Übung 1: OLS-Regression für Autopreise

1. Geben Sie für die folgenden Merkmale das jeweilige Skalenniveau und mögliche Merkmalausprägungen. Unterscheiden Sie die Merkmale ferner in diskrete und stetige und diskutieren Sie dabei Probleme der Messgenauigkeit.
   a. Gewicht
   b. Akademischer Grad (Hochschulabschluss)
   c. Augenfarbe
   d. Geschlecht
   e. Nettoeinkommen in CHF

	Merkmal	Ausprägungen	Skalenniveau	Diskret?
a				
b				
c				
d				
e				

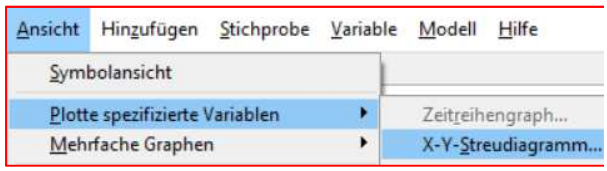
2. Welche Faktoren bestimmen den Verkaufspreis eines Gebrauchtautos. Welche Vorzeichen erwarten Sie?

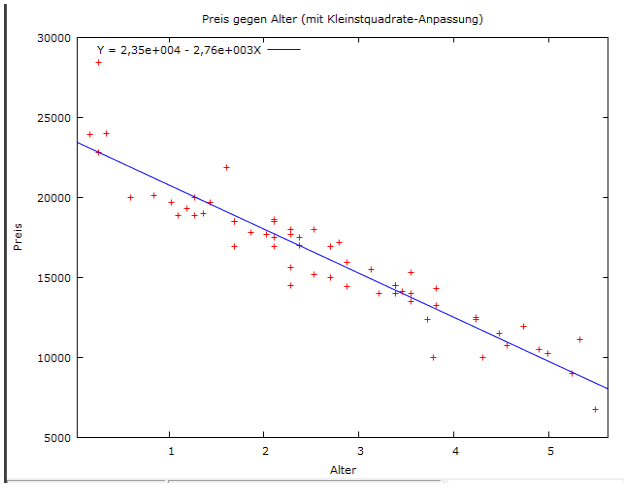
3. Welche sind davon qualitative Faktoren?

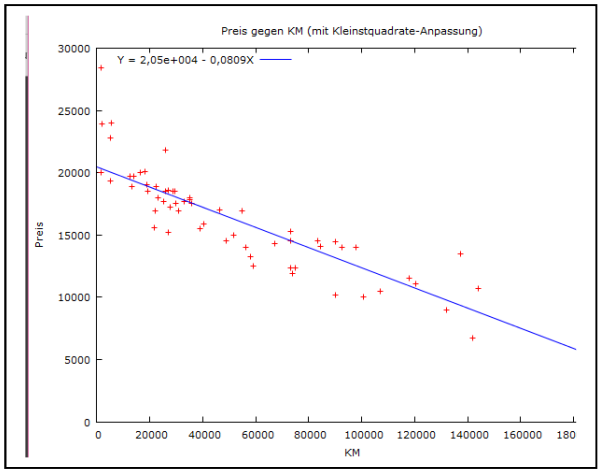
4. Erklären Sie was ein Streudiagramm ist.

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

   Hinweis: Y-Achsen-Variable: Preis







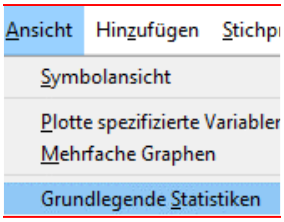
6. Was sagen diese Streudiagramme über den statistischen Zusammenhang zwischen den Autopries und den ausgewählten Variablen (Alter, KM) aus?

7. Was sind der mittlere 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



8. Interpretieren Sie den Median für den Regressor “Kilometerstand“ KM.

9. Welche wichtige Information gibt die Standardabweichung im Allgemeinen?

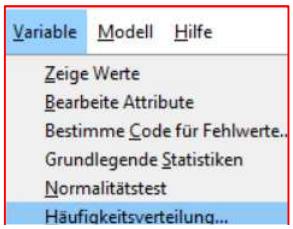
10. Was ist der Vorteil der Standardabweichung gegenüber der Varianz als Streuungsmass?

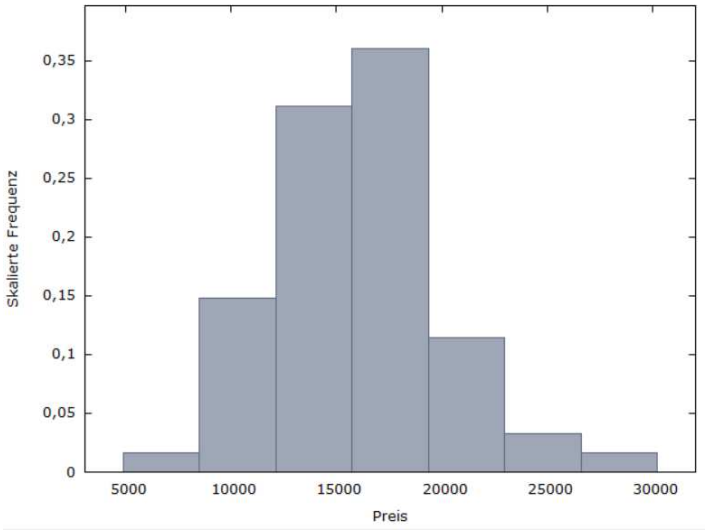
11. Welche Variable weist die geringste und höchste Standard Abweichung auf? Was können Sie über die Repräsentativität des Mittelwertes dieser Variablen sagen?

12. Interpretieren Sie die Standardabweichung für den Regressor Alter.

13. Erklären Sie was ein Histogramm ist.

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



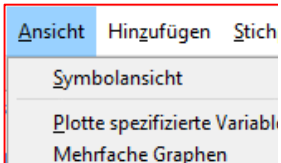


15. Welche ist die modale Klasse dieses Histogramms?

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

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

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
```

19. Erklären Sie kurz was der Variationskoeffizient ist.

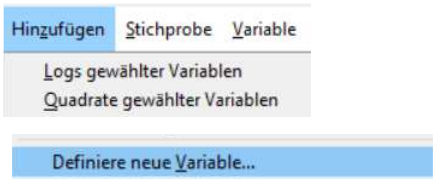
20. Was ist der Vorteil des Variationskoeffizienten gegenüber der Standardabweichung?

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

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

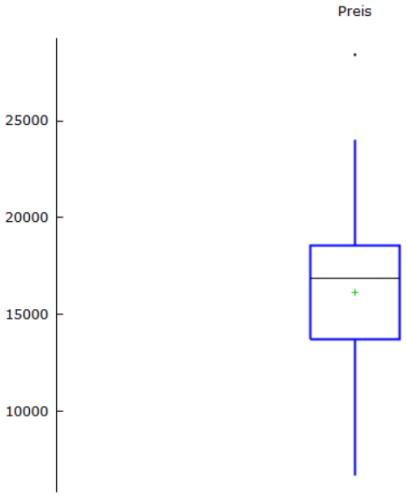
    Beispiel: KM1000 = KM / 1000

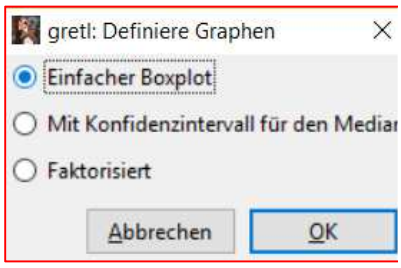


23. Vergleichen Sie die Standardabweichungen und Variationskoeffizienten für folgende Grössen: Preis – Preis100 und KM – KM1000

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

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 _____gegeben.
- Die Länge der Box entspricht _____
- Die Linie innerhalb der Box gibt die Lage _____ wieder.
- Das grüne Kreuz innerhalb der Box entspricht _____.
- Der Punkt oberhalb der Box entspricht _____.

25. Erklären Sie was Schiefe ist.

26. Erklären Sie was Kurtosis (Wölbung) ist. Wie ist der Exzess definiert?

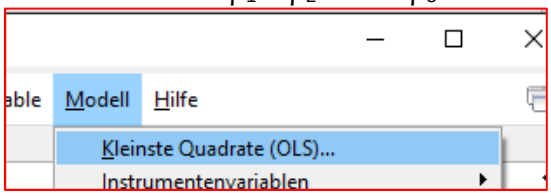
27. Analysieren Sie kurz die Wölbung und Kurtosis für folgende Variablen: Preis, Preis100, KM und KM1000

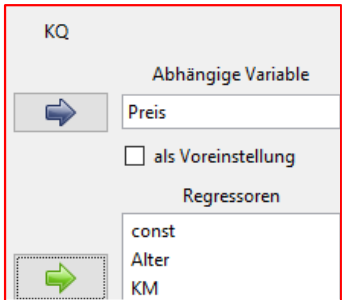
	Schiefe	Überwölbung
Preis	0,24267	0,43224
KM	1,0238	0,45949
Preis100	0,24267	0,43224
KM1000	1,0238	0,45949

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





Model 1

	Koeffizient	Std.-fehler	t-Quotient	p-Wert	
const	-15712,1	6664,31	-2,358	0,0217	**
Alter	25809,2	2218,17	11,64	6,53e-017	***

Mittel d. abh. Var.	53367,72	Stdabw. d. abh. Var.	42556,40
Summe d. quad. Res.	3,30e+10	Stdfehler d. Regress.	23643,60
R-Quadrat	0,696472	Korrigiertes R-Quadrat	0,691328
F(1, 59)	135,3810	P-Wert(F)	6,53e-17
Log-Likelihood	-699,8602	Akaike-Kriterium	1403,720
Schwarz-Kriterium	1407,942	Hannan-Quinn-Kriterium	1405,375

Model 2

	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?

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

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

32. Interpretieren Sie den Regressionskoeffizienten b_3 im Modell 2.

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	***

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

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

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.

    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^*.

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 b2 und b3 im Modell 4.

41. Prüfen Sie den Zusammenhang zwischen b_i und b_i^* für i = 1, 2,3.

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'.

44. Prüfen Sie den Zusammenhang zwischen b_i und b_i' für i = 1,2,3.

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

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

45. Erklären Sie kurz was das Bestimmtheitsmass ist.

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

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

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

49. Welche Grenzen besitzt das Bestimmtheitsmass?

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

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	0.902
Adjust. R^2	0.884	0.899

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

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

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

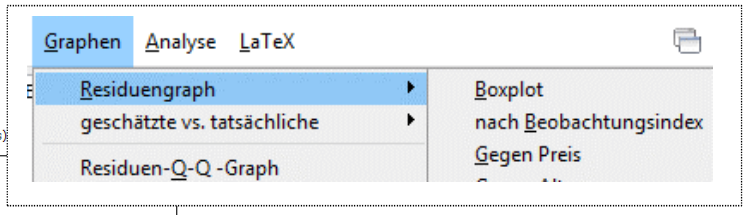
55. Erstellen Sie den Residuengraph für Regressionsmodell 2:

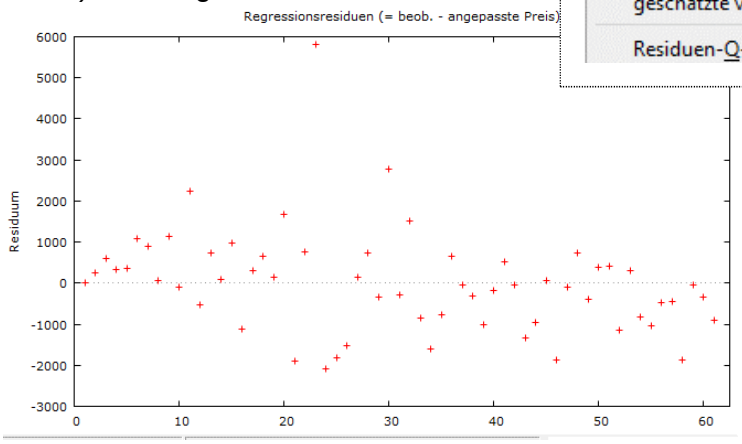
    gretl Output-Fenster: Graphen / Residuengraph

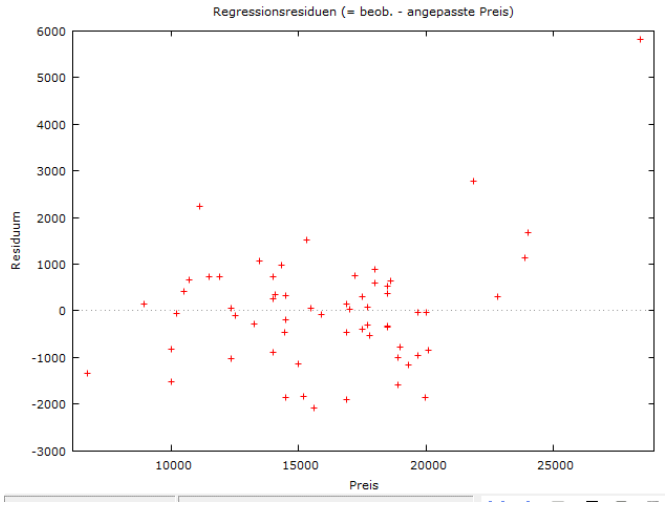
    i) Nach Beobachtungsindex

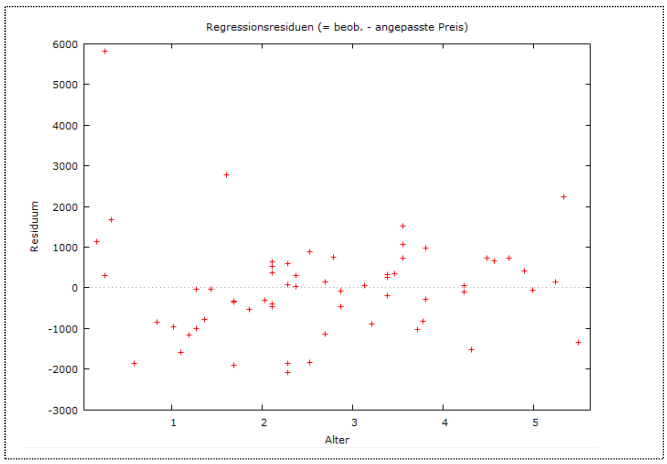
    ii) Gegen Alter

    iii) Gegen Preis







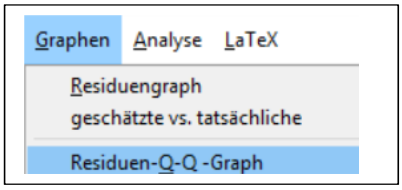


ii. Hohe Volatilität der Residuen bei geringem Alter (wenn Auto fast neu ist)

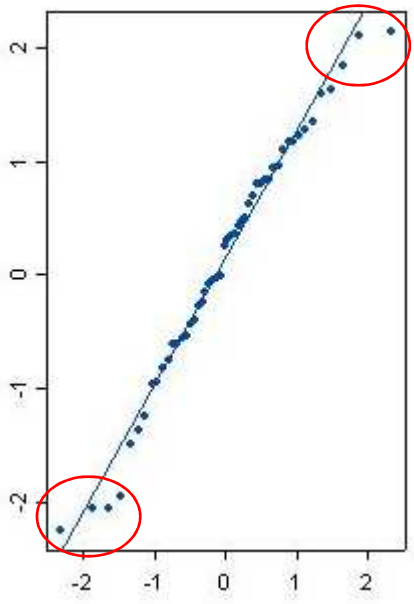
iii. Zunehmende Streuung der Residuen bei zunehmender Preis

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

gretl Output-Fenster: Graphen / Residuen QQ-Graph



57. Erstellen Sie ein QQ-Plot mittels gretl.

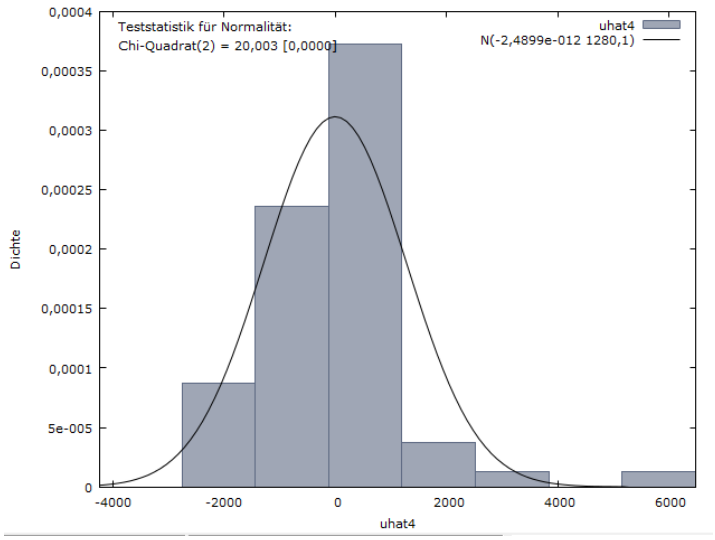


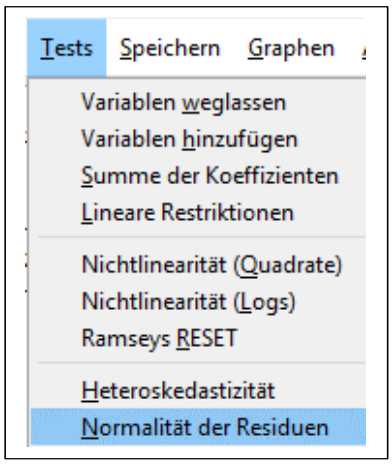
gretl Output-Fenster: Graphen / Residuen QQ-Graph

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





59. Welche Kritik können Sie an diesem Model üben?