

Diagnostik von Regressionsmodellen (1)

Bei Regressionsanalysen sollte immer geprüft werden, ob das **Modell angemessen** ist und ob die **Voraussetzungen** eines Regressionsmodells erfüllt sind.

Das Modell einer linearen OLS-Regression ist dann angemessen, wenn

1. die **Residuen** (d.h. die Schätzfehler $e_i = y_i - \hat{y}_i$) so um die Regressionsgerade streuen, dass **kein systematischer Trend** (z.B. kurvilinear) mehr in ihnen enthalten ist,
2. die standardisierten oder studentisierten Residuen (s.u.) insgesamt **normalverteilt** sind,
3. die Varianz der Residuen über den gesamten Bereich der vorhergesagten Werte gleich ist (= **Varianzhomogenität** der Residuen); wenn dies nicht der Fall ist spricht man von Heteroskedastizität der Residuen;
4. und die Regressionskoeffizienten **nicht** durch nur wenige **besonders einflussreiche Fälle** entscheidend beeinflusst werden, was ebenfalls bei einer Analyse der Residuen sichtbar werden kann.

Diagnostik von Regressionsmodellen (2)

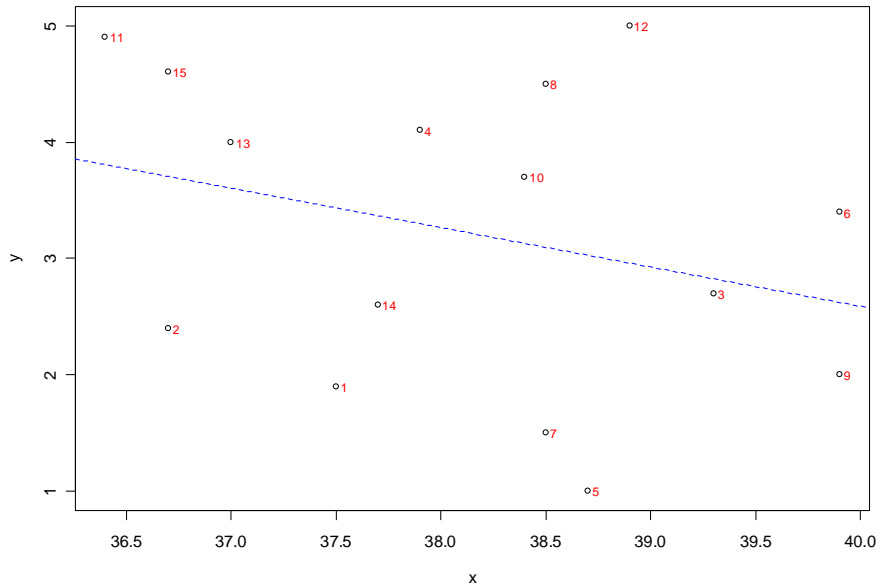
Streudiagramme der AV- und UV-Werte sind wichtige Hilfsmittel, eignen sich aber nur beschränkt zur Regressionsdiagnostik, insbesondere dann, wenn in dem Modell mehr als eine UV spezifiziert wird.

Auch für Modelle mit mehreren UVs sind **Residuenplots** sehr nützlich, in denen

1. die Residuen gegen die vorhergesagten Werte gezeichnet werden; hier lässt sich erkennen, ob systematische Trends durch die Regressionskoeffizienten möglicherweise nicht erfasst werden,
2. die standardisierten Residuen mittels sog. Q-Q-Plots (s.u.) darauf hin geprüft werden, ob sie normalverteilt sind,
3. die Wurzel aus dem Betrag der standardisierten Residuen gegen die vorhergesagten Werte gezeichnet werden; wird dabei ein systematischer Trend sichtbar, ist die Varianz der Residuen nicht homogen,
4. die standardisierten Residuen gegen Maße der „Hebelwirkung“ der UVs (*leverage*) gezeichnet werden; Fälle, die sowohl eine große *leverage* als auch einen großen Residualwert haben, haben besonderen Einfluss auf die Regressionskoeffizienten. Dies wird auch mit Hilfe von Höhenlinien der *Cooks Distanz*, die in dieselbe Grafik gezeichnet werden können, sichtbar.

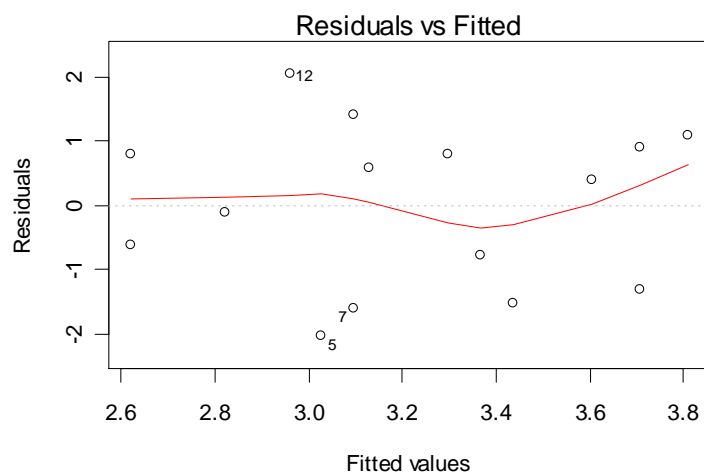
Diagnostik von Regressionsmodellen (3)

Beispiel: Einfache lineare Regression einer Y- auf eine X-Variable



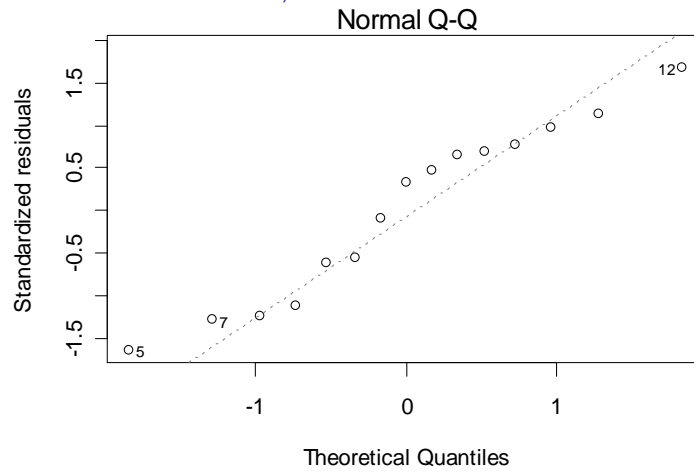
Diagnostik von Regressionsmodellen (4)

a) Plot der Residuen (Schätzfehler) gegen die vorhergesagten Werte (es ist kein systematischer, nicht erfasster Trend in den Residuen zu erkennen):

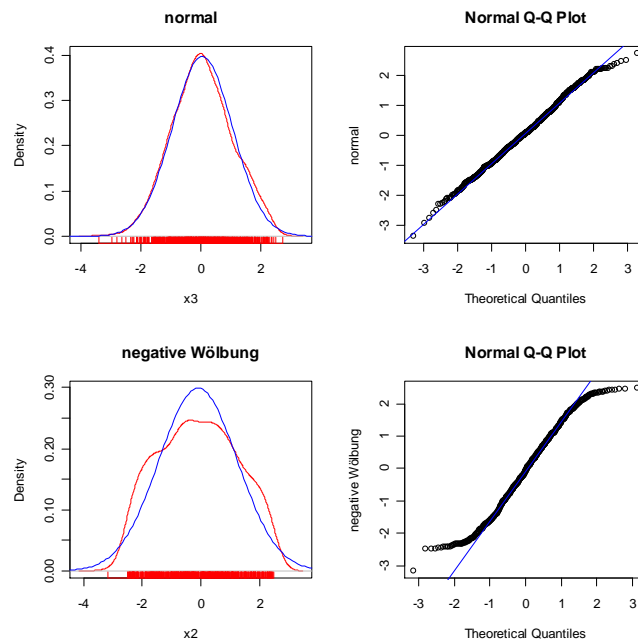


Diagnostik von Regressionsmodellen (5)

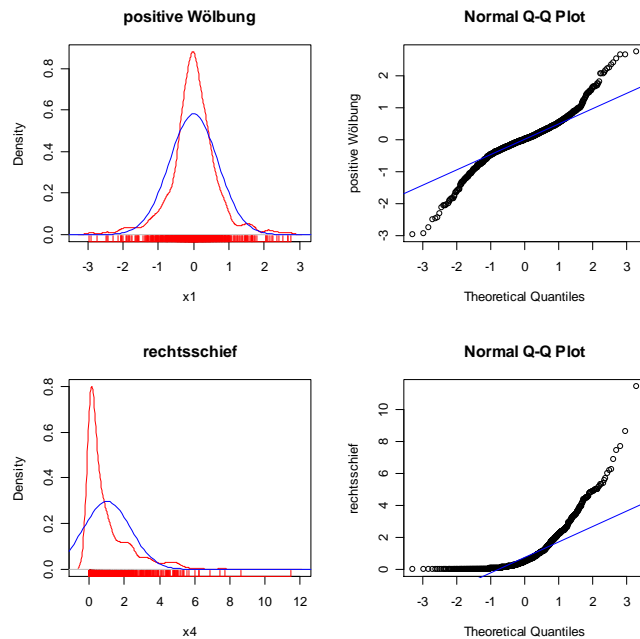
b) Normal-Quantile-Quantile-Plot; die standardisierten Residuen (s.u.) sollten normalverteilt sein, also auf der Geraden liegen (eventuell ist Fall 5 ein Fall, der aus dem Rahmen fällt):



Beispiele für Normal-Q-Q-Plots bei normalverteilten und nicht normalverteilten Daten (I)

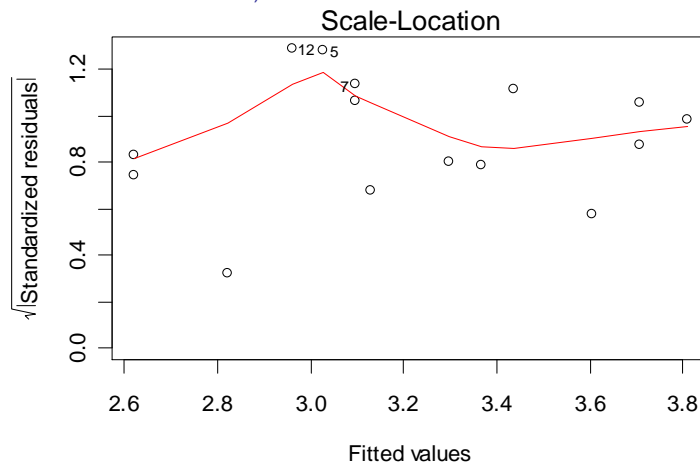


Beispiele für Normal-Q-Q-Plots bei normalverteilten und nicht normalverteilten Daten (II)



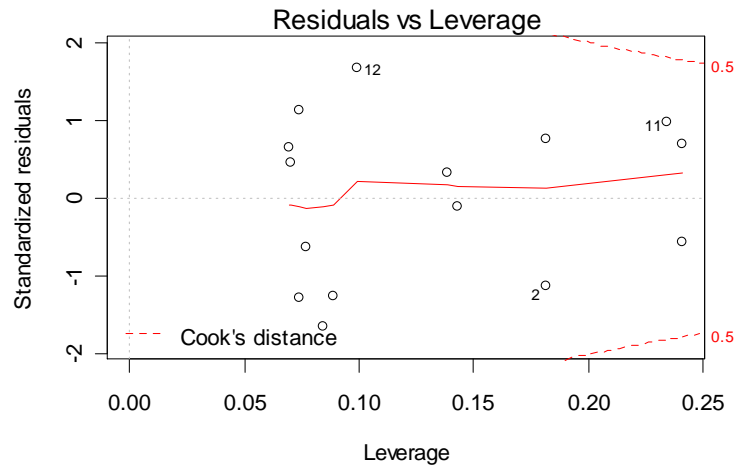
Diagnostik von Regressionsmodellen (6)

c) Wurzel aus dem Betrag der standardisierten (s.u.) Residuen gegen die vorhergesagten Werte. Die rote Linie zeigt, dass die Varianz der Residuen einigermaßen homogen ist (kein auffälliger Trend; die größten Residuen zeigen sich bei den Fällen 12 und 5):



Diagnostik von Regressionsmodellen (7)

d) Plot der standardisierten Residuen gegen *Leverage* (Maß der Extremität der Fälle hinsichtlich der UV) und *Cooks Distanz* (kein Wert überschreitet den kritischen Wert für einen extremen Einfluss auf die Regressionsgerade):



Diagnostik von Regressionsmodellen (8)

Leverage („Hebelwirkung“) misst die Extremität der Fälle hinsichtlich einer oder mehrerer **unabhängiger Variablen**. Für den Fall **einer** UV ist die Formel:

$$\text{leverage} = h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

Minimaler Wert ist $1/n$, maximaler Wert ist 1.0. Der Mittelwert bei n Fällen ist $(k+1)/n$, wobei k die Anzahl der UVs ist.

Bei großer Fallzahl und vielen UVs werden **Werte größer $2(k+1)/n$ als groß** betrachtet, in kleinen Stichproben erst bei Werten größer als $3(k+1)/n$ (bei einer UV und 11 Fällen wäre dieser Wert 0.55).

Es ist sinnvoll, nur wenige Fälle mit größter Leverage genauer zu prüfen. Fälle mit großer Leverage **können** (aber müssen nicht) die Ergebnisse von Regressionsanalysen beeinflussen; ob sie das tun, hängt zusätzlich davon ab, wie groß die jeweiligen Residuen sind.

Diagnostik von Regressionsmodellen (9)

Cooks Distanz informiert darüber, wie sehr ein einzelner Fall die Koeffizienten der Regressionsgleichung beeinflusst. Die Formel ist:

$$D_i = \frac{\sum (\hat{y} - \hat{y}_{(i)})^2}{(k+1) \cdot \hat{\sigma}_{\text{residuen}}^2}$$

wobei k der Anzahl der unabhängigen Variablen und $\hat{y}_{(i)}$ dem vorhergesagten Y -Wert ohne Fall i entspricht. D.h., es werden die vorhergesagten Werte von Y in Regressionsgleichungen mit und ohne Fall i für alle Daten des Datensatzes verglichen.

Minimaler Wert ist 0, je höher der Wert, um so größer der Einfluss eines Falls auf die Koeffizienten der Regressionsgleichung.

Als einflussreich gelten **Werte ab 1.0** oder genauer ab dem kritische Wert der F -Verteilung bei $\alpha = .50$ mit $(k+1)$ und $(n-k-1)$ Freiheitsgraden. Bei einer UV und einer Stichprobe mit 11 Fällen wäre der Wert 0.75.

Fälle, die **sowohl** eine starke Hebelwirkung (*leverage*) **als auch** ein großes standardisiertes Residuum haben (s.u.), üben einen großen Einfluss auf die Regressionskoeffizienten aus: Dies schlägt sich in einem großen Wert für Cooks Distanz nieder.

Diagnostik von Regressionsmodellen (10)

Die **standardisierten Residuen** (auch „internally studentized residuals“ genannt)* in den Grafiken b), c) und d) stellen das Verhältnis der „rohen“ Residuen $e_i = y_i - \hat{y}_i$ zur geschätzten Standardabweichung der individuellen Residuen dar. Je kleiner die Standardabweichung der individuellen Residuen, um so präziser werden sie geschätzt. Sie ergibt sich aus dem geschätzten Standardschätzfehler in der Population gewichtet an der Wurzel aus $(1 - \text{leverage})$:

$$\hat{\sigma}_{e_i} = \hat{\sigma}_{\text{residuen}} \cdot \sqrt{1 - h_i} \quad \text{mit} \quad \hat{\sigma}_{\text{residuen}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - k - 1}}$$

wobei k der Anzahl der unabhängigen Variablen im Gesamtmodell darstellt. Die Präzision der geschätzten Residuen ist also um so größer, je kleiner die „Hebelwirkung“ des Falles bezogen auf die X -Variablen und je geringer die Fehlervarianz der Regressionsgleichung ist.

Das standardisierte Residuum eines Falles i ist gleich $\frac{e_i}{\hat{\sigma}_{e_i}}$.

Der Betrag der standardisierten Residuen liegt zwischen 0 und $\sqrt{n-k-1}$.

* Studentisierte Residuen (auch „externally studentized residuals“ genannt) haben bessere statistische Eigenschaften und werden deshalb häufiger benutzt. Auf sie wird wegen ihrer Komplexität an dieser Stelle nicht eingegangen.