Varianzanalyse (ANOVA analysis of variance)

Es handelt sich um eine Einführung in das Thema, die sich auf ANOVA mit einem Faktor beschränkt.

Varianzanalyse mit einem Faktor

Die Varianzanalyse ermöglicht im Unterschied zum t-Test den Vergleich von zwei und mehr Mittelwerten.

Das Vorgehen wird im Folgenden an einem einfachen Beispiel (Quelle: HAFL) beschrieben.

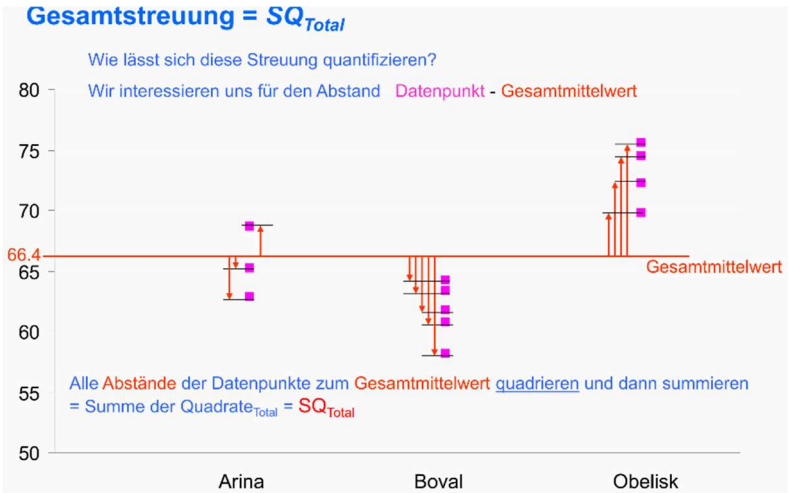
In der Tabelle sind die Erträge von drei Weizensorten Arina, Boval und Obelisk aufgeführt. Es interessiert die Frage, ob die mittleren Erträge der drei Verfahren unterschiedlich sind.

Sorte	Ertrag in dt/ha					Mittelwert
Arina	62.9	68.7	65.2			65.6
Boval	60.7	64.2	58.1	61.7	63.3	61.6
Obelisk	72.2	75.6	74.4	69.8		73.0

Faktor: Sorte
Zielvariable: Ertrag pro Parzelle
Wiederholungen: 5 pro Parzelle, also bei 3 Sorten total 15 Parzellen (3 Werte davon fehlen)

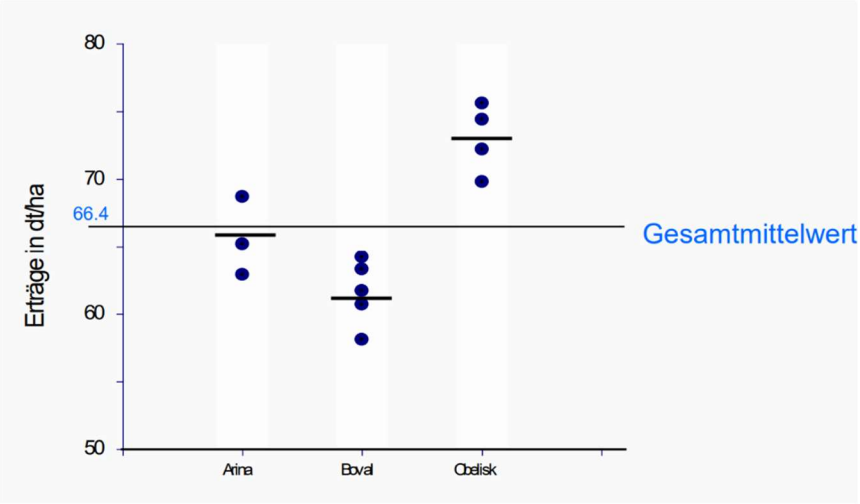
Die grundlegende Idee besteht darin, Streuungen zu vergleichen. Es gilt nämlich die sogenannte Varianzzerlegung. Als Mass für die Streuung werden die folgenden Summen SQ verwendet. Für diese gilt, dass die Gesamtstreuung als Summe der Streuung aufgrund des Faktors (Streuung zwischen den Gruppen) und der zufälligen Streuung (Streuung innerhalb der Gruppen) dargestellt werden kann,
Es gilt also (Beweis → Stochastik→Deskriptive Statistik→Lineare Regression)

$$SQ_{total} = SQ_{Faktor} + S_{Rest}$$



In der Abbildung sind die insgesamt noch 12 Erträge dargestellt. Zusätzlich sind die Mittelwerte der 3 Sorten und der Gesamtmittelwert zu erkennen.

Bei SQ_{total} ist für die Abweichungsquadrate der Gesamtmittelwert massgebend.
Bei SQ_{Faktor} die Mittelwerte der einzelnen Sorten



Berechnung von SQ_{total}

$$SQ_{total} = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(y_{ij} - \overline{y}_{total}\right)^2$$

Bezeichnungen:

y_{ij}	der j-te Ertrag der i-ten Sorte)
$\overline{y}_{total}$	der Gesamtmittelwert
i	die Nummer der Sorte, der Index für die Sorte
j	der j-te Ertrag in einer Sorte
k	Anzahl Stufen (Gruppen) des Faktors
r	Anzahl der Erträge in der Gruppe

Berechnung von **SQ_{total}** mit Excel

Sorte			Ertrag in dT/ha			Mittelwert	r
Arina	62.9	68.7	65.2			65.6	3
Boval	60.7	64.2	58.1	61.7	63.3	61.6	5
Obelisk	72.2	75.6	74.4	69.8		73.0	4
				Mittelwert total		66.4	12
Abw.quad							
Arina	12.25	5.29	1.44			18.98	
Boval	32.49	4.84	68.89	22.09	9.61	137.92	
Obelisk	33.64	84.64	64.00	11.56		193.84	
				SQ total		350.74	

$SQ_{total} = 350.74$

Berechnung von SQ_{Faktor}

Da bei SQ_{Faktor} die Abweichungen der Mittelwerte vom Gesamtmittelwert massgebend sind, kann die Berechnung erfolgen, indem man in der bisherigen Tabelle die Erträge durch den Mittelwert der einzelnen Sorten ersetzt:

$$SQ_{Faktor} = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{y}_{Faktor\_i} - \bar{y}_{total}\right)^2$$

Bezeichnung:

$\bar{y}_{Faktor\_i}$ Mittelwert der Erträge für den Faktor i (Gruppenmittelwert der Sorte i)

Berechnung von SQ_Faktor

Arina	65.6	65.6	65.6			65.6	
Boval	61.6	61.6	61.6	61.6	61.6	61.6	
Obelisk	73.0	73.0	73.0	73.0		73.0	
						66.4	
Abw.quad							
Arina	0.64	0.64	0.64			1.92	
Boval	23.04	23.04	23.04	23.04	23.04	115.20	
Obelisk	43.56	43.56	43.56	43.56		174.24	
							SQ_Faktor
					SQ_Faktor	**291.36**	**291.36**

Berechnung von SQ_{Rest}

Bei der Berechnung von SQ_{Rest} sind die Abweichungen der Erträge für die Sorte i vom Mittelwert der Sorte i massgebend;

$$SQ_{Rest} = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{Faktor\_i}\right)^2$$

Berechnung von SQ Rest

Sorte	Ertrag in dT/ha					Mittelwert	r	
Arina	62.9	68.7	65.2			65.6	3	
Boval	60.7	64.2	58.1	61.7	63.3	61.6	5	
Obelisk	72.2	75.6	74.4	69.8		73.0	4	
				Mittelwert total		66.4	12	
Abw.quad								
Arina	7.29	9.61	0.16			17.06		
Boval	0.81	6.76	12.25	0.01	2.89	22.72		
Obelisk	0.64	6.76	1.96	10.24		19.60		
								SQ Rest
					SQ Rest	59.38		**59.38**
						Kontrolle		**0.00**

Bei Kontrolle wird überprüft, ob die Streuungszerlegung erfüllt ist.

Für den Test sind die Summen SQ noch durch die entsprechenden Freiheitsgrade zu dividieren.

SQ_Total: Division durch $(n-1)$, wobei n die Anzahl der Datenpunkte ist.
SQ_Faktor: Division durch $(k-1)$, wobei k die Anzahl der Gruppen ist
SQ_Rest: Division durch $(n-k)$, wobei k die Anzahl der Gruppen ist.

Man erhält daraus die durchschnittlichen quadrierten Abweichungen DG:

$$DQ_{Total} = \frac{SQ_{Total}}{n-1} \approx \frac{350.74}{11} \approx 31.89$$
$$DQ_{Faktor} = \frac{SQ_{Faktor}}{k-1} \approx \frac{291.36}{2} \approx 145.68$$
$$DQ_{Rest} = \frac{SQ_{Rest}}{n-k} \approx \frac{59.38}{9} \approx 6.59$$

Eine entsprechende Gleichung wie bei der Varianzzerlegung gilt für die DQ nicht mehr.

Testen des Effekts

Will man testen, ob die Sorte (die Gruppen, die Stufen des Faktors, das Treatment) eine Wirkung hat, bildet man den folgenden Quotienten:

$$\frac{DQ_{Faktor}}{DQ_{Rest}} \approx \frac{145.68}{6.59} \approx 22.08$$

Es kann bewiesen werden, dass das Verhältnis dieser Varianzen F-verteilt ist mit den Freiheitsgraden (FG_{Faktor}, FG_{Rest}).
Die **F-Verteilung** (nach Ronald Aylmer Fisher), ist eine stetige Wahrscheinlichkeitsverteilung. Sie wird verwendet um zu entscheiden, ob der Unterschied zweier Stichprobenvarianzen zufällig ist oder auf unterschiedliche Gruppenmittelwerte hinweist.

Die F-Verteilung ist von der Chiquadratverteilung abgeleitet.

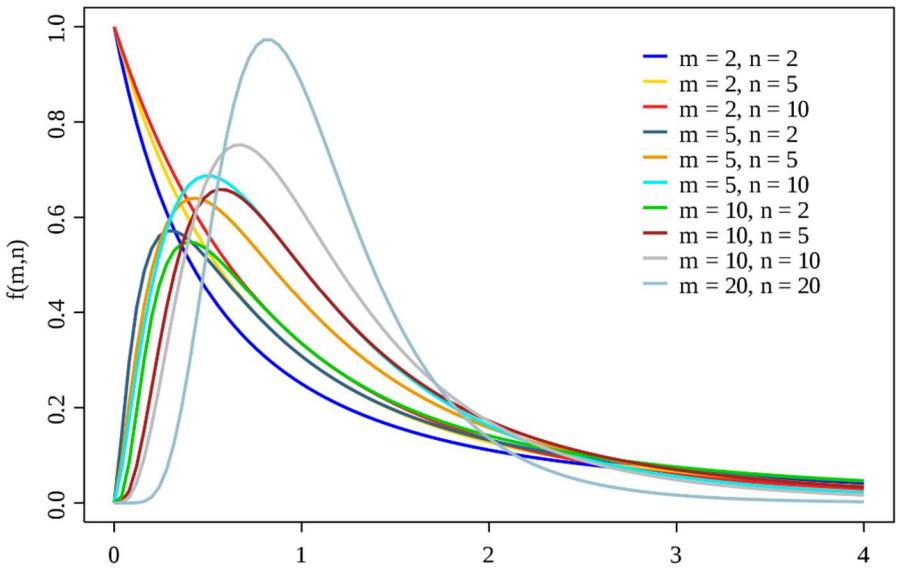
Sind X_1 und X_2 zwei Zufallsvariablen mit $X_1 \sim \chi^2_m$ und $X_2 \sim \chi^2_n$ heisst eine Zufallsvariable Y F-verteilt, wenn gilt:

$$Y := \frac{X_1/m}{X_2/n}$$

Für den Erwartungswert von Y gilt:

$E(Y) = \frac{n}{n-2}$ mit n > 2

In der Abbildung sind verschiedene Beispiele von Dichten dargestellt.
Die Quantile findet man in Tabellen.
Mass für die Streuung bei Wiederholungen eines Verfahrens.



Hypothesen:

H_A: Die Mittelwerte der verschiedenen Faktorstufen sind verschieden, d.h. mindestens zwei Mittelwerte unterscheiden sich.

H_0: Die Mittelwerte Faktorstufen unterscheiden sich nicht.

In der folgenden ANOVA-Tabelle sind die Testergebnisse zusammengefasst
4.26

Ursache der Streuung	SQ	FG	DQ	F	p-Wert
Sorte	291.36	2	145.68	22.08	0.000338
Rest	59.38	9	6.59		
Total	350.74				

```
mod1 <- lm(Ertrag~Sorte,data=Weizensorten)

anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: Ertrag
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Sorte      2 291.36 145.680   22.08 0.000338 ***
## Residuals  9  59.38   6.598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Tabelle für $p = 0{,}05$

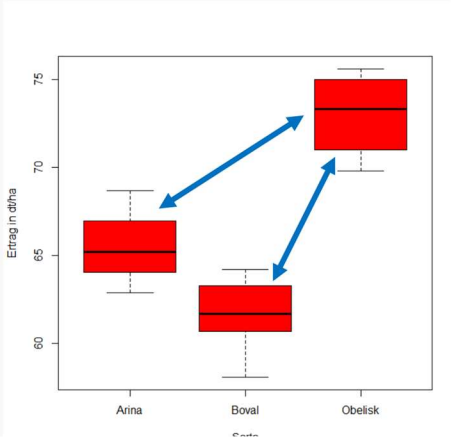
	$df1$									
$df2$	1	2	3	4	5	6	7	8	9	10
1	162	200	216	225	230	234	237	239	241	242
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,73
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75

Aus der Tabelle der F-Verteilung ergibt sich der kritische Wert
$F_{kritisch} = F_{2;9} = 4.26$
und der entsprechende p-Wert
$p = 0.000338$
Die Nullhypothese wird also verworfen. Der Faktor Sorte hat also einen signifikanten Effekt.

Noch bleibt aber unklar, welche Sorte den grösseren Ertrag hat als die anderen. Der Entscheid kann grafisch mit einem Boxplot oder mit dem Test für Multiple Vergleiche von Tukey-Kramer erfolgen.

```
TukeyHSD(aov(mod1), "Sorte")

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mod1)
##
## $Sorte
##                 diff       lwr       upr     p adj
## Boval-Arina    -4.0 -9.237385  1.237385 0.1379145
## Obelisk-Arina   7.4  1.922612 12.877388 0.0110434
## Obelisk-Boval  11.4  6.589155 16.210845 0.0002569
```



Zusammenfassung:
Bei den drei Weizensorten Arina, Boval und Obelisk hatte der Faktor Sorte einen signifikanten Einfluss auf die Erträge (ANOVA mit einem Faktor $F_{(2,9)} = 22.08$, $p < 0.001$. Der Tukey-Kramer-Test ergab, dass der Ertrag der Sorte Obelisk signifikant grösser war als der von Arina und Boval. Zwischen Arina und Boval bestand jedoch kein signifikanter Unterschied.

Ergänzungen:
Sind die Voraussetzungen nicht erfüllt, dann ist der nichtparametrische Kruskal Wallis Test eine Alternative.
Eine andere Möglichkeit besteht darin, die Daten zu transformieren:
Beispiele von Transformationen: $ln(y), sqrt(y), arcsin(y)$, ...

Lineare Regression

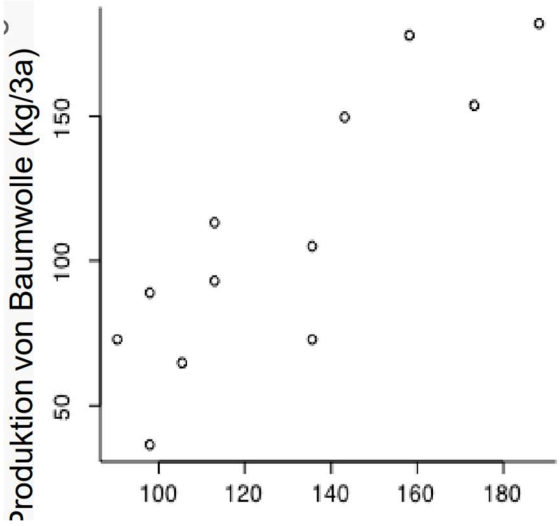
Das Thema wurde bereits im Abschnitt Beschreibende Statistik behandelt. In diesem Abschnitt wird der Zusammenhang zwischen der linearen Regression und der Varianzanalyse untersucht. Ausserdem geht es um die Fragen, wie signifikant die einzelnen Parameter sind und wie die Voraussetzungen für eine Lineare Regression überprüft werden können.

Zusammenhang der linearen Regression mit ANOVA

Wie sind die Summen in der Varianzzerlegung
$SQ_{total} = SQ_{Faktor} + S_{Rest}$
zu interpretieren?

SQ_{total}:	Summe der quadrierten Abstände zwischen der y-Koordinate eines Datenpunktes und dem Mittelwert aller y-Koordinaten.
SQ_{Faktor}	Summe der quadrierten Abstände zwischen den durch das Modell vorausgesagten y-Werten vom Mittelwert aller y Koordinaten,
S_{Rest}	Summe der quadrierten Abstände zwischen der y- Koordinate eines Datenpunktes und dem durch das Modell vorausgesagten y-Werten. Diese Abweichungen heissen Residuen.

Beispiel : (HAFL)
Baumwollproduktion
Die Produktion von 1 kg Baumwolle benötigt während der ganzen Produktionszeit sehr viel Wasser, etwa 10 000 – 20 000 Liter.



Bewässerung (cm)	Ertrag (kg/3a)
90.4	72.8
97.9	36.4
97.9	89.0
105.5	64.7
113.0	113.3
113.0	93.1
135.6	105.2
135.6	72.8
143.1	149.7
158.2	178.0
173.3	153.7
188.3	182.0

Für die Ausgleichsgerade erhält man mit R den folgenden Output:

```
reg <- lm(Ertrag~Bewaesserung,data=Baumwolle)
coef(reg)

##  (Intercept) Bewaesserung
##   -55.396511     1.273011
```

Die Ausgleichsgerade hat also die Gleichung
$y = 1.27 \cdot x - 55{,}4$ Steigung (slope) 1.27 und den y-Achsenabschnitt (Intercept) -55.4.

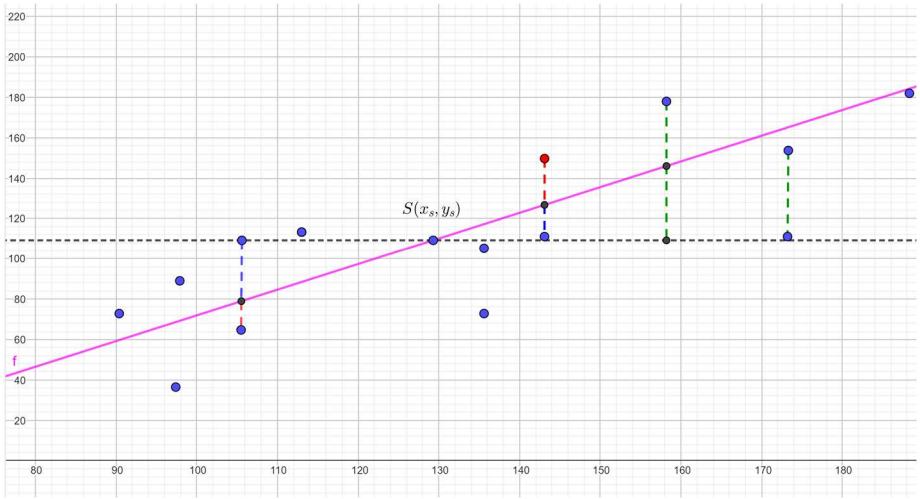
Wie kann die Regression als ANOVA interpretiert werden?

SQ_{total}(SQT):
Summe der quadrierten Abstände der Datenpunkte von der zur x-Achse parallelen Geraden v.
Sie ist ein Mass für die Gesamtstreuung (an zwei Stellen grün dargestellt)

SQ_{Modell} (SQM):
Summe der quadrierten Abstände der vom Modell vorausgesagten Datenpunkte von der zur x-Achse parallelen Geraden durch den Schwerpunkt S.
Sie ist ein Mass für die vom Modell «erklärte» Streuung (an zwei Stellen blau dargestellt)

SQ_{Rest} (SQR): D
Die Summe der quadrierten Abstände zwischen den y-Koordinaten der Datenpunkte und den entsprechenden Punkten auf der Gerade (… durch das Modell hervorgesagt). Das ist die zufällige Streuung der Datenpunkte um die Gerade (Modell
Summe der quadrierten Abstände der Datenpunkte von den durch das Modell vorausgesagten Punkten auf der Ausgleichsgeraden. Sie ist ein Mass für die vom Modell nicht erklärte, zufällige Streuung (an zwei Stellen rot dargestellt)

Im der folgenden Abbildung sind die Datenpunkte und die Regressionsgerade durch den «Schwerpunkt» S dargestellt, zusätzlich an einigen Stellen die den SQ entsprechenden Abstände-



Die Werte im Beispiel sind in der folgenden ANOVA-Tabelle dargestellt:

Ursache der Streuung	SQ	FG	DQ	F	p-Wert
Modell	17996.2	1	17996.2	28.622	0.0003
Rest	6287.6	10	628.8		
Total	24283.8	11			

```
anova(reg)

## Analysis of Variance Table
##
## Response: Ertrag
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Bewaesserung  1 17996.2 17996.2  28.622 0.0003236 ***
## Residuals    10  6287.6   628.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wie „gut“ ist die Gerade die am besten passt?

Das sogenannte Bestimmtheitsmass R^2 ist ein Mass für die Güte des Modells bzw. der Ausgleichsgeraden.

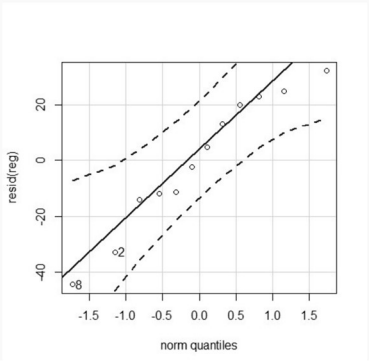
$$R^2 = \frac{SQ_{Faktor}}{SQ_{Total}} \approx \frac{17996.2}{24283.8} \approx 0.741$$

Zusammenfassung:
Das Modell erklärt einen hohen Anteil der Gesamtstreuung. Die Bewässerung hat somit einen signifikanten Einfluss auf den Baumwollertrag.
Die Gleichung der Ausgleichsgeraden $y = 1.27 \cdot x - 55.4$ kann so interpretiert werden, dass eine Vergrösserung der Bewässerung um eine Einheit ein Wachstum des Ertrags um ca 27% bedeutet.

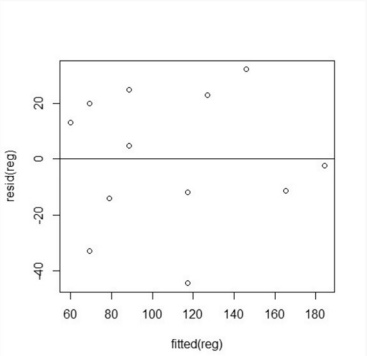
Es ist noch zu prüfen, ob die Voraussetzungen für eine lineare Regression erfüllt sind:

1.
Die Residuen sollten normalverteilt mit dem Mittelwert 0 sein und kein Muster aufweisen
.
Dies kann grafisch auch mit einem qqPlot geprüft werden. In der Abbildung sind die Quantile der Normalverteilung so transformiert, dass sie als Gerade erscheinen. Die einzelnen Quantile der Daten sind durch Punkte dargestellt. Bei normalverteilten Residuen sollten mindestens 95% innerhalb der gestrichelten Linien liegen. Beim Beispiel ist diese Bedingung erfüllt. Allerdings ist ein gewisses Muster zu erkennen (Interpretation?).

```
library(car)
qqPlot(resid(reg))
```




```
plot(resid(reg)~fitted(reg)
abline(h=0)
```



Die Frage, ob die Residuen normalverteilt sind, kann auch rechnerisch mit dem Shapiro-Wilk normality-Test überprüft werden.
Zu beachten ist, dass bei diesem Test die Nullhypothese lautet: Die Residuen sind normalverteilt.
Damit bedeutet ein p-Wert kleiner als 5%, dass die Daten nicht normalverteilt sind.

2,
Für alle x-Werte sollte die Varianz der Residuen etwa gleich bleiben.
Im folgenden Beispiel ist die Homogenität der Varianzen verletzt (Heteroskedastizität in der Abbildung).
Zur Prüfung der Homogenität der Varianzen kann etwa der Levene-Test verwendet werden.