**Aufgabe 3** (12 + 4 + 7 + 7 + 3 + 8 + 5 + 10 = 56 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 1994) und enthält folgende Variablen:

| | |
|---|---|
| `nmqm` | Nettomiete pro $m^2$ |
| `wfl` | Wohnfläche |
| `rooms` | Anzahl der Zimmer |
| `mvdauer` | Mietvertragsdauer in Jahren |
| `wohns` | Einfache Wohnlage? (Ja = 1, Nein = 0) |
| `wohnm` | Mittlere Wohnlage? (Ja = 1, Nein = 0) |
| `wohng` | Gehobene Wohnlage? (Ja = 1, Nein = 0) |
| `zh` | Zentralheizung vorhanden? (Ja = 0, Nein = 1) |
| `kueche` | Gehobene Küchenausstattung? (Ja = 1, Nein = 0) |

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

```
Call:
lm(formula = nmqm ~ wfl + rooms + mvdauer, data = mietspiegel)

Residuals:
       Min         1Q     Median         3Q        Max
-12.084446  -2.891030   0.059114   2.482189  15.335323

Coefficients:
              Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 17.49387887  0.70958115 24.65381 < 2.22e-16
wfl          0.00409398  0.01764259      ???  0.8166387
rooms       -1.32574701  0.45719596 -2.89973  0.0039791
mvdauer             ???  0.01937426 -7.51505 5.1445e-13

Residual standard error: 4.58074 on 338 degrees of freedom
Multiple R-squared:  0.215128,Adjusted R-squared:  ???
F-statistic: 30.8812 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.

Wie viele Beobachtungen gingen in die obige Schätzung ein?

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

(b) Geben Sie an, welche Regressionskoeffizienten signifikant negativ sind zum Signifikanzniveau 1%.

*Hinweis: Sie können* $\widehat{\beta}_{mvdauer} = -0.14559855$ *verwenden.*

(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 1 € reduziert. ($\alpha = 5\%$)

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.

(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 gehobene Küchenausstattung sowie eine Zentralheizung in den Modellansatz aufgenommen.

Das erweiterte Modell liefert folgenden Output:

```
Call:
lm(formula = nmqm ~ wfl + rooms + mvdauer + wohns + wohng + kueche +
    zh, data = mietspiegel)

Residuals:
       Min         1Q     Median         3Q        Max
-12.403544  -2.912047  -0.231323   2.735572  12.275988

Coefficients:
              Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 14.3288979  0.8755226 16.36611 < 2.22e-16
wfl         -0.0178379  0.0171166 -1.04214  0.2981013
rooms       -0.8612234  0.4429709 -1.94420  0.0527109
mvdauer     -0.1195687  0.0189858 -6.29779 9.5094e-10
wohns       -1.0137620  0.8547963 -1.18597  0.2364775
wohng        1.4306665  0.5450339  2.62491  0.0090655
kueche       2.7280428  0.9306825  2.93123  0.0036090
zh           3.2042655  0.6331912  5.06050 6.9184e-07

Residual standard error: 4.30272 on 334 degrees of freedom
Multiple R-squared:  0.315702,Adjusted R-squared:  0.301361
F-statistic: 22.0131 on 7 and 334 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 einfacher, mittlerer und gehobener Wohnlage differenziert. Begründen Sie, warum die Variable `wohnm` nicht zusätzlich in obigen Modellansatz aufgenommen werden kann.

Welches Problem würde bezüglich der Modellschätzung entstehen?

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

```
Call:
lm(formula = nmqm ~ wfl + rooms + mvdauer + wohns + wohng + kueche +
    zh + I(kueche * zh), data = mietspiegel)

Residuals:
       Min         1Q     Median         3Q        Max
-12.504826  -2.913239  -0.239361   2.744747  12.268970

Coefficients:
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)    14.3719535  0.8840074 16.25773 < 2.22e-16
wfl            -0.0180712  0.0171497 -1.05373  0.2927683
rooms          -0.8561883  0.4437400 -1.92948  0.0545204
mvdauer        -0.1198696  0.0190268 -6.30003 9.4180e-10
wohns          -1.0139956  0.8558952 -1.18472  0.2369733
wohng           1.4407064  0.5463796  2.63682  0.0087609
kueche          1.6033196  3.1161748  0.51452  0.6072331
zh              3.1554204  0.6470232  4.87683 1.6729e-06
I(kueche * zh)  1.2273213  3.2448267  0.37824  0.7054938

Residual standard error: 4.30825 on 333 degrees of freedom
Multiple R-squared:  0.315996,Adjusted R-squared:  0.299564
F-statistic: 19.2299 on 8 and 333 DF,  p-value: < 2.22e-16
```

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

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

- falls eine Zentralheizung installiert ist?
- falls keine Zentralheizung installiert ist?

Würden Sie auf Grundlage der folgenden Varianzinflationsfaktoren von einem Multikollinearitätsproblem bei der obigen Modellschätzung ausgehen? (Begründung!)

```
      wfl          rooms        mvdauer          wohns          wohng
  4.01824        4.00571        1.10285        1.04750        1.13136
   kueche             zh I(kueche * zh)
 12.12312        1.11586       12.16950
```

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

```
studentized Breusch-Pagan test

data:  miet_lm2
BP = 15.76, df = 7, p-value = 0.0274
```

Was wird mit diesem Test untersucht und zu welchem Ergebnis kommt man für $\alpha = 5\%$? Ändert sich die Entscheidung wenn Sie die Signifikanzniveaus 1% und 10% verwenden?

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

$$\begin{pmatrix} 0.75974 & -0.00326 & -0.07228 & -0.00451 & -0.01330 & 0.00509 & -0.16623 & -0.27744 \\ -0.00326 & 0.00027 & -0.00580 & 0.00002 & 0.00048 & -0.00256 & -0.00099 & 0.00019 \\ -0.07228 & -0.00580 & 0.18058 & -0.00095 & -0.02451 & 0.03951 & 0.05212 & -0.00764 \\ -0.00451 & 0.00002 & -0.00095 & 0.00035 & -0.00010 & -0.00182 & 0.00136 & 0.00190 \\ -0.01330 & 0.00048 & -0.02451 & -0.00010 & 0.71755 & 0.08669 & 0.00420 & -0.03734 \\ 0.00509 & -0.00256 & 0.03951 & -0.00182 & 0.08669 & 0.33279 & 0.07263 & -0.00249 \\ -0.16623 & -0.00099 & 0.05212 & 0.00136 & 0.00420 & 0.07263 & 1.17396 & -0.00299 \\ -0.27744 & 0.00019 & -0.00764 & 0.00190 & -0.03734 & -0.00249 & -0.00299 & 0.32606 \end{pmatrix}$$

Treffen Sie unter Verwendung eines geeigneten Tests eine Entscheidung bezüglich der Hypothese, dass sich eine Zentralheizung stärker auf die Quadratmetermiete auswirkt als eine gehobene Küchenausstattung. ($\alpha = 1\%$)

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 |
|---|---|---|---|---|
| 332 | 1.28411 | 1.64946 | 1.96714 | 2.33763 |
| 333 | 1.28410 | 1.64944 | 1.96711 | 2.33760 |
| 334 | 1.28409 | 1.64943 | 1.96709 | 2.33756 |
| 335 | 1.28408 | 1.64941 | 1.96707 | 2.33753 |
| 336 | 1.28408 | 1.64940 | 1.96705 | 2.33750 |
| 337 | 1.28407 | 1.64939 | 1.96703 | 2.33746 |
| 338 | 1.28406 | 1.64937 | 1.96701 | 2.33743 |
| 339 | 1.28405 | 1.64936 | 1.96699 | 2.33740 |
| 340 | 1.28405 | 1.64935 | 1.96697 | 2.33737 |
| 341 | 1.28404 | 1.64933 | 1.96695 | 2.33733 |
| 342 | 1.28403 | 1.64932 | 1.96692 | 2.33730 |

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

| $n\backslash m$ | 4 | 5 | 6 | 334 | 335 | 336 | 337 | 338 |
|---|---|---|---|---|---|---|---|---|
| 4 | 6.38823 | 6.25606 | 6.16313 | 5.63891 | 5.63888 | 5.63884 | 5.63881 | 5.63878 |
| 5 | 5.19217 | 5.05033 | 4.95029 | 4.37708 | 4.37704 | 4.37701 | 4.37697 | 4.37694 |
| 6 | 4.53368 | 4.38737 | 4.28387 | 3.68181 | 3.68177 | 3.68173 | 3.68169 | 3.68165 |
| 334 | 2.39869 | 2.24102 | 2.12575 | 1.19753 | 1.19739 | 1.19724 | 1.19710 | 1.19696 |
| 335 | 2.39861 | 2.24093 | 2.12567 | 1.19735 | 1.19721 | 1.19706 | 1.19692 | 1.19678 |
| 336 | 2.39853 | 2.24085 | 2.12559 | 1.19717 | 1.19703 | 1.19689 | 1.19674 | 1.19660 |
| 337 | 2.39845 | 2.24077 | 2.12551 | 1.19700 | 1.19685 | 1.19671 | 1.19656 | 1.19642 |
| 338 | 2.39837 | 2.24069 | 2.12543 | 1.19682 | 1.19668 | 1.19653 | 1.19639 | 1.19625 |

**Aufgabe 4** (3 + 4 + 5 + 3 + 4 + 7 = 26 Punkte)

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

Unter Verwendung der Daten des US-amerikanischen Current Population Survey für das Jahr 1985 für Männer und Frauen im Alter von 18 bis 64 Jahren soll der Zusammenhang zwischen dem Stundenlohn (`wage`) und verschiedenen Einflussgrößen untersucht werden. Als erklärende Variablen stehen zur Verfügung:

| | |
|---|---|
| `age` | Alter der befragten Person |
| `experience` | Berufserfahrung in Jahren |
| `male` | Dummy-Variable für das Geschlecht (männlich = 1, weiblich = 0) |
| `married` | Verheiratet? (Ja = 1, Nein = 0) |

(a) Im ersten Schritt wird nur die Berufserfahrung der befragten Personen als echter Regressor verwendet, wodurch sich folgender **R**-Output ergibt:

```
Call:
lm(formula = log(wage) ~ experience, data = CPS1985)

Residuals:
       Min         1Q     Median         3Q        Max
-2.0875478 -0.3838890  0.0090399  0.3699066  1.8135211

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1.97737771 0.03986096 49.60688  < 2e-16
experience  0.00459042 0.00183748  2.49821 0.012783

Residual standard error: 0.525167 on 532 degrees of freedom
Multiple R-squared:  0.0115953,Adjusted R-squared:  0.0097374
F-statistic: 6.24107 on 1 and 532 DF,  p-value: 0.0127828
```

Können Sie eine Aussage über die Signifikanz des Erklärungsansatzes treffen ($\alpha = 5\%$)? Geben Sie sowohl die Null- als auch die Gegenhypothese an!

(b) Im nächsten Schritt werden nun auch alle übrigen Regressoren sowie die quadrierte Berufserfahrung in den Modellansatz aufgenommen, wodurch folgender **R**-Output entsteht:

```
Call:
lm(formula = log(wage) ~ experience + I(experience^2) + age +
    married + male, data = CPS1985)

Residuals:
       Min         1Q     Median         3Q        Max
-2.2572541 -0.2957040  0.0133629  0.2940488  2.1483400

Coefficients:
                    Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)     -0.190374951  0.166142496 -1.14585    0.25237
experience      -0.056494155  0.010147123 -5.56750 4.1182e-08
```

```
I(experience^2) -0.000503935  0.000123157 -4.09180 4.9497e-05
age              0.090215344  0.008029195 11.23591 < 2.22e-16
married          0.050050865  0.043283329  1.15635    0.24806
male             0.254659928  0.038785017  6.56594 1.2388e-10

Residual standard error: 0.44513 on 528 degrees of freedom
Multiple R-squared:  0.295246,Adjusted R-squared:  0.288573
F-statistic: 44.2396 on 5 and 528 DF,  p-value: < 2.22e-16
```

Vergleichen Sie die Koeffizienten von `experience` in den Outputs aus den Aufgabenteilen (a) und (b). Was fällt Ihnen auf und womit könnten Sie diese Beobachtung erklären?

(c) Erläutern Sie kurz, was man unter *interner Validität* und *externer Validität* versteht. Nehmen Sie Stellung zu der *internen Validität* des Modells aus Aufgabenteil (a). Begründen Sie Ihre Antwort.

(d) Besteht zwischen dem logarithmierten Stundenlohn und der Berufserfahrung ein signifikanter nicht-linearer Zusammenhang? Begründen Sie Ihre Antwort.

(e) Verifizieren oder entkräften Sie die Behauptung, dass Männer bei gleicher Berufserfahrung, gleichem Alter und gleichem Beziehungsstatus signifikant mehr verdienen als Frauen. (Begründung!)

Geben Sie die geschätzte prozentuale Lohndifferenz an.

(f) Verwenden Sie einen geeigneten Test zur Überprüfung der Hypothese, dass die Störtermvarianz für die Gruppe der über 40-jährigen größer ist als für die Gruppe der unter 40-jährigen. ($\alpha = 1\%$)

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.

Verwenden Sie zur Bearbeitung der Aufgabenstellung folgende Outputs:

```
Call:
lm(formula = log(wage) ~ experience + I(experience^2) + age +
    married + male, data = CPS1985, subset = age <= 40)

Residuals:
      Min        1Q    Median        3Q       Max
-1.158467 -0.270224  0.000603  0.273937  2.191436

Coefficients:
                  Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)     -0.35587622  0.20500040 -1.73598    0.083455
experience      -0.03224576  0.01972226 -1.63499    0.102957
I(experience^2) -0.00164127  0.00072540 -2.26258    0.024279
age              0.09494362  0.01009774  9.40246  < 2.22e-16
married          0.00742994  0.04931935  0.15065    0.880340
male             0.20320617  0.04570756  4.44579 0.000011802

Residual standard error: 0.423701 on 347 degrees of freedom
Multiple R-squared:  0.318749,Adjusted R-squared:  0.308932
F-statistic: 32.4713 on 5 and 347 DF,  p-value: < 2.22e-16
```

```
Call:
lm(formula = log(wage) ~ experience + I(experience^2) + age +
    married + male, data = CPS1985, subset = age > 40)

Residuals:
       Min         1Q     Median         3Q        Max
-2.2671165 -0.3167679  0.0546708  0.3519452  0.9696302

Coefficients:
                    Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)      0.2139881342  0.6582171640  0.32510   0.745492
experience      -0.0818947502  0.0385738745 -2.12306   0.035154
I(experience^2) -0.0000182335  0.0005222989 -0.03491   0.972191
age              0.0861453663  0.0140715885  6.12194 5.9104e-09
married          0.0764784983  0.0902920821  0.84701   0.398145
male             0.3345209518  0.0728822326  4.58988 8.4372e-06

Residual standard error: 0.479884 on 175 degrees of freedom
Multiple R-squared:  0.279014,Adjusted R-squared:  0.258415
F-statistic: 13.5447 on 5 and 175 DF,  p-value: 3.58517e-11
```

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

| $n\backslash m$ | 5 | 10 | 175 | 347 | 523 | 528 |
|---|---|---|---|---|---|---|
| 5 | 10.96702 | 10.05102 | 9.08318 | 9.05214 | 9.04148 | 9.04128 |
| 10 | 5.63633 | 4.84915 | 3.96922 | 3.93949 | 3.92926 | 3.92906 |
| 175 | 3.12348 | 2.42368 | 1.42351 | 1.36759 | 1.34667 | 1.34627 |
| 347 | 3.07031 | 2.37227 | 1.34849 | 1.28438 | 1.25934 | 1.25885 |
| 523 | 3.05234 | 2.35488 | 1.32150 | 1.25330 | 1.22596 | 1.22542 |
| 528 | 3.05201 | 2.35456 | 1.32099 | 1.25270 | 1.22531 | 1.22477 |

*sowie gegebenenfalls die folgende Tabelle mit 0.01-Quantilen einiger $F(m,n)$-Verteilungen:*

| $n\backslash m$ | 5 | 10 | 175 | 347 | 523 | 528 |
|---|---|---|---|---|---|---|
| 5 | 0.09118 | 0.17742 | 0.32016 | 0.32570 | 0.32762 | 0.32765 |
| 10 | 0.09949 | 0.20622 | 0.41260 | 0.42154 | 0.42465 | 0.42471 |
| 175 | 0.11009 | 0.25194 | 0.70249 | 0.74157 | 0.75672 | 0.75701 |
| 347 | 0.11047 | 0.25384 | 0.73122 | 0.77859 | 0.79789 | 0.79827 |
| 523 | 0.11060 | 0.25450 | 0.74257 | 0.79407 | 0.81569 | 0.81612 |
| 528 | 0.11060 | 0.25451 | 0.74279 | 0.79438 | 0.81605 | 0.81648 |