Übungsserie 1: Lösungen

OLS-Regression für Autopreise

1. Geben Sie für die folgenden Merkmale das jeweilige Skalenniveau und mögliche Merkmalsausprägungen. Unterscheiden Sie die Merkmale ferner in diskrete und stetige und diskutieren Sie dabei Probleme der Messgenauigkeit.

	Merkmal	Ausprägungen	Skalenniveau	Diskret?
a	Gewicht	60Kg; 90 Kg	metrisch	stetig
b	Ak. Grad	Bachelor, Master; Ph.D.	ordinal	diskret
c	Augenfarbe	Blau, brauch, grau	nominal	diskret
d	Geschlecht	Männlich, weiblich	nominal	diskret
e	Nettoeinkommen in CHF	6000; 10‘000	metrisch	stetig

Nominalskala: Qualitative Merkmalsausprägung → kann nicht angeordnet werden

Ordinalskala: Rang, Merkmalsausprägungen können angeordnet, aber nicht gerechnet werden.

Metrische Skala: Quantitativ, sowohl Anordnung als auch Rechnen ist möglich

2. Welche Faktoren bestimmen den Verkaufspreis eines Gebrauchtautos. Welche Vorzeichen erwarten Sie?

   Marke / Modell / Typ / Kilometerstand (-) / Zustand des Autos (+ je besser) usw.

3. Welche sind davon qualitative Faktoren?

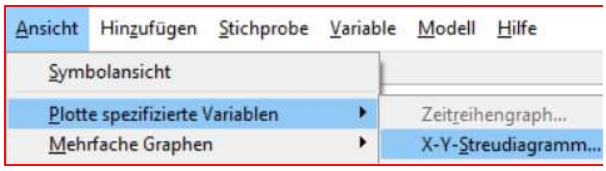
   Zustand des Autos, Marke, Autotyp usw.

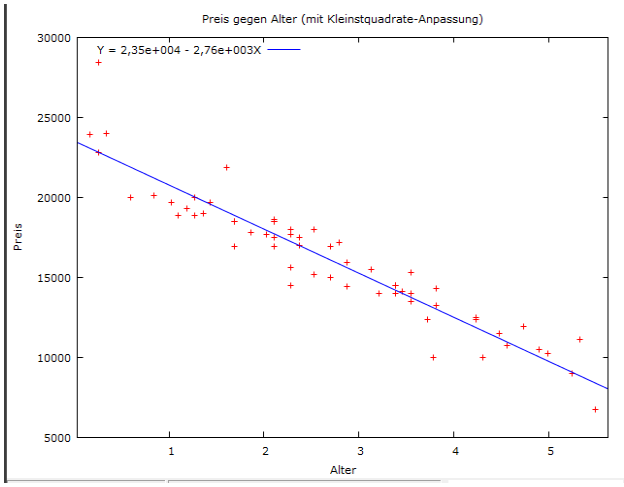
4. Erklären Sie was ein Streudiagramm ist.

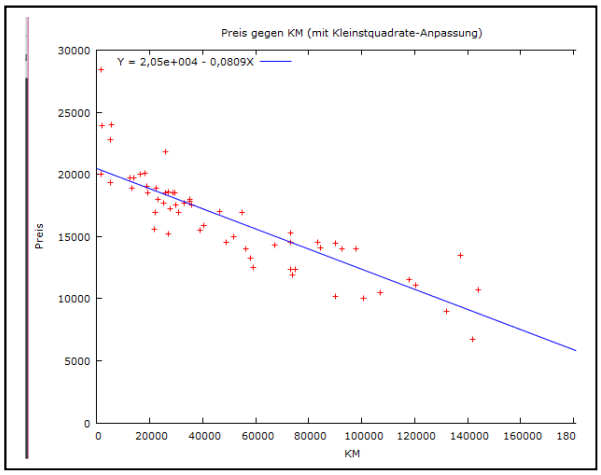
   Ein Streudiagramm (E: scatter plot) ist die graphische Darstellung von beobachteten Wertepaaren zweier statistischer Merkmale (Preis, Alter) oder (Preis, KM) usw. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktwolke ergibt.

5. Erstellen Sie folgende Streudiagramme:
   i. Preis gegen Alter
   ii. Preis gegen Kilometerstand

   Hinweis: Y-Achsen-Variable: Preis







6. Was sagen diese Streudiagramme über den statistischen Zusammenhang zwischen den Autopreis und den ausgewählten Variablen (Alter, KM) aus?

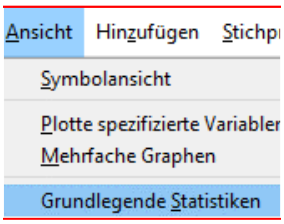
   Es besteht einen negativen linearen Zusammenhang zwischen Autopreis und Alter sowie zwischen Autopreis und Kilometerstand → die Steigung der Regressionsgerade ist negativ.

7. Was ist der Mittelwert für die Variablen Kilometerstand, Preis und Alter von Gebrauchtautos in dieser Stichprobe?

   gretl Hauptfenster: Ansicht / Grundlegende Statistiken

	arith. Mittel	Median	Minimum	Maximum
Preis	16140,	16900,	6700,0	28400,
Alter	2,6766	2,5300	0,17000	5,4900
KM	53368,	35900,	1500,0	1,8800e+005

	Std. Abw.	Var'koeff.	Schiefe	Überwölbung
Preis	4029,8	0,24968	0,24267	0,43224
Alter	1,3761	0,51412	0,15308	-0,72132
KM	42556,	0,79742	1,0238	0,45949



   Der durchschnittliche Preis beträgt CHF 16‘140, das durchschnittliche Auto ist 2.6 Jahre alt und hat einen Kilometerstand von km 53‘368.

8. Interpretieren Sie den Median für die Variabler KM “Kilometerstand“.

   50% der Gebrauchtautos haben einen Kilometerstand über km 35‘900 und 50% darunter.

9. Welche wichtige Information gibt die Standardabweichung im Allgemeinen?

   Die Standardabweichung ist vor allem ein Mass dafür, wie **repräsentativ** der Mittelwert eines Datensatzes für die jeweiligen Daten ist. Sie gibt Auskunft darüber, ob der Mittelwert einer Verteilung einen geeigneten Erwartungswert darstellt. Niedrige Standardabweichungen implizieren eine gute Repräsentativität des Mittelwertes, hohe wiederum eine schlechte Repräsentativität.

10. Was ist der Vorteil der Standardabweichung gegenüber der Varianz als Streuungsmass?

    Vorteil: Sie hat die gleiche Messeinheit wie die ursprünglichen Messwerte.

    Beispiel: Wenn die Zahl der Kinder in einem Haushalt untersucht wird, so ist die Einheit der Varianz ein Quadratkind, die Einheit der Standardabweichung aber wieder ein Kind.

11. Welche Variable weist die geringste und höchste Standardabweichung auf? Was können Sie über die Repräsentativität des Mittelwertes dieser Variablen sagen?

Kilometerstand weist die höchste und Alter die kleinste Standardabweichung auf. Das Durchschnittsalter von 2.67 Jahren ist repräsentativ, hingegen der Mittelwert für den Kilometerstand nicht, da die Standardabweichung relativ gross ist.

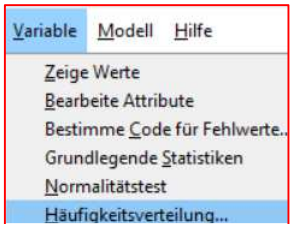
12. Interpretieren Sie die Standardabweichung für die Variable Alter.

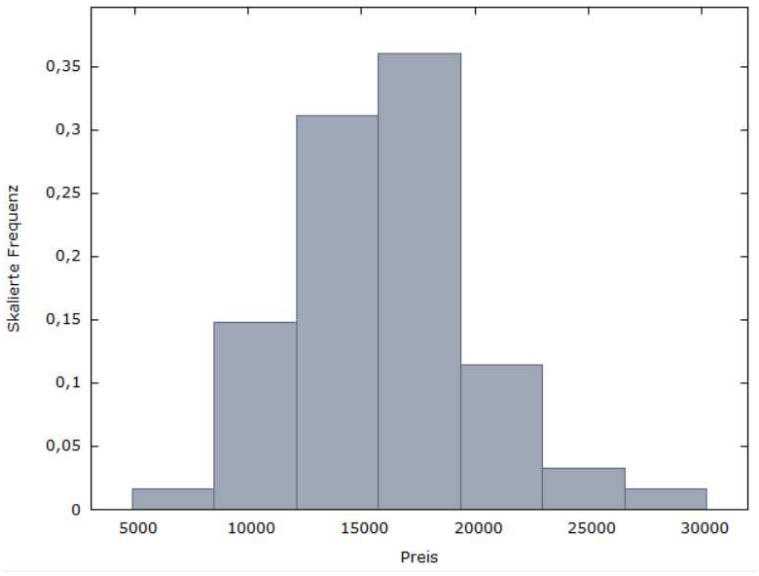
Die Standardabweichung für Alter (s_{Alter}) beträgt 1.37 Jahre. Die durchschnittliche Abweichung vom Mittelwert (2.67J) beträgt 1.37 Jahre = mittlere Abweichung vom Mittelwert. In diesem Fall ist das Durchschnittsalter eher repräsentativ für die Gebrauchtautos in der Stichprobe.

13. Erklären Sie was ein Histogramm ist.

Ein **Histogramm** ist eine graphische Darstellung der Häufigkeitsverteilung von Daten, welche in Klassen (*bins*) eingeteilt sind, die eine konstante oder variable Breite haben können. Es werden direkt nebeneinanderliegende Rechtecke von der Breite der jeweiligen Klasse gezeichnet, deren Flächeninhalte die (relativen oder absoluten) Klassenhäufigkeiten darstellen. Die Höhe jedes Rechtecks stellt dann die (relative oder absolute) Häufigkeitsdichte dar, also die (relative oder absolute) Häufigkeit dividiert durch die Breite der entsprechenden Klasse.

14. Erstellen Sie das Histogramm für die Variable *Autopreis*






```
Häufigkeitsverteilung für Preis, Beob. 1-61
Zahl der Klassen = 7, Mittel = 16140,2, St'Abw. = 4029,83

       Intervall          Mitte    Häufigkeit    rel.     kum.

          < 8508,3     6700,0          1      1,64%    1,64%
 8508,3 - 12125,      10317,           9     14,75%   16,39% *****
 12125, - 15742,      13933,          19     31,15%   47,54% ***********
 15742, - 19358,      17550,          22     36,07%   83,61% ************
 19358, - 22975,      21167,           7     11,48%   95,08% ****
 22975, - 26592,      24783,           2      3,28%   98,36% *
       >= 26592,      28400,           1      1,64%  100,00%
```

15. Welche ist die modale Klasse dieses Histogramms?

Der Modus (Modalwert) ist definiert als die am häufigsten vorkommende Merkmalausprägung.
Modale Kasse = [CHF 15‘742 – CHF 19‘358] → Klasse mit der grössten Histogrammhöhe

16. Erklären Sie den Hauptvorteil des Korrelationskoeffizienten gegenüber der Kovarianz.

Die Kovarianz macht eine Aussage über die **Richtung** des linearen Zusammenhangs zweier Variablen, aber nicht über deren Stärke! Der Korrelationskoeffizient ist eine normierte Kovarianz.

17. Welche Korrelationen erwarten Sie zwischen den Variablen (Preis, Alter, KM)?

corr(Preis, Alter): negative Korrelation → je älter das Gebrauchtauto, desto billiger das Auto

corr(Preis, KM): negative Korrelation → je höher der Kilometerstand, desto billiger das Auto

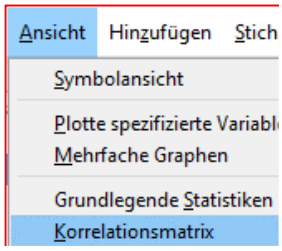
corr(KM, Alter): positive Korrelation → je älter das Gebrauchtauto, desto höher der Kilometerstand

18. Analysieren Sie die Korrelation zwischen Preis, KM und Alter mittels gretl. Lassen sich Ihre Erwartungen bestätigen? Welches Variablen-Paar weist die höchste Korrelation auf? Ist dieses Ergebnis plausibel?

gretl Hauptfenster: Ansicht / Korrelationsmatrix

```
Korrelationskoeffizienten, benutze die Beobachtungen 1 - 61
5% kritischer Wert (zweiseitig) = 0,2521 für n = 61

          Preis         Alter            KM
         1,0000       -0,9417       -0,8548  Preis
                       1,0000        0,8345  Alter
                                     1,0000  KM
```



Das Paar (Alter, Preis) weist die höchste Korrelation auf. Je älter das Gebrauchtauto, desto billiger.

19. Erklären Sie kurz was der Variationskoeffizient ist.

Im Gegensatz zur Varianz ist der Variationskoeffizient ein relatives Streuungsmass, welches nicht von der Masseinheit der statistischen Variable bzw. Zufallsvariable abhängt. Der Variationskoeffizient hat dementsprechend keine Einheit und wird in **%** ausgedrückt.

Formel: $\mathrm{varK}(X) = \frac{\sqrt{\mathrm{var}(X)}}{E(X)}$ wobei E(X) = Erwartungswert der Zufallsvariable X

Gilt Standardabweichung **>** Mittelwert bzw. Erwartungswert → varK **> 1**

20. Was ist der Vorteil des Variationskoeffizienten gegenüber der Standardabweichung?

Eine Variable mit grossem Mittelwert weist im Allgemeinen eine grössere Varianz auf als eine mit einem kleinen Mittelwert.

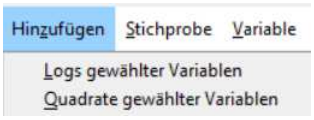
Varianz und Standardabweichung sind nicht normiert → kann nicht beurteilt werden, ob eine Varianz gross oder klein ist.

Variationskoeffizient ist von der ausgewählten Skala unabhängig und dementsprechend aussagekräftiger als die Standardabweichung.

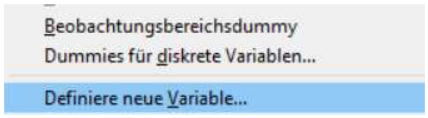
21. Welche Variable weist den grössten Variationskoeffizienten auf? Wie interpretieren Sie diese Zahl?

Die Variable „KM“ weist den höchsten Variationskoeffizienten auf. Die Standard Abweichung beträgt ca. 80% des Mittelwertes. Wenn der Variationskoeffizient höher als 100% (oder 1), ist die Streuung sehr gross.

22. Definieren Sie zwei neuen Variablen:



- Preis100: Gibt den Preis in Einheiten von CHF 100 an.
- KM1000: Gibt die km-Zahl in Einheiten von 1000 km an.



gretl Hauptfenster: Hinzufügen / Definiere neue Variable

KM1000 = KM / 1000

Preis100 = Preis / 100

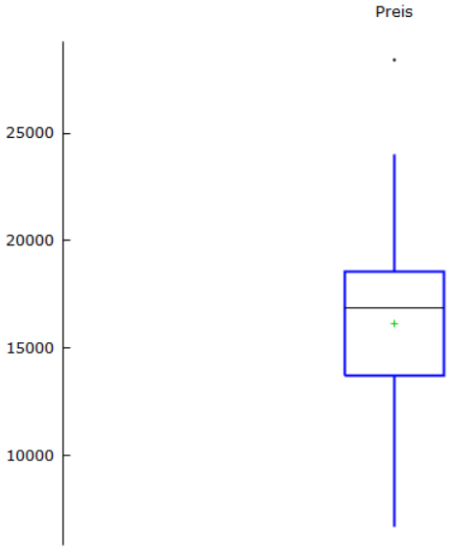
23. Vergleichen Sie die Standardabweichungen und Variationskoeffizienten für folgende Grössen: Preis – Preis100 und KM – KM1000

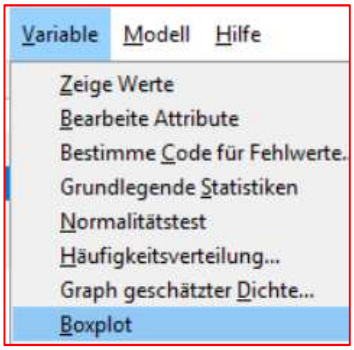
	Std. Abw.	Var'koeff.
Preis	4029,8	0,24968
Preis100	40,298	0,24968
KM	42556,	0,79742
KM1000	42,556	0,79742

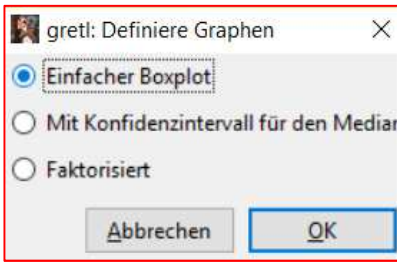
VaK ist unabhängig der benutzten Skala!

Neue Stdabw. ist 10x kleiner als die ursprüngliche!

24. Erstellen Sie einen Box-Plot für die Variable Autopreis. Welche Informationen vermitteln einen Box-Plot?







- Die untere bzw. obere Grenze der Box ist durch das untere bzw. obere Quartil gegeben → die Hälfte der beobachteten Werte liegt innerhalb der Box.
- Die Länge der Box entspricht einem Quartilsabstand $d_Q = x_{0.75} - x_{0.25}$
- Die Linie innerhalb der Box gibt die Lage des Medians (16‘900) wieder.
- Das grüne Kreuz innerhalb der Box entspricht dem Mittelwert (16‘140).
- Der Punkt oberhalb der Box entspricht dem Extremwert (28‘400).

25. Erklären Sie was Schiefe ist.

Die Schiefe (Skewness bzw. Skew) gibt die Richtung und Stärke der Schiefe (Asymmetrie) einer Wahrscheinlichkeitsverteilung. Sie zeigt an, ob und wie stark die Verteilung nach rechts (positive Schiefe) oder nach links (negative Schiefe) geneigt ist.
Alle drei Verteilungen sind rechtsschief.

26. Erklären Sie was Kurtosis (Wölbung) ist. Wie ist der Exzess definiert?
Mass für die Wölbung einer eingipfligen Verteilung. Bei Normalverteilung K = 3
Ekzess = K - 3 → Mass für die Abweichung gegenüber der Normalverteilung
Für positive Werte ist das Maximum der Häufigkeitsverteilung grösser als das einer Normalverteilung mit gleicher Varianz und umgekehrt.

27. Analysieren Sie die Wölbung und Kurtosis für folgende Variablen: Preis, Preis100, KM und KM1000

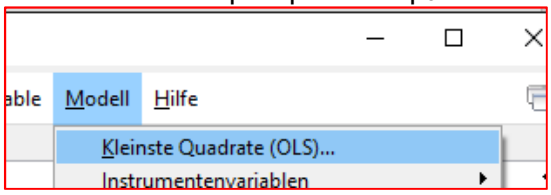
	Schiefe	Überwölbung
Preis	0,24267	0,43224
KM	1,0238	0,45949
Preis100	0,24267	0,43224
KM1000	1,0238	0,45949

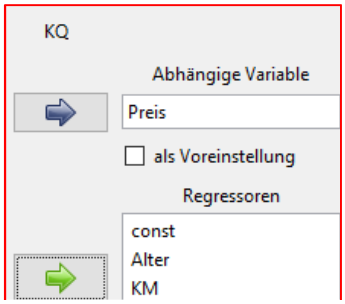
Die Variablen Preis und KM haben rechtsschiefe und spitzere Verteilungen.
Durch eine lineare Transformation einer Variable ändert sich dich Wölbung und Kurtosis nicht.

28. Schätzen Sie folgende Regressionsmodelle:

Modell 1: Preis = $\beta_1 + \beta_2$Alter + u

Modell 2: Preis = $\beta_1 + \beta_2$Alter + β_3Km + u






```
Modell 1: KQ, benutze die Beobachtungen 1-61
Abhängige Variable: Preis

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       23521,5       385,394        61,03     5,11e-055 ***
  Alter       -2757,77      128,276       -21,50     1,34e-029 ***

Mittel d. abh. Var.     16140,16   Stdabw. d. abh. Var.    4029,835
Summe d. quad. Res.     1,10e+08   Stdfehler d. Regress.   1367,299
R-Quadrat               0,886798   Korrigiertes R-Quadrat  0,884880
F(1, 59)                462,1931   P-Wert(F)               1,34e-29
Log-Likelihood         -525,9946   Akaike-Kriterium        1055,989
Schwarz-Kriterium       1060,211   Hannan-Quinn-Kriterium  1057,644
```

Model 1

Modell 2

```
Modell 4: KQ, benutze die Beobachtungen 1-61
Abhängige Variable: Preis

              Koeffizient    Std.-fehler    t-Quotient    p-Wert
  ------------------------------------------------------------------
  const       23183,6         377,445          61,42      1,76e-054 ***
  Alter       -2202,77        217,994         -10,10      2,11e-014 ***
  KM             -0,0215039     0,00704890     -3,051     0,0034    ***

Mittel d. abh. Var.      16140,16   Stdabw. d. abh. Var.     4029,835
Summe d. quad. Res.      95049375   Stdfehler d. Regress.    1280,149
R-Quadrat                0,902451   Korrigiertes R-Quadrat   0,899087
F(2, 58)                 268,2860   P-Wert(F)                4,87e-30
Log-Likelihood          -521,4558   Akaike-Kriterium         1048,912
Schwarz-Kriterium        1055,244   Hannan-Quinn-Kriterium   1051,393
```

29. Für was steht der Störterm u in einem Regressionsmodell? Warum sind die Regressionskoeffizienten mit griechischen Buchstaben bezeichnet?

    Die Regressionskoeffizienten werden mit einem griechischen Buchstaben bezeichnet (Schätzer), um deutlich zu machen, dass das Regressionsmodell die Verhältnisse in der Grundgesamtheit beschreibt.

    Die in einer Stichprobe berechneten Regressionskoeffizienten b_i sind Schätzungen für den statistischen Zusammenhang zwischen den endogenen und exogenen Variablen in der Grundgesamtheit.

    Der Störterm u repräsentiert die nicht im Modell berücksichtigten Einflüsse und mögliche Messfehler. Beide Regressionsmodelle vernachlässigen andere Einflussparameter, welche durch den Störterm verkörpert sind.

    Annahme für die Störterme: Homoskedastizität und serielle Unkorreliertheit.

30. Interpretieren Sie den Regressionskoeffizienten b_2 für beide Modelle.

    Modell 1: Preis = 23‘521.5 – 2‘757.77 Alter

    Modell 2: Preis = 23‘183.6 – 2‘202.77 Alter – 0.0215 KM

    1: Der Autopreis reduziert sich im Durchschnitt um CHF 2‘757.77 pro Jahr.

    2: Wir erwarten einen durchschnittlichen jährlichen Preisrückgang von ca. CHF 2‘202, wenn die Kilometerzahl konstant bleibt (= ceteris paribus).

31. Warum ist ein Unterschied für den Schätzer b_2 zwischen beiden Modellen zu vermerken?

    Kilometerstand und Preis sind hoch korreliert. Wenn wir den Kilometerstand nicht im Modell als Regressor berücksichtigen, kommt im Regressionskoeffizienten für das Alter im Modell A auch auf indirekter Weise der Kilometerstandseffekt zum Ausdruck. Wenn wir den Kilometerstand aber als zusätzlicher Regressor berücksichtigen, wird für diesen Effekt ‘kontrolliert’, d.h. der Regressionskoeffizient des Alters misst dann den jährlichen durchschnittlichen Wertverlust bei konstanter Kilometerzahl.

32. Interpretieren Sie den Regressionskoeffizienten b_3 im Modell 2.

    Der erwartete Wertverlust pro Kilometer beträgt ca. 2.2 Rappen, wenn das **Alter konstant** gehalten wird (= ceteris paribus).

33. Sind die Regressionskoeffizienten im Modell 2 statistisch signifikant? Betrachten Sie dabei jeweils die Sternen, die t-Werte und p-Werte.

	Koeffizient	Std.-fehler	t-Quotient	p-Wert	
const	23183,6	377,445	61,42	1,76e-054	***
Alter	-2202,77	217,994	-10,10	2,11e-014	***
KM	-0,0215039	0,00704890	-3,051	0,0034	***

- Für alle Koeffizienten gibt es drei Sterne → statistisch signifikant auf dem 1%-Niveau
- Alle t-Werte sind grösser als 2 in absoluter Wert → H_0 verwerfen
- Alle p-Werte sind kleiner als 5% → H_0 verwerfen

34. Interpretieren Sie den p-Wert für die Variable *KM*

p-Wert = P[$t_c < t_e$ = -3.051] = 3.4% t_c: kritischer Wert

35. Ermitteln Sie den erwarteten Preis eines Gebrauchtautos mit einem Alter von 4 Jahren und 50‘000 Km.

E(Preis| Alter = 4J, km = 50‘000) = 23‘183.6 – 2‘202.77(4) – 0.0215(50‘000) = CHF 13‘297.50

36. Schätzen Sie das neue Modell 3: Preis = β_1^* + β_2^* Alter + β_3^* KM1000 + u*

Abhängige Variable: Preis

	Koeffizient	Std.-fehler	t-Quotient	p-Wert	
const	23183,6	377,445	61,42	1,76e-054	***
Alter	-2202,77	217,994	-10,10	2,11e-014	***
KM1000	-21,5039	7,04890	-3,051	0,0034	***

Mittel d. abh. Var.	16140,16	Stdabw. d. abh. Var.	4029,835
Summe d. quad. Res.	95049375	Stdfehler d. Regress.	1280,149
R-Quadrat	0,902451	Korrigiertes R-Quadrat	0,899087
F(2, 58)	268,2860	P-Wert(F)	4,87e-30
Log-Likelihood	-521,4558	Akaike-Kriterium	1048,912
Schwarz-Kriterium	1055,244	Hannan-Quinn-Kriterium	1051,393

Model 3

37. Interpretieren Sie den Koeffizienten b_3 im Modell 3.

Der erwartete Wertverlust des Gebrauchtautos pro neue Einheit = tausend Kilometer beträgt ca. CHF 21.5 (bzw. 2.15 Rappen pro km), wenn das **Alter konstant** gehalten wird (= ceteris paribus).

Die Einheit des Regressors KM1000 (= KM/1000) ist in 1000 km eingegeben

Modell 2: Preis = 23‘183.6 –2‘202.77 Alter – 0.0215 KM

Modell 3: Preis = 23‘183.6 -2‘202.77 Alter – 21.5039 KM1000

38. Prüfen Sie den Zusammenhang zwischen b_3 und b_3^*.

Modell 3: b_3^* = -21.5039 = 1000b_3 = 1000 (-0.0215)

39. Schätzen Sie das neue Modell 4: Preis100 = β_1^* + β_2^* Alter + β_3^* KM + u*

Model 4

```
Abhängige Variable: Preis100

              Koeffizient    Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       231,836         3,77445        61,42      1,76e-054 ***
  Alter       -22,0277        2,17994       -10,10      2,11e-014 ***
  KM           -0,000215039   7,04890e-05    -3,051     0,0034    ***

Mittel d. abh. Var.      161,4016   Stdabw. d. abh. Var.     40,29835
Summe d. quad. Res.      9504,937   Stdfehler d. Regress.    12,80149
R-Quadrat                0,902451   Korrigiertes R-Quadrat   0,899087
F(2, 58)                 268,2860   P-Wert(F)                4,87e-30
Log-Likelihood          -240,5404   Akaike-Kriterium         487,0808
Schwarz-Kriterium        493,4134   Hannan-Quinn-Kriterium   489,5626
```

Modell 2: Preis = 23‘183.6 –2‘202.77 Alter – 0.0215 KM

Modell 4: Preis100 = 231.836 -22.0277 Alter – 0.0002150 KM

40. Interpretieren Sie die Koeffizienten b_2 und b_3 im Modell 4.

b_2: Der Autopreis sinkt durchschnittlich um 22.028 neue Einheiten = hundert Franken (= CHF 2‘202.8) pro Jahr, ceteris paribus.

b_3: Wenn der Kilometerstand um 1 km zunimmt, lässt sich ein Preisrückgang von ca. CHF 0.000215*100 = CHF 0.0215 erwarten, ceteris paribus.

41. Prüfen Sie den Zusammenhang zwischen b_i und b_i^* für i = 1, 2,3 (Modell 1 vs Modell 4).

b_1^* = 231.836 = b_1/100 = = -23‘183.6 / 100

b_2^* = - 22.027 = b_2/100 = -2‘202.77 / 100

b_3^* = –0.000215 = b_3 / 100 = –0.0215 / 100

42. Schätzen Sie das neue Modell 5: Preis100 = β_1' + β_2' Alter + β_3' KM1000 + u‘

Model 5

```
Abhängige Variable: Preis100

              Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ----------------------------------------------------------
  const       231,836        3,77445        61,42      1,76e-054 ***
  Alter       -22,0277       2,17994       -10,10      2,11e-014 ***
  KM1000       -0,215039     0,0704890      -3,051     0,0034    ***

Mittel d. abh. Var.      161,4016   Stdabw. d. abh. Var.     40,29835
Summe d. quad. Res.      9504,937   Stdfehler d. Regress.    12,80149
R-Quadrat                0,902451   Korrigiertes R-Quadrat   0,899087
F(2, 58)                 268,2860   P-Wert(F)                4,87e-30
Log-Likelihood          -240,5404   Akaike-Kriterium         487,0808
Schwarz-Kriterium        493,4134   Hannan-Quinn-Kriterium   489,5626
```

43. Interpretieren Sie den Regressionskoeffizienten b_3'.

Wenn der Kilometerstand um eine Einheit = 1000 km zunimmt, lässt sich ein Preisrückgang von ca. CHF 0.215*100 = CHF 21.5 erwarten, ceteris paribus.

44. Prüfen Sie den Zusammenhang zwischen b_i und b_i' für i = 1,2,3. (Modell 1 vs Modell5)

Modell 2: Preis = 23‘183.6 –2‘202.77 Alter – 0.0215 KM

Modell 5: Preis100 = 231.836 -22.027 Alter – 0.215 KM1000

Zusammenhänge:

b_1' = 231.836 = b_1/100 = 23‘183.6 7 /100

b_2' = - 22.027 = b_2/100 = -2‘202.77 / 100

b_3' = -0.215 = b_3 * 10 = -0.0215* 10 (=1000/100)

45. Erklären Sie kurz was das Bestimmtheitsmass ist.

Diese Kennzahl R^2 stellt den Anteil der Varianz der abhängigen Variable y dar, der durch die lineare Regression erklärt wird.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}$$

46. Interpretieren Sie das Bestimmtheitsmass für beide Extremwerte $R^2 = 0$ und $R^2 = 1$. Was ist die Implikation für die RSS und ESS?

$R^2 = 0$ → kein linearer Zusammenhang → ESS (erklärte Streuung) = 0

Wenn eine Regression ein R^2 nahe 0 besitzt, bedeutet dies, dass die gewählten Regressoren nicht gut geeignet sind, die abhängige Variable y zu erklären. In diesem Falle haben wir eine schlechte Modellanpassung ("poor model fit") oder Regressionsgüte.

Achtung! Ein R^2 nahe bei null zeigt an, dass es keinen linearen Zusammenhang zwischen der abhängigen und den unabhängigen Variablen gibt, aber ein nicht-linearer Zusammenhang (z.B. quadratischen Zusammenhang) kann vorliegen!

$R^2 = 0 \Rightarrow S_{ee} = S_{yy} \Leftrightarrow$ RSS = TSS $\Rightarrow$ die nicht erklärte Streuung entspricht der Variation der abhängigen Variable y

$R^2 = 1$ → perfekter linearer Zusammenhang → die Daten liegen auf einer Gerade
Besitzt eine Regression ein R^2 nahe 1, bedeutet dies, dass die Regressoren gut geeignet sind, die abhängige Variable y zu erklären und letztendlich vorherzusagen. Das Modell besitzt eine gute Anpassungsgüte (good model fit) oder gute Regressionsgüte.
$R^2 = 1 \Rightarrow S_{ee} = 0 \Leftrightarrow$ ESS = TSS $\Rightarrow$ alle Residuen sind null!

47. Interpretieren Sie das Bestimmtheitsmass für Modell 2. Weist dieses Modell eine gute Anpassungsgüte auf?

$R^2 = 0.90$: Ca. 90% der Gesamtvariation der Autopreise lässt sich durch das Regressionsmodell erklären → Regressionsmodell hat eine gute Anpassungsgüte.

48. Prüfen Sie die Relation für die Einfachregression (Modell 1): $r_{xy} = \pm\sqrt{R^2}$

Der Korrelationskoeffizient zwischen Autopreis und Alter beträgt r_{xy}= –0.9417 (Frage 18).

$(-0.9417)^2 = 0.88679 = R^2$ vom Modell 1

49. Welche Grenzen besitzt das Bestimmtheitsmass? Nennen Sie drei Kritikpunkte.

- Das Bestimmtheitsmass zeigt zwar die Qualität der linearen Approximation, jedoch nicht, ob das Modell richtig spezifiziert wurde. Ein falschspezifiziertes Modell kann ein hohes R^2 aufweisen, obwohl es unbrauchbar ist.
- Ein hohes R^2 erlaubt nicht immer eine gute Vorhersage der abhängigen Variable y!
- Das Hinzufügen neuer Regressoren erhöht R2, auch wenn diese irrelevant wären.

50. Hat sich das Bestimmtheitsmass für die verschiedenen Skalierungen geändert?

R^2 hat sich nicht geändert! Eine neue Skalierung der Variablen hat keinen Einfluss auf R^2.

	Modell 2	Modell 3	Modell 4
Variablen	Preis, Alter, KM	Preis, Alter, KM1000	Preis, Alter und KM
R^2	0.9024	0.9024	0.9024

51. Vergleichen Sie die adjustierten R^2-Werte für beide Modelle 1 und 2. Welches Modell würden Sie anhand dieses Kriteriums vorziehen?

	Modell 1: Alter	Modell 2: Alter und KM
R^2	0.886	0.9024
Adjust. R^2	0.884	0.899

Das adjustierte Bestimmtheitsmass hat sich vergrössert.
Modell 2 ist aufgrund des höheren adjustierten Bestimmtheitsmasses vorzuziehen.

52. Erklären Sie kurz warum R^2 durch das Hinzufügen eines weiteren Regressors nicht geringer wird.

Im ungünstigsten Fall ist der geschätzte Regressionskoeffizient des zusätzlichen Regressors nicht von null verschieden und die Streuung der Residuen bleibt unverändert. In der Regel steigt R^2, da die Streuung der Residuen geringfügig sinkt.

53. Was ist der Vorteil des adjustierten $\overline{R}^2$ gegenüber R^2?

Das adjustierte R^2 berücksichtigt sowohl die Modellanpassung als auch die Sparsamkeit des Modells und erlaubt den Vergleich zwischen Modellen mit unterschiedlicher Anzahl Regressoren.
Es besteht aus dem Wert des einfachen R^2 welcher mit einem "Strafterm" belegt wird. Daher nimmt das korrigierte R^2 in der Regel einen geringeren Wert als das einfache R^2 an und kann in manchen Fällen sogar negativ werden.

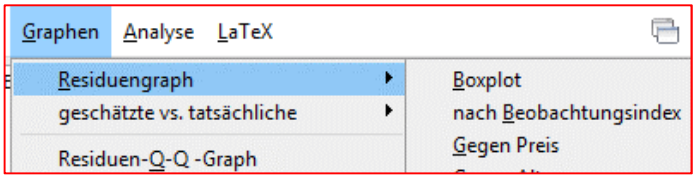
54. Erklären Sie kurz was der Strafterm ist und wie er funktioniert.

$$\overline{R}^2 = 1 - \frac{N-1}{N-k}\frac{S_{ee}}{S_{yy}}$$

Der Strafterm beträgt (N-1) / (k-1) wobei k die Anzahl Regressor darstellt. Er steigt mit der Anzahl der unabhängigen Variablen (k) → N-k ↓ und (N-1) / (N-k) steigt
Durch Hinzunahme einer neuen Variablen kann das Modell im Sinne des korrigierten R^2 nur dann verbessert werden, wenn der zusätzliche Erklärungsgehalt (S_{ee} ↓) den Strafterm mehr als ausgleicht.
Fazit: $\overline{R}^2$ ist zwar nicht direkt wie das normale R^2 als Prozentsatz an erklärter Varianz der abhängigen Variable y zu interpretieren, berücksichtigt und bestraft aber die Anzahl an unabhängigen Variablen im Modell.

55. Erstellen Sie den Residuengraph für Regressionsmodell 2:

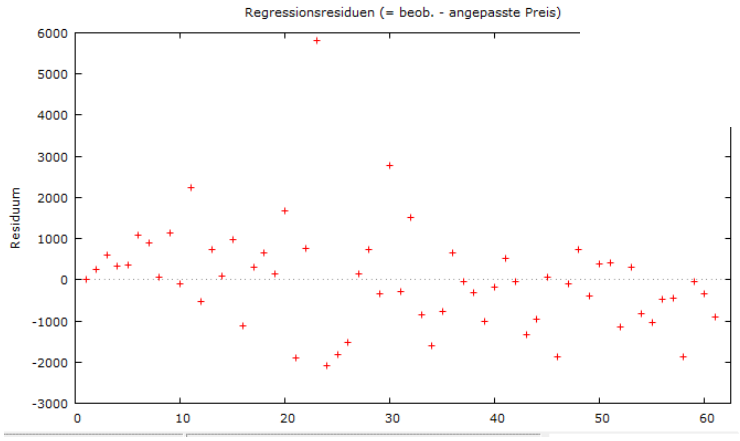
gretl Output-Fenster: Graphen / Residuengraph

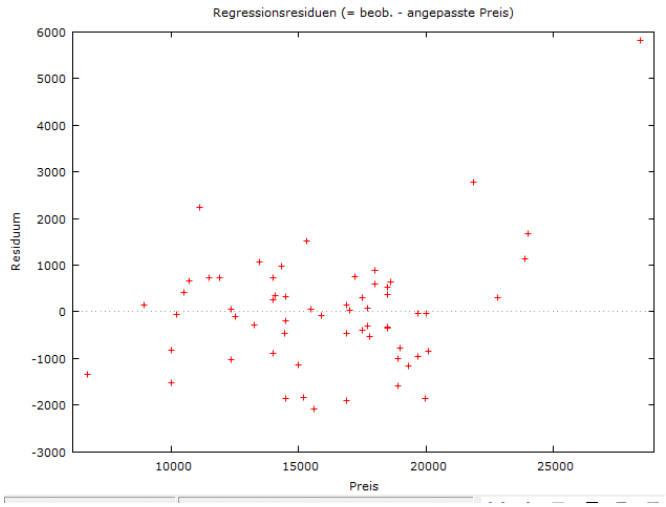


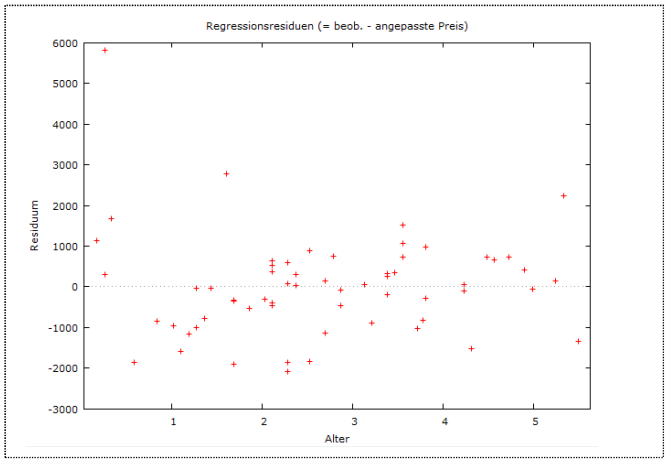
i) Nach Beobachtungsindex

ii) Gegen Alter

iii) Gegen Preis





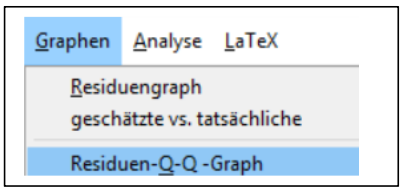


ii. Hohe Volatilität der Residuen bei geringem Alter (wenn Auto fast neu ist)

iii. Zunehmende Streuung der Residuen bei zunehmender Preis

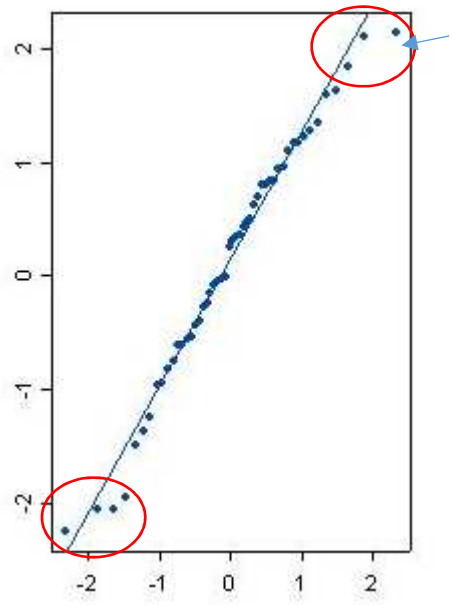
56. Erklären Sie kurz was ein QQ-Plot (Quantil-Quantil Plot) ist.

gretl Output-Fenster: Graphen / Residuen QQ-Graph



Grafisches Werkzeug, in dem die Quantile zweier statistischer Variablen gegeneinander abgetragen werden, um ihre Verteilungen zu vergleichen.

57. Erstellen Sie ein QQ-Plot mittels gretl.



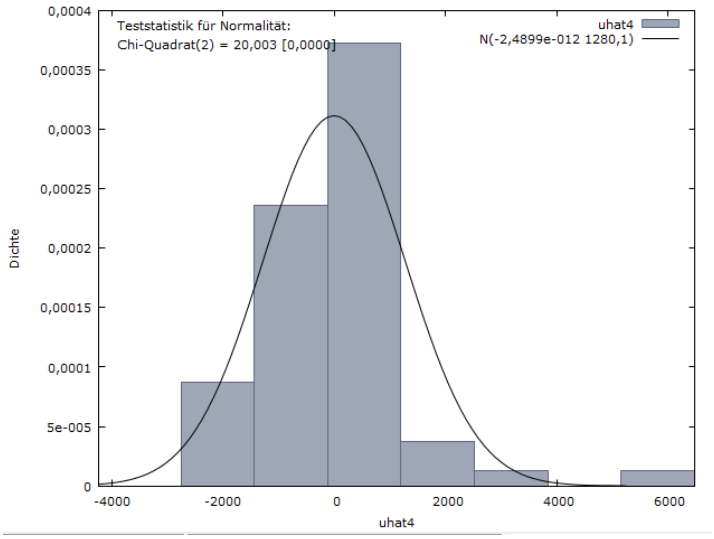
An den Extremitäten ist die Normalität der Residuen verletzt

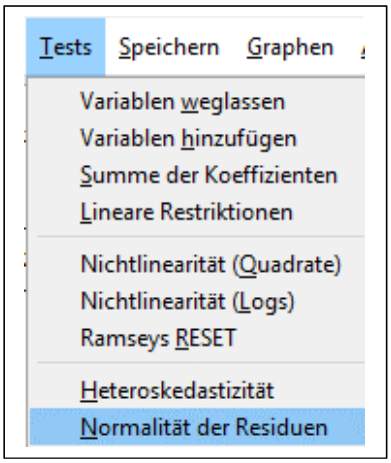
gretl Output-Fenster: Graphen / Residuen QQ-Graph

Wenn die Residuen normalverteilt sind, sollten sie auf einer Gerade liegen.

58. Testen Sie die Normalität der Residuen des Modells B.

gretl Output-Fenster: Tests / Normalität der Residuen





Nullhypothese H_0: Die Residuen sind normalverteilt.

Der Wert der JB-Statistik liegt über 4.6 → H_0 wird verworfen → die Residuen sind nicht normalverteilt

59. Welche Kritik können Sie an diesem Model üben?

Dieses Modell berücksichtigt nur zwei erklärenden Variablen. Möglicherweise sind auch qualitative Merkmale des Gebrauchtautos von Bedeutung für den Autopreis.