lineare Regression: Diagnose

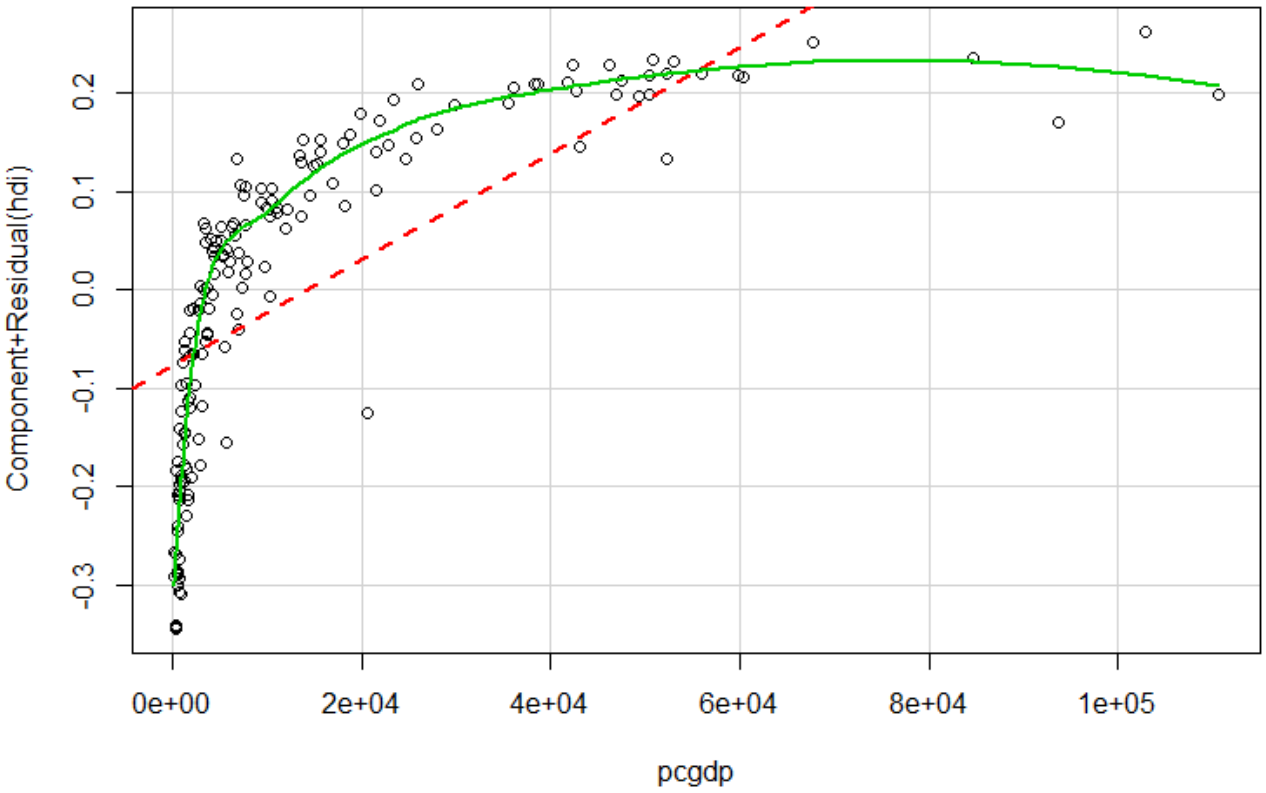
Benjamin Schlegel

05. Juni 2016

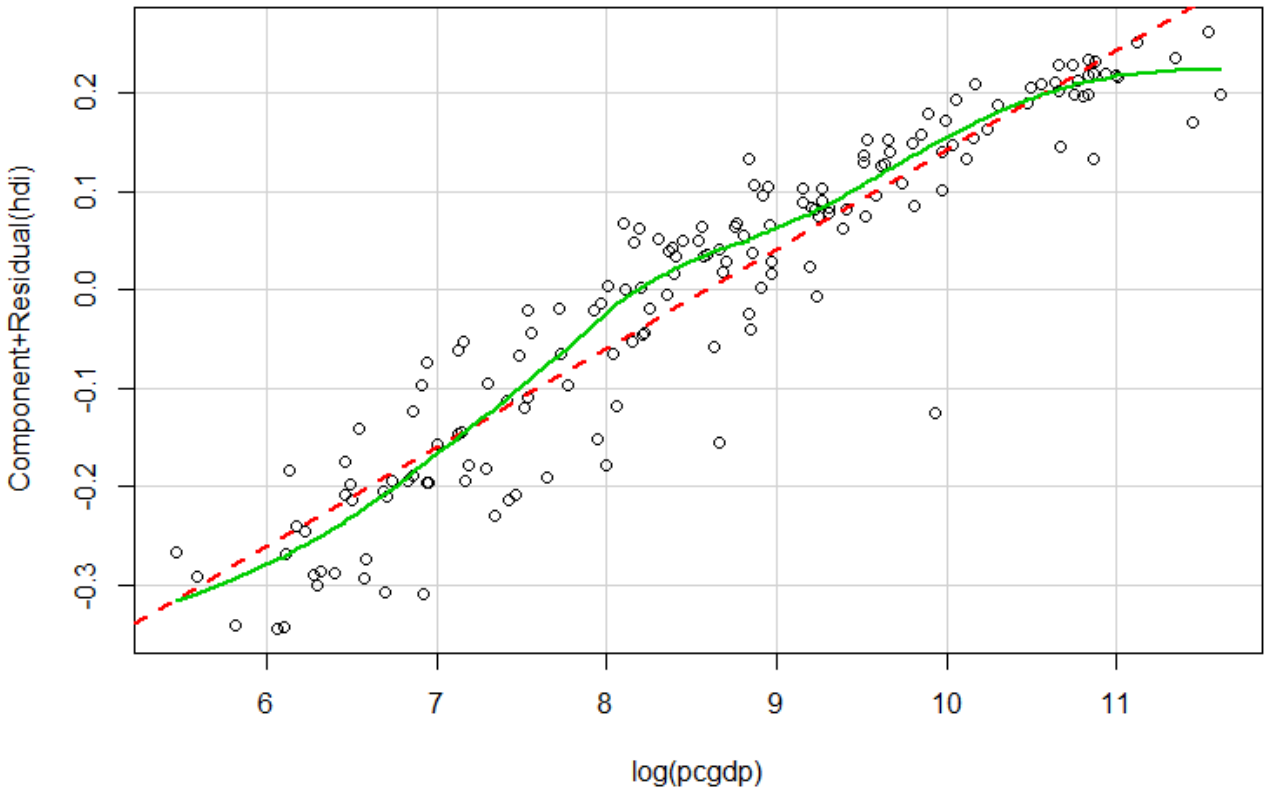
Bei einer lineare Regression können mehrere Probleme auftreten. Man kann prüfen, ob die Beziehung zwischen den Parametern linear ist, ob die Fehler normalverteilt sind, ob die Varianz konstant ist (Homoskedastizität), ob Multikollinearität herrscht oder ob Extremfälle Probleme verursachen.

Linearität

Eine Diagnosemöglichkeit ist zu schauen, ob die Beziehung zwischen der abhängigen und der unabhängigen linear ist. Dies ist wichtig, da die lineare Regression von einer linearen Beziehung in den Parametern ausgeht.

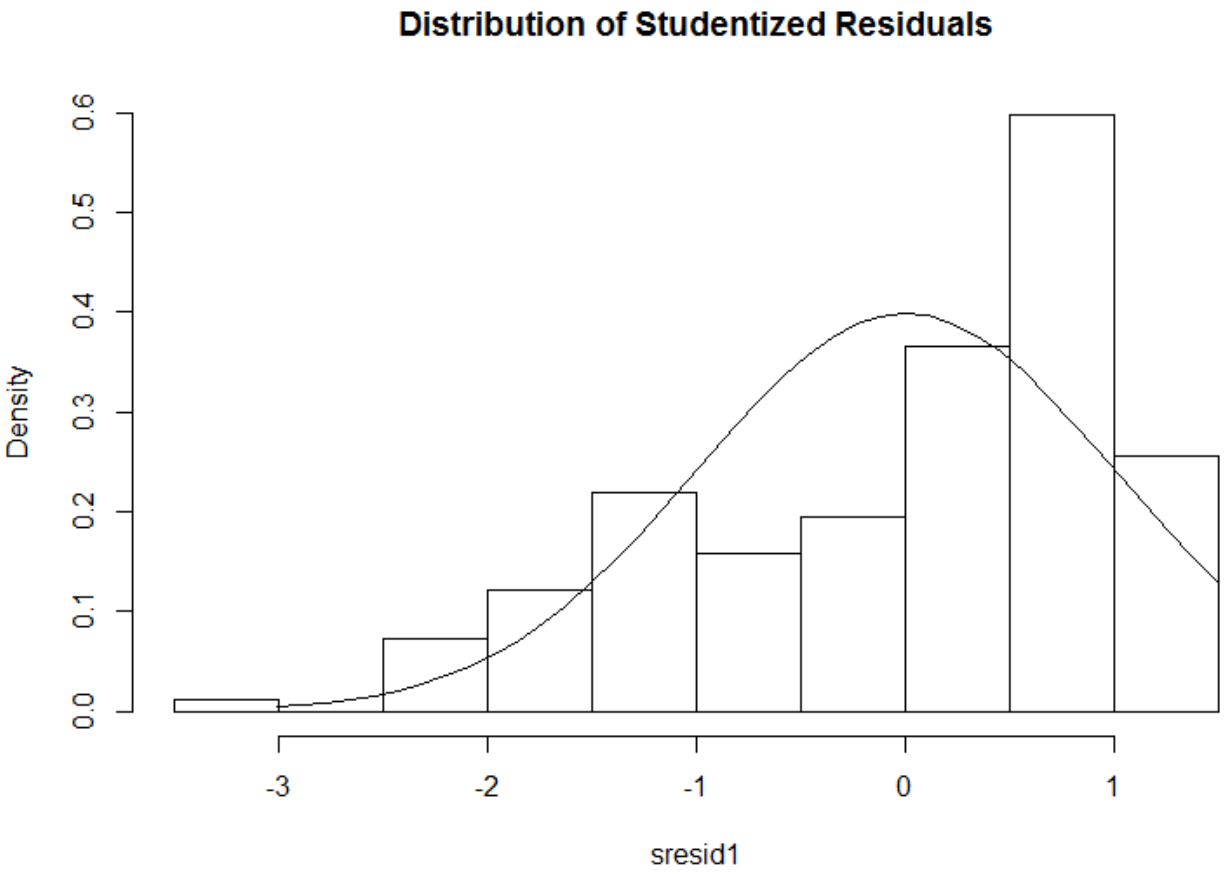


Ist die Beziehung nicht linear, kann eine Logarithmierung oder Quadrierung helfen, die Beziehung zu linearisieren. Im Beispiel in der Grafik kann beispielsweise eine Logarithmierung helfen, dann Zusammenhang linearer zu erhalten.

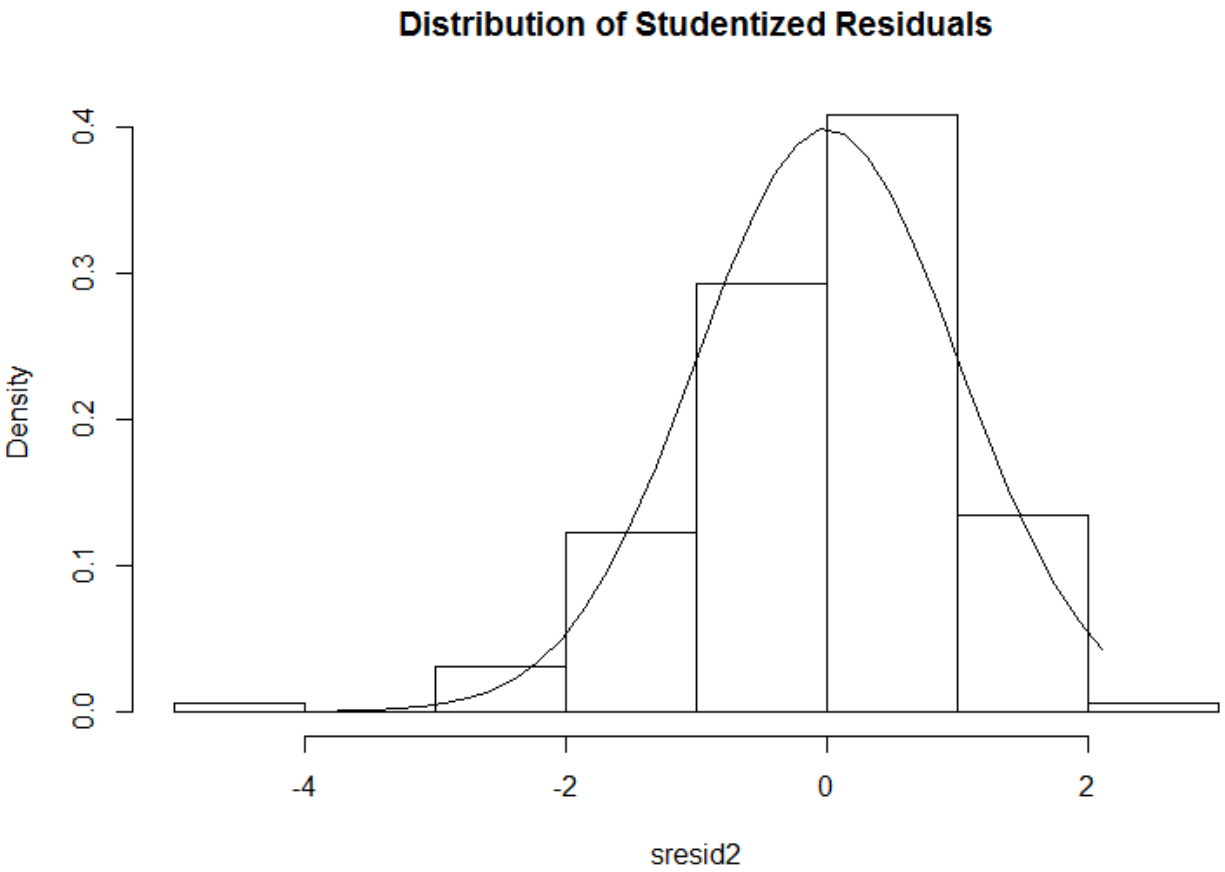


Normalverteilte Fehler

Eine wichtige Regressionsannahme ist, dass die Fehler normalverteilt sind.



Im Beispiel ist die Annahme verletzt. Damit ist die geschätzte Varianz nicht mehr zwingend die beste.

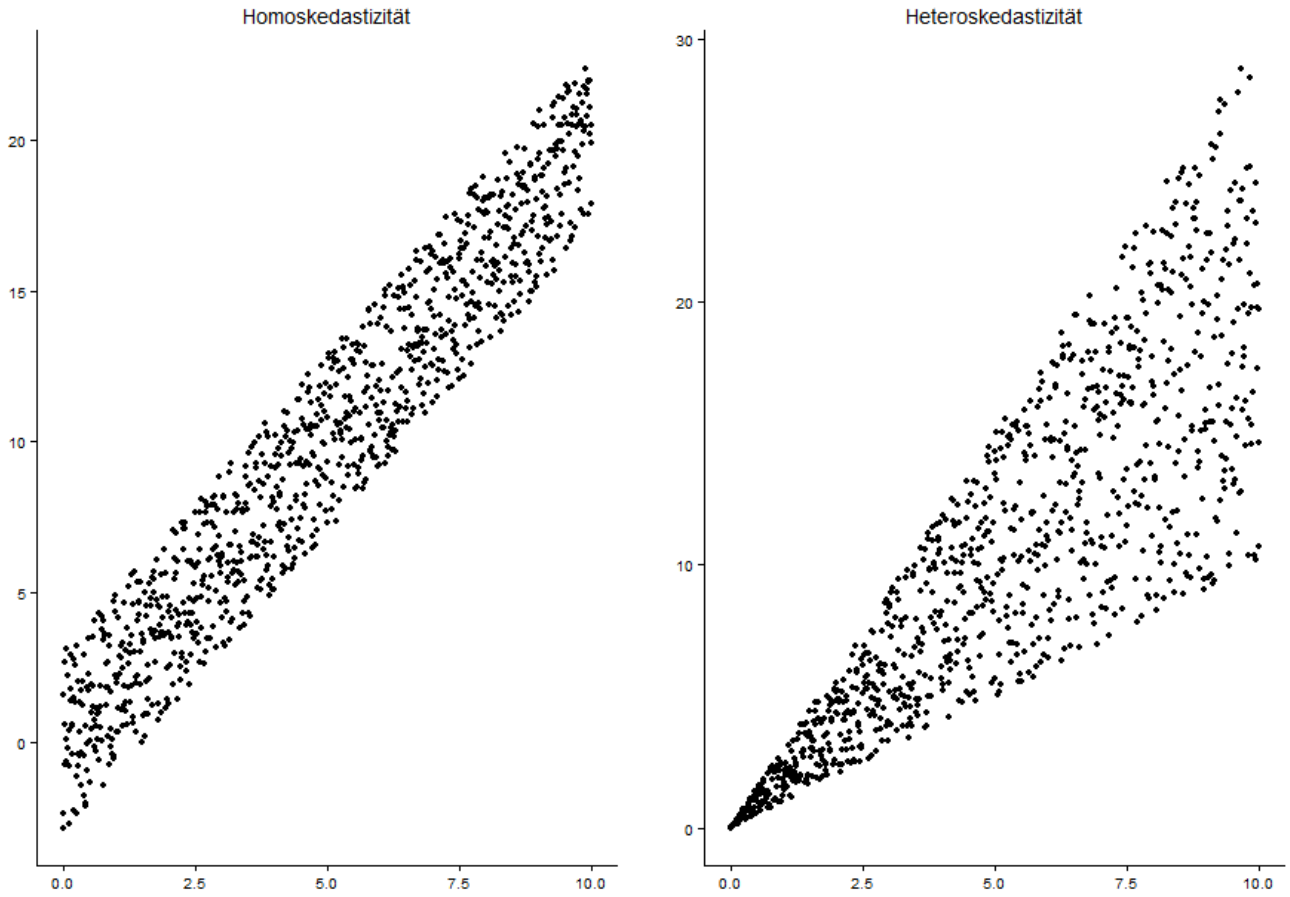


Auch hier konnte das Problem mit der Logarithmierung der unabhängigen Variable gelöst werden.

Homoskedastizität

Eine weitere wichtige Annahme der linearen Regression ist eine gleichmässig verteilte Varianz (Homoskedastizität). Ist diese Annahme verletzt (Heteroskedastizität), so werden die Standardfehler zu gross geschätzt. Heteroskedastizität kann mit dem Goldfeld-Quandt Test, Breusch-Pagan Test oder dem White Test diagnostiziert werden.

Beim Goldfeld-Quandt Test werden zwei Stichprobenhälften miteinander verglichen. Für beide Stichproben wird eine Regression gerechnet und anschliessend die Varianzen verglichen, ob sie genug ähnlich sind. Beim Breusch-Pagan Test werden die standardisierten Residuen auf die unabhängigen Variablen regressiert. Dabei sollte nur die Konstante (Achsenabschnitt) einen Wert ungleich Null haben. Sind die Koeffizienten der unabhängigen Variablen signifikant von Null unterschiedlich, herrscht Heteroskedastizität. Der White Test ist ein Spezialfall des Breusch-Pagan Tests, welcher weniger sensibel auf eine Verletzung der Normalverteilungsannahme der Residuen reagiert.



Das Problem kann mit robusten Standardfehlern nach Huber-White gelöst werden.

Multikollinearität

Es darf keine perfekte Multikollinearität herrschen. Wird vom Modell perfekte Multikollinearität festgestellt, wird in der Regel automatisch eine der Variablen aus dem Modell entfernt. Herrscht nicht perfekte aber starke Multikollinearität, ist es ein Problem, da die Schätzungen ungenau werden. Das Problem kann durch Varianz-Inflationsfaktoren diagnostiziert werden. Ist die Quadratwurzel grösser als 2, besteht ein Multikollinearitätsproblem.

Ausreisser und Hebelwirkung

Ausreisser können zu einen Problem werden, da sie die Schätzer verzerren. Ausreisser können durch Fehlermessungen entstehen oder einfach einen Sonderfall darstellen. Manchmal ist es sinnvoll, die Ausreisser aus dem Datensatz zu entfernen.

Ebenfalls zu einem Problem führen können Fälle mit einer starken Hebelkraft. Die Hebelkraft misst den Einfluss eines Wertes auf die Regression. Die Hebelkraft wird **Leverage** genannt.

Die **Cook's Distance** ist eine Mischung aus den beiden Werten und misst den Einfluss eines Wertes auf die Schätzung von OLS.