

Aufgabe 3 (12 + 4 + 7 + 7 + 1 + 5 + 11 + 5 = 52 Punkte)

Hinweis: Beachten Sie die Tabellen 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)
Sdach	Schiebedach vorhanden? (Ja = 1, Nein = 0)

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

Call:

```
lm(formula = Preis ~ Alter + Kilstand + Tuev)
```

Residuals:

Min	1Q	Median	3Q	Max
-2292.59	-472.15	-22.75	495.50	2863.13

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9056.860	389.082	23.278	< 2e-16
Alter	???	3.403	-11.127	< 2e-16
Kilstand	-9.785	1.447	???	2.18e-10
Tuev	-3.287	8.517	-0.386	0.7

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

Residual standard error: 779.6 on 168 degrees of freedom

Multiple R-squared: 0.6154, Adjusted R-squared: ???

F-statistic: 89.61 on ??? and ??? DF, p-value: < 2.2e-16

Stellen Sie das zugrunde liegende Modell in formaler Schreibweise dar und geben Sie explizit an, welche Variable hier mit Hilfe welcher Regressoren erklärt wird.

Wie viele Beobachtungen gingen in die obige Schätzung ein?

Bestimmen Sie ferner die mit ??? markierten Größen (mit den Bezeichnungen $\hat{\beta}_{\text{alter}}$, t_{kilstand} , $\overline{R}^2$, DF_1 und DF_2).

- (b) Geben Sie an, welche Regressionskoeffizienten signifikant von Null verschieden sind zum Signifikanzniveau 5%.

- (c) Testen Sie zum Signifikanzniveau 5%, ob eine Erhöhung des Alters des Autos um einen Monat den Verkaufspreis um signifikant mehr als 30 € reduziert.

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 $\hat{\beta}_{\text{alter}} = -37.866$ verwenden.

Hinweis: Verwenden Sie näherungsweise statt der benötigten Quantile der passenden t -Verteilung die entsprechenden Quantile der Standardnormalverteilung.

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

Call:

```
lm(formula = Preis ~ Alter + Kilstand + Tuev + ABS + Sdach)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2457.79	-534.64	-4.36	498.83	2721.65

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9311.171	413.047	22.543	< 2e-16 ***
Alter	-38.329	3.405	-11.258	< 2e-16 ***
Kilstand	-9.742	1.450	-6.720	2.78e-10 ***
Tuev	-5.483	8.600	-0.638	0.5246
ABS	-237.664	129.523	-1.835	0.0683 .
Sdach	-9.862	138.911	-0.071	0.9435

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

Residual standard error: 776.4 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

Testen Sie unter Zuhilfenahme geeigneter Bestimmtheitsmaße zum Signifikanzniveau von 10%, 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) Erläutern Sie kurz, warum es problematisch wäre, im Modell aus Aufgabenteil (d) zusätzlich eine Dummyvariable **Sdachno** (mit dem Wert 1, falls kein Schiebedach vorhanden ist) aufzunehmen.

- (f) Unter Zuhilfenahme der Interaktionsvariablen `Alter * Sdach` wurde ein weiteres Modell geschätzt, welches folgenden (gekürzten) Output ergibt:

Call:

```
lm(formula = Preis ~ Alter + Kilstand + Tuev + ABS + Sdach +
    I(Alter * Sdach))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10522.982	696.352	15.112	< 2e-16 ***
Alter	-49.975	6.380	-7.833	5.52e-13 ***
Kilstand	-9.526	1.438	-6.626	4.67e-10 ***
Tuev	-4.877	8.512	-0.573	0.5674
ABS	-189.882	130.049	-1.460	0.1462
Sdach	-1740.870	817.084	-2.131	0.0346 *
I(Alter * Sdach)	15.544	7.232	2.149	0.0331 *

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

Residual standard error: 768.1 on 165 degrees of freedom

Multiple R-squared: 0.6334, Adjusted R-squared: 0.6201

F-statistic: 47.51 on 6 and 165 DF, p-value: < 2.2e-16

Geben Sie eine Interpretation für den Regressionskoeffizienten zu "I(Alter * Sdach)" an.

Wie groß ist der marginale Effekt des Alters auf den Verkaufspreis,

- falls ein Schiebedach vorhanden ist?
- falls kein Schiebedach vorhanden ist?

- (g) Man vermutet, dass sich die Parameterstruktur zwischen den Gruppen mit Schiebedach und ohne Schiebedach unterscheidet. Bei einer KQ-Schätzung des Modells für den Fall ohne die Variable `sdach` erhält man den folgenden Output:

Call:

```
lm(formula = Preis ~ Alter + Kilstand + Tuev + ABS)
```

Residuals:

Min	1Q	Median	3Q	Max
-2461.42	-534.56	-5.86	496.29	2718.80

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9308.080	409.521	22.729	< 2e-16 ***
Alter	-38.343	3.389	-11.315	< 2e-16 ***
Kilstand	-9.753	1.437	-6.786	1.92e-10 ***
Tuev	-5.535	8.544	-0.648	0.5180
ABS	-238.241	128.882	-1.849	0.0663 .

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 774.1 on 167 degrees of freedom

Multiple R-squared: 0.6231, Adjusted R-squared: 0.6141

F-statistic: 69.03 on 4 and 167 DF, p-value: < 2.2e-16

Als (gekürzter) Output einer entsprechenden Schätzung für die Teilgruppe mit Schiebedach ergibt sich:

Residual standard error: 826.5 on 124 degrees of freedom
 Multiple R-squared: 0.5328, Adjusted R-squared: 0.5178
 F-statistic: 35.36 on 4 and 124 DF, p-value: < 2.2e-16

Eine entsprechende Schätzung für die Teilgruppe ohne Schiebedach liefert:

Residual standard error: 563.1 on 38 degrees of freedom
 Multiple R-squared: 0.8504, Adjusted R-squared: 0.8346
 F-statistic: 53.98 on 4 and 38 DF, p-value: 3.636e-15

Prüfen Sie mit Hilfe dieser Ergebnisse zum Signifikanzniveau $\alpha = 5\%$, ob ein Strukturbruch in den Regressionskoeffizienten (in beliebiger Form) vorliegt.

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.

- (h) 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 p-Quantilen der Standardnormalverteilung

p	0.85	0.90	0.95	0.975	0.99	0.995	0.9995
N_p	1.036	1.282	1.645	1.960	2.326	2.576	3.291

sowie die folgenden Tabellen mit 0.90- und 0.95-Quantilen einiger $F(m, n)$ -Verteilungen:

$n \backslash m$	2	3	4	5	164	165	166	167
2	9.0000	9.1618	9.2434	9.2926	9.4851	9.4852	9.4852	9.4852
3	5.4624	5.3908	5.3426	5.3092	5.1402	5.1401	5.1401	5.1400
4	4.3246	4.1909	4.1072	4.0506	3.7714	3.7713	3.7713	3.7712
5	3.7797	3.6195	3.5202	3.4530	3.1181	3.1180	3.1179	3.1178
164	2.3352	2.1175	1.9795	1.8827	1.2222	1.2218	1.2215	1.2212
165	2.3350	2.1173	1.9792	1.8825	1.2217	1.2214	1.2211	1.2208
166	2.3348	2.1171	1.9790	1.8823	1.2213	1.2210	1.2207	1.2204
167	2.3346	2.1169	1.9788	1.8821	1.2209	1.2206	1.2202	1.2199

$n \backslash m$	2	3	4	5	162	163	164	165
2	19.0000	19.1643	19.2468	19.2964	19.4896	19.4896	19.4896	19.4897
3	9.5521	9.2766	9.1172	9.0135	8.5434	8.5433	8.5432	8.5431
4	6.9443	6.5914	6.3882	6.2561	5.6504	5.6502	5.6501	5.6500
5	5.7861	5.4095	5.1922	5.0503	4.3898	4.3897	4.3895	4.3894
162	3.0518	2.6604	2.4275	2.2700	1.2959	1.2955	1.2950	1.2946
163	3.0515	2.6601	2.4271	2.2696	1.2953	1.2949	1.2944	1.2940
164	3.0511	2.6597	2.4268	2.2693	1.2947	1.2943	1.2939	1.2934
165	3.0508	2.6594	2.4264	2.2689	1.2942	1.2937	1.2933	1.2928

Aufgabe 4 (1 + 14 + 7 + 8 = 30 Punkte)

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

Mit Hilfe der Statistiksoftware **R** soll der Datensatz **CPSSWEducation** aus dem Paket **AER** untersucht werden, welcher Informationen bezüglich der Lohnverteilung von Vollzeitangestellten in den USA im Jahr 2004 zur Verfügung stellt. Als abhängige Variable soll **earnings** verwendet werden, welche den durchschnittlichen Stundenlohn der Angestellten in USD angibt. Die erklärenden Variablen sind in diesem Fall das Alter in Jahren (**age**), die Ausbildungsdauer in Jahren (**education**) sowie das Geschlecht (**female** mit weiblich = 1 und männlich = 0). Bei der Analyse der Daten entstand folgender (gekürzter) Output:

Call:

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-24.59186	9.39844	-2.617	0.00893	**
age	0.72289	0.31704	2.280	0.02267	*
education	1.57873	0.06935	22.763	< 2e-16	***
female	-3.38915	0.32659	-10.378	< 2e-16	***

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

Residual standard error: 8.609 on 2946 degrees of freedom

Multiple R-squared: 0.1623, Adjusted R-squared: 0.1615

F-statistic: 190.3 on 3 and 2946 DF, p-value: < 2.2e-16

- Entscheiden Sie zum Signifikanzniveau $\alpha = 1\%$, ob Signifikanz des Erklärungsansatzes vorliegt.
- Berechnen Sie die zweiseitigen Prognoseintervalle zum Niveau $1 - \alpha = 0.95$ für das erwartete Einkommen eines 29-jährigen Mannes mit einer Ausbildungsdauer von 12 Jahren sowie für dessen Erwartungswert. Verwenden Sie dazu die geschätzte Varianz-Kovarianzmatrix:

$$\widehat{V}(\hat{\beta}) = \begin{pmatrix} 88.3306048 & -2.9646810 & -0.0639216 & 0.0287764 \\ -2.9646810 & 0.1005144 & 0.0000061 & -0.0008303 \\ -0.0639216 & 0.0000061 & 0.0048100 & -0.0035233 \\ 0.0287764 & -0.0008303 & -0.0035233 & 0.1066582 \end{pmatrix}$$

Hinweis: Verwenden Sie näherungsweise statt der benötigten Quantile der passenden t-Verteilung die entsprechenden Quantile der Standardnormalverteilung.

- Treffen Sie unter Verwendung eines geeigneten Tests eine Entscheidung bezüglich der Hypothese, dass sich ein zusätzliches Ausbildungsjahr stärker auf den Stundenlohn auswirkt als ein zusätzliches Lebensjahr, wenn das Signifikanzniveau $\alpha = 5\%$ 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.

Hinweis: Verwenden Sie näherungsweise statt der benötigten Quantile der passenden t-Verteilung die entsprechenden Quantile der Standardnormalverteilung.

- (d) Man vermutet, dass heteroskedastische Störgrößen vorliegen könnten dahingehend, dass die Varianz der Störgrößen für die Gruppe der Frauen erheblich von der Varianz der Störgrößen für die Gruppe der Männer abweicht. Um dies zu überprüfen, wurde der Datensatz anhand der Größe `female` in zwei Teile aufgeteilt: einen Teil, in dem die Gruppe der Frauen betrachtet wird, und einen Teil, in dem das Modell für die Gruppe der Männer betrachtet wird. Für jeden Teildatensatz wurde das Modell dann separat geschätzt:

Call:

```
lm(formula = earnings ~ age + education, data = CPSSWEducation,
    subset = (female == 1))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-15.378	-4.533	-1.072	2.769	70.603

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-21.5499	12.6830	-1.699	0.0896 .
age	0.5196	0.4274	1.216	0.2244
education	1.5478	0.0972	15.924	<2e-16 ***

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

Residual standard error: 7.409 on 1199 degrees of freedom

Multiple R-squared: 0.1754, Adjusted R-squared: 0.1741

F-statistic: 127.6 on 2 and 1199 DF, p-value: < 2.2e-16

Call:

```
lm(formula = earnings ~ age + education, data = CPSSWEducation,
    subset = (female == 0))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-19.257	-5.691	-1.664	3.386	75.023

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-28.97141	13.25517	-2.186	0.0290 *
age	0.86293	0.44725	1.929	0.0538 .
education	1.59753	0.09542	16.742	<2e-16 ***

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

Residual standard error: 9.349 on 1745 degrees of freedom

Multiple R-squared: 0.1399, Adjusted R-squared: 0.139

F-statistic: 142 on 2 and 1745 DF, p-value: < 2.2e-16

Überprüfen Sie die oben genannte Vermutung mit Hilfe eines geeigneten Tests zum Signifikanzniveau $\alpha = 10\%$. Liegt damit Evidenz für Heteroskedastie in den Störgrößen vor?

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 Tabelle mit p -Quantilen der Standardnormalverteilung

p	0.85	0.90	0.95	0.975	0.99	0.995	0.9995
N_p	1.036	1.282	1.645	1.960	2.326	2.576	3.291

sowie die folgende Tabelle mit 0.99-Quantilen einiger $F(m, n)$ -Verteilungen

$n \backslash m$	3	4	5	2946	2947
3	29.45670	28.70990	28.23708	26.12909	26.12909
4	16.69437	15.97702	15.52186	13.46694	13.46694
5	12.05995	11.39193	10.96702	9.02416	9.02416
2946	3.78827	3.32554	3.02345	1.08952	1.08952
2947	3.78827	3.32553	3.02345	1.08952	1.08951

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

$n \backslash m$	1198	1199	1200	1745	1746	1747	1748	1749	1750
1198	1.09975	1.09973	1.09971	1.09168	1.09167	1.09166	1.09165	1.09164	1.09163
1199	1.09973	1.09971	1.09969	1.09166	1.09165	1.09164	1.09163	1.09162	1.09160
1200	1.09970	1.09968	1.09966	1.09163	1.09162	1.09161	1.09160	1.09159	1.09158
1745	1.09079	1.09076	1.09074	1.08196	1.08195	1.08194	1.08192	1.08191	1.08190
1746	1.09077	1.09075	1.09073	1.08195	1.08193	1.08192	1.08191	1.08190	1.08189
1747	1.09076	1.09074	1.09072	1.08193	1.08192	1.08191	1.08190	1.08189	1.08188
1748	1.09075	1.09073	1.09071	1.08192	1.08191	1.08190	1.08189	1.08187	1.08186
1749	1.09074	1.09072	1.09069	1.08191	1.08190	1.08188	1.08187	1.08186	1.08185
1750	1.09073	1.09071	1.09068	1.08190	1.08188	1.08187	1.08186	1.08185	1.08184