**Aufgabe 3** (15 + 1 + 7 + 7 + 7 + 5 = 42 Punkte)

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

Mit Hilfe der Statistiksoftware **R** soll der Datensatz `HousePrices` aus dem Paket `AER` untersucht werden, welcher Informationen bezüglich der Preise von Häusern in der Stadt Windsor in Kanada aus den Monaten Juli, August und September im Jahr 1987 enthält. Es wird der Zusammenhang zwischen dem *Verkaufspreis eines Hauses in Dollar* und der *Anzahl an Schlafzimmern*, der *Anzahl an Badezimmern*, der *Anzahl der Garagen* sowie den Dummy-Variablen *Einfahrt vorhanden ja/nein*, *Klimaanlage vorhanden ja/nein* und *Freizeitraum vorhanden ja/nein* betrachtet.

| | |
|---|---|
| `price` | Verkaufspreis in Dollar |
| `bedrooms` | Anzahl der Schlafzimmer |
| `bathrooms` | Anzahl der Badezimmer |
| `garage` | Anzahl der Garagen |
| `driveway` | Einfahrt vorhanden? (Ja = 1, Nein = 0) |
| `aircon` | Klimaanlage vorhanden? (Ja = 1, Nein = 0) |
| `recreation` | Freizeitraum vorhanden? (Ja = 1, Nein = 0) |

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

```
Call:
lm(formula = price ~ bedrooms + bathrooms + garage, data = HousePrices)

Residuals:
     Min       1Q   Median       3Q      Max
-60569.2 -13915.5  -1690.5  12376.4  88416.6

Coefficients:
            Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 15624.66    3877.49  4.02958 6.3870e-05 ***
bedrooms         ???    1323.12  4.85040 1.6125e-06 ***
bathrooms   21230.52    1955.31 10.85786 < 2.22e-16 ***
garage       8913.52        ???  8.34799 5.8133e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21059.2 on 542 degrees of freedom
Multiple R-squared:  0.381446,Adjusted R-squared:  ???
F-statistic: 111.413 on ??? and ??? DF,  p-value: < 2.22e-16
```

(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 $\hat{\beta}_{bedrooms}$, $\hat{\sigma}_{garage}$, $\overline{R^2}$, $DF_1$ und $DF_2$).

(iv) Welchen Test gibt die letzte Zeile des Outputs an? Zu welchem Ergebnis kommt dieser Test? Begründen Sie Ihre Antwort und geben Sie die zugehörigen Hypothesen an.

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

(c) Testen Sie zum Signifikanzniveau 1%, ob eine zusätzliche Garage den Verkaufspreis um signifikant mehr als 7000 Dollar erhöht.

*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: Sie können $\hat{\sigma}_{garage} = 1067.74$ verwenden.*

(d) Es wird vermutet, dass der Effekt eines zusätzlichen Badezimmers mehr als doppelt so groß ist wie der Effekt einer zusätzlichen Garage. Können Sie die geäußerte Vermutung bei einem Signifikanzniveau von $\alpha = 5\%$ bestätigen?

*Geben Sie 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 Varianz-Kovarianz-Matrix:*

$$\hat{V}(\hat{\beta}) = \begin{pmatrix} 15034951.13 & -3922023.42 & -1969655.24 & -87669.95 \\ -3922023.42 & 1750643.67 & -926545.81 & -112254.36 \\ -1969655.24 & -926545.81 & 3823253.57 & -286811.03 \\ -87669.95 & -112254.36 & -286811.03 & 1140076.57 \end{pmatrix}$$

(e) Man vermutet, dass weitere Variablen einen Einfluss auf den Verkaufspreis haben. Aus diesem Grund werden die Dummy-Variablen für das Vorhandensein einer Klimaanlage, das Vorhandensein einer Einfahrt und das Vorhandensein eines Freizeitraums in die Modellschätzung aufgenommen. Das erweiterte Modell liefert folgenden Output:

```
Call:
lm(formula = price ~ bedrooms + bathrooms + garage + aircon +
    driveway + recreation, data = HousePrices)

Residuals:
      Min        1Q    Median        3Q       Max
-46253.52 -11170.66  -1460.93   9408.44  90015.42

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  3630.364   3889.729  0.93332    0.35107
bedrooms     5575.752   1151.791  4.84094 1.6903e-06 ***
bathrooms   18208.940   1710.896 10.64293 < 2.22e-16 ***
garage       6492.831    948.575  6.84483 2.0897e-11 ***
aircon      16816.306   1742.442  9.65100 < 2.22e-16 ***
```

```
driveway     15223.419   2306.028  6.60158 9.7552e-11 ***
recreation    9269.080   2078.026  4.46052 9.9564e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18221.7 on 539 degrees of freedom
Multiple R-squared:  0.539465,Adjusted R-squared:  0.534339
F-statistic:  105.23 on 6 and 539 DF,  p-value: < 2.22e-16
```

Testen Sie unter Zuhilfenahme geeigneter Bestimmtheitsmaße zum Signifikanzniveau 5%, ob wenigstens einer der zusätzlichen Regressoren 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 den erwarteten Verkaufspreis an, wenn das Haus über 3 Schlafzimmer, 2 Badezimmer, 2 Garagen sowie eine Klimaanlage und eine Einfahrt verfügt, allerdings kein Freizeitraum vorhanden ist.

(ii) Wie ändert sich die Punktprognose, wenn ein Schlafzimmer zu einem Freizeitraum umgebaut wird?

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

| $n\backslash p$ | 0.9 | 0.95 | 0.975 | 0.99 |
|---|---|---|---|---|
| 540 | 1.28312 | 1.64768 | 1.96437 | 2.33327 |
| 541 | 1.28312 | 1.64768 | 1.96436 | 2.33326 |
| 542 | 1.28312 | 1.64767 | 1.96435 | 2.33325 |
| 543 | 1.28311 | 1.64766 | 1.96434 | 2.33323 |
| 544 | 1.28311 | 1.64766 | 1.96433 | 2.33322 |
| 545 | 1.28311 | 1.64765 | 1.96433 | 2.33321 |

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

| $n\backslash m$ | 2 | 3 | 4 | 537 | 538 | 539 | 540 |
|---|---|---|---|---|---|---|---|
| 2 | 19.0000 | 19.1643 | 19.2468 | 19.4939 | 19.4939 | 19.4939 | 19.4939 |
| 3 | 9.5521 | 9.2766 | 9.1172 | 8.5316 | 8.5316 | 8.5316 | 8.5316 |
| 4 | 6.9443 | 6.5914 | 6.3882 | 5.6348 | 5.6348 | 5.6348 | 5.6348 |
| 537 | 3.0125 | 2.6215 | 2.3885 | 1.1527 | 1.1526 | 1.1525 | 1.1525 |
| 538 | 3.0125 | 2.6215 | 2.3885 | 1.1526 | 1.1525 | 1.1525 | 1.1524 |
| 539 | 3.0124 | 2.6214 | 2.3885 | 1.1525 | 1.1524 | 1.1524 | 1.1523 |
| 540 | 3.0124 | 2.6214 | 2.3884 | 1.1524 | 1.1524 | 1.1523 | 1.1522 |

**Aufgabe 4** (7 + 12 + 12 + 4 + 5 = 40 Punkte)

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

Mit Hilfe der Statistiksoftware **R** soll der Datensatz `CPS1985` aus dem Paket `AER` untersucht werden, welcher Informationen bezüglich der Lohnverteilung von Vollzeitangestellten in den USA im Jahr 1985 zur Verfügung stellt. Die abhängige Variable `wage`, welche den durchschnittlichen Stundenlohn der Angestellten in USD angibt, soll durch die Ausbildungsdauer in Jahren `education`, die Berufserfahrung in Jahren `experience` und die Dummy-Variablen Geschlecht `female` und Familienstand `married` erklärt werden.

| | |
|---|---|
| `wage` | Stundenlohn in Dollar |
| `education` | Ausbildungsdauer in Jahren |
| `experience` | Berufserfahrung in Jahren |
| `female` | Geschlecht? (männlich = 0, weiblich = 1) |
| `married` | Verheiratet? (Ja = 1, Nein = 0) |

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

```
Call:
lm(formula = wage ~ education + experience + female + married,
    data = CPS1985)

Residuals:
   Min     1Q Median     3Q    Max
 -9.66  -2.72  -0.64   1.91  37.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.3262     1.1924   -3.63  0.00031 ***
education     0.9337     0.0790   11.82  < 2e-16 ***
experience    0.1071     0.0174    6.17  1.4e-09 ***
female       -2.3319     0.3879   -6.01  3.4e-09 ***
married       0.5410     0.4220    1.28  0.20040
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.45 on 529 degrees of freedom
Multiple R-squared:  0.255,Adjusted R-squared:  0.25
F-statistic: 45.4 on 4 and 529 DF,  p-value: <2e-16
```

Man vermutet allerdings, dass die bei der ersten Modellschätzung getroffene Annahme homoskedastischer Störgrößen nicht gerechtfertigt ist. Um dies zu überprüfen, wird eine Hilfsregression quadrierter Resiuduen $\texttt{uhat2}_i := \hat{u}_i^2$ der ursprünglichen Modellschätzung mit folgendem Ergebnis durchgeführt:

```
Call:
lm(formula = uhat2 ~ education + experience + female + married,
    data = CPS1985)

Residuals:
   Min     1Q Median     3Q    Max
 -42.3  -17.5   -9.1    0.1 1418.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.724     18.445   -1.18   0.2394
education      3.541      1.222    2.90   0.0039 **
experience     0.138      0.269    0.51   0.6073
female        -3.666      5.999   -0.61   0.5415
married       -8.428      6.528   -1.29   0.1973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.9 on 529 degrees of freedom
Multiple R-squared:  0.0199,Adjusted R-squared:  0.0125
F-statistic: 2.68 on 4 and 529 DF,  p-value: 0.0309
```

Führen Sie einen Breusch-Pagan-Test in der Variante nach Koenker durch, um das Vorliegen von Heteroskedastie in den Störgrößen zum Signifikanzniveau $\alpha = 5\%$ zu überprüfen.

*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.*

(b) Es wird vermutet, dass sich die Parameterstruktur zwischen den Gruppen der über 35-jährigen Personen und der unter 35-jährigen Personen unterscheidet. Bei einer Schätzung des Modells für die Teilgruppe der unter 35-jährigen Personen ergibt sich folgender Output:

```
Call:
lm(formula = wage ~ education + experience + female + married,
    data = CPS1985, subset = (age < 35))

Residuals:
   Min     1Q Median     3Q    Max
 -8.30  -2.47  -0.64   1.43  38.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.7192     1.9460   -2.94   0.0036 **
education     0.9296     0.1305    7.12  1.0e-11 ***
experience    0.2615     0.0647    4.04  7.1e-05 ***
female       -1.2233     0.5532   -2.21   0.0279 *
married      -0.1237     0.5843   -0.21   0.8325
```

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

Residual standard error: 4.43 on 258 degrees of freedom
Multiple R-squared:  0.186,Adjusted R-squared:  0.174
F-statistic: 14.8 on 4 and 258 DF,  p-value: 7.13e-11
```

Eine entsprechende Schätzung für die Teilgruppe der über 35-jährigen Personen liefert den folgenden Output:

```
Call:
lm(formula = wage ~ education + experience + female + married,
    data = CPS1985, subset = (age >= 35))

Residuals:
    Min      1Q  Median      3Q     Max
-10.095  -2.721  -0.402   2.413  13.688

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.2708     2.1085   -1.08    0.282
education     0.9106     0.1088    8.37  3.4e-15 ***
experience    0.0652     0.0323    2.02    0.044 *
female       -3.3573     0.5368   -6.25  1.6e-09 ***
married       0.4665     0.6397    0.73    0.467
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.37 on 266 degrees of freedom
Multiple R-squared:  0.31,Adjusted R-squared:  0.299
F-statistic: 29.8 on 4 and 266 DF,  p-value: <2e-16
```

Berechnen Sie zunächst die Residuenquadratsummen für beide Teilregressionen sowie für den Modellansatz aus der Aufgabenstellung. 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.*

(c) Es wurde der zusätzliche Regressor `I(female * experience)` in das Modell aufgenommen. Man erhält folgenden Output:

```
Call:
lm(formula = wage ~ education + experience + female + married +
    I(female * experience), data = CPS1985)

Residuals:
   Min     1Q Median     3Q    Max
-10.27  -2.58  -0.66   1.94  37.08

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.1153     1.2166   -4.20  3.1e-05 ***
education                0.9434     0.0785   12.01  < 2e-16 ***
experience               0.1520     0.0234    6.50  1.8e-10 ***
female                  -0.7355     0.6806   -1.08   0.2803
married                  0.3900     0.4226    0.92   0.3564
I(female * experience)  -0.0891     0.0313   -2.85   0.0046 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.42 on 528 degrees of freedom
Multiple R-squared:  0.267,Adjusted R-squared:  0.26
F-statistic: 38.4 on 5 and 528 DF,  p-value: <2e-16
```

(i) Lässt sich die Frage beantworten, wie groß der Effekt des Geschlechts auf den Stundenlohn in Dollar ist? Begründen Sie Ihre Entscheidung und geben Sie den Effekt gegebenenfalls an.

(ii) Berechnen Sie das zweiseitige Prognoseintervall zum Niveau $1 - \alpha = 99\%$ für den Stundenlohn eines 29-jährigen ledigen Mannes mit einer Ausbildungsdauer von 12 Jahren und einer Berufserfahrung von 10 Jahren. Verwenden Sie dazu die geschätzte Varianz-Kovarianzmatrix:

$$\hat{V}(\hat{\beta}) = \begin{pmatrix} 1.48001 & -0.08786 & -0.01443 & -0.20297 & -0.03703 & 0.00869 \\ -0.08786 & 0.00617 & 0.00054 & 0.00096 & -0.00238 & -0.00011 \\ -0.01443 & 0.00054 & 0.00055 & 0.00831 & -0.00284 & -0.00049 \\ -0.20297 & 0.00096 & 0.00831 & 0.46317 & -0.02790 & -0.01757 \\ -0.03703 & -0.00238 & -0.00284 & -0.02790 & 0.17856 & 0.00166 \\ 0.00869 & -0.00011 & -0.00049 & -0.01757 & 0.00166 & 0.00098 \end{pmatrix}$$

(d) Geben Sie zwei mögliche Konstellationen an, unter denen notwendige Annahmen für die Konsistenz und Unverzerrtheit der Koeffizientenschätzer $\hat{\beta}$ verletzt werden.

(e) Erläutern Sie kurz, was man unter perfekter Multikollinearität und imperfekter Multikollinearität versteht. Wie kann man ein Modell auf imperfekte Multikollinearität überprüfen?

*Hinweis: Verwenden Sie die folgende Tabelle mit Quantilen einiger $\chi^2(n)$-Verteilungen:*

| $n\backslash p$ | 0.9 | 0.95 | 0.975 | 0.99 |
|---|---|---|---|---|
| 2 | 4.60517 | 5.99146 | 7.37776 | 9.21034 |
| 3 | 6.25139 | 7.81473 | 9.34840 | 11.34487 |
| 4 | 7.77944 | 9.48773 | 11.14329 | 13.27670 |
| 5 | 9.23636 | 11.07050 | 12.83250 | 15.08627 |
| 6 | 10.64464 | 12.59159 | 14.44938 | 16.81189 |

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

| $n\backslash m$ | 3 | 4 | 5 | 6 | 523 | 524 | 525 | 526 |
|---|---|---|---|---|---|---|---|---|
| 3 | 29.4567 | 28.7099 | 28.2371 | 27.9107 | 26.1473 | 26.1472 | 26.1472 | 26.1471 |
| 4 | 16.6944 | 15.9770 | 15.5219 | 15.2069 | 13.4849 | 13.4849 | 13.4849 | 13.4848 |
| 5 | 12.0600 | 11.3919 | 10.9670 | 10.6723 | 9.0415 | 9.0414 | 9.0414 | 9.0414 |
| 6 | 9.7795 | 9.1483 | 8.7459 | 8.4661 | 6.9006 | 6.9005 | 6.9005 | 6.9004 |
| 523 | 3.8193 | 3.3552 | 3.0523 | 2.8365 | 1.2260 | 1.2258 | 1.2257 | 1.2256 |
| 524 | 3.8192 | 3.3551 | 3.0523 | 2.8365 | 1.2258 | 1.2257 | 1.2256 | 1.2255 |
| 525 | 3.8191 | 3.3550 | 3.0522 | 2.8364 | 1.2257 | 1.2256 | 1.2255 | 1.2254 |
| 526 | 3.8190 | 3.3550 | 3.0521 | 2.8363 | 1.2256 | 1.2255 | 1.2253 | 1.2252 |

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

| $n\backslash p$ | 0.9 | 0.95 | 0.975 | 0.99 | 0.995 |
|---|---|---|---|---|---|
| 525 | 1.28317 | 1.64776 | 1.96449 | 2.33347 | 2.58523 |
| 526 | 1.28316 | 1.64776 | 1.96448 | 2.33346 | 2.58521 |
| 527 | 1.28316 | 1.64775 | 1.96448 | 2.33344 | 2.58519 |
| 528 | 1.28316 | 1.64774 | 1.96447 | 2.33343 | 2.58517 |
| 529 | 1.28315 | 1.64774 | 1.96446 | 2.33342 | 2.58515 |
| 530 | 1.28315 | 1.64773 | 1.96445 | 2.33340 | 2.58514 |