Aufgabe 3 (14 + 3 + 7 + 7 + 16 + 5 + 7 = 59 Punkte)

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

Der Kraftstoffverbrauch von Kraftfahrzeugen lässt sich anhand mehrerer Variablen schätzen. Mit Hilfe eines Regressionsmodells soll nun der Zusammenhang zwischen dem Kraftstoffverbrauch in l/100 km und der Anzahl der Zylinder, der PS-Anzahl, dem Gewicht des Fahrzeugs (in Tonnen), der Art des Getriebes (0 = Automatik, 1 = Manuell) und der Anzahl der Vorwärtsgänge betrachtet werden. Die Daten wurden aus dem Motor Trend US-Magazin aus dem Jahr 1974 entnommen. Der hier verwendete Datensatz stammt aus dem Paket `datasets` und enthält folgende Variablen:

`Verbrauch`	Kraftstoffverbrauch (in Liter pro 100 Kilometer)
`Zylinder`	Anzahl der Zylinder
`PS`	Anzahl der PS
`Gewicht`	Gewicht (in Tonnen)
`Getriebe`	Getriebe (0 = Automatik, 1 = Manuell)
`Gaenge`	Anzahl der Vorwärtsgänge

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

```
Call:
lm(formula = Verbrauch ~ Zylinder + PS + Gewicht)

Residuals:
      Min        1Q    Median        3Q       Max
-3.776301 -1.128622  0.264741  0.922434  3.453129

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 1.1757450  1.1249006 1.04520   0.304871
Zylinder          ???  0.3468233 0.54729   0.588513
PS          0.0148237        ??? 1.98269   0.057291
Gewicht     5.6398855  1.0278431     ??? 7.3371e-06

Residual standard error: 1.58112 on 28 degrees of freedom
Multiple R-squared:  0.848707,Adjusted R-squared:  ???
F-statistic: 52.3571 on ??? and ??? DF,  p-value: 1.32216e-11
```

(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 $\widehat{\beta}_{\texttt{Zylinder}}$, $\widehat{\sigma}_{\texttt{PS}}$, $\texttt{t}_{\texttt{Gewicht}}$, $\overline{\texttt{R}^2}$, $\texttt{DF}_1$, $\texttt{DF}_2$)

(iv) Interpretieren Sie das Schätzergebnis für die Variable `Gewicht`.

(b) Geben Sie an, welche Regressionskoeffizienten (β_1,β_2,β_3) signifikant von Null verschieden sind zum Signifikanzniveau 10%.

(c) Verwenden Sie einen geeigneten Test, um eine Entscheidung darüber zu treffen, ob eine Erhöhung des Gewichts um eine Tonne den Kraftstoffverbrauch um signifikant mehr als 4 Liter pro 100 km erhöht, wenn man ein Signifikanzniveau von $\alpha = 10\%$ zu Grunde legt.

(d) Es wird vermutet, dass weitere Kriterien einen Einfluss auf den Kraftstoffverbrauch haben. Aus diesem Grund werden eine Variable für die Anzahl der Vorwärtsgänge (`Gaenge`) und eine Dummy-Variable für die Art des Getriebes (`Getriebe`) in den Modellansatz aufgenommen.

Das erweiterte Modell liefert folgenden Output:

```
Call:
lm(formula = Verbrauch ~ Zylinder + PS + Gewicht + Getriebe +
    Gaenge)

Residuals:
      Min        1Q    Median        3Q       Max
-4.022573 -0.956270  0.297081  0.729350  3.277090

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  3.3691543  3.7861546  0.88986 0.38170024
Zylinder     0.0725174  0.4190376  0.17306 0.86394681
PS           0.0173609  0.0100048  1.73526 0.09453499 .
Gewicht      5.7977305  1.3137341  4.41317 0.00015816 ***
Getriebe     0.8798927  1.0927092  0.80524 0.42798439
Gaenge      -0.6583360  0.7456817 -0.88286 0.38539978
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.61201 on 26 degrees of freedom
Multiple R-squared:  0.85397,Adjusted R-squared:  0.825887
F-statistic: 30.4091 on 5 and 26 DF,  p-value: 4.46444e-10
```

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) Es wird vermutet, dass der Effekt des Gewichts auf den Kraftstoffverbrauch sechsmal so groß ist wie der Effekt der Getriebeart. Können Sie die geäußerte Vermutung bei einem Signifikanzniveau von $\alpha = 10\%$ bestätigen?

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 Matrix $(X'X)^{-1}$*:*

$$\begin{pmatrix} 5.51645 & -0.38824 & 0.0098479 & -0.80723 & 0.137307 & -0.92347 \\ -0.38824 & 0.06757 & -0.0012560 & -0.03117 & 0.001815 & 0.05400 \\ 0.00985 & -0.00126 & 0.0000385 & -0.00167 & -0.000466 & -0.00138 \\ -0.80723 & -0.03117 & -0.0016693 & 0.66417 & 0.237981 & 0.04856 \\ 0.13731 & 0.00182 & -0.0004661 & 0.23798 & 0.459486 & -0.16654 \\ -0.92347 & 0.05400 & -0.0013833 & 0.04856 & -0.166540 & 0.21398 \end{pmatrix}$$

(f) Nimmt man die Interaktionsvariablen `I(Gewicht*Gewicht)` und `I(Gewicht*Getriebe)` in das Regressionsmodell auf, ergibt sich folgender Output:

```
Call:
lm(formula = Verbrauch ~ Zylinder + PS + Gewicht + Getriebe +
    Gaenge + I(Gewicht * Gewicht) + I(Gewicht * Getriebe))

Residuals:
      Min        1Q    Median        3Q       Max
-4.412257 -1.077968  0.364544  0.699686  2.923790

Coefficients:
                         Estimate Std. Error  t value Pr(>|t|)
(Intercept)           10.7885201 10.0950519  1.06869  0.29584
Zylinder               0.3529601  0.5196791  0.67919  0.50352
PS                     0.0139062  0.0117613  1.18236  0.24864
Gewicht               -4.6743214 11.8112967 -0.39575  0.69579
Getriebe              -3.6575716  7.3070233 -0.50056  0.62124
Gaenge                -0.4523908  0.7844590 -0.57669  0.56952
I(Gewicht * Gewicht)   2.7717951  3.0537365  0.90767  0.37308
I(Gewicht * Getriebe)  2.8954307  5.2497537  0.55154  0.58637

Residual standard error: 1.63767 on 24 degrees of freedom
Multiple R-squared:  0.860878,Adjusted R-squared:  0.820301
F-statistic: 21.2158 on 7 and 24 DF,  p-value: 7.84411e-09
```

Wie groß ist bei einem 1.5 Tonnen schweren Auto der marginale Effekt des Gewichts auf den Verbrauch,

(i) falls das Auto ein Automatik-Getriebe besitzt?

(ii) falls das Auto ein manuelles Getriebe besitzt?

(g) (i) Welches Problem kann mit Hilfe der Varianzinflationsfaktoren erkannt werden? Geben Sie die zugehörige Faustregel an.

(ii) Welche Folgen ergeben sich hinsichtlich der Schätzung bei Vorliegen eines großen Varianz-Inflations-Faktors?

(iii) Würden Sie bzgl. des Modells aus Aufgabenteil (f) niedrige oder hohe Varianzinflationsfaktoren vermuten? Begründen Sie Ihre Antwort.

(iv) Wie verändern sich die Varianz-Inflations-Faktoren, wenn Regressoren aus dem Modell entfernt werden?

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

$n\backslash p$	0.9	0.95	0.975	0.99
25	1.31635	1.70814	2.05954	2.48511
26	1.31497	1.70562	2.05553	2.47863
27	1.31370	1.70329	2.05183	2.47266
28	1.31253	1.70113	2.04841	2.46714
29	1.31143	1.69913	2.04523	2.46202
30	1.31042	1.69726	2.04227	2.45726

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

$n\backslash m$	2	3	4	25	26	27	28
2	19.0000	19.1643	19.2468	19.4558	19.4573	19.4587	19.4600
3	9.5521	9.2766	9.1172	8.6341	8.6301	8.6263	8.6229
4	6.9443	6.5914	6.3882	5.7687	5.7635	5.7586	5.7541
25	3.3852	2.9912	2.7587	1.9554	1.9472	1.9395	1.9323
26	3.3690	2.9752	2.7426	1.9375	1.9292	1.9215	1.9142
27	3.3541	2.9604	2.7278	1.9210	1.9126	1.9048	1.8975
28	3.3404	2.9467	2.7141	1.9057	1.8973	1.8894	1.8821

Aufgabe 4 (2 + 4.5 + 12.5 + 4 = 23 Punkte)

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

Der Datensatz `BudgetUK` aus dem `R`-Paket `Ecdat` beinhaltet Angaben hinsichtlich der Ausgaben und Einnahmen sowie der Anzahl der Kinder und des Alters des Familienoberhaupts von Haushalten aus dem Vereinigten Königreich aus den Jahren 1980 bis 1982. Die Ausgaben eines Haushaltes werden in Pfund (£) pro Woche angegeben und lassen sich hier in Ausgaben für Nahrungsmittel, Kraftstoff, Kleidung und Alkohol unterteilen.

`Essen`	Ausgaben für Nahrungsmittel
`Tanken`	Ausgaben für das Tanken
`Kleidung`	Ausgaben für die Kleidung
`Alkohol`	Ausgaben für Alkohol
`Gesamtausgaben`	Gesamtausgaben pro Haushalt
`Einkommen`	Einkommen pro Haushalt (in Pfund pro Woche)
`Alter`	Alter des Familienoberhaupts (in Jahren)
`Kinder`	Anzahl der Kinder

(a) Es werden zunächst drei Varianten geschätzt, jeweils für einen linearen, quadratischen sowie kubischen Zusammenhang zwischen den Ausgaben für Nahrungsmittel und dem Einkommen pro Haushalt. Die entsprechenden Ergebnisse lauten

(i) für das lineare Modell:

```
Call:
lm(formula = Essen ~ Einkommen)

Residuals:
      Min        1Q    Median        3Q       Max
-23.65950  -7.06645  -1.18645   5.38093  51.25463

Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 21.8630226  1.9405928 11.26616 < 2.22e-16 ***
Einkommen    0.0789613  0.0133875  5.89812 1.1983e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.1096 on 248 degrees of freedom
Multiple R-squared:  0.123017,Adjusted R-squared:  0.119481
F-statistic: 34.7878 on 1 and 248 DF,  p-value: 1.19832e-08
```

(ii) für das quadratische Modell:

```
Call:
lm(formula = Essen ~ Einkommen + I(Einkommen^2))

Residuals:
      Min        1Q    Median        3Q       Max
-23.76415  -7.01805  -1.15767   5.46244  51.18509

Coefficients:
                    Estimate    Std. Error  t value    Pr(>|t|)
(Intercept)    20.5473204372  4.6890535800  4.38198 0.000017399 ***
Einkommen       0.0966099569  0.0587878640  1.64337     0.10158
I(Einkommen^2) -0.0000508752  0.0001649963 -0.30834     0.75808
```

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

Residual standard error: 11.1299 on 247 degrees of freedom
Multiple R-squared:  0.123355,Adjusted R-squared:  0.116256
F-statistic:  17.378 on 2 and 247 DF,  p-value: 8.68467e-08
```

(iii) für das kubische Modell:

```
Call:
lm(formula = Essen ~ Einkommen + I(Einkommen^2) + I(Einkommen^3))

Residuals:
      Min        1Q    Median        3Q       Max
-23.83947  -7.01675  -1.18366   5.48941  51.21568

Coefficients:
                   Estimate   Std. Error  t value Pr(>|t|)
(Intercept)     2.11686e+01  1.00846e+01  2.09911 0.036827 *
Einkommen       8.41422e-02  1.88504e-01  0.44637 0.655724
I(Einkommen^2)  2.36360e-05  1.08284e-03  0.02183 0.982603
I(Einkommen^3) -1.32365e-07  1.90105e-06 -0.06963 0.944547
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.1524 on 246 degrees of freedom
Multiple R-squared:  0.123372,Adjusted R-squared:  0.112681
F-statistic: 11.5402 on 3 and 246 DF,  p-value: 4.19348e-07
```

Anhand welcher Kennzahl (aus den angegebenen Outputs) lässt sich die Anpassungsgüte von Modellen miteinander vergleichen? Welches der drei Modelle würden Sie anhand dieser Kennzahl auswählen? Begründen Sie Ihre Antwort.

(b) Als Alternative zu den polynomialen Modellen (siehe Aufgabenteil (a)) wurden weitere Modelle geschätzt. Die entsprechenden Ergebnisse lauten

(i) für Modell 1:

```
Call:
lm(formula = log(Essen) ~ log(Einkommen))

Residuals:
       Min         1Q     Median         3Q        Max
-1.2836990 -0.1928031  0.0171774  0.1980736  1.0067675

Coefficients:
                Estimate Std. Error t value   Pr(>|t|)
(Intercept)    1.9136833  0.2789916 6.85929 5.4854e-11 ***
log(Einkommen) 0.3119534  0.0574777 5.42738 1.3601e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.32122 on 248 degrees of freedom
Multiple R-squared:  0.106166,Adjusted R-squared:  0.102562
F-statistic: 29.4565 on 1 and 248 DF,  p-value: 1.36008e-07
```

(ii) für Modell 2:

```
Call:
lm(formula = Essen ~ log(Einkommen))

Residuals:
      Min        1Q    Median        3Q       Max
-23.55645  -6.99380  -0.91946   4.94749  50.67655

Coefficients:
                Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    -23.15762    9.67438 -2.39371  0.017423 *
log(Einkommen)  11.50374    1.99311  5.77175 2.333e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.1387 on 248 degrees of freedom
Multiple R-squared:  0.11842,Adjusted R-squared:  0.114865
F-statistic: 33.3131 on 1 and 248 DF,  p-value: 2.333e-08
```

(iii) für Modell 3:

```
Call:
lm(formula = log(Essen) ~ Einkommen)

Residuals:
       Min         1Q     Median         3Q        Max
-1.2684915 -0.1970757  0.0021335  0.2108026  1.0219750

Coefficients:
               Estimate  Std. Error t value   Pr(>|t|)
(Intercept) 3.138719239 0.056079766 55.9688 < 2.22e-16 ***
Einkommen   0.002110256 0.000386877  5.4546 1.1866e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.321048 on 248 degrees of freedom
Multiple R-squared:  0.107119,Adjusted R-squared:  0.103519
F-statistic: 29.7526 on 1 and 248 DF,  p-value: 1.18658e-07
```

Geben Sie anhand der Outputs an, um welches Modell bzw. um welche Spezifikation es sich jeweils handelt. Welche Modelle aus den Aufgabenteilen (a) und (b) lassen sich miteinander vergleichen? Welche(s) Modell(e) würden Sie hinsichtlich der Anpassungsgüte auswählen?

(c) Neben der Variablen `Einkommen` werden nun zusätzlich die Variablen `Alter` und `Kinder` in das Regressionsmodell aufgenommen. Die entsprechende Regression liefert folgenden Output:

```
Call:
lm(formula = Essen ~ Einkommen + Alter + Kinder)

Residuals:
      Min        1Q    Median        3Q       Max
-27.74377  -6.76856  -1.32942   4.56126  52.09754
```

```
Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 10.0922569  3.9944801 2.52655  0.0121471 *
Einkommen    0.0710282  0.0133675 5.31351 2.4089e-07 ***
Alter        0.2393811  0.0868019 2.75779  0.0062561 **
Kinder       2.7833354  1.4032866 1.98344  0.0484301 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.9079 on 246 degrees of freedom
Multiple R-squared:  0.161389,Adjusted R-squared:  0.151162
F-statistic: 15.7807 on 3 and 246 DF,  p-value: 2.03925e-09
```

Berechnen Sie anhand dieses Modells die zweiseitigen Prognoseintervalle zum Niveau $1 - \alpha = 0.9$ für die Ausgaben für Nahrungsmittel sowie für deren Erwartungswert, falls das Einkommen bei 135£ pro Woche, das Alter des Familienoberhaupts bei 35 Jahren und die Anzahl der Kinder bei 1 liegt.

Hinweis: Verwenden Sie dazu die geschätzte Varianz-Kovarianzmatrix $\hat{V}(\hat{\beta})$:

$$\begin{pmatrix} 15.9559 & -0.0151 & -0.2440 & -3.0552 \\ -0.0151 & 0.0002 & -0.0002 & -0.0013 \\ -0.2440 & -0.0002 & 0.0075 & 0.0032 \\ -3.0552 & -0.0013 & 0.0032 & 1.9692 \end{pmatrix}$$

(d) Geben Sie zwei mögliche Konstellationen an, unter denen notwendige Annahmen für die Konsistenz und Unverzerrtheit der KQ-Koeffizientenschätzer verletzt werden.

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

$n \backslash p$	0.9	0.95	0.975	0.99
245	1.28502	1.65110	1.96969	2.34166
246	1.28500	1.65107	1.96965	2.34160
247	1.28499	1.65105	1.96961	2.34154
248	1.28497	1.65102	1.96958	2.34148
249	1.28496	1.65100	1.96954	2.34142
250	1.28495	1.65097	1.96950	2.34136