Aufgabe 3 (6 + 4 + 8 + 4 + 10 + 4 + 9 + 4 + 8 = 57 Punkte)

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

Mit Hilfe eines multiplen linearen Regressionsmodells soll auf Grundlage von Daten aus dem Jahr 1975 der Stundenlohn `wage` erwerbstätiger Frauen (in [USD]) mit Hilfe verschiedener Regressoren erklärt werden.
Man vermutet zunächst, dass das Alter `age`, die bisherige Berufserfahrung `expe`, die Ausbildungsdauer `educ` (jeweils in Jahren) sowie die Dummy-Variable `coll` (mit Wert 1, falls ein College-Abschluss vorhanden ist, 0 sonst) einen Einfluss auf den Stundenlohn haben. Daher wird das Modell

$$\texttt{wage}_i = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{expe}_i + \beta_3 \texttt{educ}_i + \beta_4 \texttt{coll}_i + u_i$$

($i = 1, \ldots, n$) mit Hilfe der KQ-Methode unter Annahme homoskedastischer Störgrößen mit folgendem Ergebnis geschätzt:

```
Call:
lm(formula = wage ~ age + expe + educ + coll)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3445 -1.0965 -0.2055  0.8228  6.5911

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.16678    0.94493   0.176   0.8600
age         -0.02373    0.01409     ???   0.0930 .
expe         0.05998    0.01325   4.527 8.43e-06 ***
educ         0.31972    0.06469   4.942 1.25e-06 ***
coll         0.54958        ???   1.719   0.0866 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.617 on 321 degrees of freedom
Multiple R-squared:  0.2904,Adjusted R-squared:  ???
F-statistic: 32.84 on 4 and 321 DF,  p-value: < 2.2e-16
```

Gehen Sie zunächst von der korrekten Spezifikation des Modells aus.

(a) Bestimmen Sie die drei fehlenden Werte (im Output mit `???` gekennzeichnet) und notieren Sie diese (mit den Bezeichnungen $t_{\texttt{age}}$, $\widehat{\sigma}_{\texttt{coll}}$ und $\overline{R^2}$) in Ihr Klausurheft.

(b) Hat das Alter der Erwerbstätigen einen signifikant negativen Einfluss auf den Stundenlohn (Signifikanzniveau $\alpha = 0.05$)? Ist der Stundenlohn signifikant höher, wenn die Erwerbstätige über einen College-Abschluss verfügt, verglichen damit, dass die Erwerbstätige keinen College-Abschluss besitzt (Signifikanzniveau $\alpha = 0.05$)? Begründen Sie jeweils Ihre Antwort.

(c) Testen Sie zum Signifikanzniveau $\alpha = 0.05$, ob die erwartete Zunahme des Stundenlohns pro zusätzlichem Ausbildungsjahr mehr als 0.20 USD beträgt.

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) Als Output eines Breusch-Pagan-Tests (nach Koenker) erhalten Sie das folgende Resultat:

```
studentized Breusch-Pagan test

data:  lm(wage ~ age + expe + educ + coll)
BP = 18.1236, df = 4, p-value = 0.001167
```

Was wird mit diesem Test untersucht und wie lautet das Ergebnis der Untersuchung, wenn man ein Signifikanzniveau von $\alpha = 0.01$ zu Grunde legt?

(e) Eine heteroskedastie-konsistente Schätzung der Varianz-Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}$ mit $\widehat{\mathrm{V}}_{\mathrm{hc1}}(\widehat{\boldsymbol{\beta}})$ liefert:

$$\widehat{\mathrm{V}}_{\mathrm{hc1}}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} 0.79843185 & -0.00639849 & 0.00030246 & -0.04675981 & 0.15871063 \\ -0.00639849 & 0.00019391 & -0.00007577 & -0.00005583 & -0.00000672 \\ 0.00030246 & -0.00007577 & 0.00016617 & 0.00004975 & -0.00002209 \\ -0.04675981 & -0.00005583 & 0.00004975 & 0.00428832 & -0.01477411 \\ 0.15871063 & -0.00000672 & -0.00002209 & -0.01477411 & 0.09984689 \end{pmatrix}$$

Berechnen Sie auf dieser Grundlage ein Konfidenzintervall für $\beta_1 + \beta_2$, also für den „Netto"-Effekt eines zusätzlichen Jahres an Berufserfahrung bei Berücksichtigung der damit unausweichlich einhergehenden Erhöhung des Lebensalters, zum Konfidenzniveau $1 - \alpha = 0.90$.

(f) Man vermutet nun, dass auch die Frage, ob die Erwerbstätige in einer großen Stadt lebt (Dummy-Variable `city` mit Wert 1, falls der Lebensmittelpunkt in einer großen Stadt liegt, 0 sonst), einen Einfluss auf den Stundenlohn hat und schätzt daher (unter Annahme homoskedastischer Störgrößen) das Modell

$$\texttt{wage}_i = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{expe}_i + \beta_3 \texttt{educ}_i + \beta_4 \texttt{coll}_i + \beta_5 \texttt{city}_i + u_i$$

$(i = 1, \ldots, n)$ mit dem Ergebnis:

```
Call:
lm(formula = wage ~ age + expe + educ + coll + city)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8315 -1.0369 -0.2255  0.8383  6.2545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19187    0.91824   0.209   0.8346
age         -0.03334    0.01386  -2.406   0.0167 *
expe         0.06483    0.01292   5.018 8.69e-07 ***
educ         0.30410    0.06296   4.830 2.12e-06 ***
coll         0.49706    0.31092   1.599   0.1109
city         0.82542    0.18481   4.466 1.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.571 on 320 degrees of freedom
Multiple R-squared:  0.332,Adjusted R-squared:  0.3216
F-statistic: 31.81 on 5 and 320 DF,  p-value: < 2.2e-16
```

Können Sie die oben geäußerte Vermutung bei einem Signifikanzniveau von $\alpha = 0.01$ bestätigen?

Würden Sie auf Grundlage der folgenden Varianzinflationsfaktoren von einem Multikollinearitätsproblem bei der obigen Modellschätzung ausgehen?

```
     age     expe     educ     coll     city
1.435380 1.399360 2.739475 2.730525 1.045770
```

(g) In Erweiterung des Modells aus Teil (f) vermutet man nun, dass die Frage, ob die Erwerbstätige Mutter ist (Dummy-Variable `kids` mit Wert 1, falls die Erwerbstätige Mutter ist, 0 sonst), sowohl einen unmittelbaren Effekt auf den Stundenlohn hat als auch über eine Interaktion mit dem Regressor `educ` mittelbar auf den Stundenlohn wirkt.

Man erhält das folgende (unwesentlich gekürzte) Ergebnis der Schätzung des Modells

$$\begin{aligned}\texttt{wage}_i = \beta_0 + \beta_1\texttt{age}_i + \beta_2\texttt{expe}_i + \beta_3\texttt{educ}_i + \beta_4\texttt{coll}_i + \beta_5\texttt{city}_i \\ + \beta_6\texttt{kids}_i + \beta_7\texttt{kids}_i\texttt{educ}_i + u_i\end{aligned}$$

($i = 1, \ldots, n$) mit der KQ-Methode unter Annahme homoskedastischer Störgrößen:

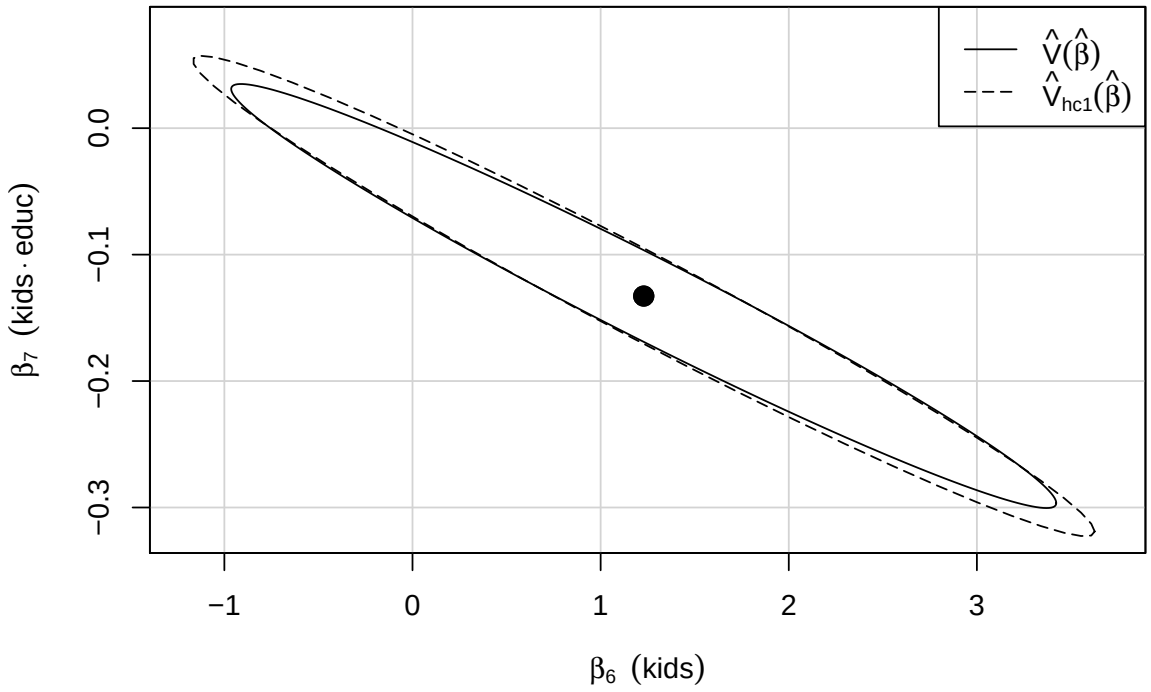
```
Call:
lm(formula = wage ~ age + expe + educ + coll + city + kids +
    I(kids * educ))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.001344   1.145471   0.001  0.99906
age           -0.041508   0.014671  -2.829  0.00496 **
expe           0.058863   0.013215   4.454 1.17e-05 ***
educ           0.376379   0.074182   5.074 6.65e-07 ***
coll           0.528048   0.311144   1.697  0.09065 .
city           0.821151   0.183536   4.474 1.07e-05 ***
kids           1.229048   1.018135   1.207  0.22827
I(kids * educ) -0.132752   0.077860  -1.705  0.08917 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.558 on 318 degrees of freedom
Multiple R-squared:  0.3472,Adjusted R-squared:  0.3328
F-statistic: 24.16 on 7 and 318 DF,  p-value: < 2.2e-16
```

Untersuchen Sie mit einem geeigneten Test, ob wenigstens einer der zusätzlichen Regressoren `kids` bzw. `kids` $\cdot$ `educ` einen signifikanten ($\alpha = 0.10$) Einfluss auf den Stundenlohn hat.

Geben Sie hierzu insbesondere 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.

(h) Können Sie die in Teil (g) zu treffende Entscheidung mit den bisher zur Verfügung stehenden Angaben auch noch fällen, wenn von Heteroskedastizität in den Störgrößen auszugehen ist?

Wie fällt Ihre Entscheidung (unter Zulassung und Berücksichtigung heteroskedastischer Störgrößen) auf der Grundlage der folgenden Konfidenzellipsen zum Konfidenzniveau $1 - \alpha = 0.90$ aus?



(i) Geben Sie auf Basis des in Teil (g) formulierten und geschätzten Modells Punktprognosen für den Stundenlohn (in [USD]) an, wenn

(i) die Erwerbstätige eine 38-jährige Mutter ohne College-Abschluss mit 10 Jahren Berufserfahrung und 12 Ausbildungsjahren ist, die in einer großen Stadt lebt,

(ii) die Erwerbstätige eine 50-jährige Frau ohne Kinder mit College-Abschluss ist, die 18 Jahre Berufserfahrung und 15 Ausbildungsjahre absolviert hat und *nicht* in einer großen Stadt lebt.

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

$n\backslash p$	0.85	0.90	0.95	0.975	0.99	0.995	0.9995
318	1.038	1.284	1.650	1.967	2.338	2.591	3.321
319	1.038	1.284	1.650	1.967	2.338	2.591	3.321
320	1.038	1.284	1.650	1.967	2.338	2.591	3.321
321	1.038	1.284	1.650	1.967	2.338	2.591	3.321
322	1.038	1.284	1.650	1.967	2.338	2.591	3.321

sowie die folgende Tabelle mit 0.90-Quantilen einiger $F(m, n)$-Verteilungen:

$n\backslash m$	2	3	318	319	320
2	9.000	9.162	9.488	9.488	9.488
3	5.462	5.391	5.137	5.137	5.137
318	2.319	2.101	1.155	1.155	1.155
319	2.319	2.101	1.155	1.155	1.154
320	2.319	2.101	1.154	1.154	1.154

Aufgabe 4 (9 + 6 + 10 = 25 Punkte)

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

Im Rahmen einer Zeitreihenanalyse wird zunächst ein multiples lineares Regressionsmodell der Gestalt

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad i = 1, \ldots, 50,$$

mit stochastisch unabhängig gemeinsam normalverteilten Störgrößen zu Grunde gelegt.

(a) Man hält einen Strukturbruch zwischen Beobachtung 25 und Beobachtung 26 für möglich. Bei einer KQ-Schätzung des Modells für den Gesamtzeitraum (unter Annahme homoskedastischer Störgrößen) erhält man den folgenden (gekürzten) Output:

```
Residual standard error: 11.87 on 47 degrees of freedom
Multiple R-squared:  0.94915,Adjusted R-squared:  0.94591
F-statistic: 292.46 on 3 and 47 DF,  p-value: < 2.22e-16
```

Als (gekürzter) Output einer entsprechenden Schätzung für den ersten Teilzeitraum ergibt sich:

```
Residual standard error: 5.7131 on 22 degrees of freedom
Multiple R-squared:  0.98707,Adjusted R-squared:  0.98531
F-statistic: 559.81 on 3 and 22 DF,  p-value: < 2.22e-16
```

Eine entsprechende Schätzung für den zweiten Teilzeitraum liefert:

```
Residual standard error: 14.85 on 22 degrees of freedom
Multiple R-squared:  0.93507,Adjusted R-squared:  0.92621
F-statistic:  105.6 on 3 and 22 DF,  p-value: 3.2461e-13
```

Prüfen Sie mit Hilfe dieser Ergebnisse zum Signifikanzniveau $\alpha = 0.05$, ob ein Strukturbruch in den Regressionskoeffizienten (in beliebiger Form) vorliegt.

Geben Sie hierzu nur den kritischen Bereich, die realisierte Teststatistik sowie die Antwort auf die Frage, ob von einem Strukturbruch auszugehen ist oder nicht, an.

(b) Man vermutet nun entgegen der Annahme in Teil (a), dass Heteroskedastie in der Art vorliegt, dass die Varianz der Störgrößen zwar innerhalb der beiden Teilzeiträume (Beobachtung 1 bis 25 bzw. Beobachtung 26 bis 50) jeweils konstant ist, zwischen den beiden Teilzeiträumen jedoch ein Unterschied in der Störgrößenvarianz besteht.

Welcher statistische Test ist zur Überprüfung dieser Vermutung geeignet? Können Sie den Test mit Hilfe der Regressionsoutputs aus Teil (a) durchführen? Falls ja, mit welchem Ergebnis (Signifikanzniveau $\alpha = 0.10$)?

Geben Sie zur Durchführung des Tests gegebenenfalls nur den kritischen Bereich, die realisierte Teststatistik sowie die Antwort auf die Frage, ob von unterschiedlichen Störgrößenvarianzen auszugehen ist oder nicht, an.

(c) Man stellt nun in der üblichen Art und Weise ein vollständiges Strukturbruchmodell mit dem Parametervektor $\boldsymbol{\beta} = (\beta_0^{(1)}, \delta_0, \beta_1^{(1)}, \delta_1, \beta_2^{(1)}, \delta_2)'$ auf und schätzt dieses Modell mit der KQ-Methode.

Im Gegensatz zu Teil (a) nimmt man nun Heteroskedastie in den Störgrößen an und berechnet zur Durchführung eines zum Test aus Teil (a) analogen Strukturbruchtests unter Verwendung der heteroskedastie-konsistenten Schätzung $\widehat{\mathrm{V}}_{\mathrm{hc1}}(\widehat{\boldsymbol{\beta}})$ der Varianz-Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}$ zunächst

$$\left(\boldsymbol{A}\widehat{\mathrm{V}}_{\mathrm{hc1}}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}'\right)^{-1} = \begin{pmatrix} 0.103 & 0.514 & 0.556 \\ 0.514 & 2.934 & 2.761 \\ 0.556 & 2.761 & 3.179 \end{pmatrix}$$

mit

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

Bekannt ist außerdem der realisierte Parameterschätzer

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0^{(1)} \\ \widehat{\delta}_0 \\ \widehat{\beta}_1^{(1)} \\ \widehat{\delta}_1 \\ \widehat{\beta}_2^{(1)} \\ \widehat{\delta}_2 \end{pmatrix} = \begin{pmatrix} 3.518 \\ -7.567 \\ 4.032 \\ 2.057 \\ 3.821 \\ 1.057 \end{pmatrix} .$$

Nutzen Sie diese (Zwischen-)Ergebnisse, um einen in dieser Situation geeigneten Test auf einen Strukturbruch in den Regressionskoeffizienten (in beliebiger Form) durchzuführen (Signifikanzniveau $\alpha = 0.05$).

Geben Sie hierzu nur den kritischen Bereich, die realisierte Teststatistik sowie die Antwort auf die Frage, ob von einem Strukturbruch auszugehen ist oder nicht, an.

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

$n\backslash m$	3	6	22	44	47
3	9.277	8.941	8.648	8.588	8.584
6	4.757	4.284	3.856	3.765	3.759
22	3.049	2.549	2.048	1.925	1.917
44	2.816	2.313	1.789	1.651	1.641
47	2.802	2.299	1.773	1.634	1.624

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

$n\backslash m$	3	6	22	44	47
3	0.108	0.210	0.328	0.355	0.357
6	0.112	0.233	0.392	0.432	0.435
22	0.116	0.259	0.488	0.559	0.564
44	0.116	0.266	0.519	0.606	0.612
47	0.116	0.266	0.522	0.609	0.616