



Swiss Statistics Meeting
Schweizer Statistiktage
Journ es suisses de
la statistique
Giornate svizzere
della statistica

Schweizer Statistiktage, Aarau, 18. Nov. 2004

Qualitative Überprüfung der Modellannahmen in der linearen Regressionsrechnung – am Beispiel der Untersuchung der Alterssterblichkeit bei Hitzeperioden in der Stadt Zürich

Daten · Analysen



Dienstleistungen

Statistik Stadt Zürich

Thomas Glauser

Datenspezialist

thomas.glauser@stat.stzh.ch

Ausgangslage



- Hohe Anzahl von Sterbefällen unter der älteren Bevölkerung im Hitzesommer 2003 in Frankreich.
- Der Zusammenhang zwischen Hitzeperioden und der Alterssterblichkeit soll für die Stadt Zürich untersucht werden. (Anfrage Stadtärztlicher Dienst Zürich)
- Daten
 - Anzahl Sommertage ($> 25^\circ$)
 - Anzahl Hitzetage ($> 30^\circ$)
 - Sterberate der über 64-Jährigen

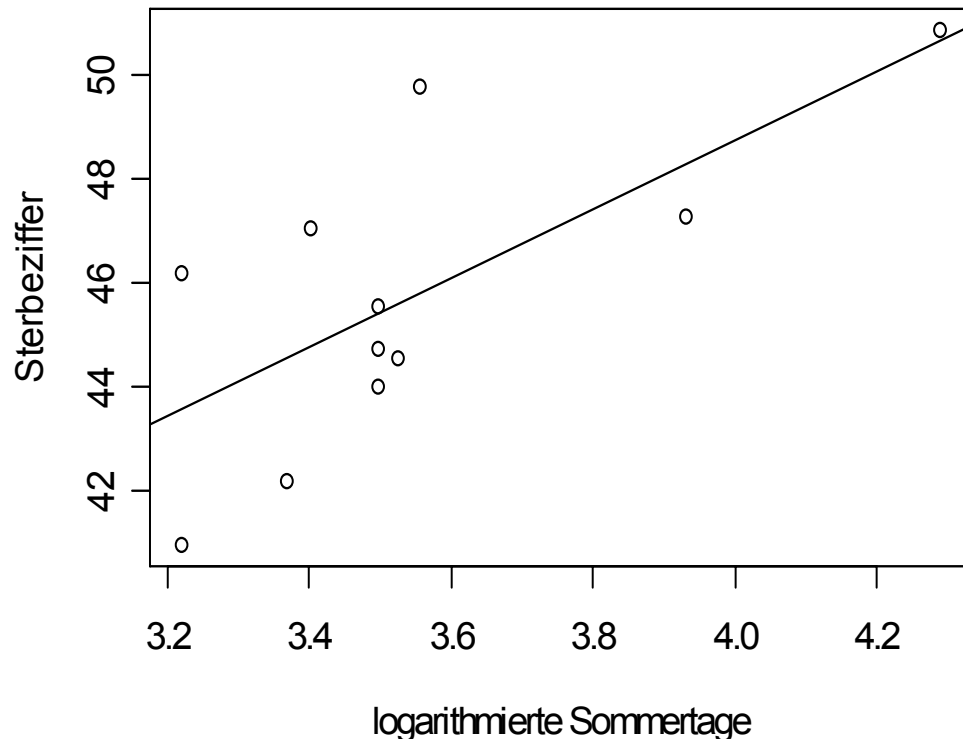


Resultat der Untersuchung

Sterbeziffer = Durchschnittliche Sterbeziffer der über 64-Jährigen in den Monaten Juni bis August

Logarithmierte Sommertage = Logarithmierte Anzahl Sommertage in den Monaten Juni bis August

○ = Pro Jahr eine Beobachtung



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.286	8.014	2.781	0.0214 *
ln_Sommertage	6.614	2.253	2.936	0.0166 *

Residual standard error: 2.227 on 9 degrees of freedom

Multiple R-Squared: 0.4893, Adjusted R-squared: 0.4325

F-statistic: 8.622 on 1 and 9 DF. p-value: 0.01658

Modell-
zusammenfassung



Ausgangslage

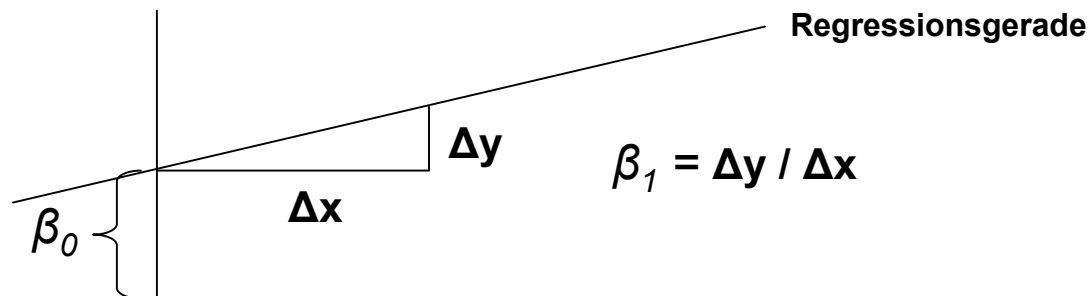
- Die lineare Regression ist ein weit verbreitetes Modell.
 - Oft wird jedoch nicht geprüft, ob die Modellannahmen auch tatsächlich zutreffen.
 - Dadurch wird die Chance verpasst, aus allfälligen Abweichungen ein besseres Modell zu entwickeln.
- ➔ Überprüfung der Modellannahmen mit der Residuen-Analyse.

Modellgleichung

- Im Modell der multiplen linearen Regression wird eine Zielvariable Y_i erklärt durch eine lineare Funktion $h(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)})$ von erklärenden Variablen. Die zufälligen Fehler werden durch den **Term E_i beschrieben**.
- Dies führt zur Modellgleichung:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)} + E_i$$

- Die Parameter sind die so genannten Koeffizienten $\beta_0, \beta_1, \beta_2, \dots, \beta_m$ der erklärenden Variablen und die Varianz σ^2 ist die Varianz der zufälligen Abweichungen E_i . Die Koeffizienten werden mit der Methode der kleinsten Quadrate (least squares) geschätzt.

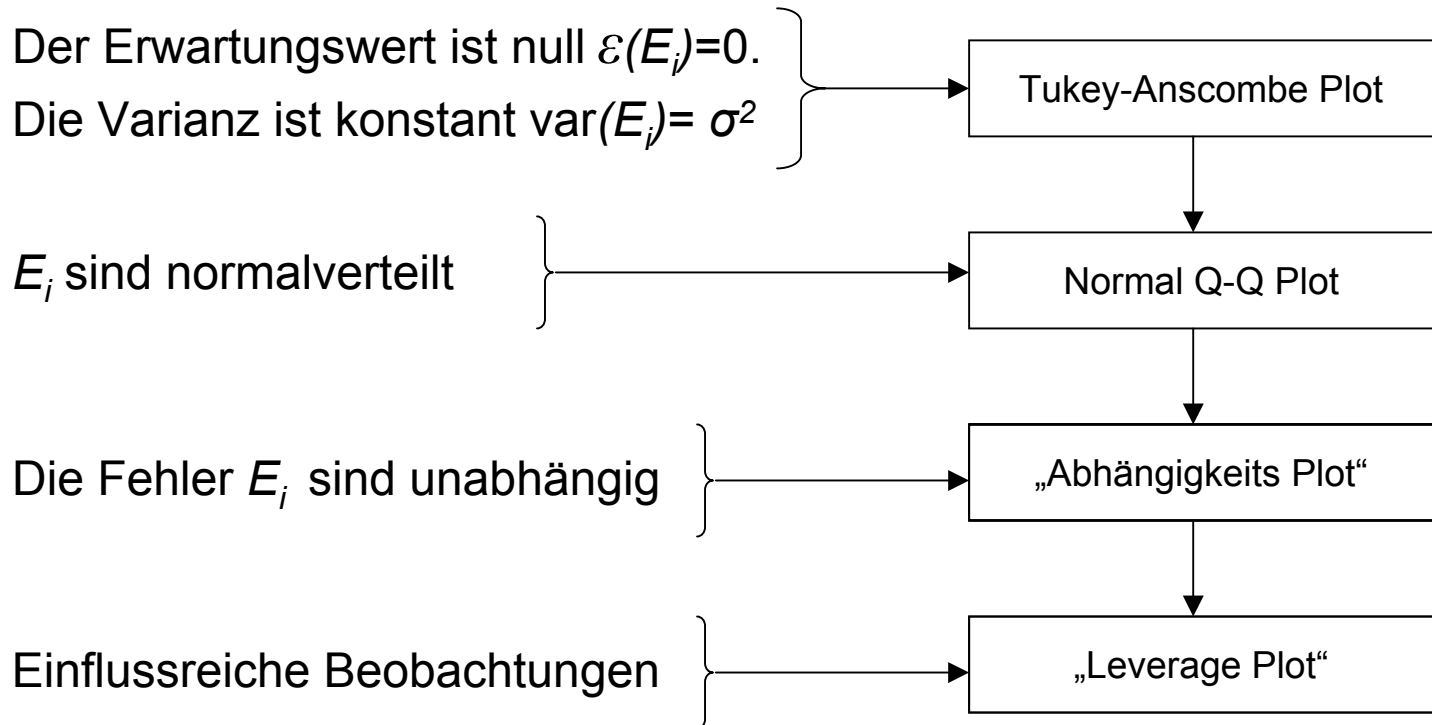




Modellannahmen

- Die Schätz- und Testverfahren, die in der linearen Regressionsrechnung verwendet werden, setzen voraus, dass die Fehler E_i i.i.d. $N(0, \sigma^2)$ verteilt sind. Dies lässt sich aufspalten in die Annahmen:
 - Ihr Erwartungswert $\mathcal{E}(E_i)=0$
 - Sie haben alle die gleiche Varianz $\text{var}(E_i)= \sigma^2$
 - Sie sind normalverteilt
 - Die Fehler E_i sind unabhängig
- Um die Qualität des Modells zu prüfen, müssen diese Modellannahmen überprüft werden. Es geht darum, allfällige Abweichungen zu entdecken und daraus ein Modell zu entwickeln, das besser zu den Daten passt. Dabei spricht man von **explorativer Datenanalyse**. Bei der Überprüfung der Annahmen werden grafische Darstellungen der Residuen verwendet, weshalb die Methode Residuen-Analyse genannt wird.

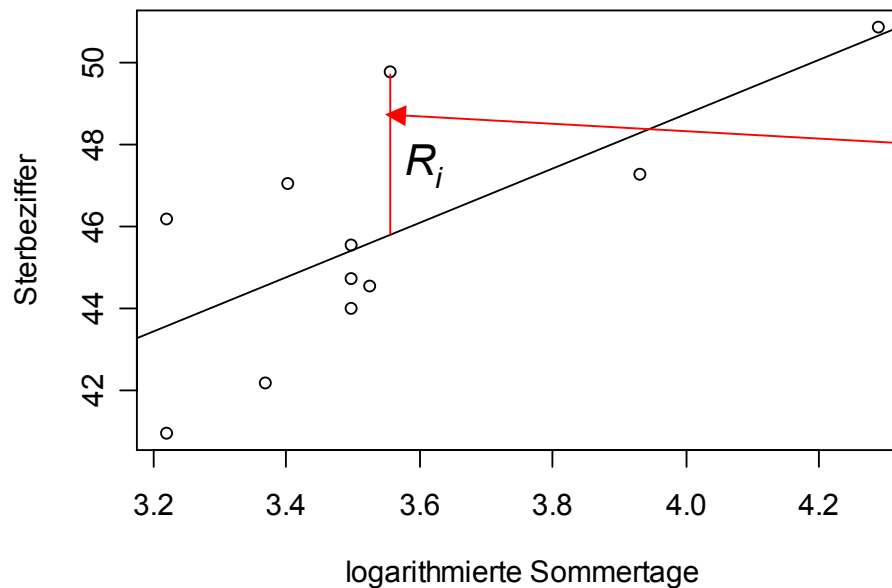
Residuenanalyse



Residuen

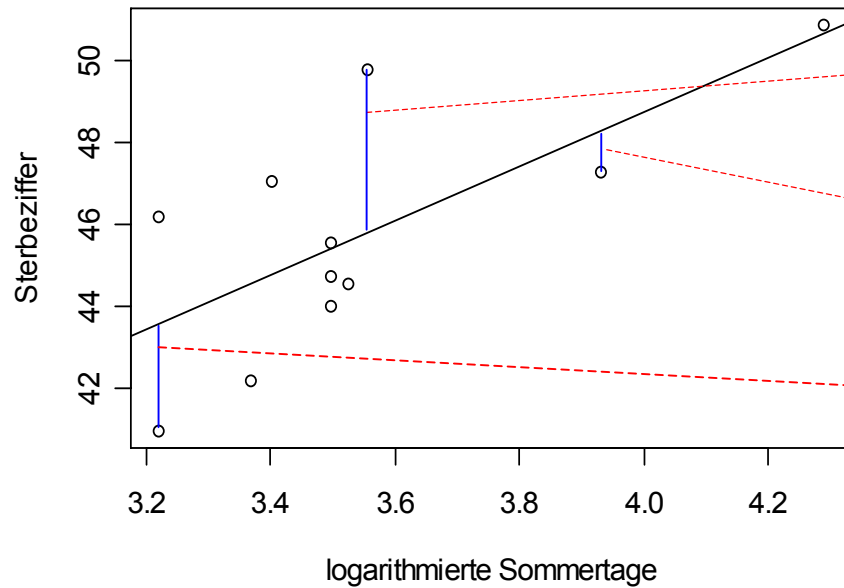
- Da man die Fehler E_i nicht kennt, werden die Residuen R_i zur Prüfung der Modellannahmen verwendet.
- Die Residuen werden mit der folgenden Formel berechnet:

$$R_i = Y_i - \hat{y}_i$$

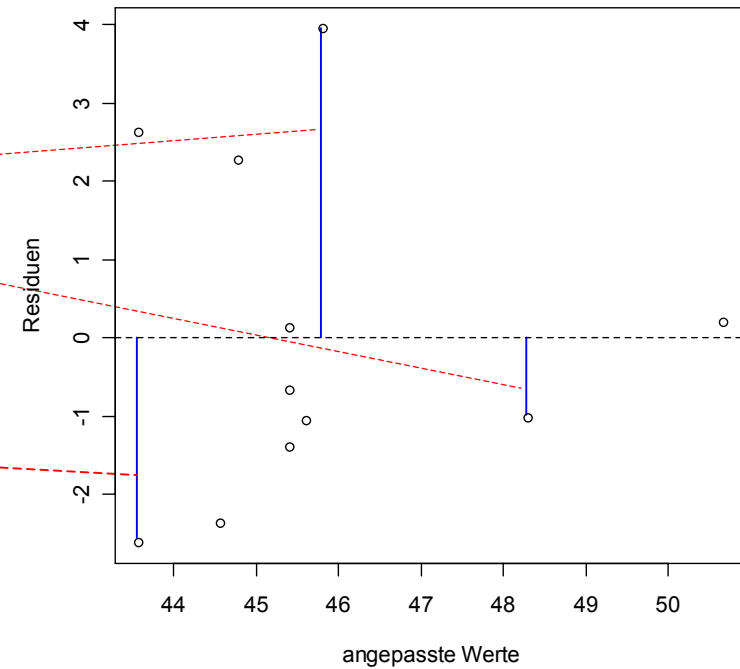


Residuen: Vertikaler Abstand zwischen Beobachtungen und Regressionsgerade

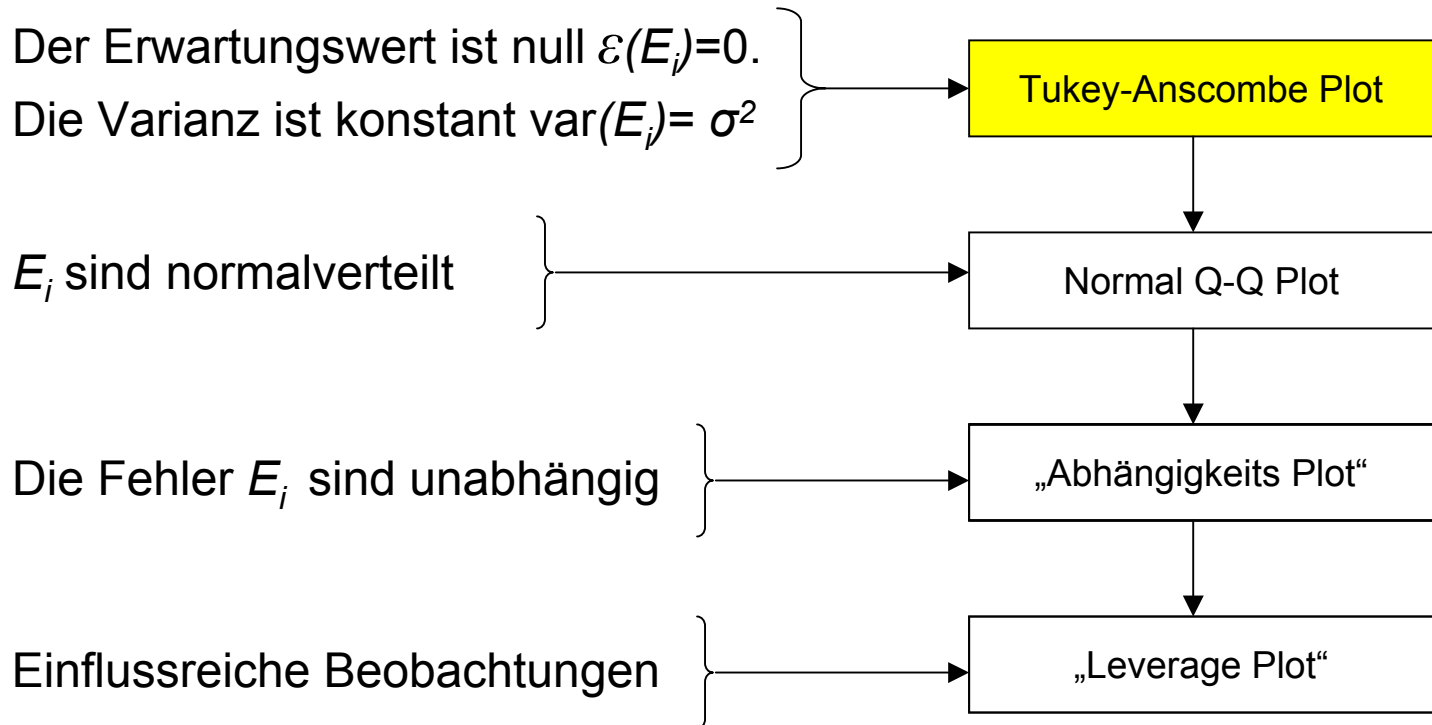
Residuen



Tukey-Anscombe Plot



Residuenanalyse

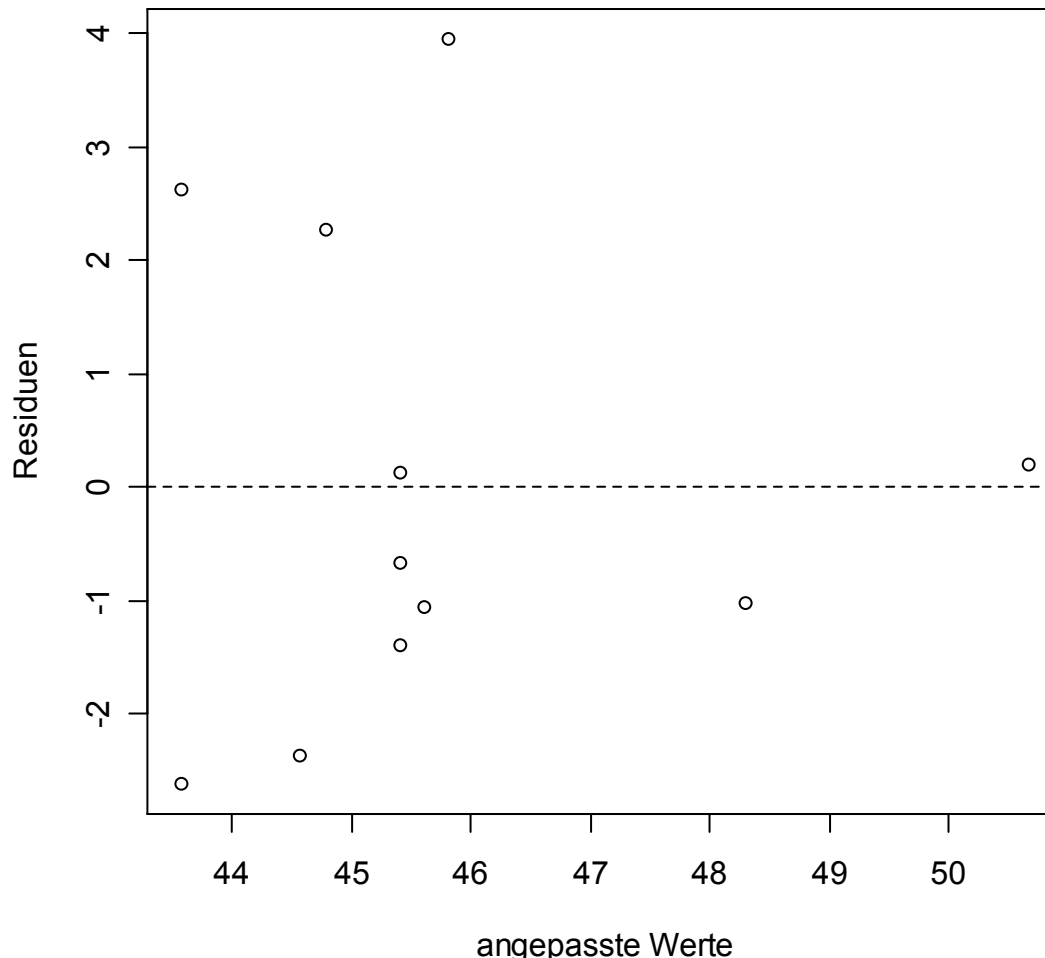


Überprüfung der Modellannahmen:

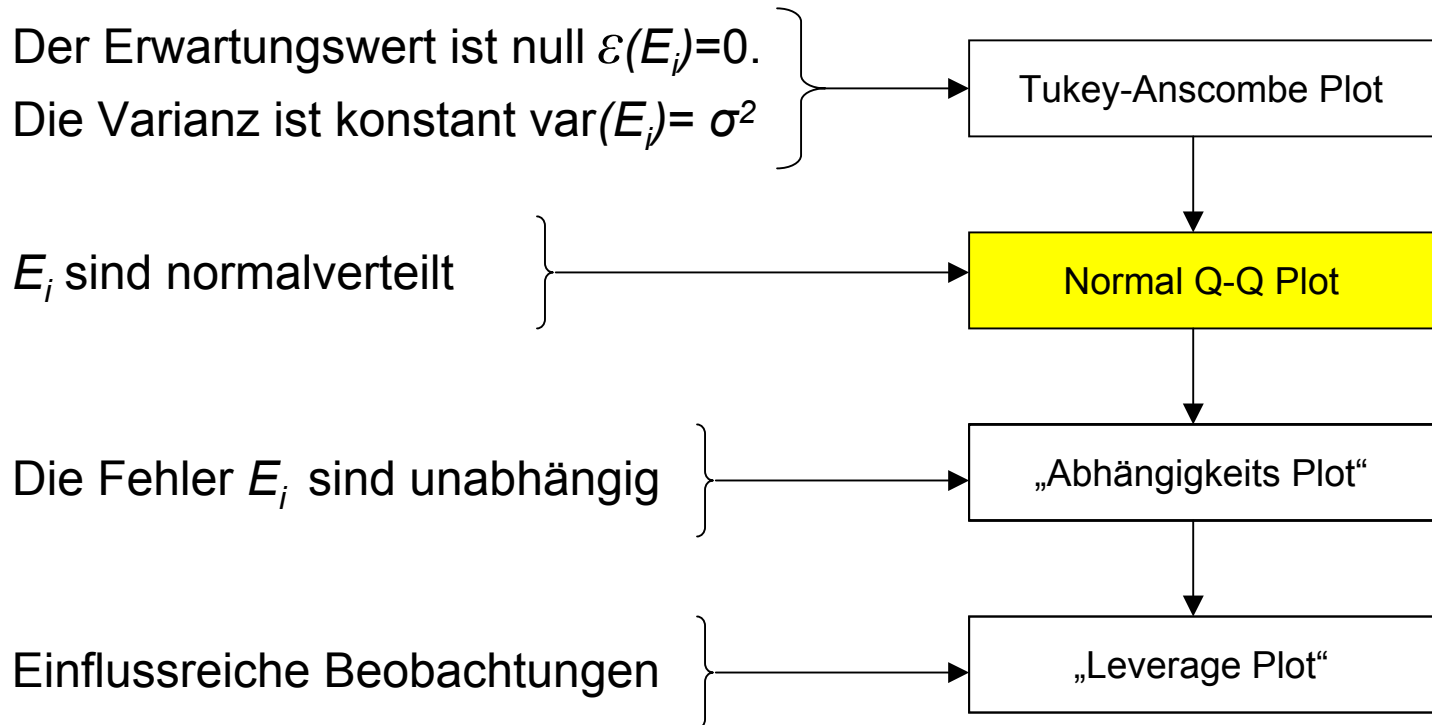
$$\mathcal{E}(R_i)=0, \text{ var}(R_i)=\textit{konst.}$$



Tukey-Anscombe Plot



Residuenanalyse

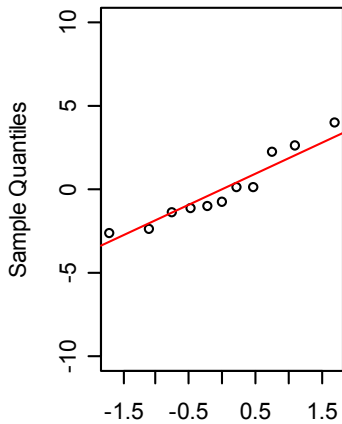


Überprüfung der Modellannahme:

$R_i \sim N(0, \sigma^2)$

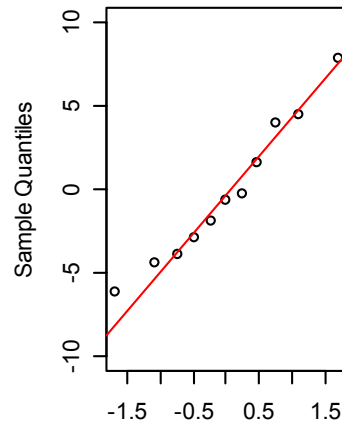


Normal Q-Q Plot



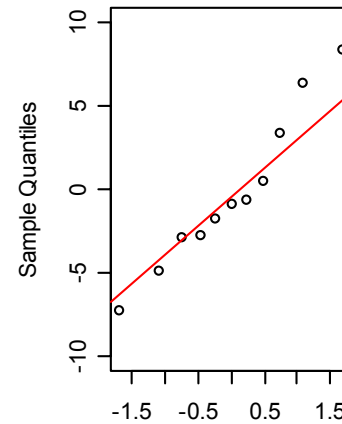
Theoretical Quantiles

Normal Q-Q Plot



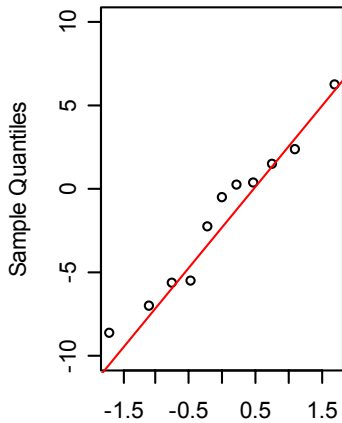
Theoretical Quantiles

Normal Q-Q Plot



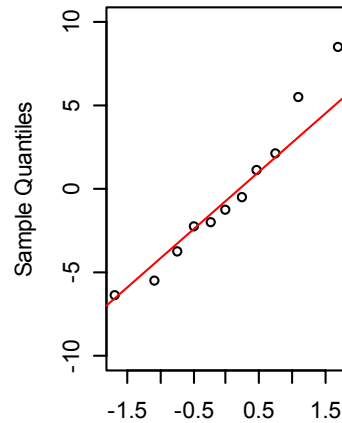
Theoretical Quantiles

Normal Q-Q Plot



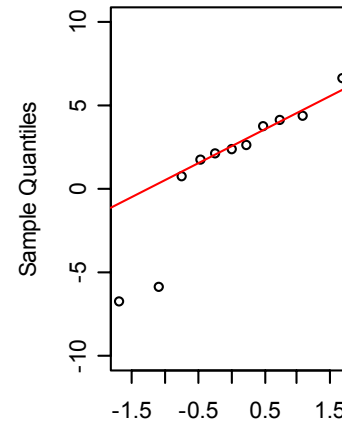
Theoretical Quantiles

Normal Q-Q Plot



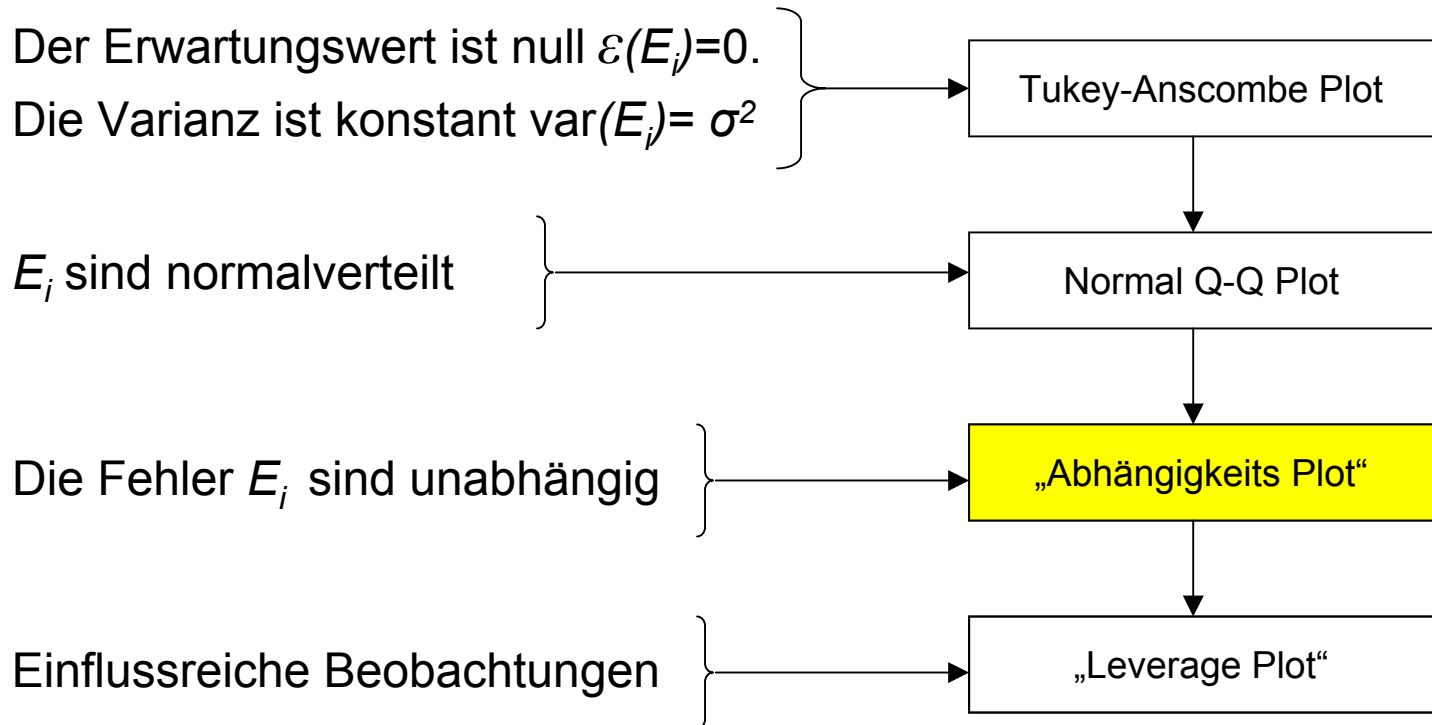
Theoretical Quantiles

Normal Q-Q Plot

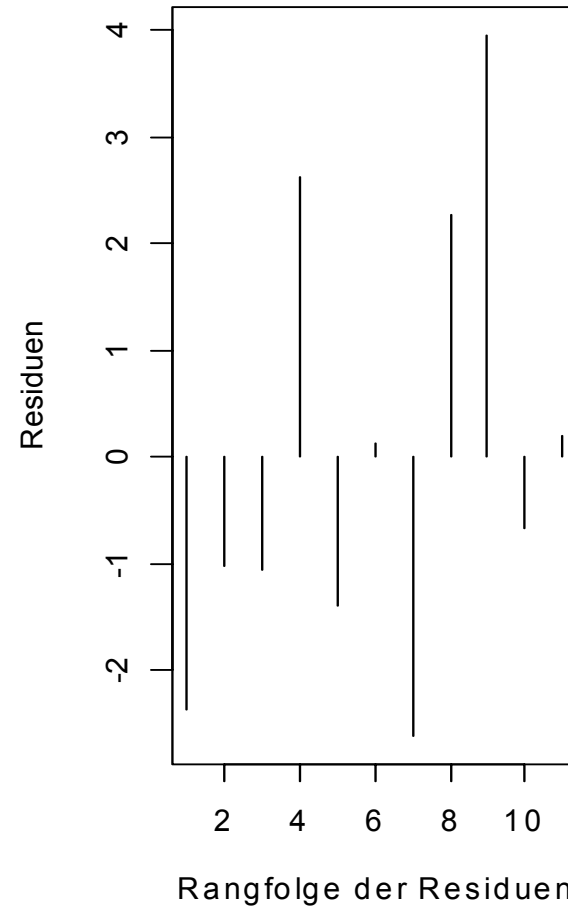
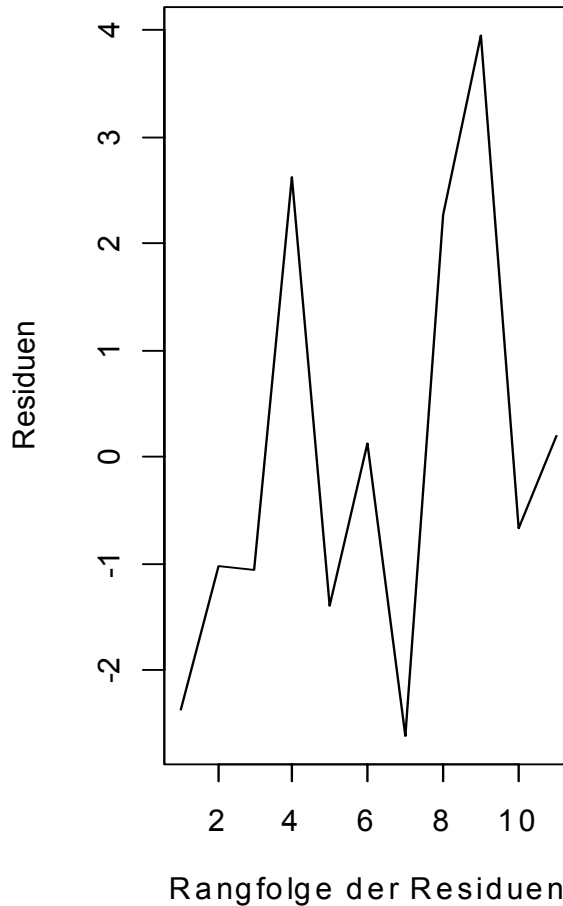


Theoretical Quantiles

Residuenanalyse



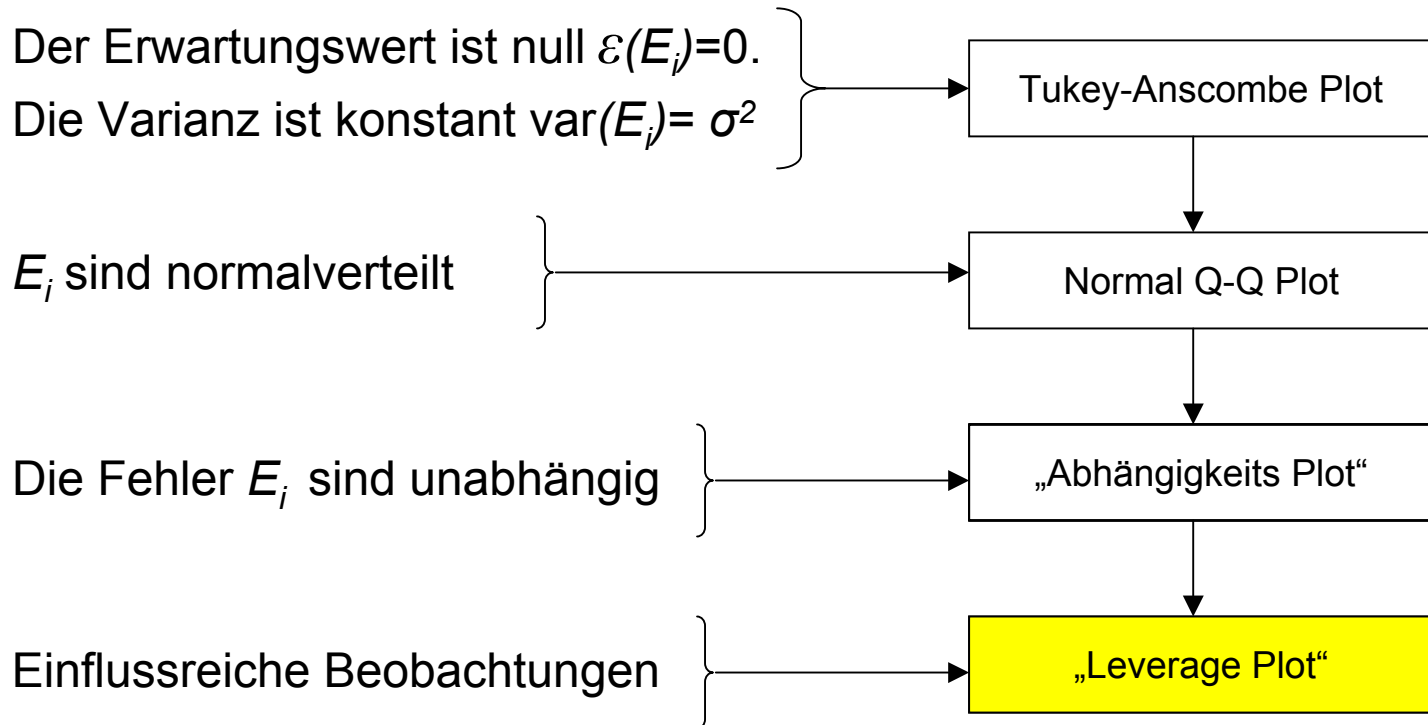
Überprüfung der Modellannahme: R_i sind unabhängig



Einflussreiche Beobachtungen

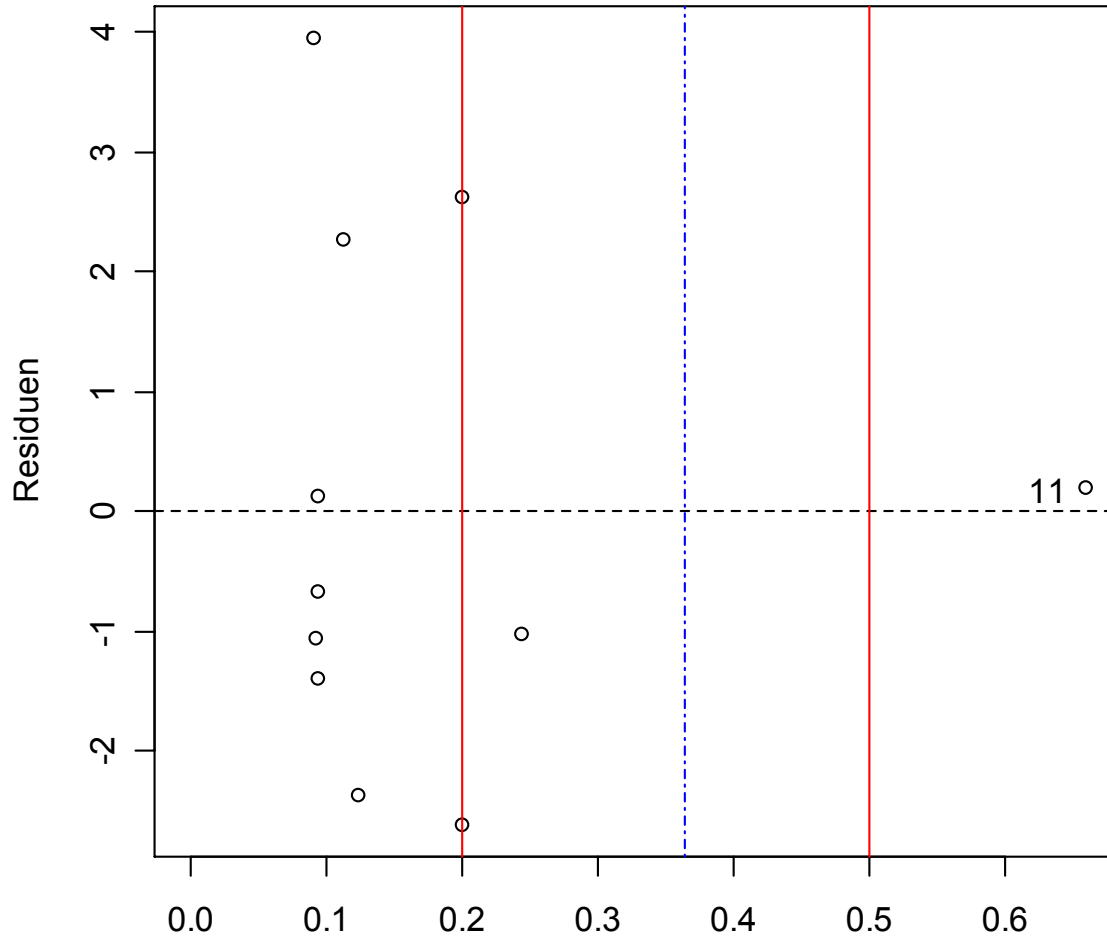
- Wir wollten die Verteilung der Zufallsfehler E_i überprüfen, haben aber die Residuen R_i benützt. Das ist jedoch nicht dasselbe.
- Falls die Fehler normalverteilt sind, so sind es die Residuen einer Kleinste-Quadrate-Schätzung auch. Aber sie haben nicht die gleiche theoretische Varianz, auch wenn die Fehler dies erfüllen; $\text{var}(R_i)$ hängt von $[x_i^{(1)}, x_i^{(2)}, \dots]$ ab! Es ist
 - $\text{var}(R_i) = (1 - H_{ii}) \sigma^2$
- Die Grössen H_{ii} heissen englisch **leverage**, was Hebelarm bedeutet.

Residuenanalyse



Überprüfung auf: Einflussreiche Beobachtungen

Erklärende Variable vs. Residuen



H_{ii} Diagonalelement der Hut Matrix

Schlussfolgerung aus der Residuenanalyse



- Aufgrund des Tukey-Anscombe Plot kann davon ausgegangen werden, dass $\mathcal{E}(E_i)=0$ ist, und dass $\text{var}(E_i)=\text{konst}$ ist.
- Die Normal Q-Q Plot zeigen, dass man von normalverteilten E_i ausgehen kann.
- In der Darstellung der Residuen nach ihrer Rangfolge ist keine Abhängigkeit zwischen den Residuen erkennbar.
- Die Darstellung der Diagonalelemente der Hut-Matrix zeigt, dass eine einflussreiche Beobachtung vorhanden ist.
- => Man sollte ein robustes Regressionsmodell verwenden, damit der Einfluss der einflussreichen Beobachtung eingeschränkt wird.

Robuste Regression

