_i$  gleich 0.5 und  $D_i$  1 anhand gestrichelter Linien eingezeichnet,

um das Erkennen einflussreicher Beobachtungen zu erleichtern.

Wie in Kapitel 3.1 angesprochen, darf unter den Regressoren keine Multikollinearität vorliegen, damit ein eindeutiger Kleinste-Quadrate-Schätzer gefunden werden kann. Wenn Multikollinearität vorliegen würde, wäre  $x_j$  linear von den anderen Regressoren abhängig. Die Schätzung der Regressionskoeffizienten wäre in diesem Fall rechnerisch nicht möglich, da  $X^T X$  singulär wäre und die generalisierte Inverse nicht bestimmt werden könnte (Hedderich und Sachs, 2015, S. 772). Um Multikollinearität zu identifizieren und zu quantifizieren, wird der Variationsinflationfaktor

$$VIF_j = \frac{1}{1 - R_j^2} \quad (17)$$

bestimmt. Ein Kollinearitätsproblem liegt bei einem  $VIF$  Wert, der größer als zehn ist vor (Fahrmeir, 2007, S. 171).

## 4 Statistische Auswertung

Um einen Überblick über die Verteilungen der Variablen zu erhalten, sind in Tabelle 2 und 3 univariate Kenngrößen zu den Variablen aufgeführt.

|                | Nettomiete (€) | Wohnfläche (qm) | Zimmeranzahl | Baujahr |
|----------------|----------------|-----------------|--------------|---------|
| arithm. Mittel | 763.06         | 71.98           | 2.70         | 1964.21 |
| Median         | 700.00         | 70.00           | 3.00         | 1957.50 |
| Minimum        | 174.75         | 15.00           | 1.00         | 1918.00 |
| Maximum        | 6000.00        | 300.00          | 8.00         | 2012.50 |
| Spannweite     | 5825.25        | 285.00          | 7.00         | 94.50   |
| IQR            | 360.46         | 30.00           | 1.00         | 25.50   |
| Standardabw.   | 338.16         | 25.74           | 0.98         | 26.51   |
| MAD            | 261.90         | 22.24           | 1.48         | 27.43   |
| Schiefe        | 2.59           | 1.35            | 0.46         | -0.18   |
| Wölbung        | 25.47          | 8.33            | 3.60         | 2.31    |

Tabelle 2: univariate Kenngrößen für metrische Variablen

Da nah beieinanderliegende Lage- (arithmetisches Mittel vs. Median) bzw. Streuungsmaße (Standardabweichung vs. MAD) auf eine symmetrische Verteilung der Variable schließen lassen (Fahrmeir, 2016, S. 60), werden den metrisch skalierten Variablen vor allem diese Kenngrößen betrachtet.

Das arithmetische Mittel (763 €) und der Median (700 €), sowie die Standardabweichung (338.16 €) und MAD (261.9 €) der Variable *Nettomiete* unterscheiden sich stark. Dies lässt auf eine asymmetrische Verteilung bzw. auf den Einfluss von Ausreißern schließen. Diese Vermutung wird in Abbildung 7 (siehe Anhang S. 18) bestätigt. Die Beobachtungs-

nummer 1975 ist besonders auffällig, da sie sich durch eine Nettomietenausprägung von 6000 € in ihrer Ausprägung nicht nur von den übrigen Beobachtungen stark unterscheidet, sondern auch weit abgelegen von den übrigen Beobachtungen ist. Ob diese Beobachtung das Ergebnis der Regressionsanalyse signifikant beeinflusst, wird in der Einflussanalyse geklärt. Die Lage- und Streuungsmaße für die metrischen Variablen *Wohnfläche*, *Zimmeranzahl* und *Baujahr* unterscheiden sich nur marginal. Jedoch ist auffällig, dass die Spannweite (285 qm) fast das zehnfache des Interquartilsabstandes (30 qm) ist. Diese Beobachtung wird vom Wölbungskoeffizienten (8.33) aufgegriffen, der zeigt, dass die Verteilung der Variable Wohnfläche leptokurtisch ist.

Die deskriptive Kenngrößen für die nominal, dichotomen Variablen zeigen, welche Zustände in den meisten der 3065 betrachteten Wohnungen vorliegen: 1085 Wohnungen befinden sich in einer guten, 110 in einer besten Lage. Damit befinden sich 1870 Wohnungen, das sind 61 % der Grundgesamtheit, in einer anderen Lagekategorie. In 99 % der Wohnungen ist die Warmwasserversorgung vom Vermieter gestellt. 93 % der Wohnungen verfügen über eine Zentralheizung. Das Bad ist in 88 % der Wohnungen nicht gefliest und normal ausgestattet. In 75 % der Wohnungen ist die Küchenausstattung normal.

Tabelle 3: Deskriptive Kenngrößen für nominal, dichotome Variablen

|                         | Gute Lage | Beste Lage | Warmwasser | Heizung | Fliese  | Bad     | Küche   |
|-------------------------|-----------|------------|------------|---------|---------|---------|---------|
| absolute Häufigkeit "0" | 1980.00   | 2955.00    | 3039.00    | 2861.00 | 380.00  | 2704.00 | 2298.00 |
| absolute Häufigkeit "1" | 1085.00   | 110.00     | 26.00      | 204.00  | 2685.00 | 361.00  | 767.00  |
| relative Häufigkeit "0" | 0.65      | 0.96       | 0.99       | 0.93    | 0.12    | 0.88    | 0.75    |
| relative Häufigkeit "1" | 0.35      | 0.04       | 0.01       | 0.07    | 0.88    | 0.12    | 0.25    |
| Modus                   | 0         | 0          | 0          | 0       | 1       | 0       | 0       |

Um die Signifikanz der metrischen wie nominalen Einflussvariablen auf den Regressanden *Nettomiete* zu untersuchen wird ein lineares Modell mit den Regressoren *Wohnfläche*, *Zimmer*, *Baujahr*, *Bezirk*, *Gute Lage*, *Beste Lage*, *Warmwasser*, *Heizung*, *Fliese*, *Bad* und *Küche* erstellt. Die Signifikanztests ergeben, dass die Variablen *Wohnfläche*, *Zimmer*, *Baujahr*, *Gute Lage*, *Beste Lage*, *Warmwasser*, *Heizung*, *Fliese*, *Bad* und *Küche* das Signifikanzniveau  $\alpha$  kleiner-gleich 0.01 einhalten und damit in das Modell mit aufgenommen werden sollten. Bei der Dummy-Variable *Bezirk* halten nur die Merkmalsausprägungen "Ludwigvorstadt-Isarvorstadt", im Folgenden *Bezirk<sub>LI</sub>* genannt, und "Maxvorstadt", im Folgenden *Bezirk<sub>M</sub>* genannt, der 25 möglichen Merkmalsausprägungen das Signifikanzniveau von  $\alpha$  gleich 0.05 ein. Das resultierende **multiple lineare Regressionsmodell 1 (M1)** lautet:

$$\begin{aligned}\hat{Y}_i^{(1)} = & -3.173 + 1.172Wohnfläche - 4.642Zimmer + 1.1649Baujahr \\ & + 1.103Bezirk_{LI} + 1.037Bezirk_M + 4.494GuteLage + 1.014BesteLage \\ & - 1.789Warmwasser - 7.861Heizung + 5.233Fliese + 3.29Bad + 8.557Küche\end{aligned}$$

Da der Großteil der Merkmalsausprägungen der Dummy-Variable *Bezirk* nicht signifikant

ist, wird nach dem "alles oder nichts"-Prinzip die gesamte Variable *Bezirk* nicht in das Modell mit aufgenommen und für die verbleibenden Regressoren wird ein lineares Modell berechnet. In diesem sind alle Regressoren mindestens zum Niveau  $\alpha = 0.01$  signifikant von 0 verschieden und das **multiple lineare Regressionsmodell 2 (M2)** lautet:

$$\begin{aligned}\hat{Y}_i^{(2)} = & -2218.7611 + 11.9284Wohnfläche - 55.5408Zimmer + 1.1074Baujahr \\ & + 82.4023GuteLage + 118.4039BesteLage - 184.5388Warmwasser \\ & - 67.2810Heizung + 52.3544Fliese - 30.6062Bad - 85.3459Küche\end{aligned}$$

Auffallend ist, dass der Regressionskoeffizient der Variable *Zimmer* negativ ist. Je größer die Variable *Zimmer* also werden würde, desto geringer würde der Regressand *Nettomiete* werden. Dies ist sachlogisch nicht begründet. Ursächlich für diesen Widerspruch wird die starke Korrelation von ca. 0.86 zwischen *Wohnfläche* und *Zimmer* vermutet, die in dem Korrelationsplot nach Spearman in Abbildung 8 (siehe Anhang S. 17) dargestellt ist. Dies begründet die Variable *Zimmer* aus dem Modell zu entfernen.

Das so berechnete **multiple lineare Regressionsmodell 3 (M3)** lautet:

$$\begin{aligned}\hat{Y}_i^{(3)} = & -2341.6908 + 10.1633Wohnfläche + 1.1553Baujahr \\ & + 87.4565GuteLage + 131.3731BesteLage - 187.5855Warmwasser \\ & - 65.2646Heizung + 53.4245Fliese + 29.3194Bad + 95.4618Küche\end{aligned}$$

In diesem Modell sind alle Regressoren mindestens zum Niveau  $\alpha = 0.01$  signifikant, die Zielvariable zu beschreiben und die Interpretation der Regressionskoeffizienten ist ebenfalls sinnvoll. Die Modellauswahl mit `ols_step_all_possible` ergibt, dass das volle Modell die höchste Güte hat:  $R_{adj}^2$  0.679, AIC 40929.48 und BIC 41505.98. Sowohl die Modellauswahlkriterien Vorwärts-Selektion sowie Rückwärts-Elimination bestätigen diese beste Auswahl der Variablen.

Um die Modellannahmen zu überprüfen werden Diagnostikplots erstellt (siehe Abbildung 9, siehe Anhang S. 18). Auf allen vier Plots fällt die Beobachtungsnummer 1975 als Ausreißer auf. Diese Beobachtung ist die selbe, deren Nettomietenausprägung in Abbildung 7 bereits als Ausreißer aufgefallen ist. Dass dieser Ausreißer ebenfalls als High Leverage Point anzusehen ist zeigen die beiden unteren Grafiken in Abbildung 9. Numerisch lässt sich der High Leverage Point durch die Cooks Distance  $D_i \approx 0.928$  sowie durch die Leverage  $h_{1945} \approx 0.038$ , die mehr als sechsmal so groß wie der Schwellenwert von  $2\frac{9}{3065} \approx 0.006$  ist, identifizieren. Daher wird diese Beobachtung aus dem Datensatz entfernt. Mit  $n = 3064$  Beobachtungen und gleichbleibender Variablenselektion wird das **multiple lineare Regressionsmodell 4 (M4)** berechnet. Es lautet:

$$\begin{aligned}\hat{Y}_i^{(4)} = & -2162.9552 + 9.8263Wohnfläche + 1.076Baujahr \\ & + 88.8216GuteLage + 111.631BesteLage - 184.1406Warmwasser \\ & - 68.0905Heizung + 54.629Fliese + 37.5936Bad + 90.5553Küche\end{aligned}$$

Im Modell M4 sind alle Regressionskoeffizienten kleiner als das Signifikanzniveau  $\alpha = 0.001$ , daher sind alle Regressoren als hochsignifikant für die Beschreibung des Regressanden anzusehen. Auch die Interpretation der Regressionskoeffizienten ist sinnvoll:  $\beta_0 = -2162.9552$  ist der *Intercept* und gibt an, wie hoch die *Nettomiete* wäre, wenn alle Regressoren den Wert 0 einnehmen würden. Hier würde dies bedeuten, dass die *Nettomiete*  $-2162.9552\text{€}$  betragen würde, wenn eine Wohnung mit *Wohnfläche* 0 Quadratmeter, 0 *Zimmern*, aus einem *Baujahr* vor  $Baujahr_{min} = 1918$  **oder** aus dem *Baujahr* 0 ??, die sich weder in einer guten noch in einer besten *Lage* befindet, die über eine zentrale *Warmwasserversorgung* und eine *Zentralheizung* verfügt, deren Bad *gefliest* ist und deren *Bad* und *Küche* normal ausgestattet ist. Wenn die Wohnfläche um ein Quadratmeter größer wird, so steigt die Schätzung der *Nettomiete* um  $\beta_1 = 9.8263\text{€}$ . Wenn das Baujahr um ein Jahr größer wird, so steigt die Schätzung der *Nettomiete* um  $\beta_2 = 1.0760\text{€}$ . Wenn sich die Wohnung in guter Lage befindet, so steigt die Schätzung der *Nettomiete* um  $\beta_3 = 88.8216\text{€}$ . Wenn sich die Wohnung in bester Lage befindet, so steigt die Schätzung der *Nettomiete* um  $\beta_4 = 111.631\text{€}$ . Wenn die Wohnung keine Warmwasserversorgung hat, so sinkt die Schätzung der *Nettomiete* um  $\beta_5 = -184.1406\text{€}$ . Wenn die Wohnung keine Zentralheizung hat, so sinkt die Schätzung der *Nettomiete* um  $\beta_6 = -68.0905\text{€}$ . Wenn das Bad nicht gefliest ist, so steigt die Schätzung der *Nettomiete* um  $\beta_7 = 54.6290\text{€}$ . Wenn die Ausstattung des Bades gehoben ist, so steigt die Schätzung der *Nettomiete* um  $\beta_8 = 37.5936\text{€}$ . Wenn die Ausstattung der Küche gehoben ist, so steigt die Schätzung der *Nettomiete* um  $\beta_9 = 90.5553\text{€}$ .

Die Validität des KQ-Schätzers  $\hat{\beta} = (\beta_0, \dots, \beta_9)$  wird durch den vollen Spaltenrang der Designmatrix durch die Tatsache, dass  $k = 9$  kleiner als  $n = 3064$  ist bestätigt.

Für das Model M4 werden Diagnostikplots erstellt, um die Einhaltung der Modellannahmen zu prüfen (Abbildung 9, siehe Anhang S. 19). Im Residualplot oben links sieht man, dass die Residuen relativ harmonisch um die Null verteilt sind. Die allermeisten Residuen zentrieren sich im Bereich circa 250 bis 1250 der angepassten Werte (wie anders formulieren), ab 1500 gibt es weniger Ausprägungen. Extreme Ausprägungen ohne benachbarte Beobachtungen gibt es allerdings nicht und die Residuen der höheren angepassten Werte liegen ebenfalls relativ gleichmäßig um die 0. Es ist jedoch erkennbar, dass die Ausprägungen der Residuen bei höher werdenden angepassten Werten auch größer werden. Insgesamt ist der Residualplot ausgeglichen und es ist von einer konstanten Varianz der Residuen auszugehen. Der QQ-Plot oben rechts zeigt, dass die Verteilung der standardisierten Residuen nicht stark von einer Normalverteilung abweicht. Die leichte Abweichung der Punktestrangs von der Winkelhalbierenden an beiden Enden zeigt an,

dass die Verteilung der standardisierten Residuen leicht platykurtisch ist, jedoch ist der große Anteil des Punktestrangs hervorzuheben, der sich zum Großteil auf der Winkelhalbierenden befindet und Anlass dazu gibt, von einer Erfüllung der Normalverteilungsannahme der Residuen auszugehen. Die beiden unteren Grafiken zur Einflussanalyse zeigen, dass die Modifikation von M3 zu M4 erfolgreich war und es jetzt keine Beobachtungen mehr gibt, die in ihrer Ausprägung und Abgelegenheit von anderen Beobachtungen das Regressionsmodell übermäßig beeinflussen. Um die Modelldiagnostik abzuschließen und den Erfolg der Modifikation des Modells M2 zu testen, werden die Varianzinflationskoeffizienten betrachtet. Wie erwartet scheint es keine Abhängigkeit zwischen den Regressoren zu geben, die drei höchsten *VIFs* haben die Variablen *Wohnfläche*(1.106), *Baujahr*(1.17) und *Heizung*(1.14), sehr weit vom Schwellenwert 10 entfernt, ab dem man von Multikollinearität spricht.

Die Modelldiagnostik hat gezeigt, dass das Modell M4 alle Modellannahmen einhält und rechtfertigt so die Wahl des Modell M4 als finales Modell zur Beschreibung des linearen Zusammenhangs zwischen der *Nettomiete* und den übrigen Variablen.

## 5 Zusammenfassung



## Literatur

- Burnham K. P. und Anderson, D. R. (2004). *Multimodel Inference: Understanding AIC and BIC in Model Selection*. SSociological Methods & Research”.
- Fahrmeir L., Heumann C. Künstler R. Pigeot I. und Tutz G. (2016). *Statistik - Der Weg zur Datenanalyse*. Springer Berlin Heidelberg.
- Fahrmeir L., Kneib T. und Lang S. (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer Berlin Heidelberg.
- Fox, John, Sanford Weisberg und Brad Price (2022). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-5. URL: <https://CRAN.R-project.org/package=carData>.
- Groß, J. (2010). *Grundlegende Statistik mit R: Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R*. Vieweg+Teubner Verlag.
- Hebbali, Aravind (2020). *olsrr: Tools for Building OLS Regression Models*. URL: <https://cran.r-project.org/web/packages/olsrr/olsrr.pdf>.
- Hedderich, J. und L. Sachs (2015). *Angewandte Statistik: Methodensammlung mit R*. Springer Berlin Heidelberg.
- Holling H. und Gediga, G. (2015). *Statistik – Testverfahren*. Hogrefe Verlag GmbH & Co. KG.
- Kohn, W. und R. Öztürk (2016). *Statistik für Ökonomen: Datenanalyse mit R und SPSS*. Springer Berlin Heidelberg.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- Sozialreferat der Landeshauptstadt München (2015). *Mietspiegel für München© 2015*. besucht am 17.11.2022. URL: <https://stadt.muenchen.de/infos/mietspiegel.html>.
- Statista Research Department (2022). *3.6 Symmetrie- und WölbungsmaßeEntwicklung der Angebotsmieten für Wohnungen in München von 2012 bis zum 2. Quartal 2022*. besucht am 17.11.2022. URL: <https://de.statista.com/statistik/daten/studie/535280/umfrage/mietpreise-auf-dem-wohnungsmarkt-in-muenchen/>.
- Toutenburg, H. (2013). *Lineare Modelle: Theorie und Anwendungen*. Physica-Verlag HD.

# Anhang

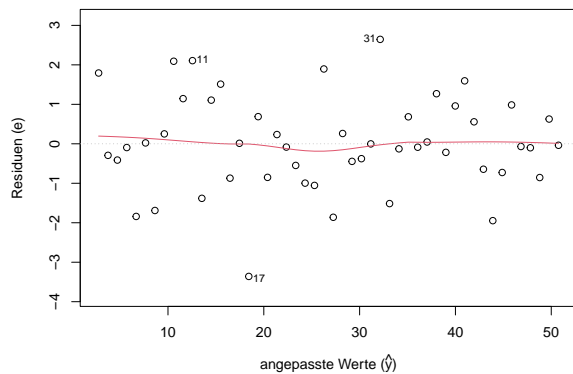


Abbildung 1: Residualplot von Beispieldaten mit homoskedastischer Residuen

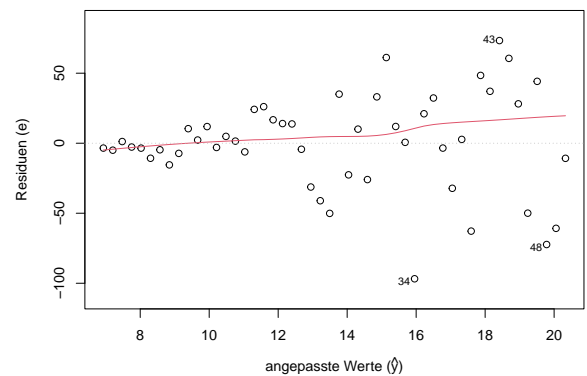


Abbildung 2: Residualplot von Beispieldaten mit heteroskedastischer Residuen

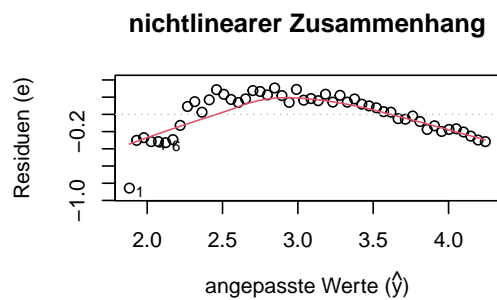
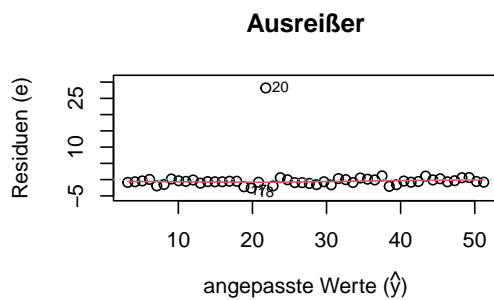
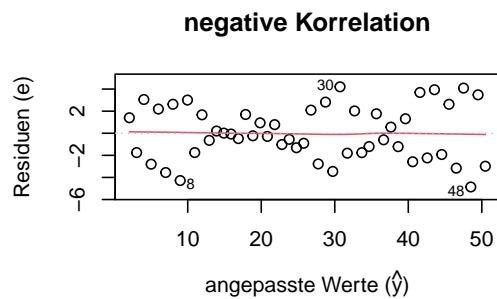
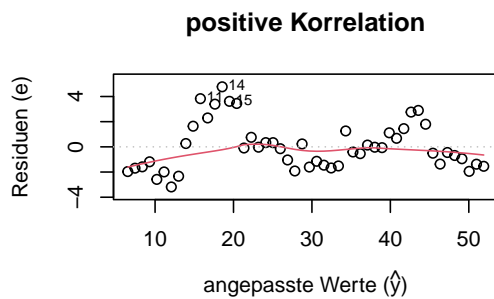


Abbildung 3: Residualplots von Beispieldaten mit unterschiedlichen Abweichungen von Modellannahmen für Residuen

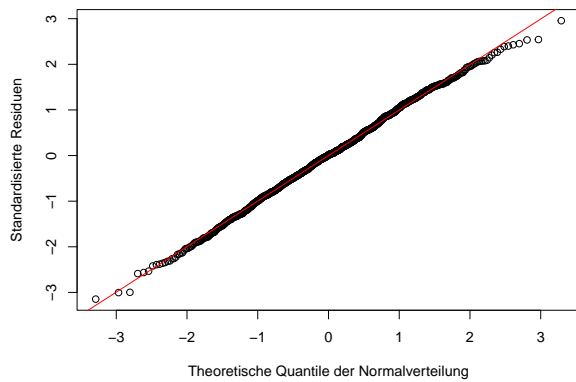


Abbildung 4: QQ-Plot (annähernd) perfekte normal verteilter standardisierter Residuen

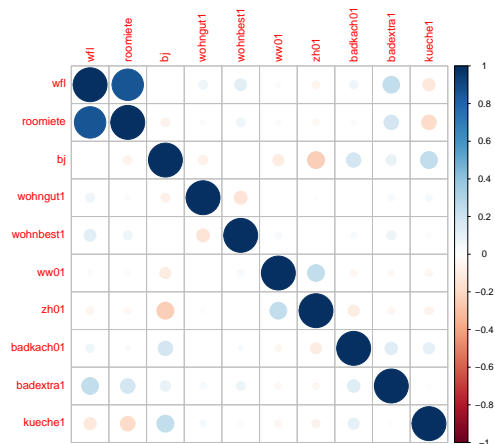


Abbildung 5: Korrelationsplot nach Spearman

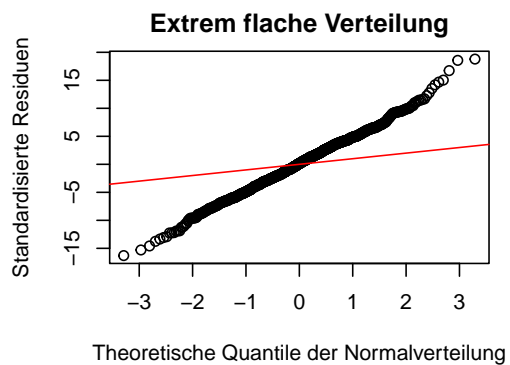
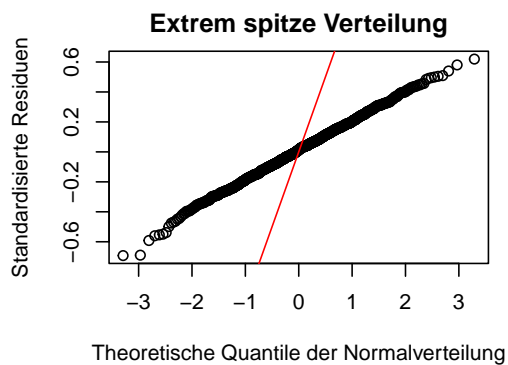
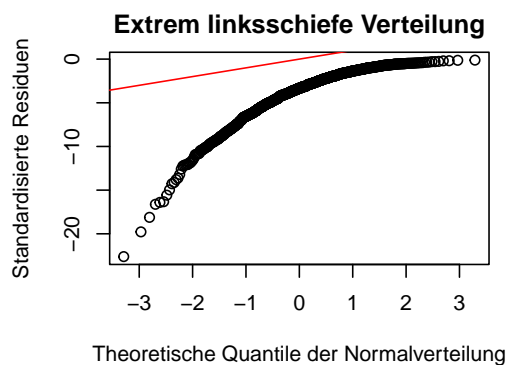
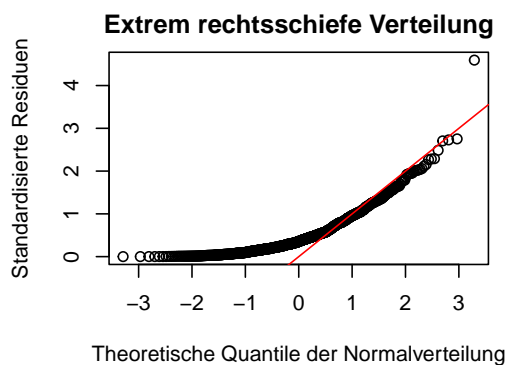


Abbildung 6: QQ-Plots von Beispieldaten mit unterschiedlichen Abweichungen von Modellannahmen für Residuen

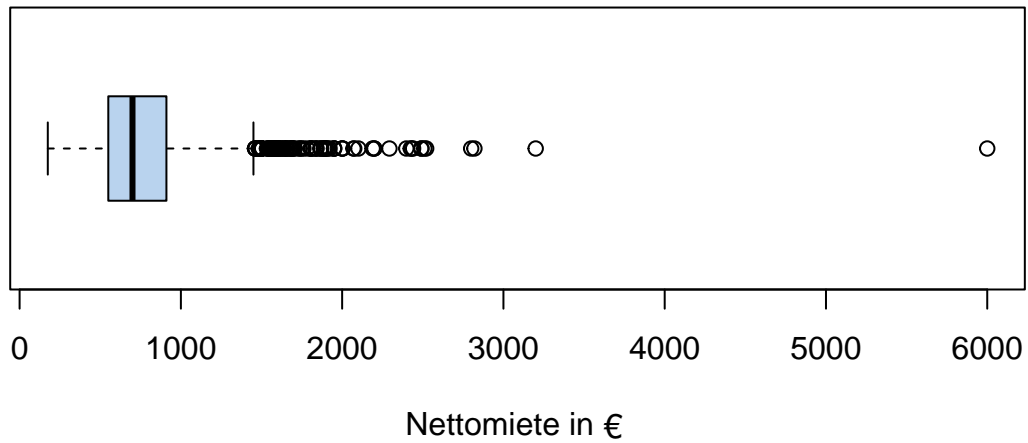


Abbildung 7: Boxplot für Variable *Nettomiete in €*

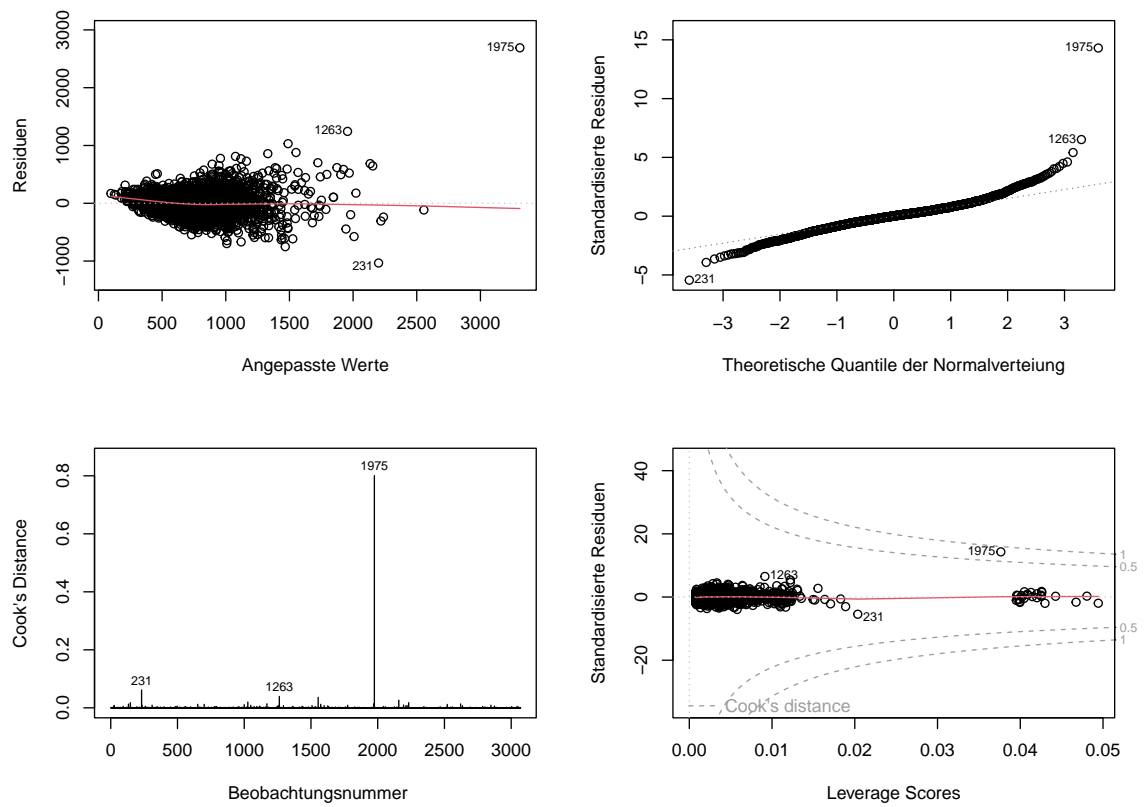


Abbildung 8: Diagnostikplots für **M3**

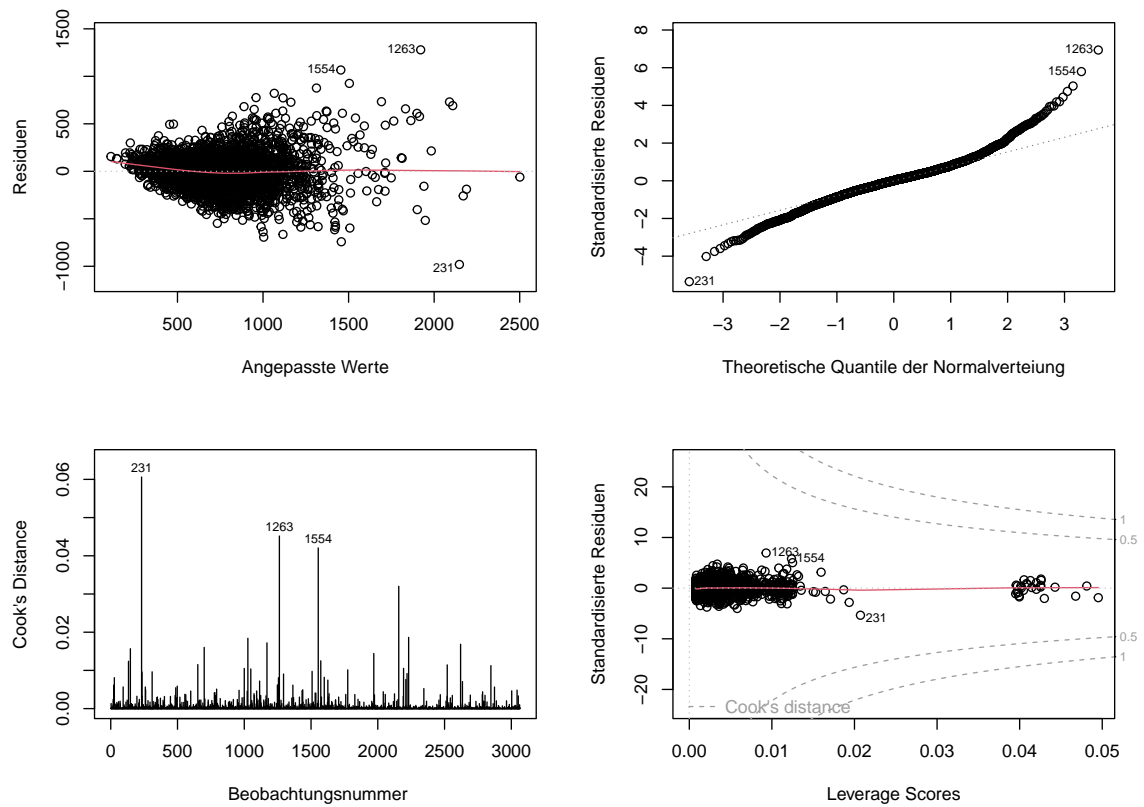


Abbildung 9: Diagnostikplots für M4