Vorlesung vom 10.05.2022

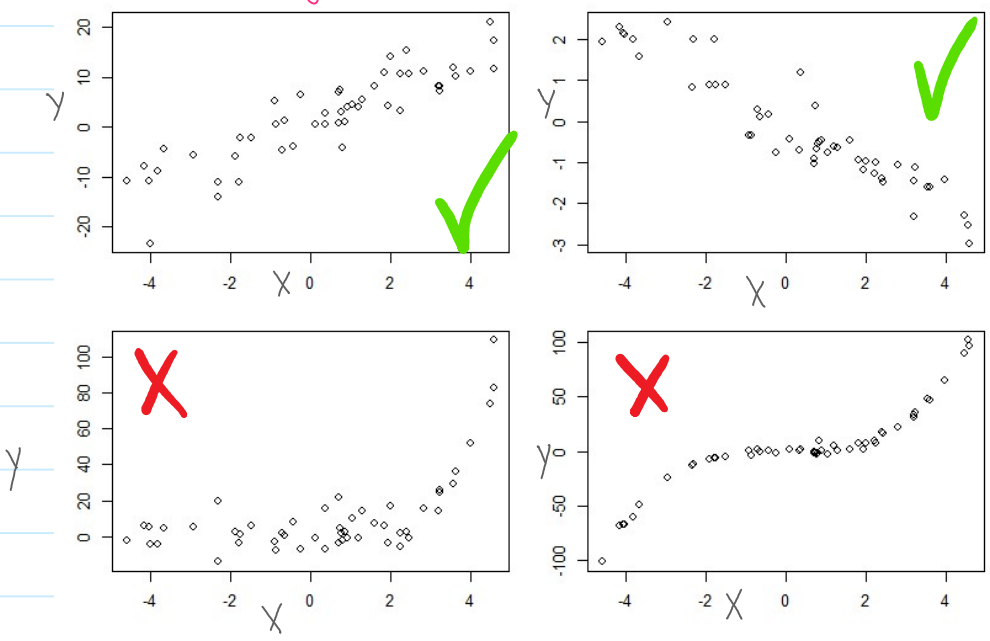
12.13 Residualanalyse

Mit dem Schätzen und Testen in einem linearen Regressionsmodell sollte man auch eine Modelldiagnose verbinden. Darunter versteht man statistische Werkzeuge, mit denen überprüft werden kann, ob die Annahmen des Standardmodells - zumindest approximativ - erfüllt sind oder deutliche Abweichungen vorliegen. Es gibt statistische Test um Modellannahmen zu überprüfen. Wir behandeln hier aber vor allem grafische Modelldiagnosemethoden, welche auf den Residuen basieren.

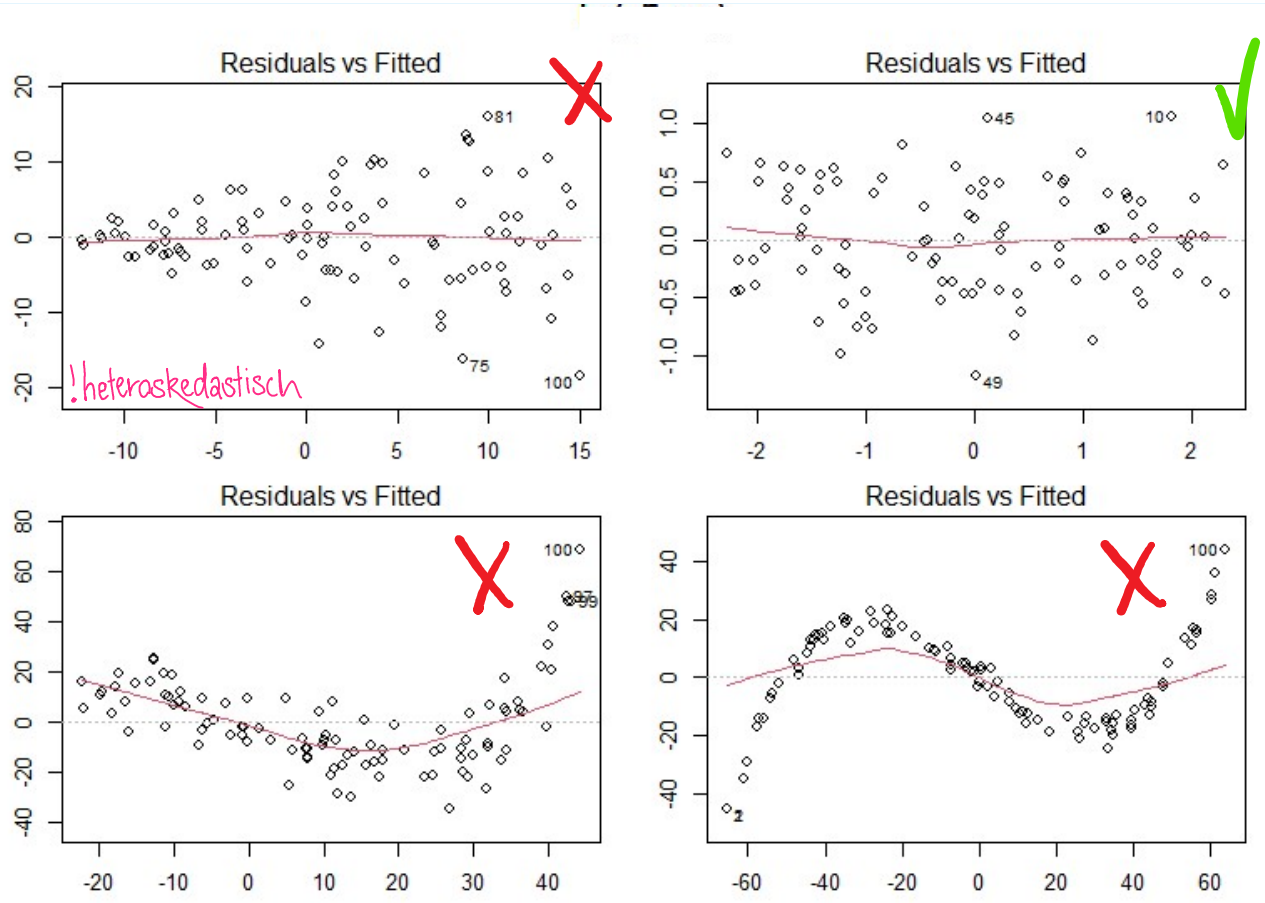
① Linearität (und Homoskedastizität)

Ob es sinnvoll ist überhaupt einen linearen Zusammenhang zwischen erklärenden Variable X und abhängiger Variable Y anzunehmen, überprüft man entweder mit Hilfe eines Streudiagramms (funktioniert nur bei der einfachen linearen Regression) oder mit Hilfe eines Residualplots.

Streudiagramm



Alternativ zum Streudiagramm, kann mit einem Residualplot überprüft werden, ob die Residuen nichtlineare Muster aufweisen. Wenn eine nichtlineare Beziehung zwischen den erklärenden Variablen und der abhängigen Variable besteht, wird dies im Residualplot sichtbar. Sind die Residuen gleichmässig um eine horizontale Linie verteilt und keine speziellen Muster ersichtbar sind, so ist dies ein guter Hinweis darauf, dass keine nichtlinearen Beziehungen vorliegen.



② Normalverteilung

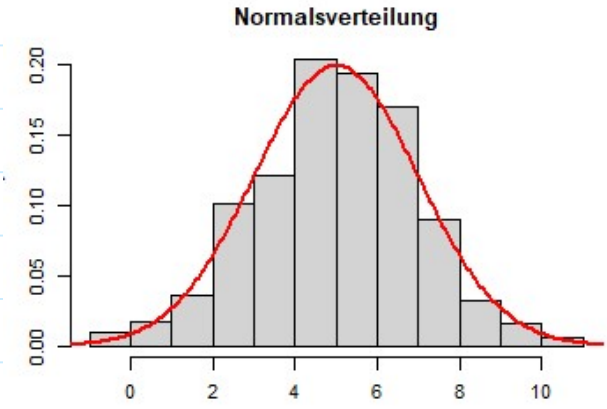
Es gibt verschiedene Optionen zu überprüfen, ob eine Normalverteilung in den Daten vorliegen könnte. Wir schauen uns eine graphische Methode mit dem Namen Quantil-Quantil-Plot (Q-Q-Plot) oder Normal-Quantil-Plot (N-Q-Plot) an. Bei einem Q-Q-Plot werden die Quantile der Häufigkeitsverteilung des Merkmals mit den Quantilen einer Normalverteilung verglichen (der Einfachheit halber, wird meist die Standardnormalverteilung verwendet).

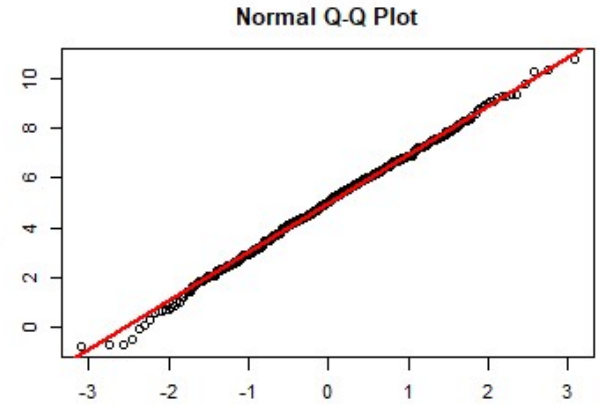
Quantil-Quantil-Plot

Seien $x_{(1)}, \ldots, x_{(n)}$ die geordnete Ordnungsstatistiken. Für $i = 1, \ldots, n$ werden die $(i - 0.5)/n$-Quantile $z_{(i-0.5)/n}$ der Standardnormalverteilung berechnet. Der Quantil-Quantil-Plot ist nun ein Streudiagramm der Punkte

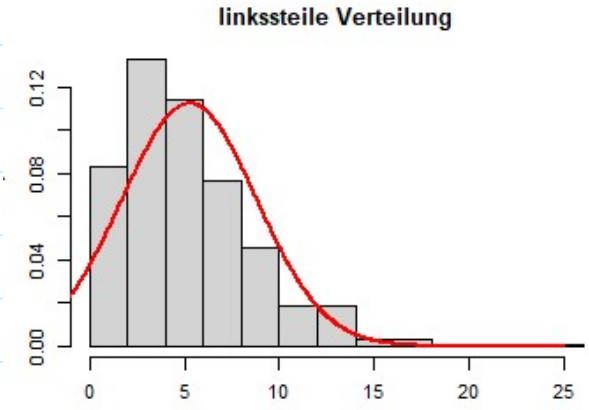
$$(z_{(1-0.5)/n}, x_{(1)}), \ldots, (z_{(n-0.5)/n}, x_{(n)}).$$

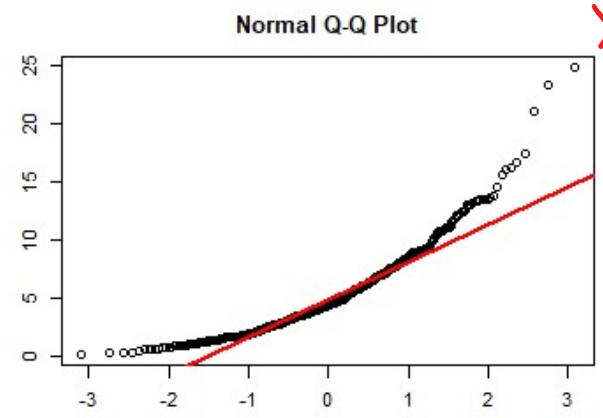
Für grosses n wird der Rechenaufwand relativ gross. Q-Q-Plots werden deshalb in statistischen Programmen am Computer erstellt.



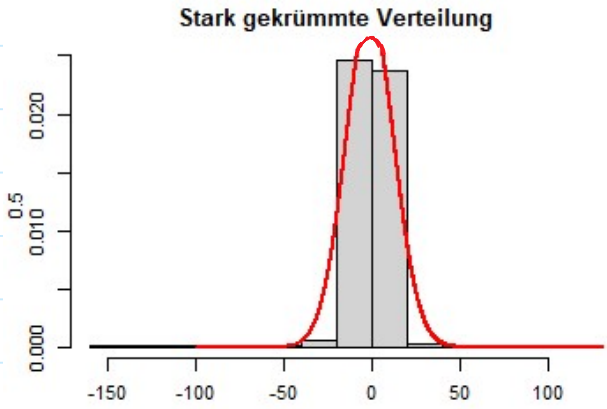


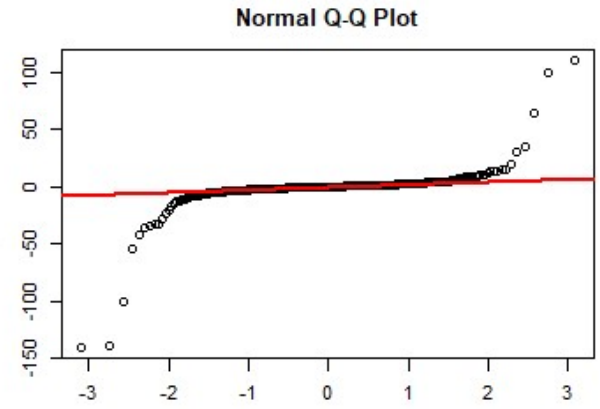
✓





✗





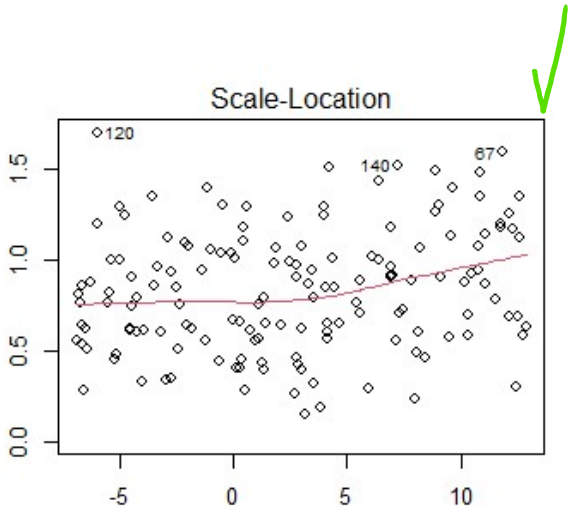
✗

③ Homoskedastizität

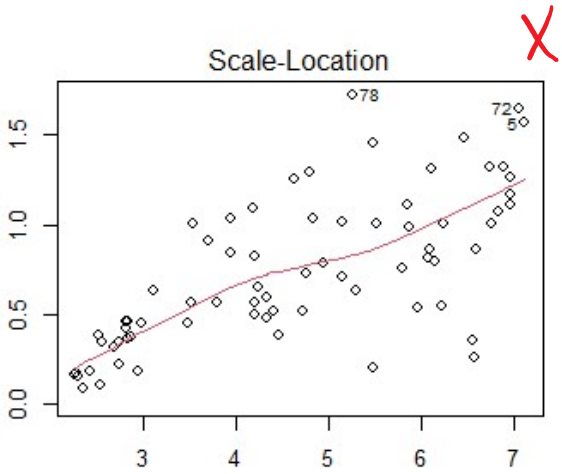
Die Annahme der konstanten Varianz der Fehler ist zentral für unsere Schlussfolgerungen. Wenn die Annahme beispielsweise nicht zutrifft, dann stimmt das Signifikanzniveau der Tests nicht.

Heteroskedastizität kann entweder mit einem Residualplot (unter ①) oder auch mit Hilfe eines Scale-Location-Plots überprüft werden.

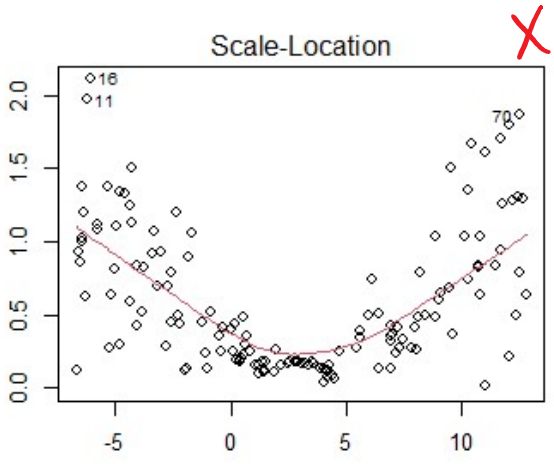
Bei Homoskedastizität erwarten wir, dass die rote Linie etwa konstant um 1 ist. Wenn es klare, nicht konstante Muster gibt, dann gibt es Hinweise auf Heteroskedastizität.



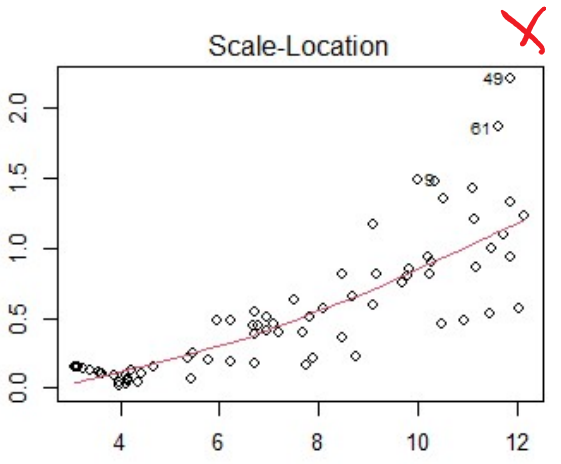
✓



✗



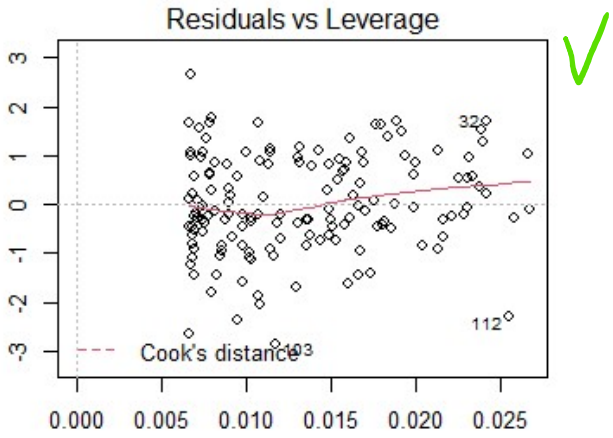
✗



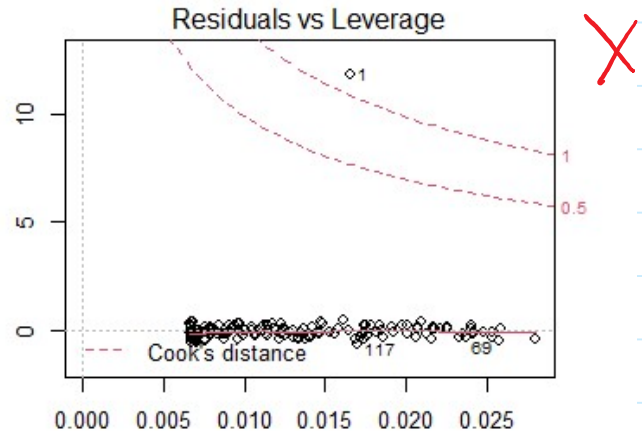
✗

④ Hebelwirkung (Leverage)

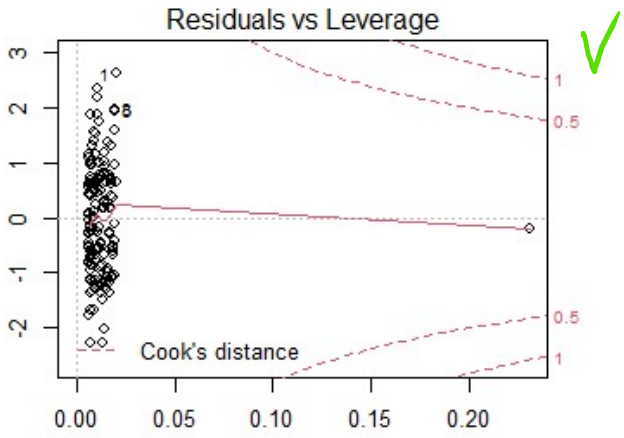
Punkte mit hoher Hebelwirkung sind Beobachtungen, die sich sehr stark auf die Schätzer der Modells auswirken. Sind Punkte mit starker Hebelwirkung vorhanden, so ist das Modell mit Vorsicht zu interpretieren. Punkte mit starker Hebelwirkung können mit einem "Leverage vs Residuals"-Plot untersucht werden.



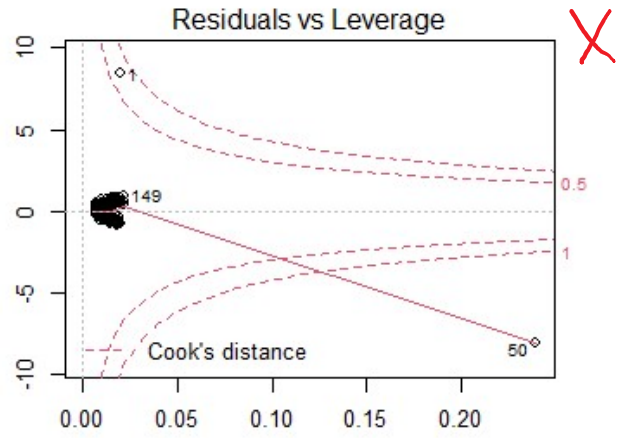
✓



✗



✓

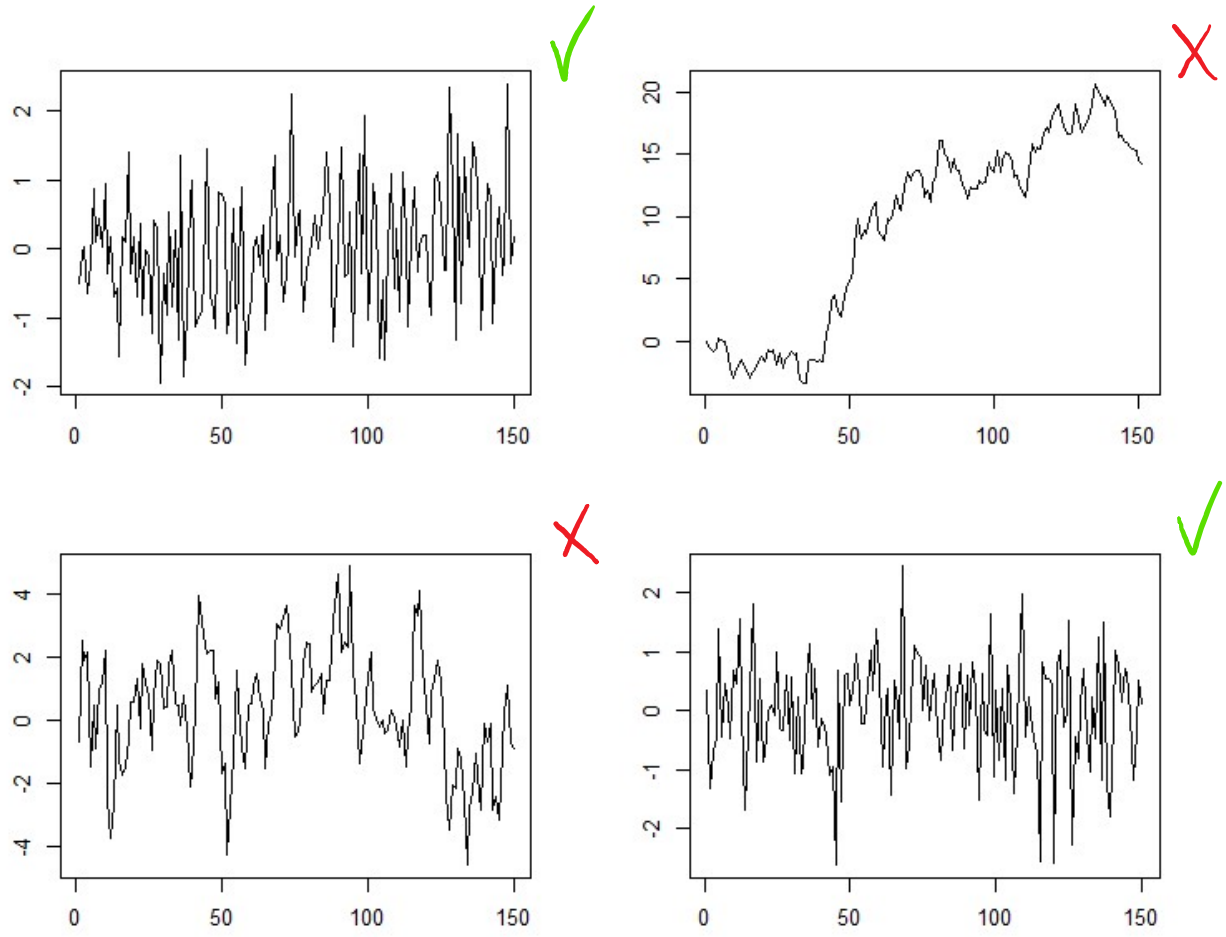


✗

⑤ Unabhängigkeit

Auch die Unabhängigkeit ist auch eine Schlüsselannahme. Wenn es starke Abhängigkeiten gibt, werden Informationen wiederholt, und infolgedessen wird die Variabilität der Schätzungen gross.

Die Menge an möglichen Abhängigkeitsstrukturen bei den Residuen ist riesig, und es gibt keine einfache Möglichkeit, sie alle zu überprüfen. Als Erstes sollte man sich überlegen, ob eine Abhängigkeitsstruktur zwischen den Messungen vorhanden sein kann. Meist kennt man sich auf dem Gebiet, welches untersucht wird ja relativ gut aus und kann daher auch einschätzen, ob die Abhängigkeitsannahme erfüllt sein könnte. Häufig wird zusätzlich das Vorkommen von Autokorrelation graphisch geprüft. Diese tritt auf, wenn eine serielle Abhängigkeit bei der Messung der Beobachtungen besteht, und ermöglicht es, zeitliche Trends zu erkennen.



Ist keine Korrelation vorhanden, weisen die Reihen kein spezifisches Verhalten der Residuen auf. Das heisst, dass die näheren Beobachtungen keine ähnlichen Werte annehmen, sondern sich ohne erkennbares Muster verändern. Andere Möglichkeiten, welche auf Autokorrelation hinweisen, sind z.B. abwechselnd kleine und grosse Residuen oder abwechseln positive und negativen Residuen.

Sind die Annahmen der (einfachen) linearen Regression nicht erfüllt, so gibt es verschiedene Möglichkeiten. Manchmal reicht es die Zielvariable zu transformieren. Ansonsten gibt es meist Abwandlungen des Standardmodells.

12.12 Prognose

Häufig ist man auch daran interessiert, zu einem neuen Wert x_0 die Realisierung der Zielvariable Y_0 zu schätzen.

Beispiel: Für die geplante Dosis eines Medikaments x_0 möchten wir den Blutdruck Y_0 Blutdruck eines Patienten kennen. Oder für x_0 eingesetzte Werbekosten und möchten wir den prognostizierten Umsatz Y_0 vorhersagen.

Nimmt man an, dass auch für Y_0 und x_0 das gleiche lineare Regressionsmodell

$$Y_0 = \alpha + \beta x_0 + \varepsilon$$

gilt wie für die Y_i, x_i, $i = 1, \dots, n$, so ist wegen $E(\varepsilon) = 0$ die Zufallsvariable

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0$$

ein vernünftiger Punktschätzer für Y_0. Über die Eigenschaften von $\hat{\alpha}$ und $\hat{\beta}$ lässt sich auch die Varianz des Prognosefehlers $Y_0 - \hat{Y}_0$ und daraus ein Vertrauensintervall für Y_0 bestimmen:

$$\left[Y_0 \pm t_{1-\alpha/2}(n-2)\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \right]$$

Man kann auch ein Konfidenzintervall für die "wahre" Regressionsgerade selbst bestimmen, und zwar mit

$$\left[\hat{\alpha} + \hat{\beta} x \pm t_{1-\alpha/2}(n-2)\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \right]$$