Regression

Bei vielen Untersuchungen werden simultan metrische Merkmale an Personen oder Objekten erhoben und man interessiert sich für einen möglichen Zusammenhang dieser Variablen. Regression ist ein Verfahren zur Untersuchung der Art eines solchen Zusammenhangs zwischen einer metrisch skalierten abhängigen Variablen (A) und einer (oder mehreren) metrischen unabhängigen Variablen (B). Die Zuordnung in abhängige und unabhängige Variable ergibt sich meist auf natürliche Weise, wenn eine Ursache - Wirkungs - Beziehung vorliegt (z.B. Größe und Gewicht von Personen) und muss ansonsten festgelegt werden . So ist das Gewicht sicher nicht Ursache für die Größe einer Person, umgekehrt ist es vorstellbar.

Typische Fragestellungen sind:
- Hängt A von B ab {A = f(B)} - einfache Regressionsanalyse ?
- Wie ändert sich A, wenn sich B ändert ?
- Gibt es außer B noch weitere Einflussgrößen auf A
  {A = f(B_1 , B_2 , ..., B_n)} - multiple Regressionsanalyse ?
- Wie ändert sich A, wenn sich mehrere Einflussgrößen B_i ändern ?

Gesucht wird eine mathematische Funktion, die den Zusammenhang zwischen den Variablen gut beschreibt. Ist eine solche Funktion (Modell) gefunden, läßt sich die eine Variable aus der oder den anderen Variablen vorhersagen. Das einfachste Modell ist die lineare Regression der Form $Y = a + b \cdot X$.

Lineare Einfachregression

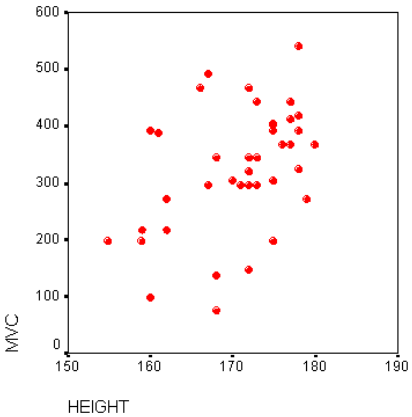
- Festlegung der abhängigen Y und unabhängigen X Variablen vorab,
- Darstellung der Meßpunkte im X - Y - Diagramm (Scatterplot) vor jeder weiteren Rechnung (!).
  Kann ein linearer Zusammenhang $Y = a + b \cdot X + e$ angenommen werden ?
  e = Abweichungen der empirischen y - Werte von der angenommenen Geraden infolge anderer Einflußgrößen und Meßfehler (Residuen).
- Auffinden einer linearen Funktion, die den Zusammenhang optimal beschreibt durch die Methode der kleinsten Quadrate ($\sum e_i^2$ = Min.).
  $\mathbf{y_i^* = a + b \cdot x_i}$ (Regressionsfunktion) mit $b = s_{xy} / s_x^2$ (Steigung der Geraden) und $a = \bar{y} - b \cdot \bar{x}$ (Achsenabschnitt). Es sind s_{xy} und s_x^2 die Covarianz und die Varianz der Beobachtungen.
- Damit schätzt man den Wert $\mathbf{y_i^*}$ (der * steht für Schätzung) der abhängigen Variablen Y für ein beliebiges x_i aus dem Beobachtungsbereich mit Hilfe des Modells.
- Die Güte der Schätzung und damit die Güte der Anpassung der Regressionsfunktion an die beobachteten Werte wird durch das Bestimmtheitsmaß R^2 beschrieben: $R^2 = (b \cdot s_x / s_y)^2$.
- Prüfung der Modellvoraussetzungen, Berechnung der Konfidenzintervalle und Prognoseintervalle.

1. Beispiel für eine lineare Einfachregression

(Datei muscle.dct von Martin Bland's Home Page: https://www-users.york.ac.uk/~mb55/)

An einer Stichprobe von 41 Männern wurden folgende Daten erhoben:
X - Variable : height "Height (cm)"
Y - Variable : mvc "Max voluntary contraction, quadriceps muscle (newtons N)"

Es interessiert die Frage, ob die Größe eines Mannes und die Stärke seiner Quadriceps - Muskelkontraktion in einem Zusammenhang stehen und ob man die Muskelkontraktion aus der Größe vorhersagen kann.



Der Scatterplot läßt einen sehr schwachen Zusammenhang vermuten. Die Berechnung des Korrelationskoeffizienten nach Pearson ergibt r = 0,42 mit p = 0,006 (signifikant) und einem sehr breiten Konfidenzintervall (WinPepi) von (0,13 ; 0,64). Annahme der Normalverteilung **beider** Variablen für Fisher' z-Transformation nicht erfüllt (Shapiro-Wilk: p = 0,006 (height) und p = 0,362 (mvc)). Mit der Kurvenanpassung von SPSS erhält man für R^2 : linear: 0,176 ; quadratisch: 0,177

Ein lineares Modell ist für diese Daten weniger gut geeignet.

Ein linearer Zusammenhang soll für die Beispielrechnung dennoch angenommen werden. Die Regressionskoeffizienten werden so bestimmt, daß die Summe der Abweichungsquadrate (SAQ) zwischen Beobachtungswert und Regressionslinie minimal ist. Mit den Residuen $e_i = (y_i - y_i^*)$ und $y_i^* = a + b \cdot x_i$ erfolgt die Bestimmung der a und b aus den partiellen Ableitungen von SAQ nach a und b. Setzt man diese gleich 0 so erhält man die Lösung aus dem sich ergebenden Gleichungssystem (siehe z.B. Bleymüller, Weißbach):
$SAQ = \sum e_i^2 = \sum (y_i - y_i^*)^2 = \sum (y_i - a - b \cdot x_i)^2 = \text{Minimum}$ (KQ - Methode)
Man erhält für eine Berechnung „per Hand“ (siehe Tabellenkalkulation unten):

$$a = \frac{\sum x_i^2 \cdot \sum y_i - \sum x_i \cdot \sum x_i y_i}{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2} \quad \text{und} \quad b = \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2} \quad \text{oder in anderer Form}$$

$$a = \bar{y} - b \cdot \bar{x} \quad \text{und} \quad b = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \quad \text{aus einer entsprechenden } \textbf{Arbeitstabelle} \text{ (siehe unten).}$$

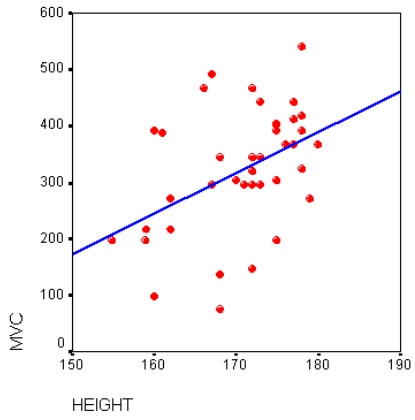
Für die Berechnung der Varianzen von a und b benötigt man die Varianz der Residuen $s_{res}^2 = 1/(n-2) \sum e_i^2$
$\mathbf{s_{res}^2 = (n-1)/(n-2) \cdot (s_y^2 - b^2 \cdot s_x^2)}$ wobei s_y^2 und s_x^2 die Varianzen der Variablen Y und X sind. In SPSS wird s_{res} auch als Standardfehler (SE) des Schätzers bezeichnet. Für die Regressionskoeffizienten erhält man z.B. mit SPSS den folgenden Ausdruck

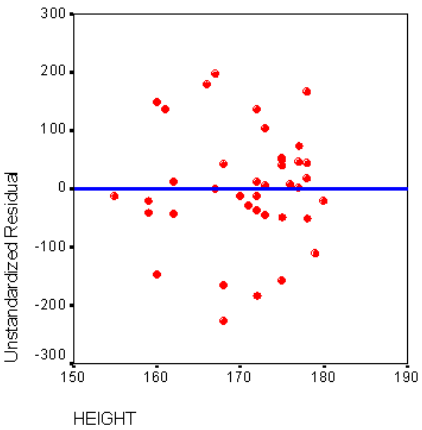
Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-907,626	426,612		-2,128	,040	-1770,530	-44,722
	HEIGHT	7,203	2,497	,419	2,885	,006	2,152	12,253

a. Abhängige Variable: MVC

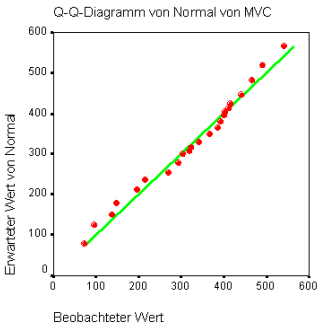
Die Regressionsfunktion lautet somit: **mvc* = - 907,63 + 7,2·height.**
Ein Mann mit einer Größe von 170 cm hat eine geschätzte Kontraktionskraft des Quadriceps von 316,4 N, beobachtet wurden 304 N, somit können -12,4 N (Residuum) nicht durch das lineare Modell erklärt werden.

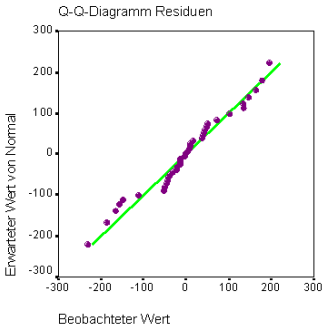




Prüfung der Modellvoraussetzungen

- Linearer Zusammenhang - grafische Prüfung und R^2.
  $R^2 = 0,176$ (sehr schwach, Modell nicht optimal)
- Normalverteilung der abhängigen Variablen MVC - grafisch Q-Q-Plot, und Shapiro-Wilk Test
  Shapiro-Wilk: p = 0,362 (Normalverteilung)





- Normalverteilung der Residuen - grafisch Q-Q-Plot, Shapiro-Wilk Test: p = 0,237 (NV)

Varianzhomogenität der Residuen - grafisch Residuen-Plot (linkes Bild).
Die Verteilung der Residuen entlang der Achse ist ähnlich. (Homoskedastizität)
Fazit: Die Modellvoraussetzungen sind als erfüllt anzusehen.

Konfidenzintervalle (C.I.) für die Regressionskoeffizienten

Mit s_y^2 = 12.583,61 und s_x^2 = 42,651 ergibt sich s_{res}^2 = 40/39 · (12.583,61 - (7,203)²·42,651) = 10.636,656 (siehe Tabelle unten).

C.I. für b:

Geschätzte Varianz von b: $\mathbf{s_b^2 = s_{res}^2 / (n-1)\cdot s_x^2}$ = 10.636,656/40·42,651 = 6,2347

Quantil der t-Verteilung $t_{0,975,39}$ = 2,0227 ; s_b = 2,497 ; b = 7,203

C.I. = (b - 2,0227·s_b ; b + 2,0227·s_b) = **(2,152 ; 12,254)** , entspricht dem obigen SPSS Ausdruck.

C.I. für a:

Geschätzte Varianz von a: $\mathbf{s_a^2 = s_b^2 \cdot \sum x_i^2 / n}$ = 6,2347·29.190,9268 = 181.996,6713

Quantil der t-Verteilung $t_{0,975,39}$ = 2,0227 ; s_a = 426,6107 ; a = - 907,626

C.I. = (a - 2,0227·s_a ; a + 2,0227·s_a) = **(-1770,531 ; -44,721)** , entspricht dem obigen SPSS Ausdruck.

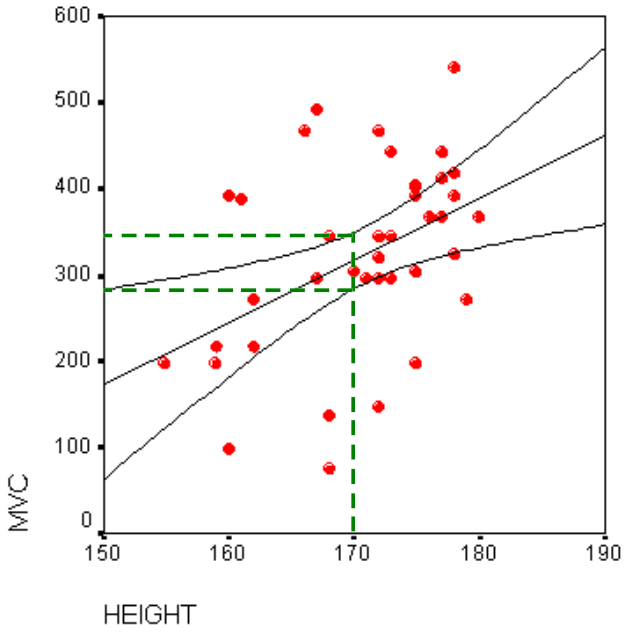
Konfidenzintervalle (C.I.) für die geschätzten Werte der abhängigen Variablen
(Konfidenzband für die Regressionsgerade)

Die Regressionsfunktion lautet: **mvc* = - 907,63 + 7,2·height .** mvc* ist ein durchschnittlicher Prognosewert. Für height = x_0 = 170 cm wird mvc* = 316,4 N geschätzt. Das Konfidenzintervall für den geschätzten Wert an der Stelle x_0 erhält man mit der Varianz der Schätzung des durchschnittlichen Prognosewertes:

$$s^2_{mvc^*} = s^2_{res} \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\cdot s_x^2} \right)$$ (Herleitung z.B. in Bleymüller, Weißbach: Statistik für Wirtschaftswissenschaftler).

aus (mvc* ± $t_{0,975,39}$·s_{mvc^*}) und mit $s^2_{mvc^*}$ = 10.636,656 · (1/41 + (170 - 170,732)² / 40· 42,651) = 262,77

C.I. = (316,4 - 2,0227·16,21 ; 316,4 + 2,0227·16,21) = **(283,61 ; 349,19)** entspricht SPSS bis auf Rundungsfehler, Bezeichnung hier (lmci ; umci).



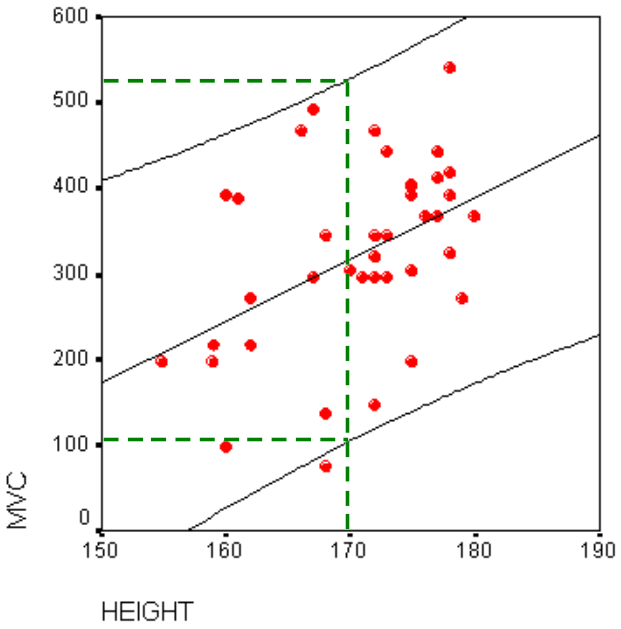
Mit 95% iger Wahrscheinlichkeit wird der wahre Mittelwert von **mvc** in der Grundgesamtheit der Männer bei einer Größe (Height) von 170 cm vom obigen C.I. überdeckt.

Konfidenzintervalle (C.I.) für die individuellen Y - Werte
(Prognoseband für die Regressionsgerade)

Während das Konfidenzband für jeden X - Wert ein Intervall angibt, in dem mit vorgegebener Wahrscheinlichkeit der mittlere Prognosewert der Grundgesamtheit liegt (ähnlich „C.I für einen geschätzten Mittelwert“), gibt das Prognoseband ein Intervall an, in dem ein vorgegebener Anteil der geschätzten individuellen Werte liegt. Es wird die Frage beantwortet, in welchem Intervall mit z.B. 95% iger WSK bei einer Größe der Männer von 170 cm ein einzelner Wert der Muskelkontraktion (MVC) zu erwarten ist.

Für die Varianz des individuellen Prognosewertes gilt: $$s^2_{mvc} = s^2_{res} \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\cdot s_x^2} \right)$$

mit s^2_{mvc} = 10.636,656 · (1 + 1/41 + (170 - 170,732)^2 / 40· 42,651) = 10.899,426 ergibt sich
P.I. = (316,4 - 2,0227·104,4 ; 316,4 + 2,0227·104,4) = **(105,23 ; 527,57)** entspricht SPSS bis auf Rundungsfehler, Bezeichnung hier (lici ; uici).



SPSS - Tabelle:

height	mvc	residuen	lmci	umci	lici	uici
170	304	-12,852	284,063	349,640	105,681	528,022

SPSS liefert diese C.I. nur für Beobachtungswerte, nicht für Zwischenwerte. Hierfür wäre das Programm **CIA** nützlich (siehe Literatur und Software).

Berechnungen mit einer Tabellenkalkulation

(Datei muscle.dct von Martin Bland's Home Page: https://www-users.york.ac.uk/~mb55/)

Nr.	height (x)	mvc (y)	x^2	x·y	Nr.	height (x)	mvc (y)	x^2	x·y
1	166	466	27556	77356	23	171	294	29241	50274
2	175	304	30625	53200	24	162	270	26244	43740
3	173	343	29929	59339	25	177	368	31329	65136
4	175	404	30625	70700	26	177	441	31329	78057
5	172	147	29584	25284	27	178	392	31684	69776
6	172	294	29584	50568	28	167	294	27889	49098
7	160	392	25600	62720	29	176	368	30976	64768
8	172	147	29584	25284	30	159	216	25281	34344
9	179	270	32041	48330	31	173	294	29929	50862
10	177	412	31329	72924	32	175	392	30625	68600
11	175	402	30625	70350	33	172	466	29584	80152
12	180	368	32400	66240	34	170	304	28900	51680
13	167	491	27889	81997	35	178	324	31684	57672
14	175	196	30625	34300	36	155	196	24025	30380
15	172	343	29584	58996	37	160	98	25600	15680
16	172	319	29584	54868	38	162	216	26244	34992
17	161	387	25921	62307	39	159	196	25281	31164
18	173	441	29929	76293	40	168	137	28224	23016
19	173	441	29929	76293	41	168	74	28224	12432
20	168	343	28224	57624	Summe	7000	13207	1196828	2267142
21	178	540	31684	96120	Mittelwert	170,73171	322,12195	29190,92683	55296,14634
22	178	417	31684	74226	Varianz	42,65122	12583,60976		

Für die Regressionskoeffizienten erhält man aus den Werten der Tabelle:

$$b = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}$$ = (55.296,14634 - 170,73171 · 322,12195) / (29.190,92683 - (170,73171)²) = 7,20295

$$a = \overline{y} - b \cdot \overline{x}$$ = 322,12195 - 7,20295 · 170,73171 = - 907,650

Aus der Tabelle: s_x^2 = 42,65122 s_y^2 = 12.583,60976 daraus
$s_{res}^2 = (n-1)/(n-2) \cdot (s_y^2 - b^2 \cdot s_x^2)$ = 40/39 · (12.583,60976 - 7,20295 · 42,65122) = 10.636,6752 daraus

$s_b^2 = s_{res}^2 / (n-1) \cdot s_x^2$ = 10.636,6752 / 40 · 42,65122 = 6,23468

$s_a^2 = s_b^2 \cdot \sum x_i^2 / n$ = 6,23468 · 29.190,92683 = 181.996,0877

Die Konfidenzintervalle berechnen sich analog dem obigen Muster.
Übereinstimmung mit obigen Werten bis auf Rundungsfehler.
Bei Verwendung eines Taschenrechners sollte möglichst mit 5 Dezimalstellen gerechnet werden.

Anhang: SPSS - Ausgaben

Deskriptive Statistiken

	Mittelwert	Standardabweichung	N
MVC	322,12	112,177	41
HEIGHT	170,73	6,531	41

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,419[a]	,176	,155	103,135

a. Einflußvariablen : (Konstante), HEIGHT

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	88510,562	1	88510,562	8,321	,006[a]
	Residuen	414833,829	39	10636,765		
	Gesamt	503344,390	40			

a. Einflußvariablen : (Konstante), HEIGHT

b. Abhängige Variable: MVC

Koeffizienten[a]

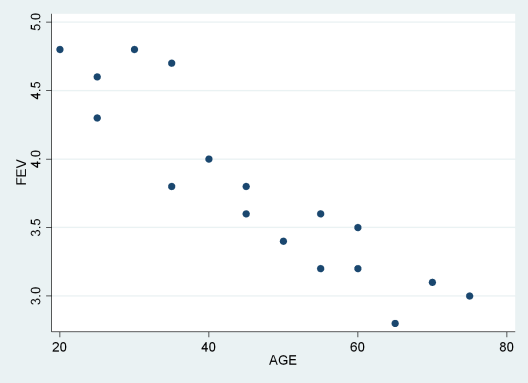
Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
		B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-907,626	426,612		-2,128	,040	-1770,530	-44,722
	HEIGHT	7,203	2,497	,419	2,885	,006	2,152	12,253

a. Abhängige Variable: MVC

2. Beispiel für eine lineare Einfachregression

An einer Stichprobe von 18 Personen wurden folgende Daten erhoben:
X - Variable : age
Y - Variable : fev "forced expiratory volume in L/s"

Es interessiert die Frage, ob das Alter Erwachsener und das forciertes Ausatmungsvolumen in 1 sek. in einem Zusammenhang stehen und ob man fev aus dem Alter vorhersagen kann.



Der Scatterplot läßt einen starken Zusammenhang vermuten. Die Berechnung des Korrelationskoeffizienten nach Pearson ergibt r = 0,92 mit p = 0,000 (signifikant) und einem Konfidenzintervall (WinPepi) von (0,79 ; 0,97). Die Annahme der Normalverteilung **beider** Variablen für Fisher' z-Transformation ist erfüllt (Shapiro-Wilk: p = 0,653 (FEV) und p = 0,905 (AGE)). Mit der Kurvenanpassung von SPSS erhält man für R^2 : linear: 0,840 ; quadratisch: 0,852
Das lineare Modell ist gut geeignet.

Die Regressionskoeffizienten erhält man z.B. aus STATA :

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
AGE	-.036129	.0039413	-9.17	0.000	-.0444842	-.0277739
_cons	5.454839	.1920743	28.40	0.000	5.047659	5.862018

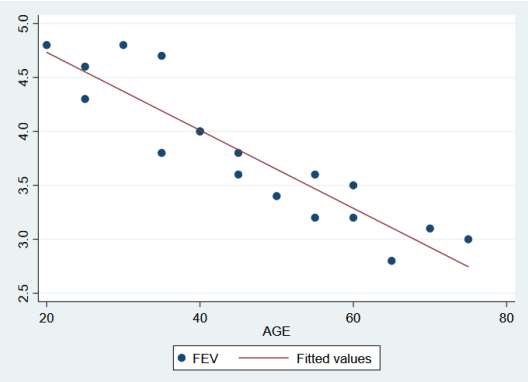
Die Regressionsfunktion ($y_i^* = a + b \cdot x_i$) lautet somit: **fev* = 5,455 - 0,036·age**

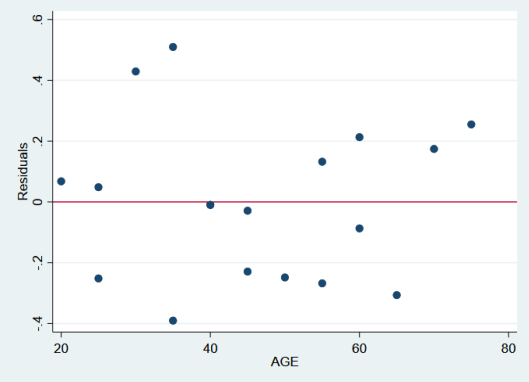
Eine Person im Alter von 60 Jahren hat ein geschätztes Ausatmungsvolumen in 1s von 3,3 L/s, das zwischen den beobachteten Werten von 3,2 L/s und 3,5 L/s liegt. Die geringen Abweichungen von 0,1 bzw. 0,2 L/s können vom Modell nicht erklärt werden.

Die **Konfidenzintervalle für die Regressionskoeffizienten** erhält man aus obiger STATA-Tabelle :
a = 5,455 (5,048 ; 5,862) und b = -0,036 (-0,044 ; -0,028)

Prüfung der Modellvoraussetzungen

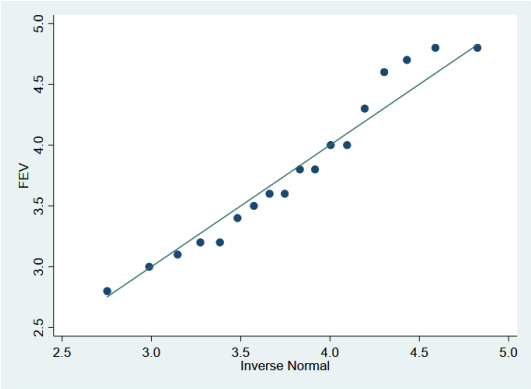
- Linearer Zusammenhang (linke Grafik): grafische Prüfung und Wert von R^2.
  R^2 = 0,840 (starker Zusammenhang, lineares Modell ist gut geeignet)



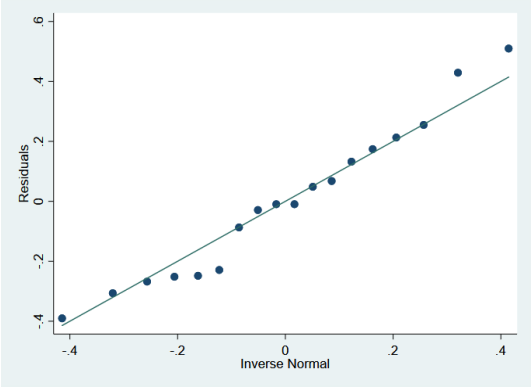


- Varianzhomogenität der Residuen (rechte Grafik): - grafische Prüfung im Residuen-Plot
  Die Verteilung der Residuen entlang der Achse ist ähnlich (Homoskedastizität, aber geringe Fallzahl).

- Normalverteilung der abhängigen Variablen FEV: grafisch Q-Q-Plot, und Shapiro-Wilk Test
  Shapiro-Wilk: p = 0,653 (Normalverteilung)



- Normalverteilung der Residuen: grafisch Q-Q-Plot und Shapiro-Wilk Test: p = 0,613 (Normalverteilung)

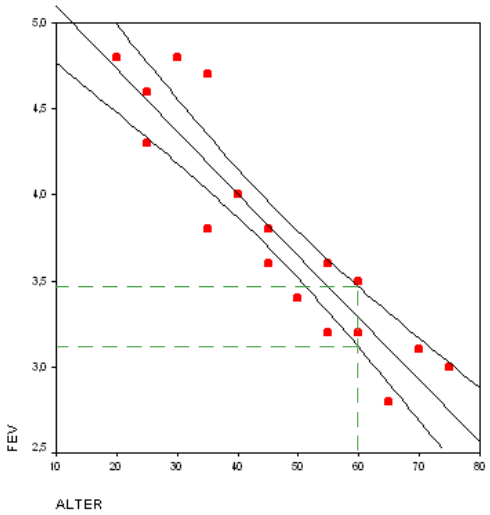


Fazit: Die Modellvoraussetzungen sind als erfüllt anzusehen.

Datensatz

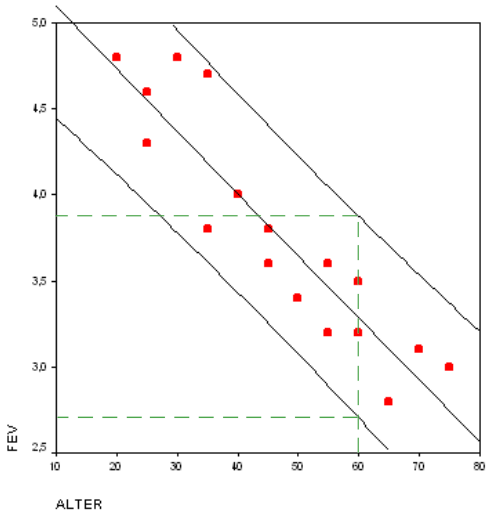
	FEV	AGE
1	3.5	60
2	3.2	55
3	2.8	65
4	3.2	60
5	3.4	50
6	3.6	45
7	3.8	35
8	4.3	25
9	3.8	45
10	3.1	70
11	3.6	55
12	4.0	40
13	3.0	75
14	4.0	40
15	4.8	20
16	4.8	30
17	4.6	25
18	4.7	35

Konfidenzband für die Regressionsgerade



Mit 95% iger Wahrscheinlichkeit wird der wahre Mittelwert von **FEV** in der Grundgesamtheit der 60-jährigen Personen vom obigen C.I. überdeckt.

Prognoseband für die Regressionsgerade



Es wird hier die Frage beantwortet, in welchem Intervall mit z.B. 95% iger WSK bei einem Alter der Personen von 60 Jahren ein einzelner Wert des forcierten Ausatmungsvolumens in 1s (FEV) zu erwarten ist.