Aufgabe 3 (12 + 4 + 7 + 2 + 2 + 7 + 15 = 49 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 für das Jahr 1985 Informationen bezüglich der Lohnverteilung von Vollzeitangestellten im Alter zwischen 18 und 64 Jahren zur Verfügung stellt. Die abhängige Variable `wage`, welche den durchschnittlichen Stundenlohn der Angestellten in USD angibt, kann durch die Ausbildungsdauer in Jahren (`education`), die Berufserfahrung in Jahren (`experience`), das Alter in Jahren (`age`) und durch die Dummy-Variablen für das Geschlecht (`female` bzw. `male`), die Mitgliedschaft in einer Gewerkschaft (`union`) und den Familienstand (`married`) erklärt werden.

`wage`	durchschnittlicher Stundenlohn
`education`	Ausbildungsdauer in Jahren
`experience`	Berufserfahrung in Jahren
`age`	Alter in Jahren
`female`	Weiblich? (Ja = 1, Nein = 0)
`male`	Männlich? (Ja = 1, Nein = 0)
`union`	Gewerkschaft? (Ja = 1, Nein = 0)
`married`	Verheiratet? (Ja = 1, Nein = 0)

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-9.53754 -3.38272 -1.07309  2.40333 37.80008

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  8.8545448  0.4745835 18.65751 < 2.22e-16
experience   0.0336646        ???  1.84724   0.065269
female      -2.1882846  0.4354507      ??? 6.8831e-07
married      0.8750473  0.4729838  1.85006   0.064862

Residual standard error: 4.99975 on 530 degrees of freedom
Multiple R-squared:  0.0588232,Adjusted R-squared:  ???
F-statistic: 11.0416 on ??? and ??? DF,  p-value: 4.84019e-07
```

(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 Bezeichungen $\hat{\sigma}_{\hat{\beta}_{\text{experience}}}$, t_{female}, $\overline{R^2}$, DF_1, DF_2).

(b) Geben Sie an, welche Regressionskoeffizienten signifikant negativ und welche signifikant positiv sind zum Signifikanzniveau 5%.

(c) Testen Sie zum Signifikanzniveau 5%, ob Verheiratete im Durchschnitt einen um mehr als 0.8 Dollar höheren Stundenlohn erhalten als Unverheiratete.

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.

(d) Verifizieren oder entkräften Sie (in Abhängigkeit gängiger Signifikanzniveaus!) die Behauptung, dass verheiratete Männer bei gleicher Berufserfahrung signifikant mehr verdienen als unverheiratete Männer. (Begründung!)

(e) Um die Unterschiede der durchschnittlichen Stundenlöhne in Bezug auf das Geschlecht darzustellen, wurde die Variable `female` in die Modellschätzung aufgenommen. Begründen Sie, warum die Variable `male` nicht zusätzlich in den Modellansatz aufgenommen werden kann.

Welches Problem würde bezüglich der Modellschätzung entstehen?

(f) Man vermutet, dass weitere Variablen einen Einfluss auf den durchschnittlichen Stundenlohn haben. Aus diesem Grund werden die Dummy-Variable für die Zugehörigkeit in einer Gewerkschaft und die Variable Ausbildungsdauer in die Modellschätzung aufgenommen. Das erweiterte Modell liefert folgenden Output:

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

Residuals:
     Min       1Q   Median       3Q      Max
-9.22488 -2.76892 -0.71335  1.84752 38.00272

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -4.4663900  1.1857636 -3.76668 0.00018404
experience   0.1017546  0.0173678  5.85879 8.2108e-09
female      -2.1471993  0.3908856 -5.49317 6.1470e-08
married      0.4678949  0.4200943  1.11379 0.26587797
education    0.9292222  0.0784974 11.83762 < 2.22e-16
union        1.4359591  0.5101818  2.81460 0.00506604

Residual standard error: 4.42222 on 528 degrees of freedom
Multiple R-squared:  0.266477,Adjusted R-squared:  0.25953
F-statistic: 38.3627 on 5 and 528 DF,  p-value: < 2.22e-16
```

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

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.

(g) Es wird weiterhin vermutet, dass der Effekt eines zusätzlichen Ausbildungsjahres mindestens doppelt so groß ist wie der Effekt der Zugehörigkeit zu einer Gewerkschaft. Können Sie die geäußerte Vermutung bei einem Signifikanzniveau von $\alpha = 1\%$ 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 Matrix $(X'X)^{-1}$:

$$(X'X)^{-1} = \begin{pmatrix} 0.07190 & -0.000510 & -0.0025909 & -0.0025807 & -0.004442 & -0.001300 \\ -0.00051 & 0.000015 & -0.0000340 & -0.0000997 & 0.000025 & -0.000050 \\ -0.00259 & -0.000034 & 0.0078130 & 0.0000087 & -0.000054 & 0.001712 \\ -0.00258 & -0.000100 & 0.0000087 & 0.0090243 & -0.000111 & -0.000678 \\ -0.00444 & 0.0000245 & -0.0000538 & -0.0001105 & 0.000315 & -0.000042 \\ -0.00130 & -0.000050 & 0.0017116 & -0.0006779 & -0.000042 & 0.013310 \end{pmatrix}$$

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

$n\backslash p$	0.9	0.95	0.975	0.99
525	1.28317	1.64776	1.96449	2.33347
526	1.28316	1.64776	1.96448	2.33346
527	1.28316	1.64775	1.96448	2.33344
528	1.28316	1.64774	1.96447	2.33343
529	1.28315	1.64774	1.96446	2.33342
530	1.28315	1.64773	1.96445	2.33340
531	1.28315	1.64773	1.96444	2.33339
532	1.28314	1.64772	1.96443	2.33338
533	1.28314	1.64772	1.96442	2.33336
534	1.28314	1.64771	1.96442	2.33335
535	1.28314	1.64771	1.96441	2.33334

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

$n\backslash m$	2	3	4	526	527	528	529	530
2	99.0000	99.1662	99.2494	99.4973	99.4973	99.4973	99.4973	99.4973
3	30.8165	29.4567	28.7099	26.1471	26.1471	26.1470	26.1470	26.1470
4	18.0000	16.6944	15.9770	13.4848	13.4848	13.4847	13.4847	13.4847
526	4.6457	3.8190	3.3550	1.2252	1.2251	1.2250	1.2249	1.2248
527	4.6456	3.8190	3.3549	1.2251	1.2250	1.2249	1.2248	1.2247
528	4.6456	3.8189	3.3548	1.2250	1.2249	1.2248	1.2247	1.2246
529	4.6455	3.8188	3.3548	1.2249	1.2247	1.2246	1.2245	1.2244
530	4.6454	3.8188	3.3547	1.2247	1.2246	1.2245	1.2244	1.2243

Aufgabe 4 (3 + 12 + 5 + 3 + 6 + 4 = 33 Punkte)

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

Der Verkaufspreis gebrauchter VW Golf-Modelle lässt sich anhand mehrerer Variablen schätzen. Mit Hilfe eines Regressionsmodells soll der Zusammenhang zwischen dem *Verkaufspreis in 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 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

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
```

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

(b) Man vermutet allerdings, dass die bei der Modellschätzung getroffene Annahme homoskedastischer Störgrößen nicht gerechtfertigt ist. Um dies zu überprüfen, wird eine Hilfsregression mit den transformierten quadrierten Residuen w_i und den Regressoren des Modells aus Aufgabenteil (a) durchgeführt, mit folgendem Ergebnis:

```
Call:
lm(formula = w ~ Alter + Kilstand + Tuev + ABS + Schiebedach)

Residuals:
      Min        1Q    Median        3Q       Max
-2.191993 -0.876530 -0.373984  0.405075 10.747980

Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.65922122  0.82757507  3.21327 0.0015767
Alter       -0.00748905  0.00682130 -1.09789 0.2738414
Kilstand    -0.00740490  0.00290458 -2.54939 0.0116967
Tuev         0.00670818  0.01723017  0.38933 0.6975328
ABS         -0.69187875  0.25951046 -2.66609 0.0084326
Schiebedach  0.72927155  0.27831959  2.62027 0.0096011

Residual standard error: 1.5556 on 166 degrees of freedom
Multiple R-squared:  0.119674,Adjusted R-squared:  0.0931579
F-statistic: 4.51329 on 5 and 166 DF,  p-value: 0.000696351
```

(i) Überprüfen Sie die oben genannte Vermutung mit Hilfe des „Original"-Breusch-Pagan-Tests zum Signifikanzniveau $\alpha = 10\%$. Liegt damit Evidenz für Heteroskedastie in den Störgrößen vor?

(ii) Wie unterscheiden sich die jeweiligen Alternativhypothesen bei Breusch-Pagan-Test und Goldfeld-Quandt-Test?

(c) Für die Variablen ergeben sich folgende Varianzinflationsfaktoren:

```
     Alter    Kilstand        Tuev         ABS     Schiebedach
  1.188356    1.191960    1.028808    1.031008        1.032339
```

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

(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?

(d) Als Alternative zu dem obigen linearen Modell wurde auch ein log-lin-Modell geschätzt:

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.67429 -0.13861  0.02287  0.14733  0.62083

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.7400370  0.1203649  22.764  < 2e-16
Alter       -0.0104348  0.0009921 -10.518  < 2e-16
Kilstand    -0.0026750  0.0004224  -6.332 2.18e-09
Tuev        -0.0012804  0.0025060  -0.511    0.610
ABS         -0.0225072  0.0377440  -0.596    0.552
Schiebedach -0.0005102  0.0404796  -0.013    0.990

Residual standard error: 0.2263 on 166 degrees of freedom
Multiple R-squared:  0.5916,Adjusted R-squared:  0.5793
F-statistic: 48.09 on 5 and 166 DF,  p-value: < 2.2e-16
```

(i) Lässt sich die Anpassungsgüte dieses Modells mit jener des Ausgangsmodells in Aufgabenteil (a) sinnvoll vergleichen? (Begründung!)

(ii) Geben Sie ferner eine Interpretation an für den zu `Alter` gehörenden Regressionskoeffizienten.

(e) Unter Zuhilfenahme der Variablen `I(Alter^2)` und `I(Alter * ABS)` wird folgendes polynomiale Modell geschätzt:

```
Call:
lm(formula = Preis ~ Alter + Kilstand + Tuev + I(Alter^2) + I(Alter *
    ABS), data = Golf)

Residuals:
     Min       1Q   Median       3Q      Max
-2.33971 -0.51100 -0.00147  0.48905  3.01492

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.3534149  1.8866341   8.138 9.03e-14
Alter          -0.1569631  0.0357682  -4.388 2.02e-05
Kilstand       -0.0097013  0.0014007  -6.926 9.07e-11
Tuev           -0.0033175  0.0083643  -0.397 0.692157
I(Alter^2)      0.0005484  0.0001632   3.360 0.000968
I(Alter * ABS) -0.0006432  0.0011243  -0.572 0.568036
```

```
Residual standard error: 0.7543 on 166 degrees of freedom
Multiple R-squared:  0.6443,Adjusted R-squared:  0.6336
F-statistic: 60.13 on 5 and 166 DF,  p-value: < 2.2e-16
```

Wie groß ist bei einem 96 Monate alten Auto der marginale Effekt des Alters auf den Verkaufspreis,

(i) falls das Auto über ABS verfügt?

(ii) falls das Auto nicht über ABS verfügt?

(f) Erläutern Sie kurz, was man unter *interner Validität* und *externer Validität* versteht.

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