

**Aufgabe 3** (14 + 5 + 3 + 7 + 7 + 4 + 12 = 52 Punkte)

*Hinweis: Beachten Sie die Tabellen mit Quantilen am Ende der Aufgabenstellung!*

Zahlreiche deutsche Städte erstellen sogenannte Mietspiegel, um Mietern, Vermietern, Mietberatungsstellen und Sachverständigen eine objektive Entscheidungshilfe in Mietfragen zur Verfügung zu stellen. Die Mietspiegel werden dabei insbesondere zur Ermittlung der ortsüblichen Vergleichsmiete (Nettomiete in Abhängigkeit von Wohnungsgröße, -ausstattung, -alter, etc.) herangezogen, wobei die Wohnlage in drei Kategorien unterteilt wird (schlechte/ gute/ beste Wohnlage). Bei der Erstellung von Mietspiegeln wird aus der Gesamtheit aller in Frage kommenden Wohnungen eine repräsentative Zufallsstichprobe gezogen und die interessierenden Daten werden von Interviewern anhand von Fragebögen ermittelt. Der hier verwendete Datensatz bezieht sich auf die Stadt München (Jahr 2003) und enthält folgende Variablen.

nmqm	Nettomiete pro $m^2$
wfl	Wohnfläche
rooms	Anzahl der Zimmer
bj	Baujahr
wohnschlecht	Schlechte Wohnlage? (Ja = 1, Nein = 0)
wohngut	Gute Wohnlage? (Ja = 1, Nein = 0)
wohnbest	Beste Wohnlage? (Ja = 1, Nein = 0)
warmwasser	Warmwasserversorgung vorhanden? (Ja = 1, Nein = 0)
zentralheizung	Zentralheizung vorhanden? (Ja = 1, Nein = 0)
badextra	Zusatzausstattung im Bad vorhanden? (Ja = 1, Nein = 0)
kueche	Gehobene Küchenausstattung? (Ja = 1, Nein = 0)

(a) Zunächst wurde ein lineares Modell geschätzt, das folgenden Output ergab:

Call:

```
lm(formula = nmqm ~ wfl + rooms + badextra + kueche, data = mietspiegel)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-6.525002	-1.668738	0.238541	1.582498	7.133229

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	10.2919619	0.3451742	29.81671	< 2.22e-16
wfl	-0.0122955	???	-1.34733	0.178646
rooms	-0.4837984	0.2212662	-2.18650	0.029365
badextra	???	0.4266865	2.11530	0.035030
kueche	1.8186494	0.4430650	???	0.000049209

Residual standard error: 2.29979 on 395 degrees of freedom

Multiple R-squared: 0.126306, Adjusted R-squared: ???

F-statistic: 14.2759 on ??? and ??? DF, p-value: 6.79971e-11

- (i) Stellen Sie das zugrunde liegende Modell in formaler Schreibweise dar und geben Sie explizit an, welche Variable hier mit Hilfe welcher Regressoren erklärt wird.
- (ii) Wie viele Beobachtungen gingen in die obige Schätzung ein?
- (iii) Bestimmen Sie ferner die mit ??? markierten Größen (mit den Bezeichnungen  $\hat{\beta}_{\text{badextra}}$ ,  $\hat{\sigma}_{\text{wfl}}$ ,  $t_{\text{kueche}}$ ,  $\overline{R^2}$ ,  $DF_1$ ,  $DF_2$ )
- (b) Geben Sie an, welche Regressionskoeffizienten signifikant von Null verschieden sind zum Signifikanzniveau 10%.
- (c) Treffen Sie eine Aussage über die Signifikanz des Erklärungsansatzes. Geben Sie sowohl die Null- als auch die Gegenhypothese an!
- (d) Verwenden Sie einen geeigneten Test, um eine Entscheidung zu treffen, ob eine gehobene Küchenausstattung die Nettomiete pro Quadratmeter in dieser Mietwohnung um signifikant mehr als 1.00€ erhöht. ( $\alpha = 5\%$ )
- Geben Sie hierzu die Hypothesen, die Teststatistik mit ihrer Verteilung unter  $H_0$ , den kritischen Bereich, die realisierte Teststatistik sowie die Testentscheidung an. Beantworten Sie auch explizit die oben formulierte Fragestellung.*
- (e) Es wird vermutet, dass weitere Kriterien einen Einfluss auf die Nettomiete pro Quadratmeter haben. Aus diesem Grund werden Dummy-Variablen für die Wohnlage (**wohnschlecht** und **wohnbest**), eine Warmwasserversorgung (**warmwasser**) und eine Zentralheizung (**zentralheizung**) sowie die Interaktionsvariable **wfl \* rooms** in den Modellansatz aufgenommen.

Das erweiterte Modell liefert folgenden Output:

Call:

```
lm(formula = nmqm ~ wfl + rooms + badextra + kueche + wohnschlecht +
    wohnbest + warmwasser + zentralheizung + I(wfl * rooms),
    data = mietspiegel)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-7.018148	-1.520290	0.052915	1.434932	7.285902

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	8.87713268	0.88771383	9.99999	< 2.22e-16	***
wfl	-0.03760209	0.01318860	-2.85111	0.00458831	**
rooms	-0.97144206	0.30352726	-3.20051	0.00148420	**
badextra	0.63731641	0.39893719	1.59754	0.11095622	
kueche	1.61902814	0.41881534	3.86573	0.00012971	***
wohnschlecht	-0.60591601	0.22451083	-2.69883	0.00726093	**
wohnbest	1.28004874	0.82677592	1.54824	0.12237524	
warmwasser	0.98613573	0.69600043	1.41686	0.15732192	
zentralheizung	2.35689519	0.46034020	5.11990	4.8122e-07	***
I(wfl * rooms)	0.00837570	0.00330257	2.53611	0.01159839	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.13468 on 390 degrees of freedom  
 Multiple R-squared: 0.256781, Adjusted R-squared: 0.23963  
 F-statistic: 14.9716 on 9 and 390 DF, p-value: < 2.22e-16

Testen Sie unter Zuhilfenahme geeigneter Bestimmtheitsmaße zum Signifikanzniveau 1%, ob wenigstens eine der zusätzlichen Variablen tatsächlich relevant ist.

*Geben Sie hierzu die Hypothesen, die Teststatistik mit ihrer Verteilung unter  $H_0$ , den kritischen Bereich, die realisierte Teststatistik sowie die Testentscheidung an. Beantworten Sie auch explizit die oben formulierte Fragestellung.*

- (f) (i) Geben Sie auf Basis des in Teil (e) formulierten und geschätzten Modells eine Punktprognose für die erwartete Nettomiete pro Quadratmeter an, wenn die Wohnung 100 Quadratmeter groß ist, über 4 Zimmer, eine Zentralheizung und eine Warmwasserversorgung verfügt. In dieser Wohnung, die eine gute Wohnlage aufweist, befindet sich keine gehobene Badausstattung und auch keine gehobene Küchenausstattung.
- (ii) Wie ändert sich die Punktprognose, wenn die Badausstattung und die Küchenausstattung verbessert werden, sodass eine gehobene Bad- und Küchenausstattung vorliegen?
- (g) Man vermutet, dass sich die Parameterstruktur zwischen den Wohnungen mit einer Wohnfläche von weniger als 100 Quadratmetern und den Wohnungen mit einer Wohnfläche von mehr als 100 Quadratmetern unterscheidet. Bei einer Schätzung des Modell für die Teilgruppe der Wohnungen mit einer Wohnfläche von weniger als 100 Quadratmetern ergibt sich folgender (gekürzter) Output:

Call:

```
lm(formula = nmqm ~ wfl + rooms + badextra + kueche + wohnschlecht +
    wohnbest + warmwasser + zentralheizung + I(wfl * rooms),
    data = mietspiegel, subset = (wfl < 100))
```

Residual standard error: 2.14969 on 339 degrees of freedom  
 Multiple R-squared: 0.269123, Adjusted R-squared: 0.249719  
 F-statistic: 13.8696 on 9 and 339 DF, p-value: < 2.22e-16

Eine entsprechende Schätzung für die Teilgruppe der Wohnungen mit einer Wohnfläche von mindestens 100 Quadratmetern liefert den folgenden (gekürzten) Output:

Call:

```
lm(formula = nmqm ~ wfl + rooms + badextra + kueche + wohnschlecht +
    wohnbest + warmwasser + zentralheizung + I(wfl * rooms),
    data = mietspiegel, subset = (wfl >= 100))
```

Residual standard error: 1.89241 on 41 degrees of freedom  
 Multiple R-squared: 0.363751, Adjusted R-squared: 0.224087  
 F-statistic: 2.60447 on 9 and 41 DF, p-value: 0.0177328

Berechnen Sie zunächst die Residuenquadratsummen für beide Teilregressionen sowie für den Modellansatz aus Teil (e). Verwenden Sie danach einen geeigneten Test, um die dargestellte Vermutung zu überprüfen ( $\alpha = 1\%$ ).

Geben Sie hierzu die Hypothesen, die Teststatistik mit ihrer Verteilung unter  $H_0$ , den kritischen Bereich, die realisierte Teststatistik sowie die Testentscheidung an. Beantworten Sie auch explizit die oben formulierte Fragestellung.

*Hinweis: Verwenden Sie die folgende Tabelle mit Quantilen einiger  $t(n)$ -Verteilungen:*

$n \backslash p$	0.9	0.95	0.975	0.99
395	1.28370	1.64872	1.96599	2.33583
396	1.28369	1.64871	1.96597	2.33580
397	1.28369	1.64870	1.96596	2.33578
398	1.28368	1.64869	1.96594	2.33575
399	1.28368	1.64868	1.96593	2.33573
400	1.28367	1.64867	1.96591	2.33571

*sowie die folgende Tabelle mit 0.99-Quantilen einiger  $F(m, n)$ -Verteilungen:*

$n \backslash m$	5	6	9	10	380	381	382	390	391
5	10.9670	10.6723	10.1578	10.0510	9.0494	9.0493	9.0492	9.0486	9.0486
6	8.7459	8.4661	7.9761	7.8741	6.9083	6.9082	6.9081	6.9075	6.9075
9	6.0569	5.8018	5.3511	5.2565	4.3383	4.3382	4.3382	4.3376	4.3375
10	5.6363	5.3858	4.9424	4.8491	3.9369	3.9368	3.9367	3.9361	3.9361
380	3.0657	2.8496	2.4542	2.3678	1.2701	1.2700	1.2698	1.2684	1.2682
381	3.0655	2.8495	2.4541	2.3676	1.2699	1.2697	1.2696	1.2682	1.2680
382	3.0654	2.8494	2.4539	2.3675	1.2697	1.2695	1.2693	1.2679	1.2678
390	3.0644	2.8484	2.4530	2.3666	1.2680	1.2678	1.2676	1.2662	1.2660
391	3.0643	2.8483	2.4529	2.3664	1.2677	1.2676	1.2674	1.2660	1.2658

**Aufgabe 4** (15 + 4 + 4 + 7 = 30 Punkte)

*Hinweis: Beachten Sie die Tabelle mit Quantilen am Ende der Aufgabenstellung!*

Der Verkaufspreis gebrauchter VW Golf-Modelle lässt sich anhand mehrerer Variablen erklären. Mit Hilfe eines Regressionsmodells soll der Zusammenhang zwischen dem *Verkaufspreis in Tausend Euro* und dem *Alter des Autos in Monaten*, der *Kilometerleistung in 1000 km*, der *Anzahl der Monate bis zum nächsten TÜV-Termin* sowie den Dummy-Variablen *ABS vorhanden ja/nein* und *Schiebedach vorhanden ja/nein* betrachtet werden. Der hier verwendete Datensatz stammt aus dem Buch von Fahrmeir, Kneib und Lang (2007) und enthält folgende Variablen:

Preis	Verkaufspreis in 1000 Euro
Alter	Alter des Autos in Monaten
Kilstand	Kilometerleistung in 1000 km
Tuev	Anzahl der Monate bis zum nächsten TÜV-Termin
ABS	ABS vorhanden? (Ja = 1, Nein = 0)
Schiebedach	Schiebedach vorhanden? (Ja = 1, Nein = 0)

- (a) Im ersten Schritt erfolgt die Schätzung mit allen angegebenen Variablen unter Annahme homoskedastischer Störgrößen. Man erhält folgenden Output:

Call:

```
lm(formula = Preis ~ Alter + Kilstand + Tuev + ABS + Schiebedach,
    data = Golf)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.45779 -0.53464 -0.00436  0.49883  2.72165
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.311171    0.413047  22.543 < 2e-16 ***
Alter       -0.038329    0.003405 -11.258 < 2e-16 ***
Kilstand    -0.009742    0.001450  -6.720 2.78e-10 ***
Tuev        -0.005483    0.008600  -0.638  0.5246
ABS         -0.237664    0.129523  -1.835  0.0683 .
Schiebedach -0.009862    0.138911  -0.071  0.9435
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7764 on 166 degrees of freedom

Multiple R-squared: 0.6231, Adjusted R-squared: 0.6118

F-statistic: 54.9 on 5 and 166 DF, p-value: < 2.2e-16

Es wird vermutet, dass der Effekt des Antiblockiersystems (ABS) auf den Verkaufspreis dreimal so groß ist wie der Effekt der Anzahl der Monate bis zum nächsten TÜV-Termin. Können Sie die geäußerte Vermutung bei einem Signifikanzniveau von  $\alpha = 10\%$  bestätigen?

Geben Sie hierzu die Hypothesen, die Teststatistik mit ihrer Verteilung unter  $H_0$ , den kritischen Bereich, die realisierte Teststatistik sowie die Testentscheidung an. Beantworten Sie auch explizit die oben formulierte Fragestellung.

Hinweis: Verwenden Sie die folgende Matrix  $(X'X)^{-1}$ :

$$\begin{pmatrix} 0.2830218 & -0.0017802 & -0.0000837 & -0.0021942 & -0.0286433 & -0.0100331 \\ -0.0017802 & 0.0000192 & -0.0000031 & 0.0000009 & 0.0000581 & -0.0000450 \\ -0.0000837 & -0.0000031 & 0.0000034 & -0.0000004 & -0.0000017 & -0.0000353 \\ -0.0021942 & 0.0000009 & -0.0000004 & 0.0001227 & 0.0002713 & -0.000167 \\ -0.0286433 & 0.0000581 & -0.0000017 & 0.0002713 & 0.0278301 & -0.0018737 \\ -0.0100331 & -0.0000451 & -0.0000353 & -0.0001669 & -0.0018737 & 0.0320105 \end{pmatrix}$$

- (b) Eine Untersuchung des Modells aus der vorhergehenden Teilaufgabe mit Hilfe des Breusch-Pagan-Tests ergab folgenden Output:

Breusch-Pagan test

```
data:  lm(Preis ~ Alter + Kilstand + Tuev + ABS + Schiebedach,
        data = Golf)
```

BP = 27.3, df = 5, p-value = 0.0000498

Erläutern Sie kurz, welche Fragestellung mit diesem Test untersucht wird und geben Sie die Hypothesen in der ausführlichen Form an. Wie lautet das Ergebnis der Untersuchung, wenn man ein Signifikanzniveau von  $\alpha = 0.01$  zu Grunde legt?

- (c) Erläutern Sie kurz, was man unter *interner Validität* und *externer Validität* versteht.
- (d) (i) Zur Diagnose welchen Phänomens wird der Varianz-Inflations-Faktor verwendet? Geben Sie die zugehörige Faustregel an.
- (ii) Geben Sie die Berechnungsformel für den Varianz-Inflations-Faktor an und erläutern Sie die darin auftretende(n) Größe(n).
- (iii) Welche Folgen ergeben sich hinsichtlich der Schätzung bei Vorliegen eines großen Varianz-Inflations-Faktors?
- (iv) Wie verändern sich die Varianz-Inflations-Faktoren, wenn Regressoren aus dem Modell entfernt werden?

Hinweis: Verwenden Sie die folgende Tabelle mit Quantilen einiger  $t(n)$ -Verteilungen:

$n \backslash p$	0.9	0.95	0.975	0.99
165	1.28670	1.65414	1.97445	2.34916
166	1.28667	1.65408	1.97436	2.34902
167	1.28664	1.65403	1.97427	2.34888
168	1.28661	1.65397	1.97419	2.34875
169	1.28658	1.65392	1.97410	2.34862
170	1.28655	1.65387	1.97402	2.34848