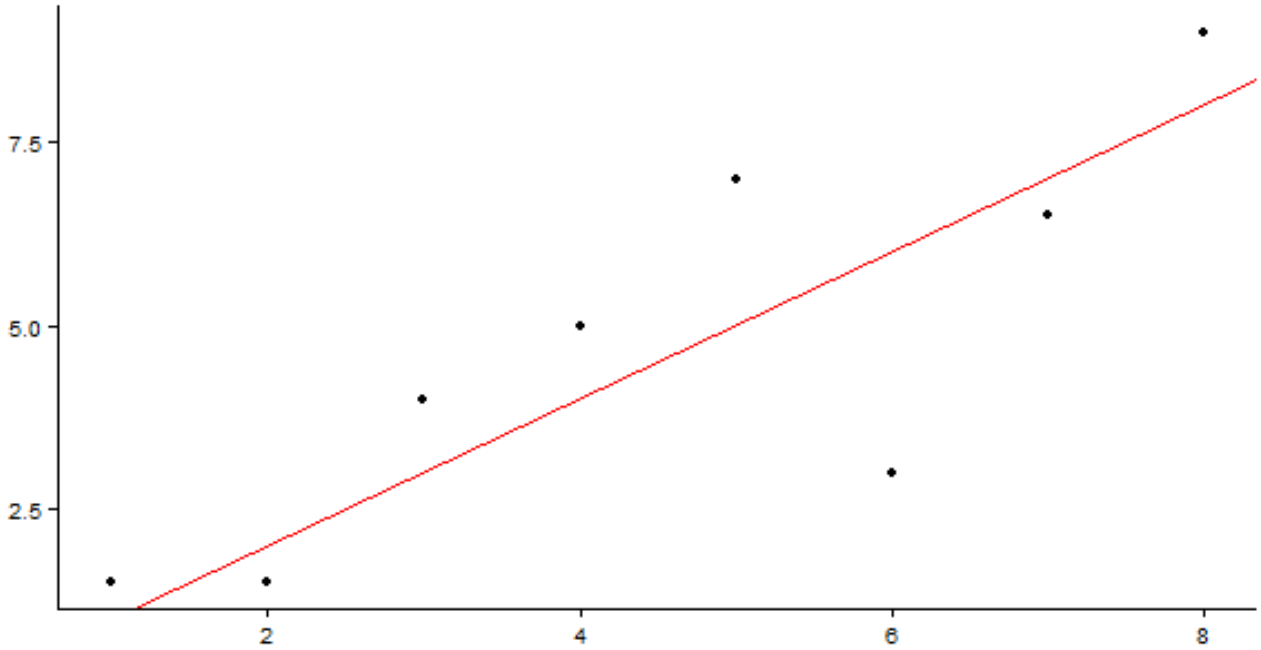
lineare Regression

Benjamin Schlegel

21. März 2016

Mit einer Regressionsanalyse wird versucht, eine abhängige Variable durch eine oder mehrere unabhängige Variablen zu erklären. Mit einer linearen Regression wird eine lineare Abhängigkeit angenommen, welche mit einer Geraden gezeichnet werden kann. Die lineare Regression sollte nur verwendet werden, wenn die abhängige Variable intervallskaliert oder ratioskaliert ist.



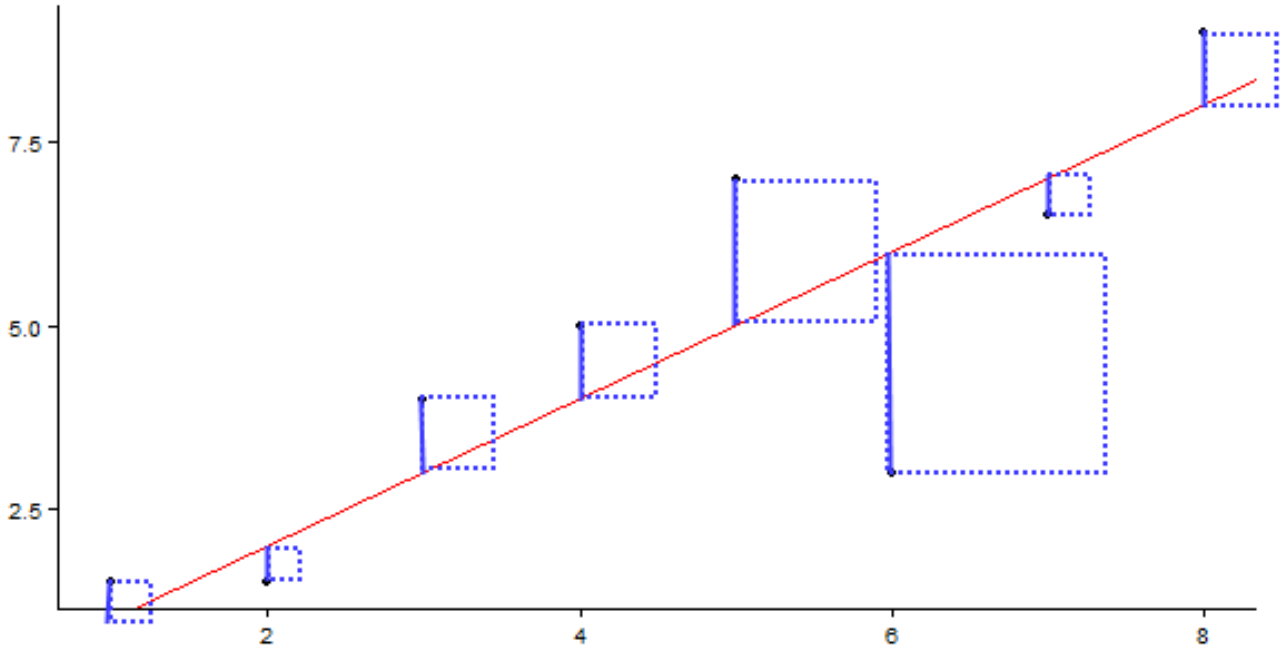
Die einfache lineare Regression besteht aus einer unabhängigen Variable. Der Zusammenhang kann mit der Formel \(y = ax + b\) dargestellt werden. Da der Zusammenhang jedoch fast nie perfekt ist, wird zusätzlich ein Fehlerterm in die Formel einbezogen. Üblich ist folgende Schreibweise:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad i = 1, ..., n$$

\(\beta_0\) ist der Achsenabschnitt und \(\beta_1\) die Steigung. Beides sind Konstanten (genannt Koeffizienten), weshalb sie kein \(i\) enthalten. \(y_i\) und \(x_i\) sind die tatsächlichen Werte der Punkte (y: abhängige Variable, x: unabhängige Variable). \(\varepsilon_i\) sind die Fehler (Residuen), welche die Abweichung der Punkte von der Ideallinie enthalten.

Bei einer linearen Regression können die Koeffizienten direkt interpretiert werden. Ein \(\beta_1\) von 1.56 bedeutet, dass pro zusätzliche Einheit bei x das y um 1.56 ansteigt.

Das Ziel einer Regression ist es, die Residuen möglichst klein zu halten. Die Residuen sind die Abweichungen der vorausgesagten y-Werte zu den effektiven y-Werten. Beim OLS (ordinary least squares) werden die Quadrate der Residuen minimiert. Folgende Grafik veranschaulicht das Prinzip:



Multivariate Regression

Statt nur einer unabhängigen Variable kann eine Regression auch mehrere unabhängige Variablen haben. Die Formel sieht dann folgendemassen aus:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \varepsilon_i, \quad i = 1, ..., n$$

Auch bei der multivariaten linearen Regression können die Koeffizienten direkt interpretiert werden. Hier ist jedoch zu beachten, dass die Werte nur dann direkt gelten, wenn die anderen unabhängigen Variablen konstant gehalten werden (genannt ceteris paribus).

Annahmen

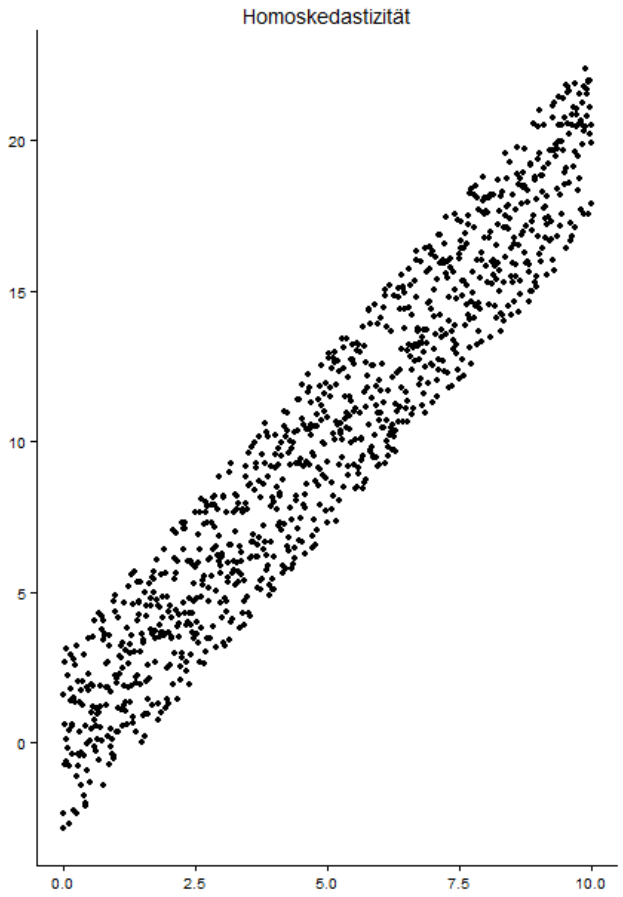
In der linearen Regression gibt es vier wichtige Annahmen. Zuerst werden die Annahmen aufgelistet, um anschliessend näher auf die einzelnen einzugehen.

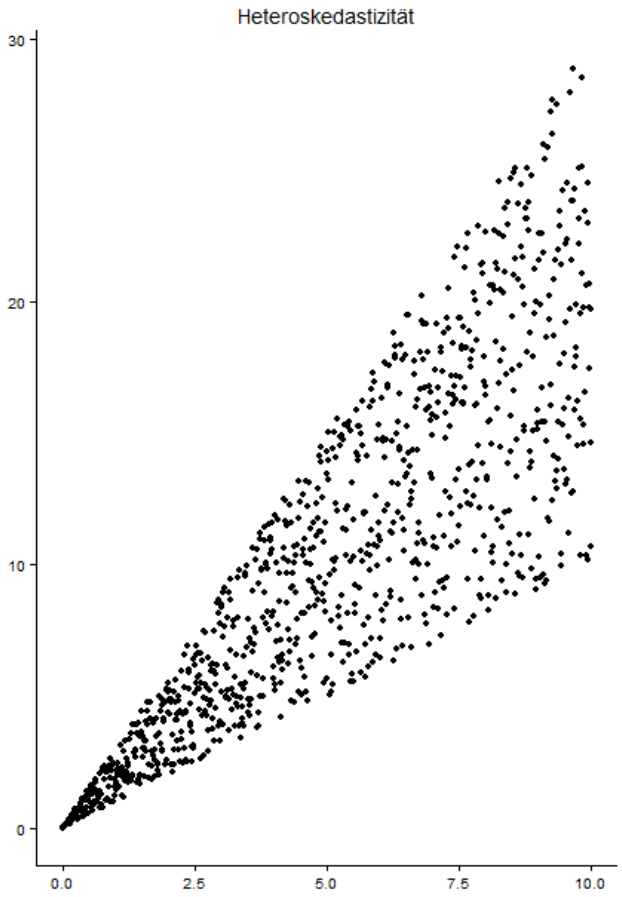
- Der Erwartungswert der Residuen ist Null.
- Die Fehlerterme sind unkorreliert.
- Es herrscht Homoskedastizität.
- Die Fehlerterme sind normalverteilt.

Die erste Annahme geht davon aus, dass die Fehlerterme im Durchschnitt Null ergeben (genannt **Erwartungswert**). Es gibt sowohl Abweichungen gegen unten als auch gegen oben. Im Schnitt heben sich diese jedoch auf. Deshalb werden auch die Quadratterme der Residuen minimiert und nicht die Residuen selber, da bei Quadratwerten sich negative und positive Abweichungen nicht aufheben.

Die zweite Annahme ist eine sehr wichtige Annahme. Sie besagt, dass die Residuen nicht miteinander **korrelieren**. Dies tun sie dann, wenn wichtige Variablen nicht ins Modell aufgenommen wurden.

Die **Homoskedastizität** besagt, dass die Varianz der Fehlerterme gleichmässig verteilt ist. Ist dies nicht der Fall spricht man von Heteroskedastizität. Liegt Heteroskedastizität vor, ist OLS nicht mehr effizient und es müssen andere Verfahren angewandt werden.





Die Fehler sind **normalverteilt**.