Aufgabe 3 (11 + 5 + 7 + 7 + 3 + 6 + 3 + 10 = 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. 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)
`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 + bj, data = mietspiegel)

Residuals:
      Min        1Q    Median        3Q       Max
-6.534591 -1.550865  0.029162  1.479532  7.327686

Coefficients:
               Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -47.71357471   7.05914107 -6.75912 3.0882e-11
wfl                 ???   0.00658401  1.02044     0.3079
rooms        -0.72540788   0.16306542 -4.44857 1.0166e-05
bj            0.02939252          ???  8.20764 1.2104e-15

Residual standard error: 2.19575 on 651 degrees of freedom
Multiple R-squared:  0.167681,Adjusted R-squared:  ???
F-statistic: 43.7174 on ??? and ??? DF,  p-value: < 2.22e-16
```

Stellen Sie das zugrunde liegende Modell in formaler Schreibweise dar und geben Sie explizit an, welche Variable hier mit Hilfe welcher Regressoren erklärt wird.

Bestimmen Sie ferner die mit `???` markierten Größen (mit den Bezeichnungen $\widehat{\beta}_{\texttt{wfl}}$, $\widehat{\sigma}_{\texttt{bj}}$, $\overline{R^2}$, $\texttt{DF}_1$ und $\texttt{DF}_2$).

(b) Geben Sie an, wie viele Beobachtungen in die obige Schätzung eingingen und welche Regressionskoeffizienten signifikant von Null verschieden sind zum Signifikanzniveau 5%.

(c) Verwenden Sie einen geeigneten Test, um eine Entscheidung zu treffen, ob ein zusätzlicher Raum in einer Mietwohnung deren Nettomiete pro Quadratmeter um signifikant mehr als 0.5 € reduziert. ($\alpha = 1\%$)

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.

(d) Es wird vermutet, dass weitere Kriterien einen Einfluss auf die Nettomiete pro Quadratmeter haben. Aus diesem Grund werden Dummy-Variablen für die Wohnlage, eine Zusatzausstattung im Bad sowie eine gehobene Küchenausstattung in den Modellansatz aufgenommen.

Das erweiterte Modell liefert folgenden Output:

```
Call:
lm(formula = nmqm ~ wfl + rooms + bj + wohnschlecht + wohnbest +
    badextra + kueche, data = mietspiegel)

Residuals:
      Min        1Q    Median        3Q       Max
-6.760139 -1.452811 -0.029276  1.401765  7.158687

Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)  -45.96562977   6.92049001 -6.64196 6.5669e-11
wfl           -0.00385785   0.00665687 -0.57953 0.56243423
rooms         -0.58284334   0.16033917 -3.63507 0.00029998
bj             0.02880374   0.00351493  8.19469 1.3475e-15
wohnschlecht  -0.69300477   0.17389583 -3.98517 7.5104e-05
wohnbest       1.28043200   0.72455891  1.76719 0.07766807
badextra       0.82340649   0.29156751  2.82407 0.00488781
kueche         1.07848227   0.34212462  3.15231 0.00169472

Residual standard error: 2.13268 on 647 degrees of freedom
Multiple R-squared:  0.219634,Adjusted R-squared:  0.211191
F-statistic: 26.0141 on 7 and 647 DF,  p-value: < 2.22e-16
```

Testen Sie unter Zuhilfenahme geeigneter Bestimmtheitsmaße zum Signifikanzniveau 5%, 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.

(e) In der Stadt München wird zwischen schlechter, guter und bester Wohnlage differenziert. Begründen Sie, warum die Variable `wohngut` nicht zusätzlich in obigen Modellansatz aufgenommen werden kann.

Welche Folgen hätte dies für die Durchführbarkeit der Modellschätzung?

(f) Es wurde der zusätzliche Regressor `badextra * kueche` in das Modell aufgenommen.

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

Residuals:
      Min        1Q    Median        3Q       Max
-6.747857 -1.444874 -0.056039  1.431223  7.175582

Coefficients:
                         Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)          -45.72782001   6.92244118 -6.60574 8.2709e-11
wfl                   -0.00347429   0.00666445 -0.52132 0.60232460
rooms                 -0.59348825   0.16059154 -3.69564 0.00023794
bj                     0.02867781   0.00351606  8.15622 1.8037e-15
wohnschlecht          -0.69376883   0.17386391 -3.99030 7.3549e-05
wohnbest               1.25801388   0.72469831  1.73591 0.08305586
badextra               0.93659840   0.30862411  3.03475 0.00250404
kueche                 1.26518009   0.38071414  3.32318 0.00094024
I(badextra * kueche)  -0.95510606   0.85509825 -1.11695 0.26442892

Residual standard error: 2.13227 on 646 degrees of freedom
Multiple R-squared:  0.221139,Adjusted R-squared:  0.211493
F-statistic:  22.927 on 8 and 646 DF,  p-value: < 2.22e-16
```

Wie werden Regressoren in der Art der hier zusätzlich aufgenommenen erklärenden Variablen bezeichnet?

Hat dieser Regressor einen signifikant negativen Einfluss zum Signifikanzniveau $\alpha = 10\%$? (Begründung!)

Wie groß ist der Effekt einer gehobenen Küchenausstattung auf die Nettomiete pro Quadratmeter,

- falls Zusatzaustattung im Bad vorhanden ist?
- falls keine Zusatzaustattung im Bad vorhanden ist?

(g) Ein Breusch-Pagan-Test (nach Koenker) produziert folgenden Output für den Modellansatz aus Aufgabenteil (d):

```
studentized Breusch-Pagan test

data:  miet_lm2
BP = 25.15, df = 7, p-value = 0.000715
```

Was wird mit diesem Test untersucht und zu welchem Ergebnis kommt man für $\alpha = 1\%$?

(h) Für die Koeffizientenschätzer aus Aufgabenteil (d) wird folgende Varianz-Kovarianz-Matrix unter der Annahme heteroskedastischer Störgrößen geschätzt:

$$\begin{pmatrix} 53.37695 & -0.00900 & 0.04521 & -0.02696 & 0.08085 & 0.39111 & 0.24054 & -0.07567 \\ -0.00900 & 0.00005 & -0.00092 & 0.00000 & 0.00021 & -0.00008 & -0.00045 & -0.00013 \\ 0.04521 & -0.00092 & 0.02548 & -0.00002 & -0.00469 & 0.00019 & 0.00416 & -0.00044 \\ -0.02696 & 0.00000 & -0.00002 & 0.00001 & -0.00005 & -0.00021 & -0.00012 & 0.00004 \\ 0.08085 & 0.00021 & -0.00469 & -0.00005 & 0.03148 & 0.01819 & 0.00818 & 0.00354 \\ 0.39111 & -0.00008 & 0.00019 & -0.00021 & 0.01819 & 0.23285 & 0.01169 & -0.00606 \\ 0.24054 & -0.00045 & 0.00416 & -0.00012 & 0.00818 & 0.01169 & 0.07899 & -0.01080 \\ -0.07567 & -0.00013 & -0.00044 & 0.00004 & 0.00354 & -0.00606 & -0.01080 & 0.10783 \end{pmatrix}$$

Treffen Sie unter Verwendung eines geeigneten Tests eine Entscheidung bezüglich der Hypothese, dass sich eine gehobene Küchenausstattung stärker auf die Quadratmetermiete auswirkt als Zusatzausstattung im Bad. ($\alpha = 10\%$)

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
646	1.28286	1.64722	1.96364	2.33213
647	1.28286	1.64721	1.96364	2.33213
648	1.28286	1.64721	1.96363	2.33212
649	1.28286	1.64720	1.96363	2.33211
650	1.28286	1.64720	1.96362	2.33210
651	1.28285	1.64720	1.96361	2.33209
652	1.28285	1.64719	1.96361	2.33208
653	1.28285	1.64719	1.96360	2.33207
654	1.28285	1.64719	1.96360	2.33206
655	1.28285	1.64718	1.96359	2.33205

sowie die folgende Tabelle mit 0.95-Quantilen einiger $F(m, n)$-Verteilungen:

$n\backslash m$	4	5	6	647	650	651
4	6.38823	6.25606	6.16313	5.63367	5.63365	5.63364
5	5.19217	5.05033	4.95029	4.37124	4.37121	4.37120
6	4.53368	4.38737	4.28387	3.67556	3.67553	3.67552
647	2.38570	2.22795	2.11258	1.13818	1.13802	1.13797
650	2.38564	2.22789	2.11251	1.13799	1.13784	1.13778
651	2.38562	2.22787	2.11249	1.13793	1.13777	1.13772

Aufgabe 4 (3 + 12 + 7 + 4 + 4 = 30 Punkte)

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

Mit Hilfe der Statistiksoftware **R** soll der Datensatz `CPSSW9298` aus dem Paket `AER` untersucht werden, welcher Informationen bezüglich der Lohnverteilung von Vollzeitangestellten im Alter zwischen 25 und 34 in den Jahren 1992 und 1998 zur Verfügung stellt. Als abhängige Variable soll `earnings` verwendet werden, welche den durchschnittlichen Stundenlohn der Arbeiter repräsentiert. Erklärende Variablen sollen aus den binären Variablen für das Geschlecht (`female,male`), dem Alter (`age`) und den zwei Dummy-Variablen für den höchsten erreichten Abschluss (`bachelor,highschool`) gewählt werden.
Folgender Output entstand bei der Analyse der Daten:

```
Call:
lm(formula = earnings ~ age + male + highschool, data = CPSSW9298)

Residuals:
       Min         1Q     Median         3Q        Max
-14.555839  -3.587976  -0.743809   2.479997  29.960221

Coefficients:
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept)  5.104382   2.204451   2.31549   0.020852
age          0.308970   0.073596   4.19819 3.0090e-05
male         2.001435   0.409582   4.88653 1.2522e-06
highschool  -4.954357   0.417603 -11.86379 < 2.22e-16

Residual standard error: 5.61088 on 760 degrees of freedom
Multiple R-squared:  0.179033,Adjusted R-squared:  0.175792
F-statistic: 55.2459 on 3 and 760 DF,  p-value: < 2.22e-16
```

(a) Können Sie eine Aussage über die Signifikanz des Erklärungsansatzes treffen?

Geben Sie sowohl die Null- als auch die Gegenhypothese an!

(b) Als ein Kritikpunkt der Schätzung wird angeführt, dass man die Daten der Jahre 1992 und 1998 nicht zusammen für eine Modellschätzung verwenden sollte, da zwischen diesen Zeitpunkten eine zu große zeitliche Differenz besteht. Aus diesem Grund werden zwei separate Schätzungen durchgeführt, zum einem mit den Daten aus dem Jahr 1992 und zum anderen mit denen aus dem Jahr 1998, was folgende zwei Outputs liefert:

```
Call:
lm(formula = earnings ~ age + male + highschool, data = CPSSW9298,
    subset = (year == 1992))

Residuals:
       Min         1Q     Median         3Q        Max
-12.027399  -3.126277  -0.496281   2.294283  25.467890

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  4.439573   2.646078  1.67779  0.0940910
age          0.315802   0.088792  3.55665  0.0004160
```

```
male         1.333018   0.476314  2.79861  0.0053558
highschool  -4.747075   0.488492 -9.71782 < 2.22e-16

Residual standard error: 5.01625 on 444 degrees of freedom
Multiple R-squared:  0.19021,Adjusted R-squared:  0.184738
F-statistic: 34.7634 on 3 and 444 DF,  p-value: < 2.22e-16

Call:
lm(formula = earnings ~ age + male + highschool, data = CPSSW9298,
    subset = (year == 1998))

Residuals:
       Min         1Q     Median         3Q        Max
-16.011735  -4.287725  -0.723183   3.211047  28.877043

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  6.782565   3.681251  1.84246 0.06635631
age          0.277157   0.122003  2.27173 0.02378367
male         2.797149   0.713804  3.91865 0.00010946
highschool  -5.295262   0.721091 -7.34340  1.826e-12

Residual standard error: 6.23229 on 312 degrees of freedom
Multiple R-squared:  0.178761,Adjusted R-squared:  0.170865
F-statistic:  22.638 on 3 and 312 DF,  p-value: 2.75189e-13
```

Berechnen Sie zunächst die Residuenquadratsummen für beide Teilregressionen sowie für den Modellansatz aus der Aufgabenstellung. Verwenden Sie danach einen geeigneten Test, um die vorgebrachte Kritik zu überprüfen. ($\alpha = 5\%$)

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.

(c) Für diesen Aufgabenteil werden nur die Daten für das Jahr 1998 verwendet. Ein weiterer Kritikpunkt des Modells bestehe nun in der Annahme homoskedastischer Störgrößen.

Verwenden Sie einen geeigneten Test zur Überprüfung der Hypothese, dass die Störtermvarianz für die Gruppe der über 30-jährigen größer ist als für die Gruppe der unter 30-jährigen. ($\alpha = 5\%$)

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.

Verwenden Sie zur Bearbeitung der Aufgabenstellung folgende Outputs:

```
Call:
lm(formula = earnings ~ age + male + highschool, data = CPSSW98,
    subset = (age >= 30))

Residuals:
       Min         1Q     Median         3Q        Max
-16.138441  -4.204824  -0.626228   3.373663  28.575615
```

```
Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  8.595416  11.669434  0.73658 0.46238999
age          0.201950   0.363894  0.55497 0.57964108
male         3.517639   1.042154  3.37536 0.00091234
highschool  -5.045332   1.063707 -4.74316  4.417e-06

Residual standard error: 6.81641 on 171 degrees of freedom
Multiple R-squared:  0.163596,Adjusted R-squared:  0.148923
F-statistic: 11.1489 on 3 and 171 DF,  p-value: 1.01414e-06

Call:
lm(formula = earnings ~ age + male + highschool, data = CPSSW98,
    subset = (age < 30))

Residuals:
       Min         1Q     Median         3Q        Max
-11.586405  -3.961868  -0.513433   2.582776  18.704910

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  2.398865   8.674095  0.27656   0.782539
age          0.465489   0.321003  1.45011   0.149314
male         1.904943   0.955186  1.99432   0.048102
highschool  -5.512482   0.950368 -5.80037 4.3706e-08

Residual standard error: 5.47497 on 137 degrees of freedom
Multiple R-squared:  0.207262,Adjusted R-squared:  0.189902
F-statistic: 11.9396 on 3 and 137 DF,  p-value: 5.40526e-07
```

(d) Folgender Output für Daten aus dem Jahr 1998 wurde unter der Annahme homoskedastischer Störterme produziert.

```
Call:
lm(formula = earnings ~ age + male + highschool, data = CPSSW98)

Residuals:
       Min         1Q     Median         3Q        Max
-16.011735  -4.287725  -0.723183   3.211047  28.877043

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  6.782565   3.681251  1.84246 0.06635631
age          0.277157   0.122003  2.27173 0.02378367
male         2.797149   0.713804  3.91865 0.00010946
highschool  -5.295262   0.721091 -7.34340  1.826e-12

Residual standard error: 6.23229 on 312 degrees of freedom
Multiple R-squared:  0.178761,Adjusted R-squared:  0.170865
F-statistic:  22.638 on 3 and 312 DF,  p-value: 2.75189e-13
```

Welche Teile dieses Outputs verlieren ihre Gültigkeit, wenn von Heteroskedastie in den Störgrößen ausgegangen wird? Sie können sich hierbei auf die Spalten `Estimate`, `Std. Error`, `t value` und `Pr(>|t|)` beschränken!

(e) Wählen Sie einen geeigneten Output, um den erwarteten durchschnittlichen Stundenlohn einer 28-jährigen Frau mit Highschool-Abschluss im Jahr 1998 zu berechnen.

Hinweis: Verwenden Sie die folgende Tabelle mit 0.95-Quantilen einiger $F(m,n)$-Verteilungen:

$n\backslash m$	4	8	137	171	756	760
4	6.38823	6.04104	5.65440	5.64919	5.63287	5.63284
8	3.83785	3.43810	2.96209	2.95529	2.93389	2.93386
137	2.43775	2.00661	1.32583	1.30977	1.25360	1.25350
171	2.42450	1.99290	1.30379	1.28694	1.22696	1.22686
756	2.38371	1.95063	1.23023	1.20962	1.12718	1.12702
760	2.38365	1.95057	1.23011	1.20949	1.12699	1.12683

sowie gegebenenfalls die folgende Tabelle mit 0.05-Quantilen einiger $F(m,n)$-Verteilungen:

$n\backslash m$	4	8	137	171	756	760
4	0.15654	0.26056	0.41021	0.41246	0.41951	0.41952
8	0.16553	0.29086	0.49835	0.50178	0.51265	0.51267
137	0.17685	0.33760	0.75425	0.76699	0.81285	0.81293
171	0.17702	0.33838	0.76349	0.77703	0.82671	0.82680
756	0.17753	0.34084	0.79770	0.81502	0.88717	0.88732
760	0.17753	0.34085	0.79776	0.81509	0.88730	0.88745