Aufgabe 3 (14 + 2 + 7 + 7 + 3 + 5 + 9 + 11 = 58 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 *Grundstücksgröße in square feet*, der *Anzahl an Schlafzimmern*, der *Anzahl an Badezimmern*, der *Anzahl der Garagen* sowie den Dummy-Variablen *Einfahrt vorhanden ja/nein* und *Klimaanlage vorhanden ja/nein* betrachtet.

price	Verkaufspreis in Dollar
lotsize	Grundstücksgröße in square feet
bedrooms	Anzahl der Schlafzimmer
bathrooms	Anzahl der Badezimmer
driveway	Einfahrt vorhanden? (Ja = 1, Nein = 0)
aircon	Klimaanlage vorhanden? (Ja = 1, Nein = 0)

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

Call:

```
lm(formula = price ~ lotsize + bedrooms + bathrooms,
    data = HousePrices)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-55955.40	-12467.40	-1145.39	10201.79	89459.56

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2418.29262	3779.41242	-0.63986	0.52253
lotsize	???	0.38797	13.94749	< 2.22e-16
bedrooms	5826.80241	1206.57078	4.82923	1.7859e-06
bathrooms	19750.21013	???	11.06403	< 2.22e-16

Residual standard error: 19191.6 on 542 degrees of freedom

Multiple R-squared: 0.486292, Adjusted R-squared: ???

F-statistic: 171.025 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 $\hat{\beta}_{lotsize}$, $\hat{\sigma}_{bathrooms}$, $\overline{R^2}$, DF_1 und DF_2).
- (iv) Wofür stehen die Abkürzungen 1Q und 3Q im Summary-Output der Residuen? Geben Sie den Mittelwert der Residuen an.

- (b) Geben Sie an, welche Regressionskoeffizienten signifikant von Null verschieden sind zum Signifikanzniveau 5%.
- (c) Testen Sie zum Signifikanzniveau 1%, ob ein zusätzliches Schlafzimmer den Verkaufspreis um signifikant mehr als 4000 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.

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

Call:

```
lm(formula = price ~ lotsize + bedrooms + bathrooms + aircon +
    driveway, data = HousePrices)
```

Residuals:

Min	1Q	Median	3Q	Max
-42670	-11246	-1106	9104	81539

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7677.399	3765.915	-2.04	0.042
lotsize	4.210	0.371	11.34	< 2e-16
bedrooms	5279.415	1095.556	4.82	1.9e-06
bathrooms	18003.082	1621.648	11.10	< 2e-16
aircon	15762.027	1661.429	9.49	< 2e-16
driveway	12018.193	2232.710	5.38	1.1e-07

Residual standard error: 17300 on 540 degrees of freedom

Multiple R-squared: 0.583, Adjusted R-squared: 0.579

F-statistic: 151 on 5 and 540 DF, p-value: <2e-16

Testen Sie unter Zuhilfenahme geeigneter Bestimmtheitsmaße zum Signifikanzniveau von 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.

- (e) Unter Zuhilfenahme der Interaktionsvariablen `lotsize * driveway` wird ein weiteres Modell geschätzt, welches folgenden Output ergibt:

Call:

```
lm(formula = price ~ lotsize + bedrooms + bathrooms + aircon +
    driveway + I(lotsize * driveway), data = HousePrices)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-43727 -11157  -1068   9071  81282
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2539.48	6999.85	0.36	0.717
lotsize	1.38	1.68	0.82	0.412
bedrooms	5327.12	1093.89	4.87	1.5e-06
bathrooms	17882.66	1620.16	11.04	< 2e-16
aircon	15788.42	1658.44	9.52	< 2e-16
driveway	1069.69	6708.74	0.16	0.873
I(lotsize * driveway)	2.97	1.72	1.73	0.084

Residual standard error: 17300 on 539 degrees of freedom

Multiple R-squared: 0.585, Adjusted R-squared: 0.58

F-statistic: 127 on 6 and 539 DF, p-value: <2e-16

Wie groß ist der marginale Effekt der Grundstücksgröße auf den Verkaufspreis,

- falls eine Einfahrt vorhanden ist?
- falls keine Einfahrt vorhanden ist?

- (f) Geben Sie auf Basis des in Teil (e) formulierten und geschätzten Modells eine Punktprognose für den erwarteten Verkaufspreis an, wenn das Grundstück 1000 square feet groß ist, über jeweils 2 Schlaf- und Badezimmer sowie eine Klimaanlage und eine Einfahrt verfügt. Wie ändert sich die Punktprognose, wenn ein Schlafzimmer und ein Badezimmer dazugebaut werden?

- (g) Treffen Sie unter Verwendung eines geeigneten Tests anhand des Modells aus Teil (e) eine Entscheidung bezüglich der Hypothese, dass sich die Anzahl der Badezimmer stärker auf den Verkaufspreis auswirkt als die Anzahl der Schlafzimmer. ($\alpha = 1\%$)

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: Sie können $\widehat{\text{Cov}}(\hat{\beta}_{\text{bedrooms}}, \hat{\beta}_{\text{bathrooms}}) = -606762.1$ verwenden.

- (h) Berechnen Sie anhand des Modells aus (d) das zweiseitige Prognoseintervall zum Niveau $1 - \alpha = 0.95$ für den erwarteten Verkaufspreis eines 500 square feet großen Hauses mit zwei Schlafzimmern und einem Badezimmer, in dem keine Klimaanlage und keine Einfahrt vorhanden sind. Verwenden Sie dazu die folgende geschätzte Varianz-Kovarianz-Matrix $\hat{V}(\hat{\beta})$:

$$\begin{pmatrix} 14182118.65 & -285.55 & -2681874.60 & -1112003.69 & 669685.94 & -3482775.39 \\ -285.55 & 0.14 & -34.10 & -72.54 & -100.77 & -229.62 \\ -2681874.60 & -34.10 & 1200243.49 & -607076.18 & -160614.07 & 151304.12 \\ -1112003.69 & -72.54 & -607076.18 & 2629741.47 & -295212.49 & -2125.15 \\ 669685.94 & -100.77 & -160614.07 & -295212.49 & 2760347.15 & -197299.68 \\ -3482775.39 & -229.62 & 151304.12 & -2125.15 & -197299.68 & 4984993.48 \end{pmatrix}$$

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

$n \setminus p$	0.9	0.95	0.975	0.99
535	1.28314	1.64771	1.96441	2.33334
536	1.28313	1.64770	1.96440	2.33333
537	1.28313	1.64770	1.96439	2.33331
538	1.28313	1.64769	1.96438	2.33330
539	1.28312	1.64769	1.96437	2.33329
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

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

$n \setminus m$	2	3	4	540	541	542	543	544
2	19.0000	19.1643	19.2468	19.4939	19.4939	19.4939	19.4939	19.4939
3	9.5521	9.2766	9.1172	8.5316	8.5316	8.5315	8.5315	8.5315
4	6.9443	6.5914	6.3882	5.6348	5.6348	5.6348	5.6347	5.6347
540	3.0124	2.6214	2.3884	1.1522	1.1522	1.1521	1.1520	1.1519
541	3.0124	2.6214	2.3884	1.1521	1.1521	1.1520	1.1519	1.1519
542	3.0124	2.6213	2.3884	1.1521	1.1520	1.1519	1.1518	1.1518
543	3.0123	2.6213	2.3883	1.1520	1.1519	1.1518	1.1518	1.1517
544	3.0123	2.6213	2.3883	1.1519	1.1518	1.1518	1.1517	1.1516

Aufgabe 4 (4 + 5 + 3 + 3 + 5 + 4 = 24 Punkte)

Mit Hilfe der Statistiksoftware **R** soll der Datensatz **CPSSWEducation** aus dem Paket **AER** untersucht werden, welche Informationen bezüglich der Lohnverteilung von Vollzeitangestellten in den USA im Jahr 2004 zur Verfügung stellt. Die abhängige Variable **earnings**, welche den durchschnittlichen Stundenlohn der Angestellten in USD angibt, soll zunächst durch die Ausbildungsdauer in Jahren (**education**) erklärt werden.

- (a) Es werden zunächst drei Varianten geschätzt, jeweils für einen linearen, quadratischen sowie kubischen Zusammenhang zwischen dem durchschnittlichen Stundenlohn und der Ausbildungsdauer. Man erhält folgende (gekürzte) Outputs:

- für das lineare Modell:

```
Call:
lm(formula = earnings ~ education, data = CPSSWEducation)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.1343714  0.9592543 -3.26751  0.0010974 **
education    1.4669255  0.0697828 21.02130 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.76936 on 2948 degrees of freedom
Multiple R-squared:  0.130357, Adjusted R-squared:  0.130062
F-statistic: 441.895 on 1 and 2948 DF, p-value: < 2.22e-16
```

- für das quadratische Modell:

```
Call:
lm(formula = earnings ~ education + I(education^2),
    data = CPSSWEducation)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9633943  3.3299284  1.79085  0.0734203 .
education     0.0488692  0.5019461  0.09736  0.9224476
I(education^2) 0.0535411  0.0187682  2.85275  0.0043644 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.75876 on 2947 degrees of freedom
Multiple R-squared:  0.132752, Adjusted R-squared:  0.132163
F-statistic: 225.552 on 2 and 2947 DF, p-value: < 2.22e-16
```

- für das kubische Modell:

Call:

```
lm(formula = earnings ~ education + I(education^2)
+ I(education^3), data = CPSSWEducation)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	30.02395185	9.69887781	3.09561	0.0019825	**
education	-6.50846562	2.53306956	-2.56940	0.0102363	*
I(education^2)	0.61311322	0.21271054	2.88238	0.0039754	**
I(education^3)	-0.01515213	0.00573737	-2.64095	0.0083110	**

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

Residual standard error: 8.7499 on 2946 degrees of freedom

Multiple R-squared: 0.1348, Adjusted R-squared: 0.133919

F-statistic: 152.998 on 3 and 2946 DF, p-value: < 2.22e-16

- (i) Welches der drei geschätzten Modelle würden Sie anhand des korrigierten Bestimmtheitsmaßes auswählen?
- (ii) Lassen die verschiedenen Schätzergebnisse auf ein Problem imperfekter Multikollinearität bezüglich des Regressors `education` im quadratischen Modell schließen? Begründen Sie Ihre Antwort.
- (b) Unter Zuhilfenahme von zwei weiteren Regressoren, `age` (Alter in Jahren) und `female` (mit weiblich = 1 und männlich = 0), wurde ein weiteres Modell geschätzt:

Call:

```
lm(formula = earnings ~ age + education + female,
data = CPSSWEducation)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-24.5918618	9.3984363	-2.61659	0.0089266	**
age	0.7228854	0.3170401	2.28011	0.0226725	*
education	1.5787339	0.0693547	22.76320	< 2.22e-16	***
female	-3.3891540	0.3265857	-10.37753	< 2.22e-16	***

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

Residual standard error: 8.60946 on 2946 degrees of freedom

Multiple R-squared: 0.16235, Adjusted R-squared: 0.161497

F-statistic: 190.327 on 3 and 2946 DF, p-value: < 2.22e-16

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

- (ii) Verifizieren oder entkräften Sie die Behauptung, dass Frauen bei gleicher Berufserfahrung und gleichem Alter signifikant weniger verdienen als Männer. (Begründung!)
- (c) Als Alternative zu den obigen polynomialen Modellen wurde auch ein log-lin-Modell geschätzt:

Call:

```
lm(formula = log(earnings) ~ age + education + female,
    data = CPSSWEducation)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.68207937	0.50371052	1.35411	0.17581
age	0.02687637	0.01699181	1.58173	0.11382
education	0.09506958	0.00371707	25.57647	< 2e-16 ***
female	-0.18754039	0.01750341	-10.71451	< 2e-16 ***

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

Residual standard error: 0.461425 on 2946 degrees of freedom

Multiple R-squared: 0.192648, Adjusted R-squared: 0.191825

F-statistic: 234.321 on 3 and 2946 DF, p-value: < 2.22e-16

- (i) Lässt sich die Anpassungsgüte dieses Modells mit jener der polynomialen Modelle sinnvoll vergleichen? (Begründung!)
- (ii) Geben Sie ferner eine Interpretation an für den zu **education** gehörenden Regressionskoeffizienten.
- (d) Ein Breusch-Pagan-Test (nach Koenker) lieferte folgenden Output für den Modellansatz aus Aufgabenteil (c):

studentized Breusch-Pagan test

data: CPSSWEducation

BP = 10.9818, df = 3, p-value = 0.011824

Was wird mit diesem Test untersucht? Zu welchem Ergebnis kommt man für $\alpha = 1\%$? Geben Sie bei Ihrer Begründung explizit die Hypothesen an.

- (e) (i) Zur Diagnose welchen Phänomens wird der Varianz-Inflations-Faktors verwendet? Geben Sie die zugehörige Faustregel an.
- (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?
- (f) Geben Sie drei mögliche Konstellationen an, unter denen notwendige Annahmen für die Konsistenz und Unverzerrtheit der Koeffizientenschätzer $\hat{\beta}$ verletzt werden.