Lösung 16.1

Zunächst ein paar allgemeine Anmerkungen zu S-Plus und zu diesem Lösungsvorschlag:

1. Nach dem Start von S-Plus unter Windows erscheint der S-Plus Kommandoeditor mit dem Kommandoprompt `>`. Danach können die Kommandos eingegeben werden.

2. Kommandos sowie Ergebnisse, die S-Plus im Kommandofenster anzeigt, werden in dieser Lösung in dieser Schrift dargestellt: `Kommando()`. Objekte, die in S-Plus durch `<-` zugewiesen werden, werden in dieser Schrift dargestellt: *objektname*.

3. Paßt ein Kommando nicht in eine Zeile oder ist es fehlerhaft, so kann es in der nächsten Zeile fortgesetzt werden. In diesem Fall erscheint in der neuen Zeile ein anderer Prompt `+`. Alte Kommandos können mit den Pfeiltasten in die aktuelle Kommandozeile geholt werden.

4. S-Plus unterscheidet zwischen Groß- und Kleinschreibung.

5. Über den Punkt "Hilfe" im Menü gelangt man in die S-Plus Online-Hilfe. Alternativ kann die Hilfe zu einem bestimmten Kommando direkt über den Befehl `help(Kommando)` aufgerufen werden.

Einlesen der Daten:

Zum Einlesen eines ASCII-Files steht der Befehl `read.table()` zur Verfügung. Falls die erste Zeile des Datensatzes wie im vorliegenden Fall die Variablennamen enthält, so verwendet S-Plus durch den Parameter `header=T` diese Zeile automatisch zur Benennung der Variablen. Zur Illustration einmal das Einlesen ohne diesen Parameter, wenn sich die Daten zum Beispiel im Verzeichnis c:/compaufg befinden, wobei man beachten sollte, daß unter Windows die Verzeichnisstruktur durch den Doppel-Backslash gegliedert ist:

```
> miete<-read.table("c:\\compaufg\\miete.txt")
```

Der Datensatz ist nun als Data Frame Objekt `miete` in S-Plus verfügbar. Durch folgenden Befehl kann man sich die ersten 5 Zeilen der ersten drei im Datensatz enthaltenen Variablen ansehen:

```
> miete[1:5,1:3]
        V1       V2       V3
1    nmiete    flaeche   bad0
2    693.29       50      0
3    736.6        70      0
4    732.23       50      0
5   1295.14       55      0
```

Wie man sieht ist die erste Zeile fälschlicherweise mit den Variablennamen belegt, weshalb die Daten durch folgenden Befehl eingelesen werden sollten:

```
> miete<-read.table("c:\\compaufg\\miete.txt",header=T)
> miete[1:5,1:3]
    nmiete   flaeche  bad0
1   693.29      50     0
2   736.60      70     0
3   732.23      50     0
4  1295.14      55     0
5   394.97      46     0
```

Um auf die Variablen direkt mit ihrem Namen zugreifen zu können, bietet S-Plus die Funktion `attach()` an. Nachdem ein Datensatz "attachedïst, kann auf jede Variable mit ihrem Namen zugegriffen werden.

```
> attach(miete)
> mean(nmiete)
   830.3258    # arithmetisches Mittel Nettomiete
```

(a)

Für die metrischen Variablen *nmiete*, *flaeche*, *mvdauer* und *nmqm* bieten sich das Histogramm, der Box-Plot, der Kerndichteschätzer sowie ein Normal-Quantilplot an. Anhand der Variable *nmiete* werden die dazu benötigten Funktionsaufrufe für S-Plus exemplarisch durchgeführt und sind in Abbildung 1 dargestellt. Die Funktionsaufrufe lauten:

```
> par(mfrow=c(2,2))
> hist(nmiete,ylab="Anteil",xlab="Nettomiete",prob=T)
> boxplot(nmiete,ylab="Nettomiete")
> plot(density(nmiete),type="l",xlab="Nettomiete",ylab="Anteil")
> qqnorm(nmiete,xlab="Quantile der Standardnormalverteilung",
+ ylab="Nettomiete")
> qqline(nmiete)
```

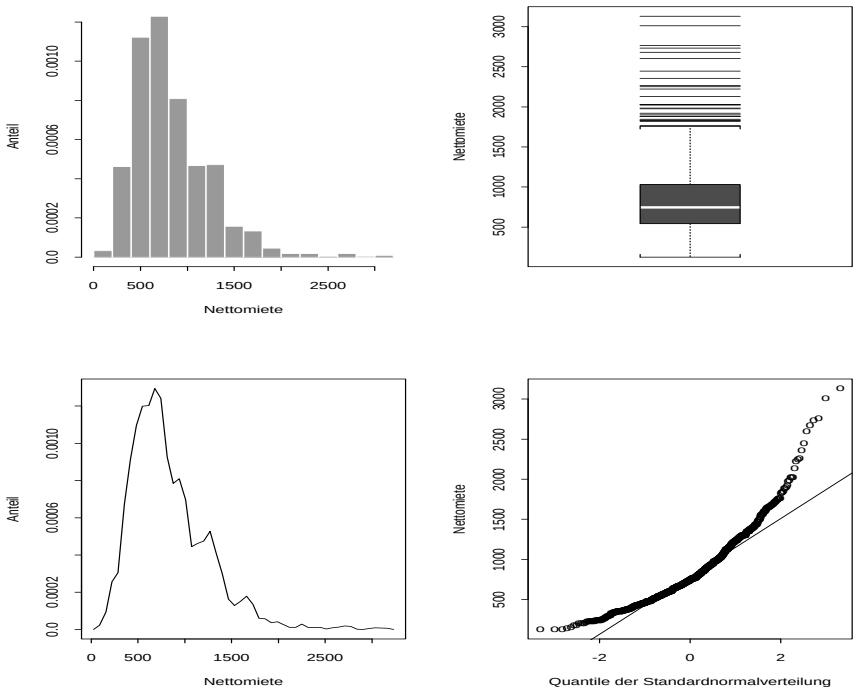

Abbildung 1: Graphische Veranschaulichung für die Variable *nmiete*

Der Befehl `par(mfrow=c(2,2))` erstellt dabei vor Berechnung der einzelnen Graphiken eine 2×2-Matrix, in die die Graphiken dann eingefügt werden und wurde für alle folgenden Abbildungen verwendet.
Das Histogramm für die Nettomiete zeigt eine linkssteilen Verteilung, wobei noch einige Ausreißer - besonders teure Wohnungen - zu sehen sind. Diese Ausreißer sind besonders gut im Box-Plot zu erkennen, in dem die Schiefe

ebenfalls zum Ausdruck kommt. Auch die Kerndichte-Schätzung und der Normal-Quantilplot zeigen eine linkssteile, deutlich von der Normalverteilung abweichende Verteilung für die Variable *nmiete* an. Die Interpretation für die Variable $flaeche$ verläuft analog.

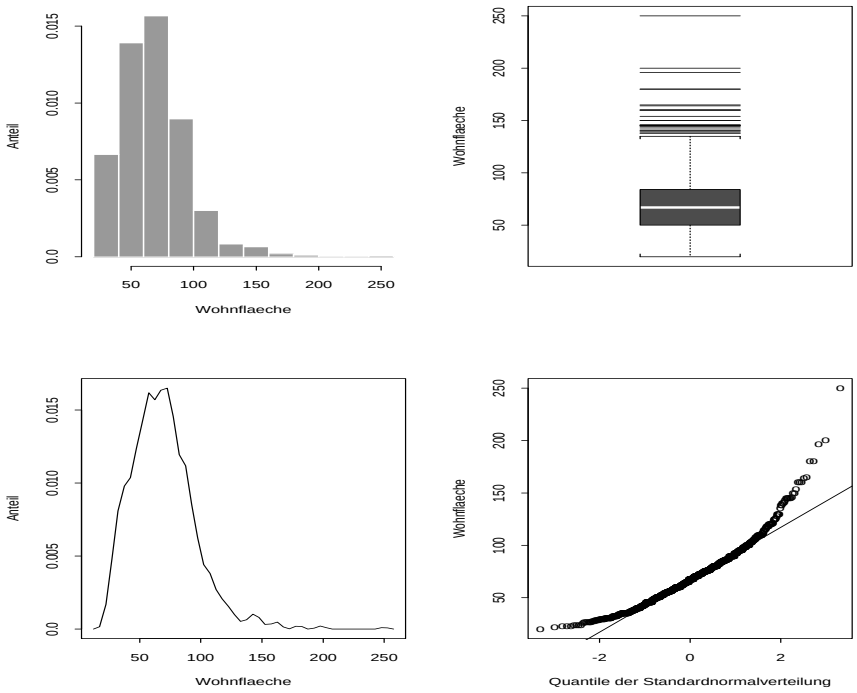

Abbildung 2: Graphische Veranschaulichung für die Variable $flaeche$

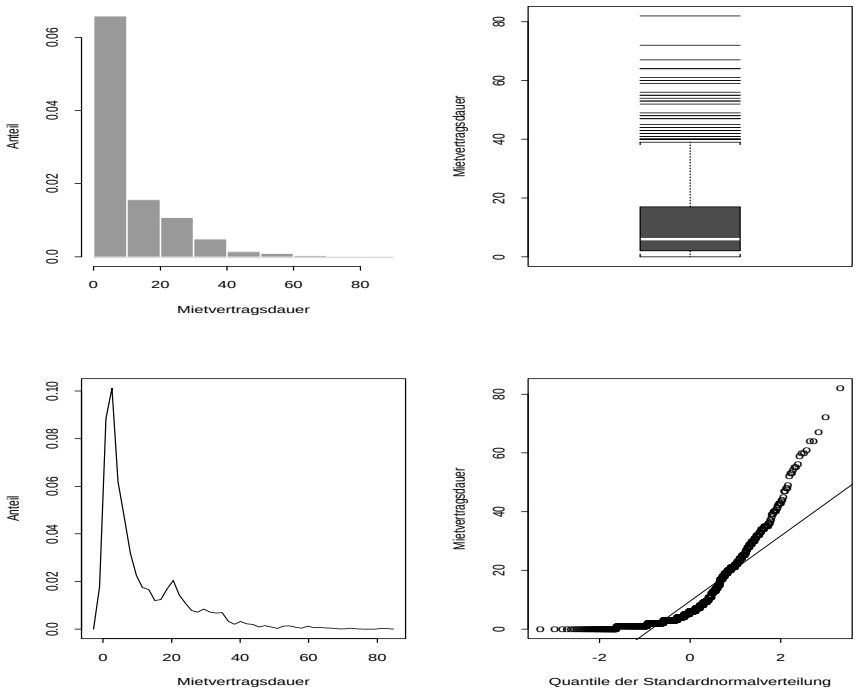

Abbildung 3: Graphische Veranschaulichung für die Variable *mvdauer*

Die Verteilung der Mietvertragsdauer ist extrem linkssteil, wobei Mietvertragsdauern von mehr als 40 Jahren sehr selten auftreten.

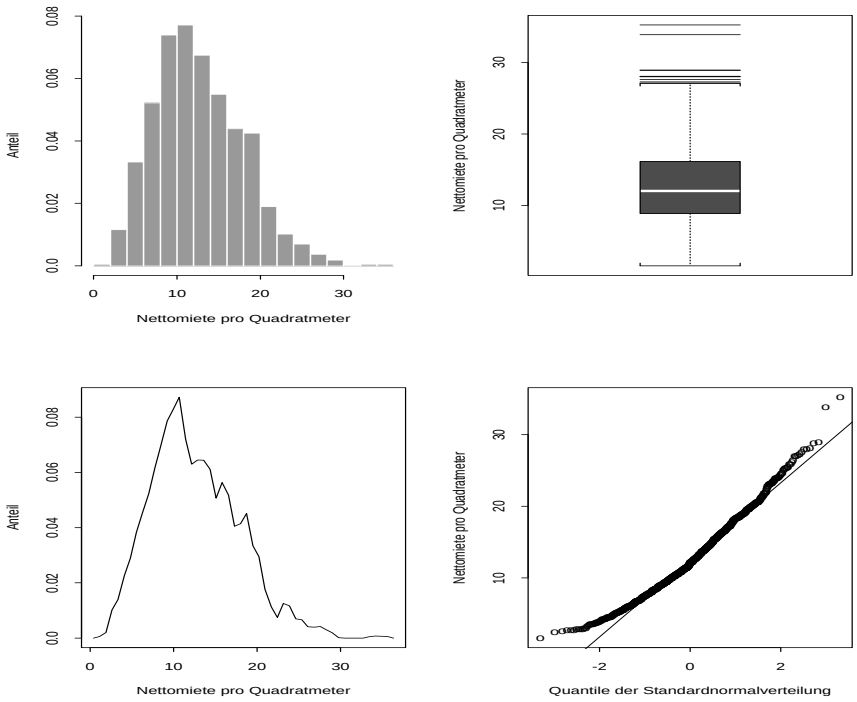

Abbildung 4: Graphische Veranschaulichung für die Variable *nmqm*

Das Histogramm der Nettomiete pro Quadratmeter zeigt eine leicht linkssteile Verteilung, die allerdings - wie der Normal-Quantilplot veranschaulicht - deutlich näher an der Normalverteilung liegt als bei der Nettomiete. Anhand des Box-Plots erkennt man außerdem noch, daß es insgesamt weniger Ausreißer gibt.

Die Anzahl der Klassen für ein Histogramm kann man durch den optionalen Parameter `nclass` variieren. So können durch folgende Kommandos Histogramme mit 5, 10, 25 bzw. 50 Klassen für die Nettomiete erstellt werden:

```
> par(mfrow=c(2,2))
> hist(nmiete,nclass=5,prob=T,ylab="Anteil")
> hist(nmiete,nclass=10,prob=T,ylab="Anteil")
> hist(nmiete,nclass=25,prob=T,ylab="Anteil")
> hist(nmiete,nclass=50,prob=T,ylab="Anteil")
```

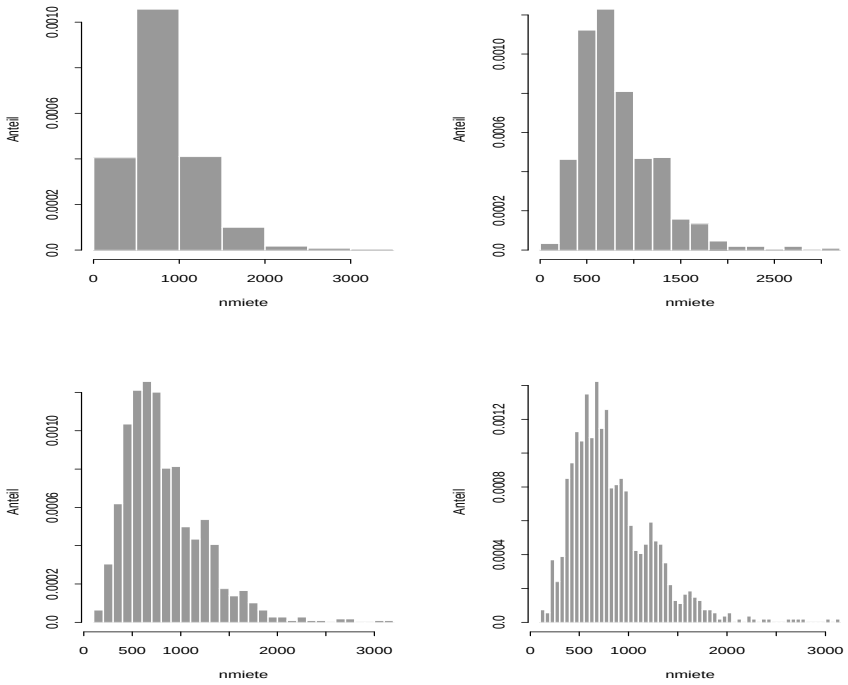

Abbildung 5: Histogramme für die Variable *nmiete*

Es wird deutlich, wie sich mit der Klassenbreite der optische Eindruck verändert. Um Fehlinterpretationen zu vermeiden, ist es deshalb zweckmäßig, neben Histogrammen auch Kerndichte-Schätzungen zu betrachten. Hier variiert die Glattheit der Schätzung zwar mit der Bandbreite, der optische Gesamteindruck ändert sich allerdings nicht so stark. Insgesamt ist auch an den folgenden Kerndichte-Schätzungen die linkssteile Verteilung der Nettomiete sehr deutlich zu sehen.

Zur Variation der Bandbreite h und des Kerntyps eines Kerndichteschätzers stehen die optionalen Parameter `width` und `type` der `density()`-Funktion zur Verfügung. So wurden im folgenden Beispiel Gauß-Kerne, die die Standardeinstellung darstellen, mit den Bandbreiten $h = 50, 100, 200, 300$ für die Nettomiete berechnet:

```
> par(mfrow=c(2,2))
> plot(density(nmiete,window="g",width=50),type="l",
+ xlab="Nettomiete (h=50)",ylab="Anteil")
> plot(density(nmiete,window="g",width=100),type="l",
+ xlab="Nettomiete (h=100)",ylab="Anteil")
> plot(density(nmiete,window="g",width=200),type="l",
+ xlab="Nettomiete (h=200)",ylab="Anteil")
> plot(density(nmiete,window="g",width=300),type="l",
+ xlab="Nettomiete (h=300)",ylab="Anteil")
```

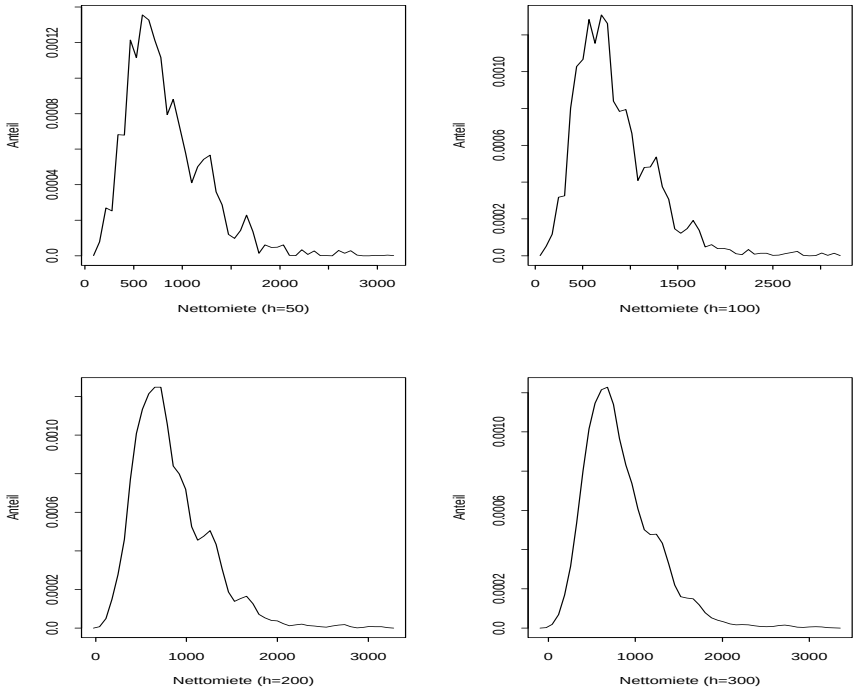

Abbildung 6: Bandbreiten-Variation des Gauss-Kerns für die Variable *nmiete*

Die Rechtecks-Kerne mit denselben Bandbreiten erhält man durch den Parameter `window="r"`:

```
> par(mfrow=c(2,2))
> plot(density(nmiete,window="r",width=50),type="l",
+ xlab="Nettomiete (h=50)",ylab="Anteil")
> plot(density(nmiete,window="r",width=100),type="l",
+ xlab="Nettomiete (h=100)",ylab="Anteil")
> plot(density(nmiete,window="r",width=200),type="l",
+ xlab="Nettomiete (h=200)",ylab="Anteil")
> plot(density(nmiete,window="r",width=300),type="l",
+ xlab="Nettomiete (h=300)",ylab="Anteil")
```

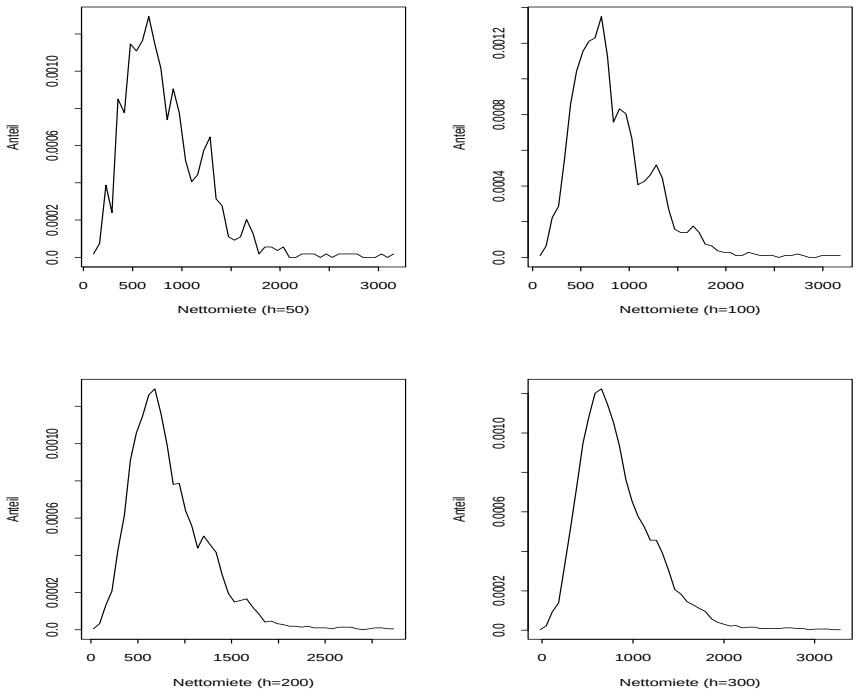

Abbildung 7: Bandbreiten-Variation des Rechteck-Kerns für die Variable *nmiete*

Man erkennt, daß die Wahl der Bandbreite h den größten Einfluß auf die Gestalt des Kerndichteschätzers besitzt, während sich die Wahl des Kerns kaum auf die Gestalt der Schätzung auswirkt.

Für die restlichen Merkmale, die in kategorialer Form vorliegen, bietet sich das Stabdiagramm an. Dazu müssen die kategorialen Daten zunächst durch die Funktion `factor()` als solche definiert werden. Die Benenung der Kategorien erfolgt dabei durch den Parameter `labels`. Anschließend kann durch die Funktion `hist()`, angewandt auf die faktoriellen Variablen, ein Säulendiagramm erstellt werden. Der optionale Parameter `prob=T` dient dazu, ein Säulendiagramm der relativen Häufigkeiten darzustellen. Die Funktionsaufrufe für die Variablen *bad*0,*zh*, *wohn* und *bjkat* lauten:

```
> hist(factor(bad0),xlab="Bad",ylab="Anteil",prob=T)
> hist(factor(zh),xlab="Zentralheizung",ylab="Anteil",prob=T)
> hist(factor(wohn),xlab="Wohnlage",ylab="Anteil",prob=T)
> hist(factor(bjkat),xlab="Baujahr-Kategorie",ylab="Anteil",prob=T)
```

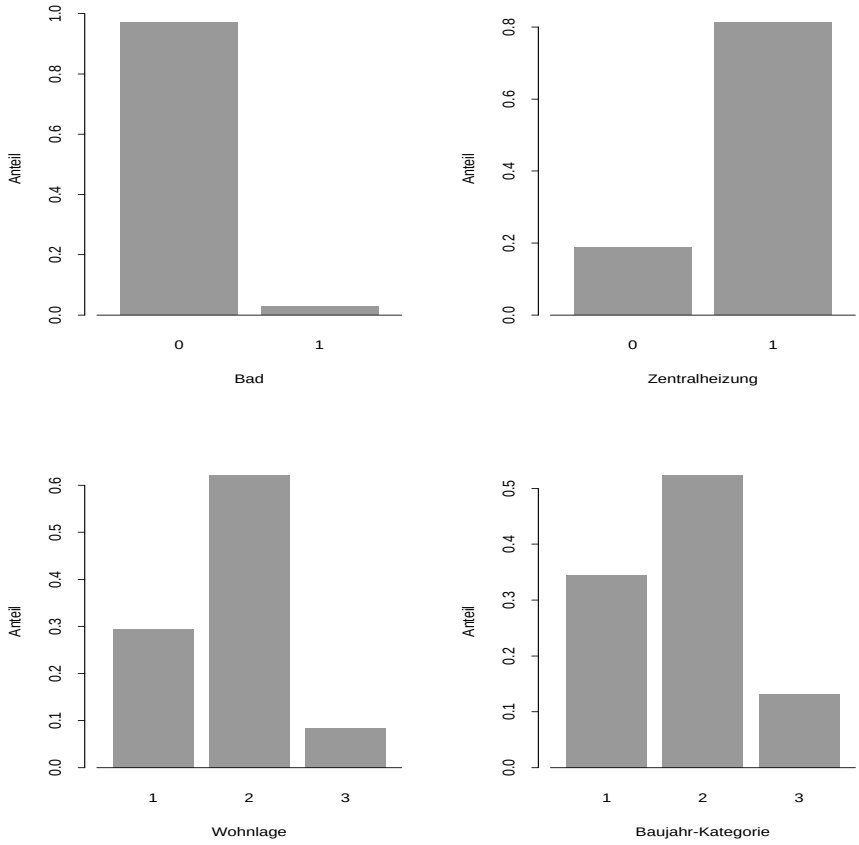

Abbildung 8: Stabdiagramme für die Variablen *bad*0, *zh*, *wohn*, *bjkat*

Die folgende Abbildung zeigt das Ergebnis für die restlichen kategorialen Variablen $ww0$, $badkach$, $fenster$, $kueche$ und adr:

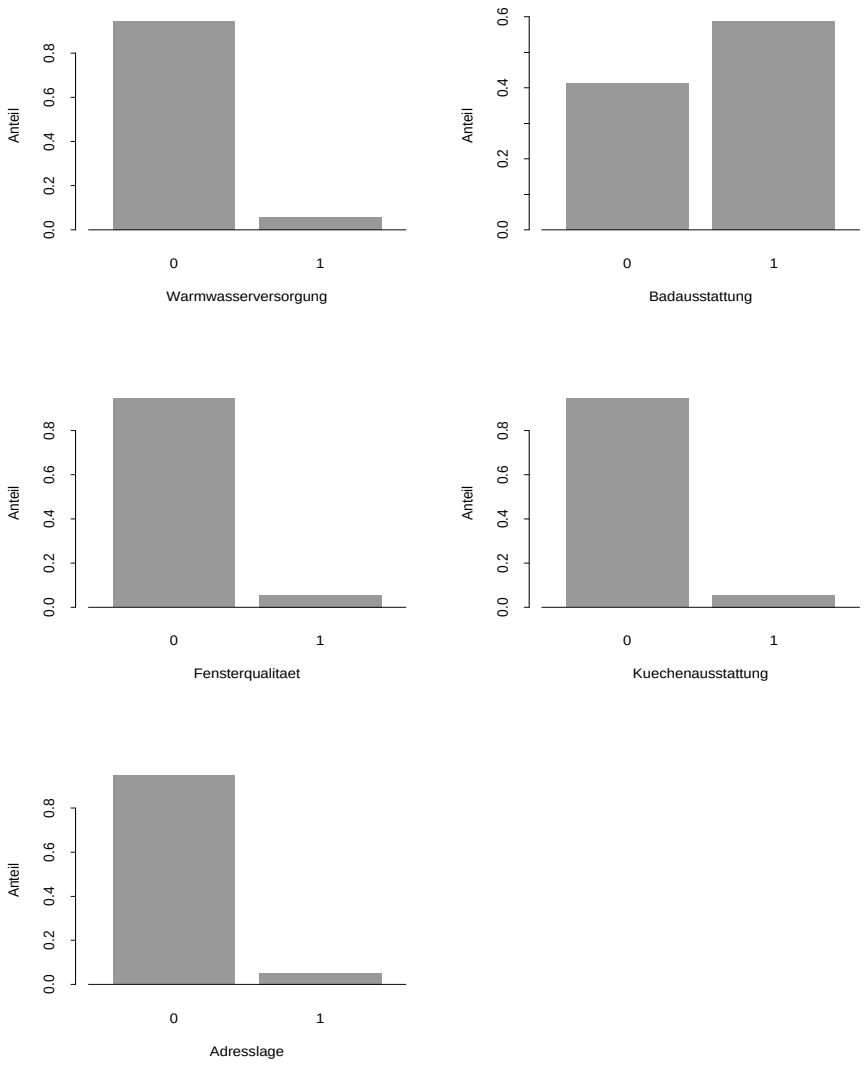

Abbildung 9: Stabdiagramme für die Variablen $ww0$, $badkach$, $fenster$, $kueche$, adr

(b)

Für metrische Merkmale erhält man durch die `summary()`-Funktion die 5-Punkte-Zusammenfassung, wobei Minimum, Maximum, Median und Mittelwert auch einzeln als Funktionen verfügbar sind. Die Funktionsaufrufe lauten:

```
> summary(nmiete)
  Min. 1st Qu. Median  Mean 3rd Qu. Max.
 127.1   543.6    746 830.3    1030 3130
> min(nmiete)
    127.06
> median(nmiete)
    746.01
> mean(nmiete)
    830.3258
> max(nmiete)
    3130
```

Um bestimmte Quantile zu berechnen, verwendet man die `quantile()`-Funktion. Zur Berechnung der Varianz und Standardabweichung die Funktionen `var()` und `sqrt()`. So lassen sich zum Beispiel die 5 %, 10 %, 90 %, und 95 % Quantile der Nettomiete sowie deren Varianz und Standardabweichung durch folgende Aufrufe berechnen:

```
> quantile(nmiete,c(0.05,0.1,0.9,0.95))
      5%      10%     90%      95%
   340.306  401.105  1350   1599.978
> var(nmiete)
  166736.4
> sqrt(var(nmiete))
  408.3337
```

Quantile lassen sich auch direkt berechnen durch

```
> quantile(nmiete,0.25)
      25%
   543.5725
> quantile(nmiete,0.75)
     75%
    1030
```

Die folgende Tabelle fasst diese Ergebnisse für alle metrischen Merkmale zusammen:

	nmiete	*flaeche*	*mvdauer*	*nmqm*
Minimum	127.10	20.0	0	1.57
5 %-Quantil	340.31	32.0	0	5.01
10 %-Quantil	401.11	37.0	1	6.23
25 %-Quantil	543.57	50.3	2	8.86
Median	746.01	67.0	6	12.04
Mittelwert	830.33	69.2	11	12.65
75 %-Quantil	1030.00	84.0	17	16.13
90 %-Quantil	1350.00	100.0	28	19.62
95 %-Quantil	1599.98	114.2	35	21.88
Maximum	3130.00	250.0	82	35.24
Varianz	166736.40	703.2	146	27.61
Standardabw.	408.33	26.5	12	5.25

Die Kennzahlen weisen ebenfalls auf Eigenschaften der Vertilungen hin, die schon aus den Histogrammen und Box-Plots aus Teilaufgabe (a) deutlich wurden. So liegt zum Beispiel bei der Nettomiete der Median deutlich unter dem Mittelwert, was auf eine deutlich linkssteile Verteilung mit einigen Ausreißern schließen läßt. Bei der Nettomiete pro Quadratmeter besitzen Median und Mittelwert sehr ähnliche Werte, was auf die symmetrische Verteilung dieser Variable hinweist.
Die Quantile der Variablen *mvdauer* zeigen, daß bei 5 % der untersuchten Wohnungen die Mietvertragsdauer weniger als ein Jahr betrug. Bei 25 % wurde eine Laufzeit von unter zwei Jahren beobachtet. Dies hängt damit zusammen, daß gemäß den gesetzlichen Bestimmungen für die Durchführung des Mietspiegels nur solche Wohnungen relevant sind, bei denen in den letzten vier Jahren ein neuer Vertrag oder eine Mieterhöhung zustande gekommen ist.

Wendet man `summary()` auf diskrete Merkmale an, wird lediglich die Anzahl der einzelnen Merkmalsausprägungen berechnet. Durch Division mit der Anzahl der Beobachtungen (n=1082) kann man daraus die relativen Häufigkeiten der einzelnen Ausprägungen erhalten. Für die Variablen *bad*0 und *wohn* lauten die Aufrufe:

```
> summary(factor(bad0))/1082
       0           1
   0.9713494   0.02865065
> summary(factor(wohn))/1082
       1           2           3
   0.2948244   0.6219963   0.0831793
```

Folgende Tabelle fasst die Ergebnisse zusammen:

	*bad*0	*zh*	*ww*0	*badkach*	*fenster*	*kueche*
0	0.97	0.19	0.94	0.41	0.95	0.91
1	0.03	0.81	0.06	0.59	0.05	0.09

Für die dreikategorialen Merkmale *adr*, *wohn* und *bjkat* ergibt sich folgendes Ergebnis:

	adr	*wohn*	*bjkat*
1	0.02	0.29	0.34
2	0.96	0.63	0.53
3	0.02	0.08	0.13

(c)
Die Erstellung der neuen Variablen *nmlog* erfolgt durch folgenden Aufruf:

```
> nmlog<-log(nmiete)
```

Zur graphischen Veranschaulichung wurden analog zu Teilaufgabe (a) folgende Kommandos verwendet:

```
> hist(nmlog,ylab="Anteil",xlab="ln(Nettomiete)",
+ prob=T)
> boxplot(nmlog,ylab="ln(Nettomiete)")
> plot(density(nmlog),type="l",xlab="ln(Nettomiete)",
+ ylab="Anteil")
> qqnorm(nmlog,xlab="Quantile der Standardnormalverteilung",
+ ylab="ln(Nettomiete)")
> qqline(nmlog)
```

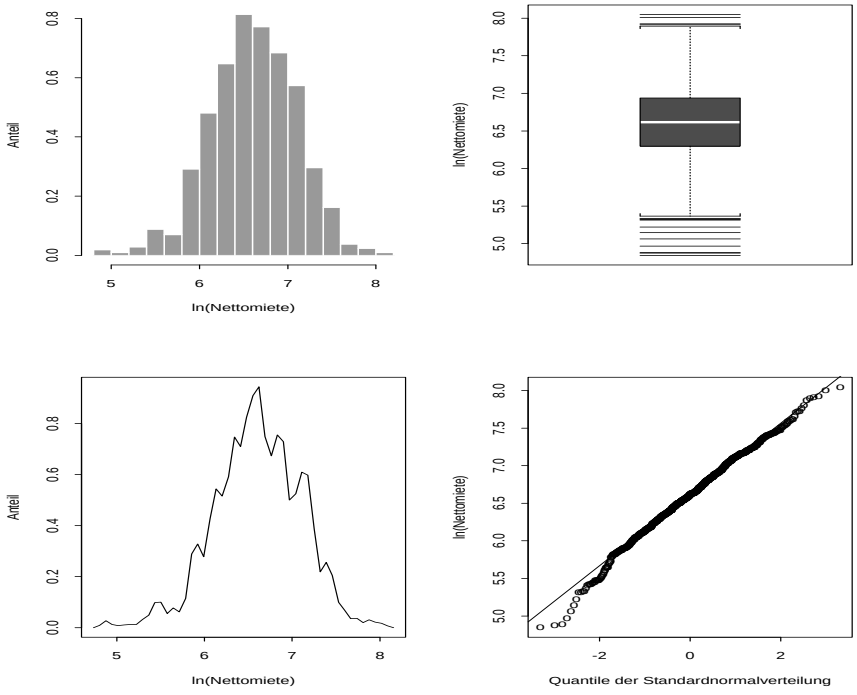

Abbildung 10: Graphische Veranschaulichung für die Variable *nmlog*

Geeignete Kennzahlen erhält man analog zu Teilaufgabe (b) durch folgende Aufrufe:

```
> summary(nmlog)
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 4.845   6.298  6.615 6.607   6.937 8.049
> quantile(nmlog,c(0.05,0.1,0.9,0.95))
       5%       10%       90%       95%
   5.829844  5.994215   7.20786   7.377745
> var(nmlog)
   0.2378607
> sqrt(var(nmlog))
   0.487709
```

Zusammengefasst:

	nmlog
Minimum	4.845
5 %-Quantil	5.830
10 %-Quantil	5.994
25 %-Quantil	6.298
Median	6.615
Mittelwert	6.607
75 %-Quantil	6.937
90 %-Quantil	7.208
95 %-Quantil	7.378
Maximum	8.049
Varianz	0.238
Standardabw.	0.488

Die Ergebnisse und Graphiken zeigen, daß die logarithmierte Nettomiete eher der Normalverteilung entspricht als die Nettomiete. Mittelwert und Median sind nun fast identisch; Box-Plot und Normal-Quantil-Plot weisen nicht auf eine deutliche Abweichung von Symmetrie und Normalverteilung hin.

(d)

Zum Erstellen der Streudiagramme benötigt man folgende Aufrufe:

```
> plot(flaeche,nmiete,xlab="Wohnflaeche",ylab="Nettomiete")
> plot(mvdauer,nmiete,xlab="Mietvertragsdauer",ylab="Nettomiete")
```

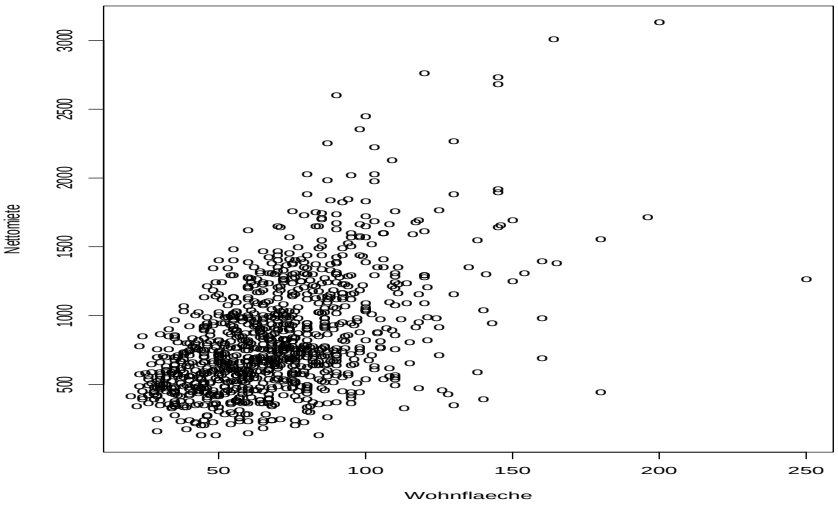

Abbildung 11: Streudiagramm Nettomiete vs. Wohnfläche

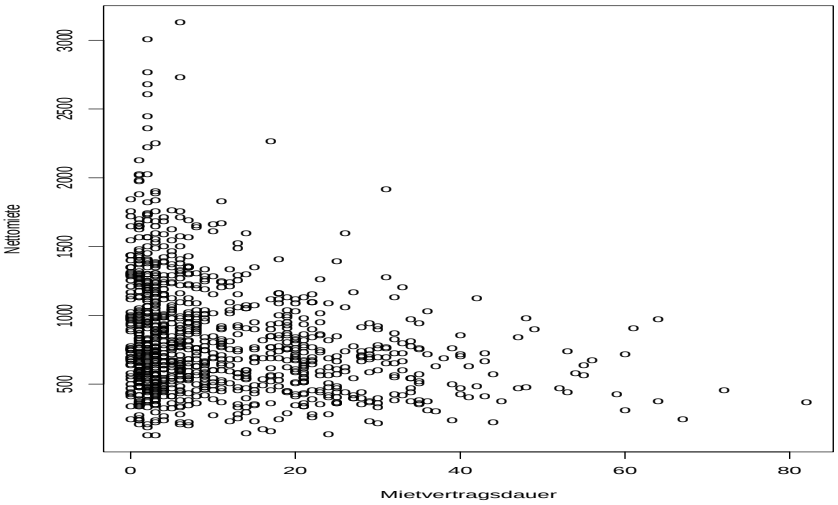

Abbildung 12: Streudiagramm Nettomiete vs. Mietvertragsdauer

Beide Streudiagramme lassen auf einen Zusammenhang zwischen den geplotteten Variablen schließen. Wie zu erwarten ist im ersten Streudiagramm

zu erkennen, daß mit steigender Wohnfläche auch die Nettomiete steigt, wobei noch einige bemerkenswerte Ausreißer auszumachen sind. Auffällig ist außerdem, daß die Streuung der Nettomieten mit der Wohnfläche zunimmt. Das zweite Streudiagramm läßt auf einen leicht negativen Zusammenhang der Merkmale *mvdauer* und *nmiete* schließen.
Die genaue Stärke und Richtung der aus den Streudiagrammen abgelesenen Zusammenhänge läßt sich anhand der Korrelationen ablesen. Dazu berechnet man die empirischen Korrelationskoeffizienten durch die Funktion `cor()`:

```
> cor(nmiete,flaeche)
   0.511144
> cor(nmiete,mvdauer)
   -0.2503123
```

Hier zeigt sich, daß zwischen Nettomiete und Wohnfläche ein deutlich positiver Zusammenhang besteht (0,511). Zwischen der Mietvertragsdauer und der Nettomiete besteht im Gegensatz dazu ein schwach negativer Zusammenhang (-0,250). Diese Kennzahlen bestätigen also die optischen Eindrücke, die man aus den Streudiagrammen bekommt. Offen bleibt hier allerdings die Frage nach der Kausalität, d.h. also ob mit steigender Mietvertragsdauer die Nettomiete wirklich sinkt, oder ob nicht einfach nur günstige Wohnungen länger angemietet werden.

(e)

Die Boxplots erhält man durch folgende Aufrufe, wobei die `split()` Funktion zur Aufteilung der Daten nach einer bestimmten Kategorie dient:

```
> boxplot(split(nmiete,bad0),xlab="Bad")
> boxplot(split(nmiete,zh),xlab="Zentralheizung")
```

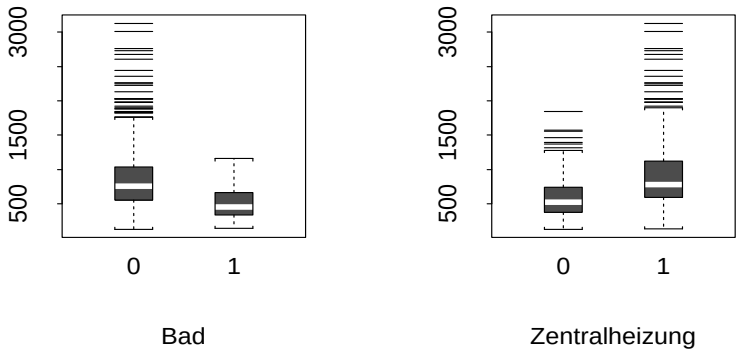

Abbildung 13: Boxplots für die Variablen *bad*0 und *zh*

Man erkennt, daß Wohnungen mit Bad bzw. Zentralheizung eine deutlich andere Verteilung besitzen als Wohnungen ohne Bad bzw. Zentralheizung. Für die Wohnungen mit Bad ist die Streuung der Nettomiete höher als für Wohnungen ohne Bad, wie man aus der breiteren Box und der höheren Ausreißerzahl erkennen kann. Außerdem sind Wohnungen ohne Bad wohl etwas billiger. Die Interpretation der übrigen Box-Plots erfolgt in analoger Weise.

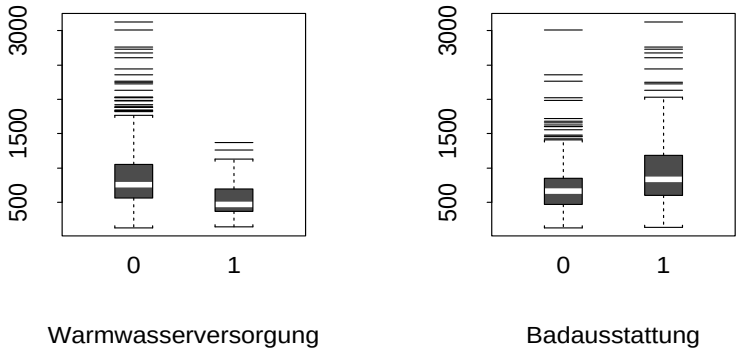

Abbildung 14: Boxplots für die Variablen $ww0$ und *badkach*

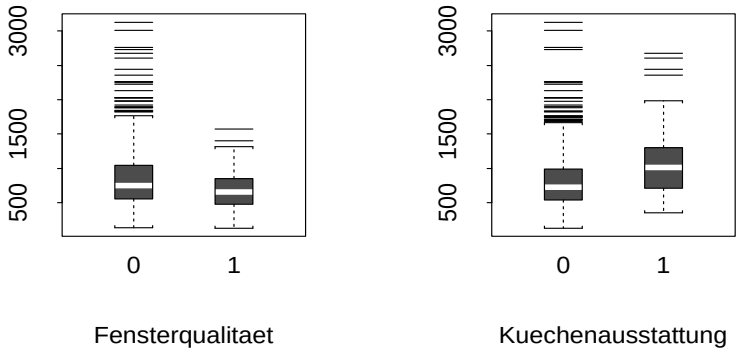

Abbildung 15: Boxplots für die Variablen $fenster$ und *kueche*

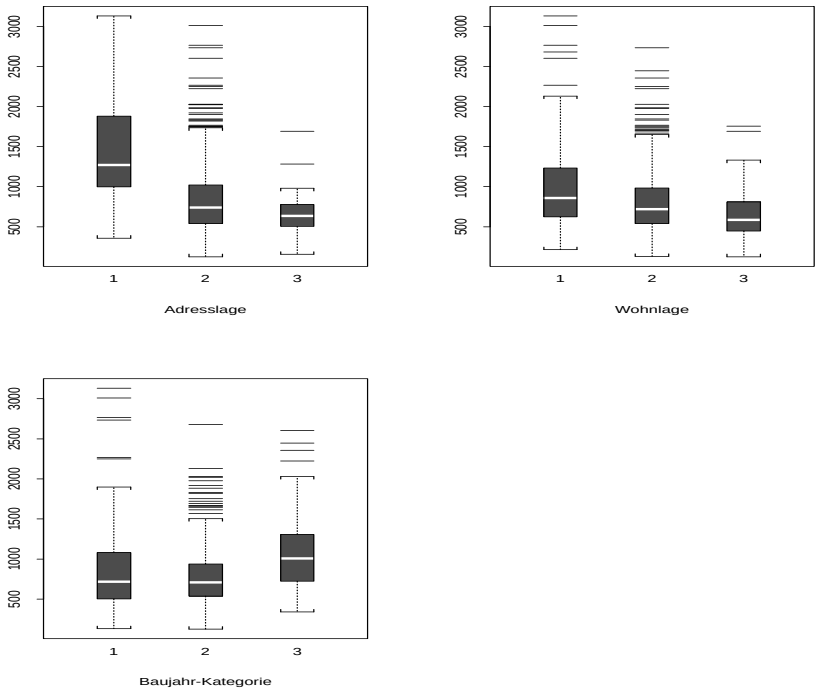

Abbildung 16: Boxplots für die Variablen *adr*, *wohn* und *bjkat*

(f)

Zum Mittelwertsvergleich verwendet man die `t.test()`-Funktion, die auch den entsprechenden t-Test durchführt. Für den Mittelwertsvergleich für die Variable *bad*0 lautet der Aufruf:

```
> t.test(nmiete[bad0==0],nmiete[bad0==1])

        Standard Two-Sample t-Test

   data:  nmiete[bad0 == 0] and nmiete[bad0 == 1]
   t = 4.6273, df = 1080, p-value = 0
   alternative hypothesis: true difference in means is not equal to 0
   95 percent confidence interval:
   196.4740 485.7748
   sample estimates:
   mean of x mean of y
   840.0993  498.9748
```

Man sieht, daß die mittlere Nettomiete bei den Wohnungen mit Bad ungefähr 840,- DM bei den Wohnungen ohne Bad lediglich 499,-DM beträgt. Der t-Wert beträgt 4.6273 woraus ein p-Wert von Null folgt. Die Hypothese, daß die Mittelwerte der beiden Gruppen gleich sind, wird also abgelehnt. Die Mittelwerte sowie t- und p-Werte der entsprechenden Tests für die anderen diskreten Variablen sind in folgender Tabelle zusammengefasst:

Variable	Mittelwert(Gruppe 0)	Mittelwert(Gruppe 1)	t-Wert	p-Wert
*bad*0	840.10	498.97	4.6273	0.0000
zh	597.91	883.68	−9.3197	0.0000
*ww*0	846.99	546.56	5.6164	0.0000
badkach	706.27	917.32	−8.6494	0.0000
$fenster$	838.63	683.65	2.8211	0.0049
kueche	803.41	1088.92	−6.8623	0.0000

Alle p-Werte sind kleiner als 1 %, so daß also alle Mittelwerte der jeweiligen Gruppen signifikant verschieden sind.

(g)
Analog zu Teilaufgabe (d) erhält man die Streudiagramme durch folgende Kommandos:

```
> plot(flaeche,nmqm,xlab="Wohnflaeche",
+ ylab="Nettomiete pro Quadratmeter")
> plot(mvdauer,nmqm,xlab="Mietvertragsdauer",
+ ylab="Nettomiete pro Quadratmeter")
```

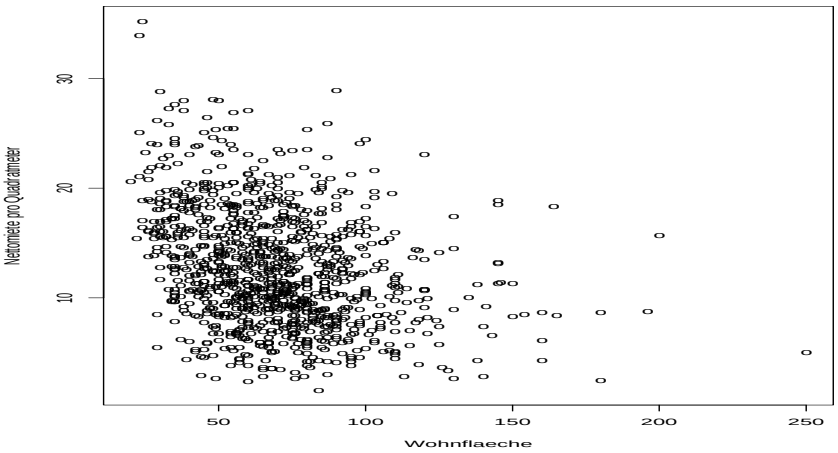

Abbildung 17: Streudiagramm Nettomiete pro qm vs. Wohnfläche

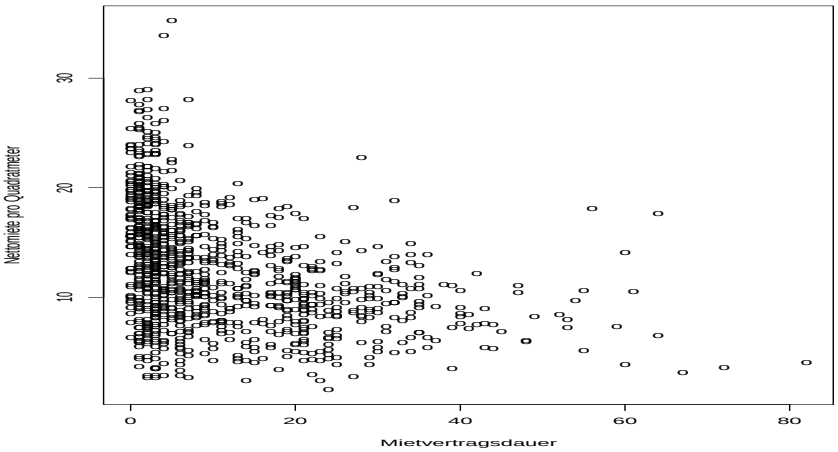

Abbildung 18: Streudiagramm Nettomiete pro qm vs. Mietvertragsdauer

Interpretationen der Streudiagramme lassen sich ähnlich wie in Teilaufgabe (e) vornehmen. Auffällig ist nun, daß ein negativer Zusammenhang zwischen Wohnfläche und Nettomiete pro Quadratmeter besteht, d.h. also, daß für größere Wohnungen

niedrigere Quadratmeterpreise zu entrichten sind. Dies zeigt sich auch am negativen Korrelationskoeffizienten, der analog zu Teilaufgabe (d) durch folgenden Aufruf berechnet wird:

```
> cor(nmqm,flaeche)
   -0.3156023
> cor(nmqm,mvdauer)
   -0.3928142
```

Die für jede Kategorie getrennten Boxplots erhält man analog zu Teilaufgabe (e) durch die Kommandos

```
> boxplot(split(nmqm,bad0),xlab="Bad")
> boxplot(split(nmqm,zh),xlab="Zentralheizung")
```

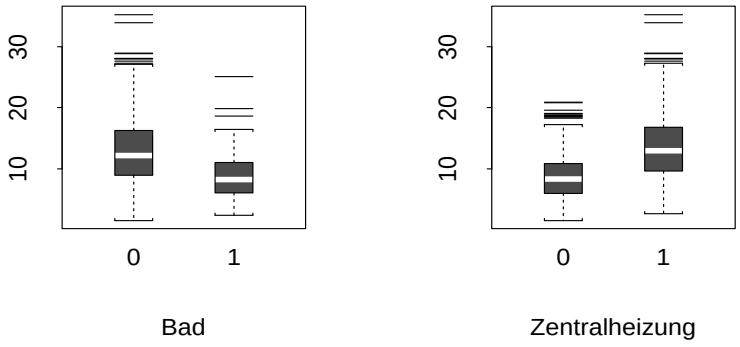

Abbildung 19: Boxplots für die Variablen *bad*0 und *zh*

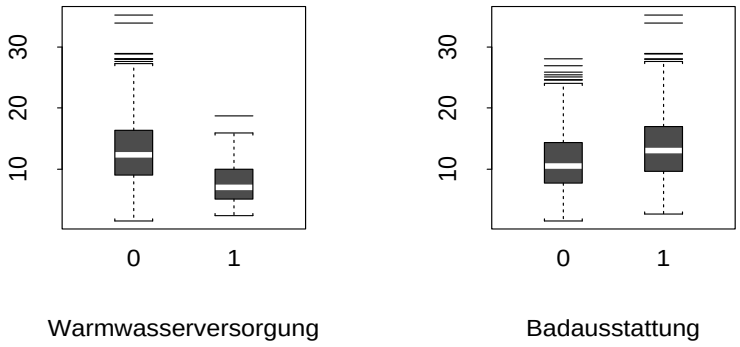

Abbildung 20: Boxplots für die Variablen *ww*0 und *badkach*

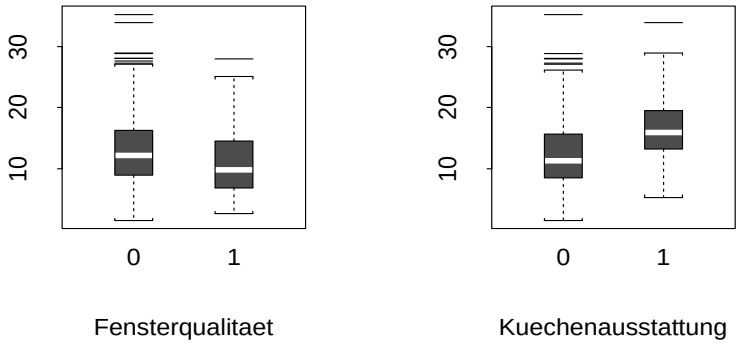

Abbildung 21: Boxplots für die Variablen *fenster* und *kueche*

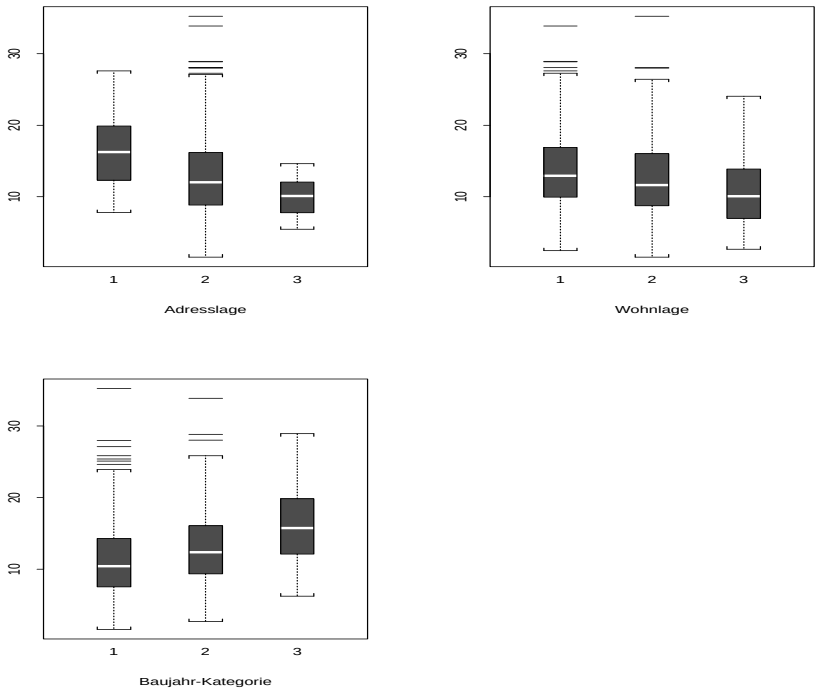

Abbildung 22: Boxplots für die Variablen *adr*, *wohn* und *bjkat*

Analog zu Teilaufgabe (f) erhält man die Mittelwertsvergleiche mit entsprechendem t-Test durch

```
> t.test(nmqm[bad0==0],nmqm[bad0==1])

        Standard Two-Sample t-Test

   data:  nmqm[bad0 == 0] and nmqm[bad0 == 1]
   t = 3.6515, df = 1080, p-value = 0.0003
   alternative hypothesis: true difference in means is not equal to 0
   95 percent confidence interval:
    1.608379 5.344711
   sample estimates:
    mean of x mean of y
     12.74622  9.269677
```

Die Ergebnisse für die anderen Variablen sind in folgender Tabelle zusammengefasst:

Variable	Mittelwert(Gruppe 0)	Mittelwert(Gruppe 1)	t-Wert	p-Wert
$bad0$	12.75	9.27	3.6515	0.0003
zh	8.98	13.49	−11.6776	0.0000
$ww0$	12.94	7.61	7.8576	0.0000
$badkach$	11.42	13.51	−6.5452	0.0000
$fenster$	12.75	10.78	2.7854	0.0054
$kueche$	12.22	16.75	−8.5659	0.0000

Auch beim Mittelwertsvergleich der Nettomiete pro Quadratmeter sind alle Mittelwertsunterschiede signifikant.

(h)

Zunächst werden die dreikategorialen Merkmale adr, $wohn$ und $bjkat$ durch Dummy-Variablen kodiert:

```
> adrsch<-rep(0,length(adr))
> adrgut<-rep(0,length(adr))
> adrgut[adr==1]<-1
> adrsch[adr==3]<-1
> wohnsch<-rep(0,length(wohn))
> wohngut<-rep(0,length(wohn))
> wohngut[wohn==1]<-1
> wohnsch[wohn==3]<-1
> bjkatalt<-rep(0,length(bjkat))
> bjkatneu<-rep(0,length(bjkat))
> bjkatalt[bjkat==1]<-1
> bjkatneu[bjkat==3]<-1
```

Als Referenzkategorie wurde dabei jeweils die mittlere Adress- bzw. Wohnlage gewählt, die vorliegt, wenn beide Dummy-Variablen den Wert Null annehmen. Für die Baujahr-Kategorie wurde das mittlere Baualter als Referenz gewählt. Der `rep()`-Befehl erzeugt zunächst einen Null-Vektor der Länge `length(adr)`. Anschließend wird im Dummy-Vektor an den entsprechenden Stellen eine 1 eingefügt. Als Ergebnis erhält man folgende Dummy-Variablen:

$$adrgut = \begin{cases} 1 & , \text{ gute Adresslage} \\ 0 & , \text{ sonst} \end{cases}$$

$$adrsch = \begin{cases} 1 & , \text{ schlechte Adresslage} \\ 0 & , \text{ sonst} \end{cases}$$

$$wohngut = \begin{cases} 1 & , \text{ gehobene Wohnlage} \\ 0 & , \text{ sonst} \end{cases}$$

$$wohnsch = \begin{cases} 1 & , \text{ einfache Wohnlage} \\ 0 & , \text{ sonst} \end{cases}$$

$$bjkatalt = \begin{cases} 1 & , \text{ Baujahr vor 1948} \\ 0 & , \text{ sonst} \end{cases}$$

$$bjkatneu = \begin{cases} 1 & , \text{ Baujahr ab 1978} \\ 0 & , \text{ sonst} \end{cases}$$

Die bisherigen Ergebnisse legen nahe, alle Variablen als erklärende Größen ins Modell aufzunehmen. Die Mittelwertsdifferenzen der binären Merkmale $bad0$, zh, $ww0$, $badkach$, $fenster$, und $kueche$ waren alle signifikant (siehe Teilaufgabe (f). Die Boxplots der dreikategorialen Merkmale adr, $wohn$ und $bjkat$ lassen darauf schließen,

daß auch diese Größen einen Einfluß auf die Nettomiete haben könnten (siehe Teilaufgabe (e)). Auch die stetigen Variablen *flaeche* und *mvdauer* sollten aufgrund ihrer Korrelationskoeffizienten zumindest überpfüft werden (siehe Teilaufgabe (d)). Ein lineares Regressionsmodell erhält man durch die `lm()`-Funktion. Dabei wird die Regressionsgleichung als Argument an diese Funktion übergeben, wobei links des $\sim$ Zeichens die abhängige Variable, rechts des $\sim$ Zeichens die Einflußgrößen stehen. Zu beachten ist außerdem, daß alle Dummy-Variablen der dreikategorialen Merkmale als erklärende Größen auf der rechten Seite stehen sollten. Der entsprechende Funktionsaufruf lautet:

```
> nmiete.lm<-lm(nmiete~flaeche+mvdauer+bad0+zh+ww0+badkach+fenster
+ kueche+adrgut+adrsch+wohngut+wohnsch+bjkatalt+bjkatneu)
> summary(nmiete.lm)
Call: lm(formula=nmiete~flaeche+mvdauer+bad0+zh+ww0+badkach+fenster
                       +kueche+adrgut+adrsch+wohngut+wohnsch
                       +bjkatalt + bjkatneu)
Residuals:
   Min     1Q Median  3Q  Max
 -1111 -173.9 -15.28 166 1378

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  165.2672   37.1376     4.4501    0.0000
    flaeche    7.8664    0.3600    21.8515    0.0000
    mvdauer   -7.4388    0.7865    -9.4586    0.0000
       bad0  -37.4227   57.6630    -0.6490    0.5165
         zh  147.9028   28.4522     5.1983    0.0000
        ww0 -116.3706   44.7022    -2.6032    0.0094
    badkach   51.9377   19.9483     2.6036    0.0094
    fenster  -52.3547   40.3508    -1.2975    0.1947
     kueche  153.7026   31.7035     4.8481    0.0000
     adrgut  295.4256   65.4720     4.5122    0.0000
     adrsch -156.1065   69.5594    -2.2442    0.0250
    wohngut   72.3855   20.7374     3.4906    0.0005
    wohnsch  -83.9030   37.9701    -2.2097    0.0273
   bjkatalt   19.7506   21.7933     0.9063    0.3650
   bjkatneu  165.0079   28.8960     5.7104    0.0000

Residual standard error: 293.6 on 1067 degrees of freedom
Multiple R-Squared: 0.4896
F-statistic: 73.11 on 14 and 1067 degrees of freedom, the p-value is 0
```

Die `summary()`-Funktion zeigt dabei die Regressionsergebnisse inklusive der Parameterschätzungen und entsprechender Teststatistiken an, die dem Objekt $nmiete1.lm$ zugeordnet wurden. Die Parameterschätzungen sind in der Spalte `Value` abzulesen; die t- und p-Werte der Parametertests in den letzten beiden Spalten. Das Bestimmtheitsmaß beträgt für dieses Modell 0.4896, die F-Statistik für den Overall-F-Test F=73.11 bei p=14 und $n - p - 1$=1067 Freiheitsgraden, was einen p-Wert von näherungsweise Null ergibt.

Aufgrund der hohen p-Werte sollte man die Variablen $bad0$ und $fenster$ aus dem Modell entfernen. Die Dummy-Variable $bjkatalt$ sollte trotz des hohen p-Wertes im Modell verbleiben, da mehrkategoriale Merkmale entweder durch alle Dummy-Variablen oder durch keine repräsentiert werden müssen. Zur Überprüfung des Baualter-Effekts müsste ein simultaner Parametertest für $\beta_{bjkatalt}$ und $\beta_{bjkatneu}$ durchgeführt werden, der allerdings nicht besprochen wurde. Das resultierdende Modell lautet dann:

```
> nmiete.lm<-lm(nmiete~flaeche+mvdauer+zh+ww0+badkach+kueche
+ adrgut+adrsch+wohngut+wohnsch+bjkatalt+bjkatneu)
> summary(nmiete.lm)
Call: lm(formula=nmiete~flaeche+mvdauer+zh+ww0+badkach+kueche
                       +adrgut+adrsch+wohngut+wohnsch
                       +bjkatalt+bjkatneu)
Residuals:
   Min     1Q Median  3Q  Max
 -1111 -173.8 -13.31 166 1379

Coefficients:
               Value Std. Error   t value  Pr(>|t|)
(Intercept)  153.7158   36.0663     4.2620    0.0000
    flaeche    7.8838    0.3579    22.0275    0.0000
    mvdauer   -7.3971    0.7859    -9.4118    0.0000
         zh  153.6154   28.1293     5.4610    0.0000
        ww0 -120.3965   44.0269    -2.7346    0.0063
    badkach   53.2480   19.8453     2.6832    0.0074
     kueche  156.0243   31.6627     4.9277    0.0000
     adrgut  295.8017   65.4726     4.5179    0.0000
     adrsch -153.3910   69.4618    -2.2083    0.0274
    wohngut   74.0641   20.6902     3.5797    0.0004
    wohnsch  -86.0659   37.8047    -2.2766    0.0230
   bjkatalt   19.9702   21.5632     0.9261    0.3546
   bjkatneu  166.1700   28.8310     5.7636    0.0000

Residual standard error: 293.6 on 1069 degrees of freedom
Multiple R-Squared: 0.4886
F-statistic: 85.12 on 12 and 1069 degrees of freedom, the p-value is 0
```

Das Bestimmtheitsmaß R^2 ist lediglich um 1 % auf 0.4886 gesunken, was immer noch auf einen guten Erklärungswert hindeutet. Der zum globalen F-Test zugehörige p-Wert beträgt näherungsweise Null, was ebenfalls für dieses Modell spricht. Die t- und p-Werte zu den einzelnen Regressoren zeigen, daß alle signifikant von Null verschieden sind außer der Dummy-Variable $bjkatalt$, die aber aus oben genannten Gründen im Modell bleiben sollte. Es zeigt sich also, daß die Variablen $flaeche$, $mvdauer$, zh, $ww0$, $badkach$, $kueche$, adr, $wohn$ und $bjkat$ Einfluß auf die Nettomiete haben.
Zur Interpretation der Parameterschätzungen läßt sich festhalten: Die Nettomiete steigt mit der Wohnfläche, während sie mit längerer Mietvertagsdauer sinkt. Neuere Wohnungen sind teurer als ältere Wohnungen, wie man aus dem positiven Parameter des Baujahres ablesen kann. Eine Zentralheizung wird mit einem Zuschlag von

154.-DM berechnet, ein gekacheltes Bad kostet 53.-DM zusätzlich zur Basismiete. Eine gehobene Küchenausstattung wird mit 156.-DM berechnet, während ein Fehlen der Warmwasserversorgung die Nettomiete um 120.-DM senkt. Die Wohnlage besitzt einen Einfluß auf die Miethöhe, wobei eine Wohnung in gehobener Wohnlage 74.-DM mehr kostet, als eine vergleichbare Wohnung in durchschnittlicher Lage; bei einfacher Wohnlage zahlt man dagegen 86.-DM weniger. Auch die Adresslage und das Baualter besitzen einen Einfluß; die entsprechenden Parameter sollten aber selbsterklärend sein.
Das Modell muß aber inhaltlich inhaltlich in folgender Hinsicht kritisch beurteilt werden: Die Höhe der Zu- oder Abschläge für Bad, Zentralheizung, etc. ist völlig unabhängig von der Wohnfläche. So erhält zum Beispiel ein kleines Appartement den gleichen Zuschlag von 154,- DM für eine Zentralheizung wie eine große 5-Zimmer-Wohnung. Wie läßt sich aber ein adäquates Modell finden, das Zu- und Abschläge pro Quadratmeter liefert? Eine Möglichkeit liegen darin, die Nettomiete pro Quadratmeter als abhängige Variable zu verwenden, was im folgenden Abschnitt untersucht wurde.

(i)

Auch hier erscheint die Aufnahme aller Variablen ins Modell sinnvoll, wie die Ergebnisse aus Teilaufgabe (g) zeigen. Es ergibt sich dann folgendes Modell:

```
> nmqm1.lm<-lm(nmqm~flaeche+mvdauer+bad0+zh+ww0+badkach+fenster
+ kueche+adrgut+adrsch+wohngut+wohnsch+bjkatalt+bjkatneu)
> summary(nmqm1.lm)

Call: lm(formula=nmqm~flaeche+mvdauer+bad0+zh+ww0+badkach+fenster
                     +kueche+adrgut+adrsch+wohngut+wohnsch
                     +bjkatalt+bjkatneu)
Residuals:
    Min     1Q  Median    3Q   Max
 -12.88 -2.882 -0.1715 2.615 18.59

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  14.9303   0.5248    28.4485   0.0000
    flaeche  -0.0630   0.0051   -12.3811   0.0000
    mvdauer  -0.1114   0.0111   -10.0248   0.0000
       bad0  -1.1363   0.8149    -1.3944   0.1635
         zh   2.6332   0.4021     6.5489   0.0000
        ww0  -0.9677   0.6317    -1.5318   0.1259
    badkach   0.6017   0.2819     2.1344   0.0330
    fenster  -0.7984   0.5702    -1.4001   0.1618
     kueche   2.3502   0.4480     5.2456   0.0000
     adrgut   2.9475   0.9252     3.1857   0.0015
     adrsch  -2.8807   0.9830    -2.9305   0.0035
    wohngut   1.0228   0.2931     3.4901   0.0005
    wohnsch  -1.2725   0.5366    -2.3714   0.0179
   bjkatalt   0.5507   0.3080     1.7882   0.0740
   bjkatneu   2.1856   0.4084     5.3523   0.0000

Residual standard error: 4.15 on 1067 degrees of freedom
Multiple R-Squared: 0.3844
F-statistic: 47.59 on 14 and 1067 degrees of freedom, the p-value is 0
```

Aufgrund der p-Werte werden $bad0$, $ww0$ und $fenster$ aus dem Modell entfernt, was folgendes Ergebnis liefert:

```
> nmqm.lm<-lm(nmqm~flaeche+mvdauer+zh+badkach+kueche+adrgut+adrsch
+ wohngut+wohnsch+bjkatalt+bjkatneu)
> summary(nmqm.lm)

Call: lm(formula=nmqm~flaeche+mvdauer+zh+badkach+kueche
                     +adrgut+adrsch+wohngut+wohnsch
                     +bjkatalt+bjkatneu)
Residuals:
    Min     1Q  Median    3Q   Max
 -12.83 -2.889 -0.1515 2.594 18.64

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  14.4458   0.4920    29.3639   0.0000
    flaeche  -0.0630   0.0051   -12.4543   0.0000
    mvdauer  -0.1114   0.0111   -10.0235   0.0000
         zh   3.0084   0.3707     8.1151   0.0000
    badkach   0.6923   0.2795     2.4772   0.0134
     kueche   2.4073   0.4482     5.3713   0.0000
     adrgut   2.9817   0.9268     3.2173   0.0013
     adrsch  -2.8140   0.9831    -2.8625   0.0043
    wohngut   1.0309   0.2928     3.5211   0.0004
    wohnsch  -1.3281   0.5346    -2.4843   0.0131
   bjkatalt   0.5212   0.3053     1.7075   0.0880
   bjkatneu   2.1862   0.4082     5.3562   0.0000

Residual standard error: 4.157 on 1070 degrees of freedom
Multiple R-Squared: 0.3804
F-statistic: 59.72 on 11 and 1070 degrees of freedom, the p-value is 0
```

Die Anpassung ist also insgesamt schlechter als bei der Modellierung der Nettomiete. Die Interpretation der Parameterschätzungen ist analog zu Teilaufgabe (h), wobei die hier berechneten Zu- und Abschläge wohnflächenbezogen sind. Das Vorhandensein einer Zentralheizung schlägt sich beispielsweise mit einem Zuschlag von cirka 3,- DM pro Quadratmeter nieder. Interessant ist hierbei, daß die Nettomiete pro Quadratmeter bei zunehmender Wohnfläche und Mietvertragsdauer sinkt. Anders als bei Modell *nmiete.lm* hat bei der Modellierung der Variable *nmqm* die Warmwasserversorgung keinen signifikanten Einfluß.
Die schlechtere Anpassung ist zu einem großen Teil darauf zurückzuführen, daß der Zusammenhang zwischen *nmqm* und *flaeche* zwar negativ ist, allerdings wohl nicht linear fallend, wie man leicht am Streudiagramm aus Teilaufgabe (g) erkennt. Auch der rein lineare Zusammenhang zwischen *nmqm* und *mvdauer* scheint nicht unbedingt gesichert zu sein, so daß eine verbesserte Modellierung an dieser Stelle ansetzen müßte (vgl. Teilaufgabe (k)).

(j)

Die Residualplots für das Modell *nmiete.lm* erhält man durch folgende Aufrufe:

```
> qqnorm(nmiete.lm$resid,ylab="Residuen",
+ xlab="Quantile der Standardnormalverteilung")
> qqline(nmiete.lm$resid)
> plot(nmiete.lm$fitted,nmiete.lm$resid,ylab="Residuen",
+ xlab="gefittete Werte")
> abline(h=0)
```

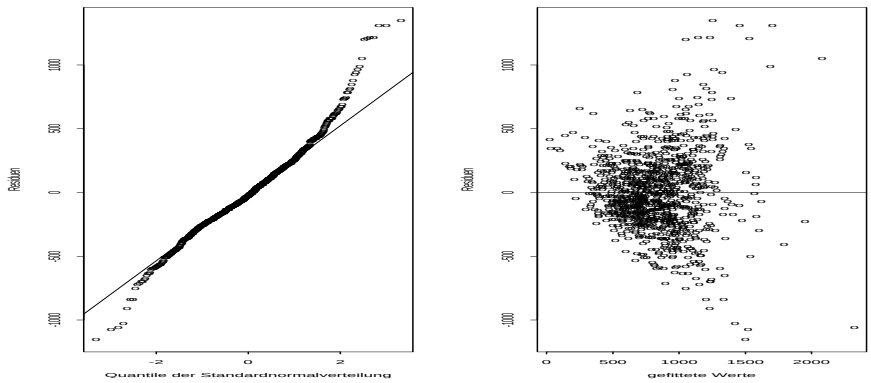

Abbildung 23: Residualplots für das Modell *nmiete.lm*

Die Residualplots für das Modell *nmqm.lm* erhält man durch folgende Aufrufe:

```
> qqnorm(nmqm.lm$resid,ylab="Residuen",
+ xlab="Quantile der Standardnormalverteilung")
> qqline(nmqm.lm$resid)
> plot(nmqm.lm$fitted,nmqm.lm$resid,ylab="Residuen",
+ xlab="gefittete Werte")
> abline(h=0)
```

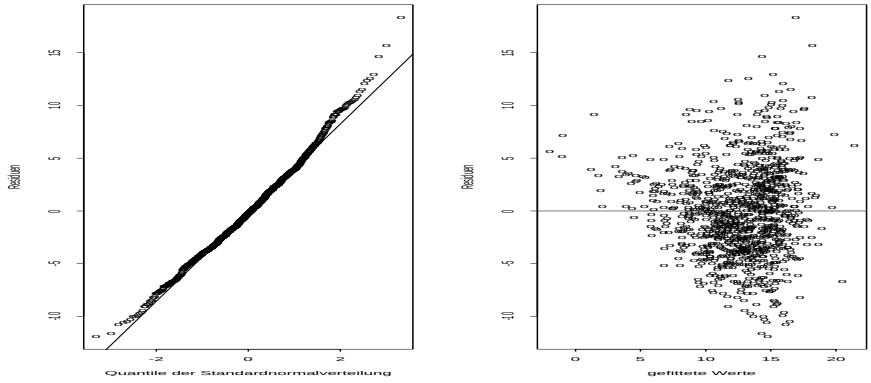

Abbildung 24: Residualplots für das Modell *nmqm.lm*

Bei beiden Modellen ist deutlich die Heteroskedastizität der Residuen erkennbar; in den Residualplots ist die Trichterform sehr ausgeprägt. An den NQ-Plots erkennt

man, daß für das Modell *nmiete.lm* die Residuen zwar eine symmetrische Verteilung besitzen, diese aber stark gekrümmt sein dürfte. Der NQ-Plot für das Modell *nmqm.lm* deutet auf eine linkssteile Verteilung hin.

(k)

Da bei der Modellierung der Nettomiete inhaltliche Probleme bei der Interpretation entstehen (vgl. Teilaufgabe (h)), sollte man sich zunächst auf die Modellierung der Nettomiete pro Quadratmeter beschränken. Wie in Teilaufgabe (i) erwähnt, scheint hauptsächlich eine feinere Modellierung der Wohnfläche sowie der Mietvertragsdauer angebracht.
Für die Mietvertragsdauer wird ein Polynom dritten Grades verwendet, was durch die `poly()`-Funktion erzielt wird. Die Wohnfläche wird als Kehrwert ins Modell aufgenommen, was sich aus dem optischen Eindruck des Streudiagramms (Abb. 17) als Möglichkeit zur Verbesserung ergibt. Dies kann durch die `I()`-Funktion erreicht werden, so daß sich das endgültige Modell durch folgenden Aufruf berechnen lässt:

```
> nmqme.lm<-lm(nmqm~I(1/flaeche)+poly(mvdauer,3)+zh+badkach+kueche
+ adrgut+adrsch+wohngut+wohnsch+bjkatalt+bjkatneu)
> summary(nmqme.lm)

Call: lm(formula=nmqme~I(1/flaeche)+poly(mvdauer,3)+zh+badkach+kueche
                     +adrgut+adrsch+wohngut+wohnsch+bjkatalt+bjkatneu)
Residuals:
    Min      1Q  Median    3Q   Max
 -12.32 -2.605 0.02633 2.432 15.52

Coefficients:
                      Value Std. Error   t value  Pr(>|t|)
      (Intercept)    4.9942    0.5187     9.6276    0.0000
     I(1/flaeche)  252.3396   18.7462    13.4609    0.0000
poly(mvdauer, 3)1  -41.7972    4.2785    -9.7692    0.0000
poly(mvdauer, 3)2   22.6987    4.1189     5.5108    0.0000
poly(mvdauer, 3)3  -13.4058    4.0279    -3.3282    0.0009
               zh    2.8724    0.3559     8.0698    0.0000
          badkach    0.7389    0.2697     2.7392    0.0063
           kueche    2.1169    0.4320     4.9008    0.0000
           adrgut    2.5393    0.8872     2.8623    0.0043
           adrsch   -2.5191    0.9448    -2.6661    0.0078
          wohngut    0.8865    0.2805     3.1600    0.0016
          wohnsch   -1.5719    0.5140    -3.0580    0.0023
         bjkatalt    0.1171    0.2943     0.3978    0.6908
         bjkatneu    2.2022    0.3958     5.5633    0.0000

Residual standard error: 3.991 on 1068 degrees of freedom
Multiple R-Squared: 0.4299
F-statistic: 61.96 on 13 and 1068 degrees of freedom, the p-value is 0
```

Man sieht, daß die Anpassung für dieses Modell mit einem Bestimmtheitsmaß von 0,43 besser ist als bei Modell *nmqm.lm*, das in Teilaufgabe (i) berechnet

wurde. Auch der negative Zusammenhang zwischen Nettomiete pro Quadratmeter und Wohnfläche bleibt trotz des positiven Regressionskoeffizienten erhalten, da die Wohnfläche im Modell *nmqme.lm* als Kehrwert enthalten ist. Ein Blick auf die folgenden Residualplots zeigt auch, daß die Normalverteilungsannahme und die Homoskedastizität für die Residuen in diesem verbesserten Modell stärker erfüllt zu sein scheinen als bei Modell *nmqm.lm* aus Teilaufgabe (i).

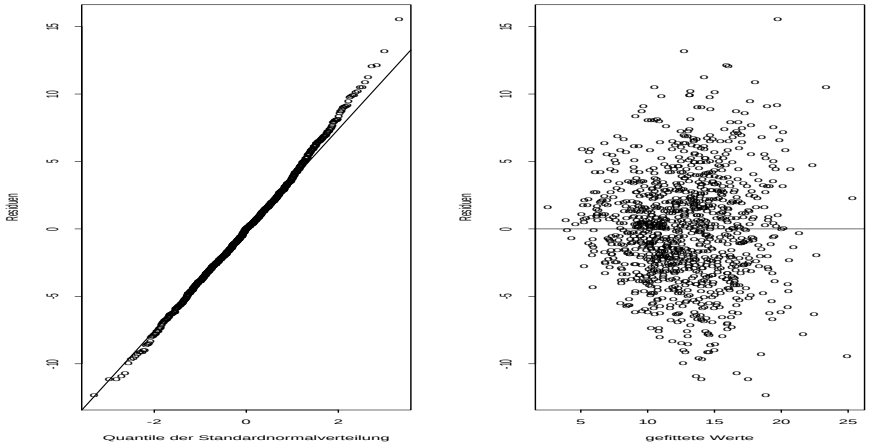

Abbildung 25: Residualplots für das Modell *nmqme.lm*

Als nächsten Schritt könnte man anstatt der Nettomiete pro Quadratmeter als abhängiger Variable die logarithmierte Nettomiete verwenden. Ein Blick auf das Streudiagramm zwischen logarithmierte Nettomiete und Wohnfläche, das hier nicht dargestellt ist aber leicht durch die `plot()`-Funktion erstellt werden kann, zeigt, daß ein positiver linearer Einfluß der Wohnfläche angebracht erscheint, weshalb das folgende Modell die Variable *flaeche* als linearen Term enthält. Die Wohnfläche wird wieder als Polynom dritten Grades modelliert. Als Modell ergibt sich dann:

```
> nmlog.lm<-lm(nmlog~flaeche+poly(mvdauer,3)+zh+ww0+badkach+kueche
+ adrgut+adrsch+wohngut+wohnsch+bjkatalt+bjkatneu)
> summary(nmlog.lm)

Call: lm(formula=nmlog~flaeche+poly(mvdauer,3)+zh+ww0+badkach
                      +kueche+adrgut+adrsch+wohngut+wohnsch
                      +bjkatalt+bjkatneu)
Residuals:
    Min      1Q  Median      3Q     Max
 -1.526 -0.1772 0.04316 0.2386 0.9629

Coefficients:
                     Value Std. Error   t value  Pr(>|t|)
      (Intercept)    5.7110    0.0425   134.4840    0.0000
          flaeche    0.0091    0.0004    20.3887    0.0000
poly(mvdauer, 3)1   -3.2504    0.3815    -8.5190    0.0000
poly(mvdauer, 3)2    1.2727    0.3693     3.4464    0.0006
poly(mvdauer, 3)3   -1.2804    0.3643    -3.5146    0.0005
```

```
              zh     0.2393     0.0344     6.9546     0.0000
             ww0    -0.1935     0.0540    -3.5853     0.0004
         badkach     0.0678     0.0243     2.7958     0.0053
          kueche     0.1502     0.0388     3.8756     0.0001
          adrgut     0.1706     0.0799     2.1355     0.0329
          adrsch    -0.1369     0.0848    -1.6142     0.1068
         wohngut     0.0754     0.0252     2.9851     0.0029
         wohnsch    -0.1689     0.0462    -3.6595     0.0003
        bjkatalt    -0.0019     0.0267    -0.0716     0.9430
        bjkatneu     0.1755     0.0354     4.9592     0.0000

Residual standard error: 0.3581 on 1067 degrees of freedom
Multiple R-Squared: 0.4678
F-statistic: 67 on 14 and 1067 degrees of freedom, the p-value is 0
```

Die Anpassung für dieses Modell ist besser als bei Modell *nmqme.lm*, allerdings sind die Zuschläge für die binären Variablen *zh*, *ww*0, *badkach*, *adr*, *wohn* und *bjkat* nicht wohnflächenabhängig (vgl. Teilaufgabe (i)). Ein Blick auf die Residualplots zeigt, daß die Normalverteilungsannahme für die Residuen kritisch erscheint, während die Homoskedastizität erfüllt zu sein scheint.

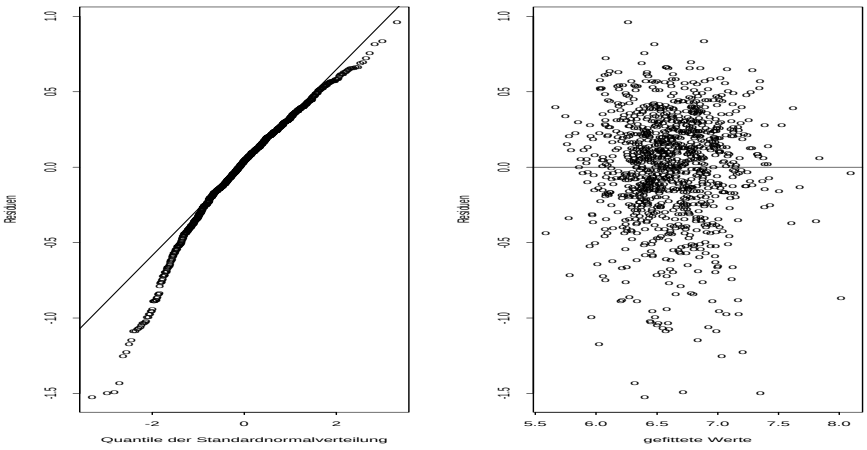

Abbildung 26: Residualplots für das Modell *nmlog.lm*