

4 Residuen-Analyse

4.1 Problemstellung

- a Die eingeführten Schätz- und Testmethoden beruhen auf **Modellannahmen**: Für die **Fehler** wurde $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$ (unabhängig) angenommen. Das kann man aufspalten:
- Der Erwartungswert der E_i ist $\mathcal{E}\langle E_i \rangle = 0$,
 - sie haben alle die gleiche theoretische Varianz $\text{var}\langle E_i \rangle = \sigma^2$,
 - sie sind normalverteilt
 - sie sind unabhängig,
- Für die Regressionsfunktion muss jeweils eine bestimmte Formel angesetzt werden, die nur einige Parameter $\beta^{(j)}$ offen lässt. Im oben besprochenen Sinne (3.2.w) wird **Linearität** vorausgesetzt. Wenn die Formel nicht die Form hat, die für die Daten „eigentlich gilt“, ist für die Fehler Annahme (a) verletzt.
- b Diese Voraussetzungen zu überprüfen, ist meistens wesentlich. Es geht dabei nicht in erster Linie um eine Rechtfertigung, sondern um die Möglichkeit, aus allfälligen Abweichungen ein **besseres Modell** entwickeln zu können. Das kann bedeuten, dass
- Variable transformiert werden,
 - zusätzliche Terme, beispielsweise Wechselwirkungen, ins Modell aufgenommen werden,
 - für die Beobachtungen Gewichte eingeführt werden,
 - allgemeinere Modelle und statistische Methoden verwendet werden.
- c Die Chancen der Modell-Verbesserung wahrzunehmen, entspricht der Grundhaltung der **explorativen Datenanalyse**. Es geht hier nicht um präzise mathematische Aussagen, Optimalität von statistischen Verfahren oder um Signifikanz, sondern um Methoden zum kreativen Entwickeln von Modellen, die die Daten gut beschreiben. Wir kommen gleich noch etwas konkreter auf die Bedeutung der Überprüfung von Voraussetzungen zurück (4.2.e).
- d Die Residuenanalyse bedient sich einiger grafischer Darstellungen und allenfalls auch einiger formaler Tests. Diese können **Symptome** dafür finden, dass ein Modell die Daten nicht genau beschreibt. Symptome können sich zu Syndromen zusammenfügen, die auf bekannte „Krankheiten“ hinweisen und die wirksame „Therapie“ klar machen. Schwierig wird es, wenn mehrere Aspekte des Modells falsch sind und sich deshalb mehrere Syndrome überlagern. Dann kann es schwierig werden, aus den verschiedenen Symptomen auf die „richtigen“ Verbesserungen des Modells zu schließen. Die Entwicklung eines Modells braucht dann Intuition, Erfahrung und Kreativität – und gute **Diagnose-Instrumente**, nämlich solche, die möglichst spezifisch sind für die Verletzung einzelner Voraussetzungen oder für die Wirksamkeit bestimmter Modellveränderungen (vergleiche 4.2.j).

- e Die Mittel zur Überprüfung von Voraussetzungen werden hier für die multiple lineare Regression mit normalverteilten Fehlern dargestellt. Die meisten Ideen sind in der **Varianzanalyse** direkt anwendbar und lassen sich auch auf andere Regressionsmodelle übertragen und sind damit grundlegend für weiteren Kapitel.

4.2 Residuen und angepasste Werte

- a In der einfachen Regression können die Voraussetzungen – mit Ausnahme der Unabhängigkeit (d) – anhand eines Streudiagramms der Zielgrösse gegen die erklärende Variable beurteilt werden. Für die multiple Regression entsteht eine ebenso anschauliche Darstellung, wenn auf der horizontalen Achse die **angepassten Werte** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \hat{\beta}_2 x_i^{(2)} + \dots + \hat{\beta}_m x_i^{(m)}$ verwendet werden, wie das schon in 3.1.h getan wurde. Was sagt uns diese Abbildung über die einzelnen Voraussetzungen?

- b (a) **Regressionsfunktion:**

▷ Die Gerade passt im Beispiel recht gut zum „**Verlauf** der Punkte“. Wenn man genau hinsieht, haben die Punkte etwas rechts von der Mitte (\hat{y}_i zwischen 0.4 und 0.7) die Tendenz, ein wenig höher zu liegen, während die Punkte rechts und links häufiger unterhalb der Geraden anzutreffen sind.

Eine leicht gekrümmte Kurve würde etwas besser zu den Daten passen. Das deutet darauf hin, dass der Erwartungswert der Zielgrösse durch die verwendete Regressionsfunktion nicht genau beschrieben wird und deshalb $\mathcal{E}\langle E_i \rangle \neq 0$ ist.

- c (b) **Gleiche Varianzen:**

▷ Die **Streubreite** der Punkte um die Gerade ist einigermaßen gleichmässig – bis auf einen oder zwei Punkte, die man als „**Ausreisser**“ bezeichnen kann, einen bei $\hat{y}_i \approx 0.73$, der nach unten abweicht, und einen bei $\hat{y}_i \approx 0.6$, der etwas zu hoch liegt. Diese extremen Punkte verletzen eher die Voraussetzung der Normalverteilung (c) als die der gleichen Varianzen (b).

Eine typische Abweichung von der Voraussetzung der gleichen Varianzen führt dazu, dass die Streubreite der Punkte für grössere angepasste Werte grösser wird, im Diagramm also die Punkte gegen rechts „trichterförmig“ auseinanderlaufen – oder umgekehrt, was seltener vorkommt (vergleiche 4.4.b). Wenn die Varianzen der Fehler verschieden sind, aber nichts mit den Werten der Regressionsfunktion zu tun haben, werden wir das in dieser Figur nicht sehen.

* Die Voraussetzung der gleichen Varianzen wird mit dem Zungenbrecher **Homoskedastizität**, jede Abweichung davon mit **Heteroskedastizität** bezeichnet.

- d (c) **Verteilung der Fehler:** Die Abweichungen von der Geraden sind die **Residuen** $R_i = Y_i - \hat{y}_i$. Sie streuen einigermaßen **symmetrisch** um die Gerade. Die beiden „Ausreisser“ haben wir schon kommentiert. Sie deuten auf eine „langschwänzige“ Verteilung hin. Auf die Beurteilung der Verteilung der Fehler kommen wir noch zurück (4.3.a).

- e Die hier festgestellten Abweichungen von den Voraussetzungen sind ohne Weiteres zu tolerieren. So die **Beurteilung** des Autors. Das ist eine reichlich unwissenschaftliche Aussage! Und in welchem Sinne „zu tolerieren“? Das ist nicht präzise zu fassen. Hier einige Überlegungen dazu:
- Bei exakter Gültigkeit der Voraussetzungen gibt es in den Daten immer wieder scheinbare Abweichungen – wie ja bei strikt durchgeführten Tests in 5% der Fälle signifikante Effekte auftreten, wenn die Nullhypothese exakt gilt. Mit Erfahrung lässt sich etwa abschätzen, wie gross solche **zufälligen Abweichungen** etwa werden können. Wir werden gleich noch diskutieren, wie man die zufälligen Abweichungen präziser fassen kann.
 - Selbst wenn in irgendeinem Sinn signifikante Abweichungen von den Voraussetzungen vorliegen, kann die Anwendung der im vorhergehenden Kapitel besprochenen Methodik immer noch zu genügend korrekten Resultaten führen. Solche Beurteilungen beruhen auf dem Wissen und der Erfahrung über die **Auswirkungen von Abweichungen auf einzelne Resultate** der Methoden, wie Verteilungen von Schätzungen, P-Werte von Tests und Ähnlichem.
 - Wie wichtig präzise Aussagen der statistischen Methoden sind, hängt von der **wissenschaftlichen Fragestellung** ab. Wenn es um eine präzise Schätzung des Effekts einer erklärenden Variablen auf die Zielgrösse in einem gut fundierten Modell geht, sind die Voraussetzungen kritischer, als wenn es darum geht, in einer Vielzahl von möglichen erklärenden Variablen die wichtigen von den unwichtigen zu trennen.

Nach diesen allgemeinen Bemerkungen zurück zum Konkreten! Wir wollen die einzelnen Voraussetzungen noch genauer untersuchen, mit besser geeigneten grafischen Darstellungen.

- f Die Betrachtungen zum Streudiagramm der beobachteten und angepassten Werte (3.1.h) lassen sich noch präziser fassen, wenn wir die Abbildung etwas abändern: Statt der beobachteten Werte Y_i tragen wir in vertikaler Richtung die **Residuen** R_i ab. Das hilft vor allem dann, Abweichungen deutlicher zu sehen, wenn die Punkte in 3.1.h wenig um die Gerade streuen, wenn also die multiple Korrelation oder das Bestimmtheitsmass R^2 hoch ist und die Residuen deshalb klein werden. Die so entstehende Darstellung heisst nach den Autoren, die sie als unverzichtbaren Bestandteil der Residuenanalyse propagiert haben, **Tukey-Anscombe-Diagramm** (Abbildung 4.2.f). In dieser Darstellung sollten die Punkte gleichmässig um die Nulllinie $R = 0$ streuen.
- g In Abbildung 4.2.f ist eine fallende Gerade eingezeichnet, die Punkte zusammenfasst, für die die Zielgrösse Y konstant (gleich dem Mittelwert der Y_i) ist. Sie wird sich als **Referenzlinie** als nützlich erweisen (4.4.m), wird aber von Programmen (bisher) nicht gezeichnet.
- Wir wollen nun die Voraussetzungen nochmals mit diesem neuen Diagramm prüfen.
- h **(a) Regressionsfunktion:** Eine Kurve in 3.1.h wird zu einer entsprechenden, „flach gelegten“ Kurve in 4.2.f. Von Auge können wir zwar Muster in solchen Darstellungen recht gut erkennen, aber es erweist sich oft als nützlich, eine mögliche Kurve einzuzeichnen. Man erhält sie mit einer geeigneten **Glättungsmethode**.

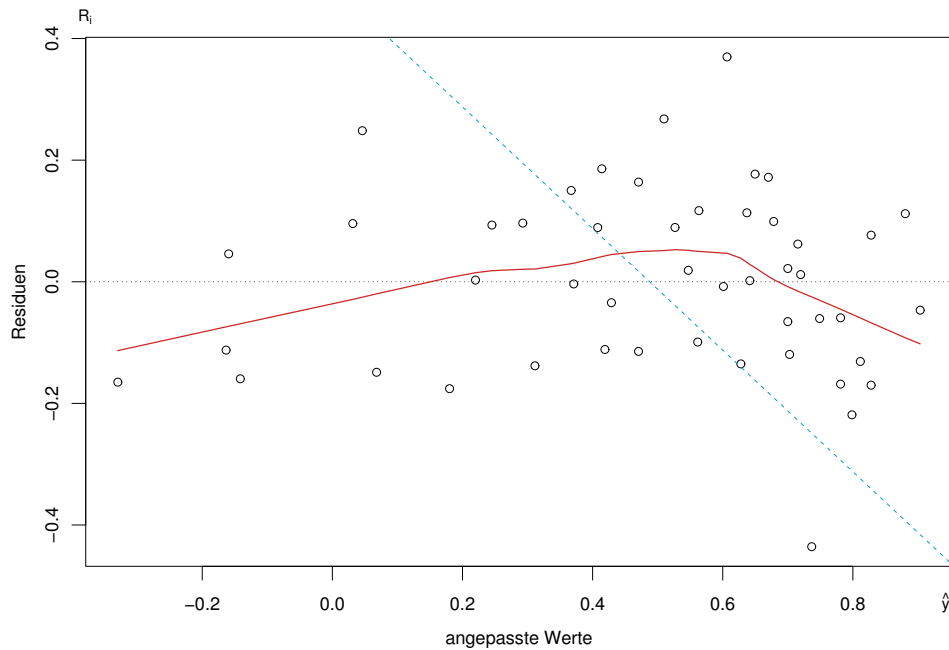


Abbildung 4.2.f: Tukey-Anscombe-Diagramm für das Beispiel der Sprengungen, mit einer Glättung und der Referenzgeraden $Y = \bar{Y}$

- i Die Voraussetzung (a) lautet ja: $\mathcal{E}\langle E_i \rangle = 0$. Wenn wir nun einige Beobachtungen mit ähnlichem \hat{y}_i zusammennehmen, also einen vertikalen Streifen in Abbildung 4.2.f herausgreifen, sollte der Mittelwert der Residuen R_i ungefähr 0 ergeben. Man kann einen solchen Streifen mit vorgegebener Breite h wählen und den Mittelwert der Residuen in der Mitte des Streifens in vertikaler Richtung einzeichnen (Abbildung 4.2.i). Variiert man nun die Position des Streifens, entlang der horizontalen Achse, so erhält man das **gleitende Mittel** (*running mean*). Diese Beschreibung soll nur die Grundidee des Glättens mit der wohl einfachsten Idee erklären. Das Verfahren kann leicht verbessert werden und sollte deshalb nicht verwendet werden. Genaueres zu Glättungsmethoden bringt der Block über „Nichtparametrische Regression“.
- j Wenn Ausreisser vorhanden sind, dann sollte sich die Glättung davon nicht beirren lassen! Einverstanden?

In einem realen Beispiel ist immer damit zu rechnen, dass **mehrere Voraussetzungen unerfüllt** bleiben. Methoden, die einzelne Voraussetzungen beurteilen lassen, auch wenn andere verletzt sind, erweisen sich als besonders nützlich. Sie erlauben es, die geeigneten Verbesserungen zu finden; eine spezifische Diagnose ermöglicht die Wahl der wirksamen Therapie.

Methoden, die auf die Verletzung bestimmter Voraussetzungen wenig reagieren, heißen **robuste Methoden**, vergleiche 4.5.d. Das gleitende Mittel reagiert stark auf einen Ausreisser, ist also in diesem Sinne nicht robust. Wir verwenden deshalb die robuste Glättungsmethode „lowess“.

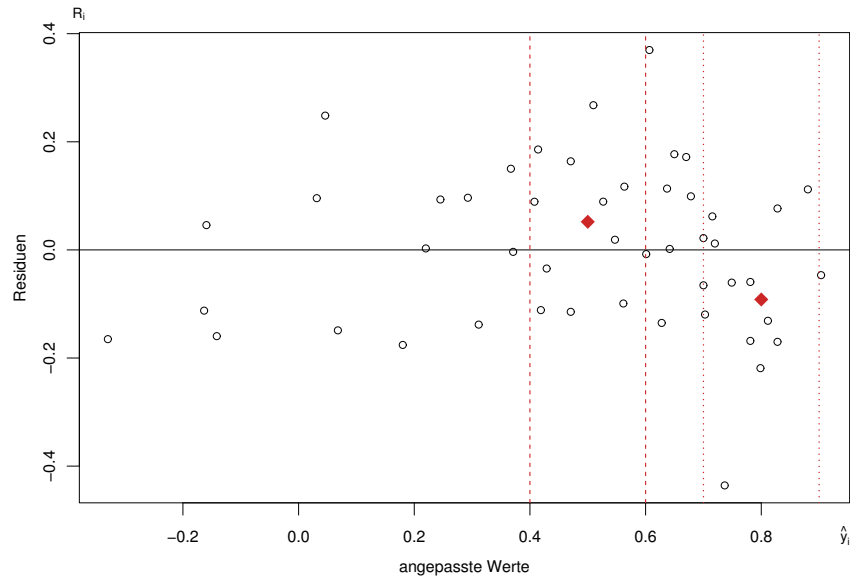


Abbildung 4.2.i: Bestimmung des gleitenden Mittels: Mittelwerte für zwei vertikale Streifen.

- k Die Glättung in Abbildung 4.2.f zeigt die Abweichung von der Linearität, die wir in Abbildung 3.1.h von Auge festgestellt haben (4.2.b), deutlich. Ist eine solche **Krümmung aufgrund des Zufalls** möglich? Oder handelt es sich um eine echte Abweichung, die wir durch die Verbesserung des Modells zum Verschwinden bringen sollten?

Es liesse sich ein formeller Test angeben, der die entsprechende Nullhypothese prüft – Näheres im Kapitel über Nichtparametrische Regression. Wir wollen hier eine informelle Methode benutzen, die sehr allgemein nützlich ist. Das Stichwort heisst **Simulation**, (vergleiche 2.2.e).

Schritt (1): Man erzeugt Beobachtungen, die dem Modell entsprechen, mit Zufallszahlen. Genauer: Es werden n standard-normalverteilte Zufallszahlen E_i^* erzeugt und daraus $Y_i^* = \hat{y}_i + \hat{\sigma}E_i^*$ bestimmt.

Schritt (2): Man führt die Regressionsrechnung mit den im Datensatz gegebenen erklärenden Variablen und den neu erzeugten Werten Y_i^* der Zielgrösse durch, berechnet die Glättung für das Tukey-Anscombe-Diagramm und zeichnet sie ins Diagramm der Daten oder in eine separate Darstellung ein.

Schritt (rep): Man wiederholt diese beiden Schritte n_{rep} Mal.

Die erzeugten Kurven entstehen aufgrund von zufälligen Schwankungen. Die Modellwerte folgen ja exakt einem linearen Modell – dem aus den Daten geschätzten multiplen *linearen* Regressionsmodell. Nun benutzt man wieder die Fähigkeit des Auges zur Mustererkennung, um informell zu beurteilen, ob die Kurve im ursprünglichen Tukey-Anscombe-Diagramm „extremer“ aussieht als die simulierten. Dabei sollte man nicht nur darauf achten, ob die ursprüngliche Glättung „in der Bandbreite“ der simulierten Kurven bleibt. Es kann auch die Form der Abweichung untypisch sein.

- 1 In Anlehnung ans Testen auf dem Niveau $5\% = 1/20$ wurde von Davies (1995) empfohlen, die durch die ursprünglichen Beobachtungen gegebene Glättung durch $n_{rep} = 19$ simulierte Kurven zu ergänzen. Ein informeller grafischer Test besteht dann darin, die 20 Kurven auf gleiche Weise (ohne die Residuen) darzustellen und unbeteiligte Personen aufzufordern, die auffälligste auszusuchen. Wenn das die Kurve ist, die den Beobachtungen entspricht, gilt die Abweichung als signifikant.

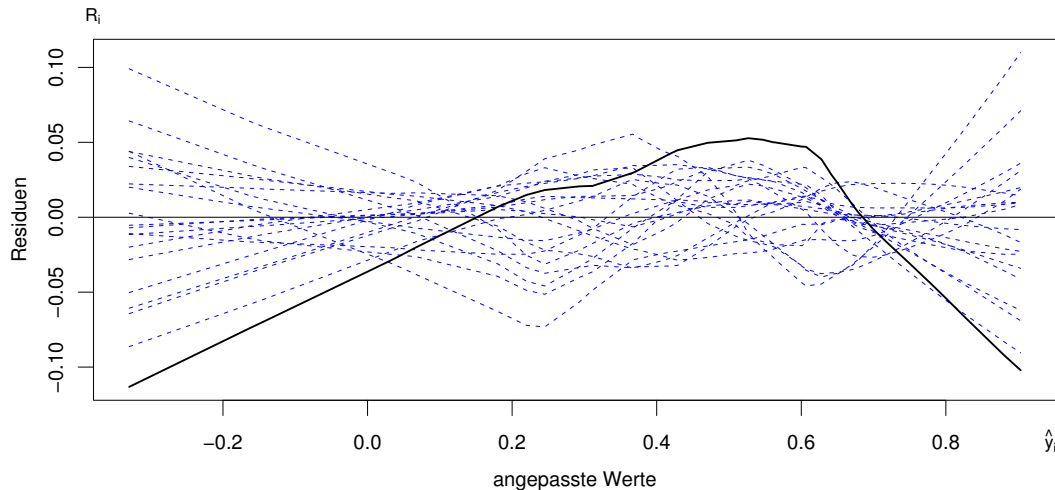


Abbildung 4.2.1: Die Glättung für die Residuen im Tukey-Anscombe-Diagramm (—) mit 19 simulierten Glättungskurven (- - -)

In Abbildung 4.2.1 wurden die Residuen weggelassen, damit das Bild einfacher wird. Es zeigt sich deutlich, dass die Glättung am linken und rechten Rand zufällig stärker streut als in der Mitte, was auch intuitiv zu erwarten ist. Die Glättung der Residuen der beobachteten Daten erscheint so oder so als die am stärksten gekrümmte Kurve. Damit kann die Abweichung als signifikant gelten.

m^* Statt der einzelnen Kurven kann man ein „Streuband“ einzeichnen, das zu jedem Wert von \hat{y} angibt, in welchem Bereich in vertikaler Richtung eine zufällige Glättungskurve liegen würde. Dazu sollte n_{rep} wesentlich grösser gewählt werden als 20, damit die Quantile mit vernünftiger Genauigkeit ermittelt werden können. Die Formen der zufälligen Kurven gehen dabei verloren. Zudem ist die Interpretation eines solchen Streifens nicht ganz einfach: Macht man daraus eine Testregel, die die Nullhypothese akzeptiert, wenn die beobachtete Kurve ganz im Streifen liegt, dann ist die Irrtumswahrscheinlichkeit höher als das Niveau, das man zur Bestimmung des Streubandes gewählt hat. Die Bestimmung eines „simultanen“ Streubandes mit vorgegebener Irrtumswahrscheinlichkeit ist schwierig.

n^* Für die Simulation von Fehlern E_i kann man statt der vorausgesetzten Normalverteilung auch die empirische Verteilung der Residuen R_i verwenden. Das ist die Idee der **Bootstrap**-Methode, die hier nicht näher besprochen wird.

Schritt (2) kann man wesentlich vereinfachen: Man rechnet nur die Glättung der simulierten Fehler aus und stellt sie dar. (Allenfalls multipliziert man die Fehler mit dem Faktor $\sqrt{1-p/n}$, siehe 4.3.g oder verwendet die empirische Verteilung der „halb-standardisierten“ Residuen $R_i/\sqrt{1-H_{ii}}$, siehe 4.3.i.) Das vernachlässigt zwar eine Quelle der Zufälligkeit der Kurve, wird aber für praktische Zwecke genau genug sein.

- o **(b) Gleiche Varianzen:** Ganz analog zu diesen Ideen kann man die Voraussetzung der gleichen Varianzen prüfen, indem man zusätzlich zu einem gleitenden Mittel eine „**gleitende Standardabweichung**“ nach oben und unten abträgt. Die Standardabweichung reagiert noch stärker auf Ausreisser und sollte deshalb noch dringender durch eine robustere Schätzung ersetzt werden. Eine einfache Möglichkeit besteht darin, die für die Glättung benützte Methode (lowess) auf die Absolutwerte $|R_i|$ der Residuen anzuwenden.

Das Programmsystem R liefert ein Streudiagramm der wurzel-transformierten $|R_i|$ gegen die angepassten Werte \hat{y}_i (Abbildung 4.2.o), das englisch *scale-location plot* genannt wird und wir **Streuungs-Diagramm** nennen wollen. Die Kurve fällt leicht, aber eine so milde Abweichung wäre, auch wenn sie sich als signifikant herausstellen sollte, unbedeutend.

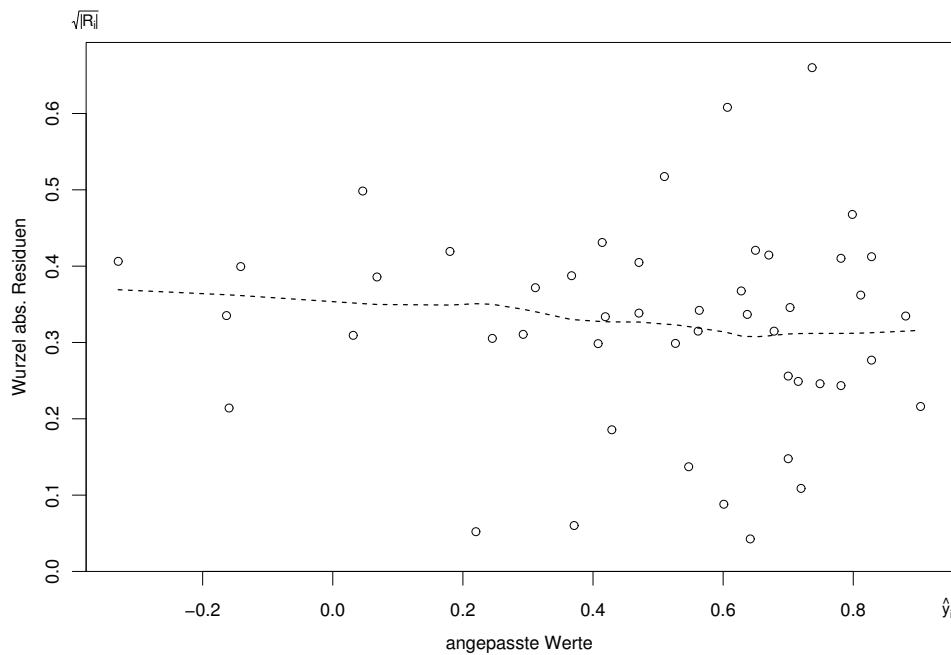


Abbildung 4.2.o: Wurzel-transformierte absolute Residuen $|R_i|$ gegen angepasste Werte im Beispiel der Sprengungen

- p* Die Glättung der (wurzel-transformierten) absoluten Residuen ergibt allerdings ein Streuungsmass, das auch für unendlich viele normalverteilte Beobachtungen nicht gleich der Standardabweichung ist. Es empfiehlt sich, einen entsprechenden Korrekturfaktor einzuführen. Da man nicht an der Streuung an sich, sondern nur an ihrer allfälligen Variation für verschiedene Bereiche von angepassten Werten interessiert ist, kann man darauf auch verzichten.

4.3 Verteilung der Fehler

- a Die Annahme der Normalverteilung ((c) in 4.1.a) kann man unter anderem grafisch überprüfen. Allerdings kennen wir die Fehler E_i nicht – aber wenigstens die **Residuen**. Das Histogramm der Residuen kann grafisch mit der geeigneten Normalverteilung verglichen werden (Abbildung 4.3.a). Diese ist durch den Erwartungswert 0 und die empirische Varianz der Residuen festgelegt.

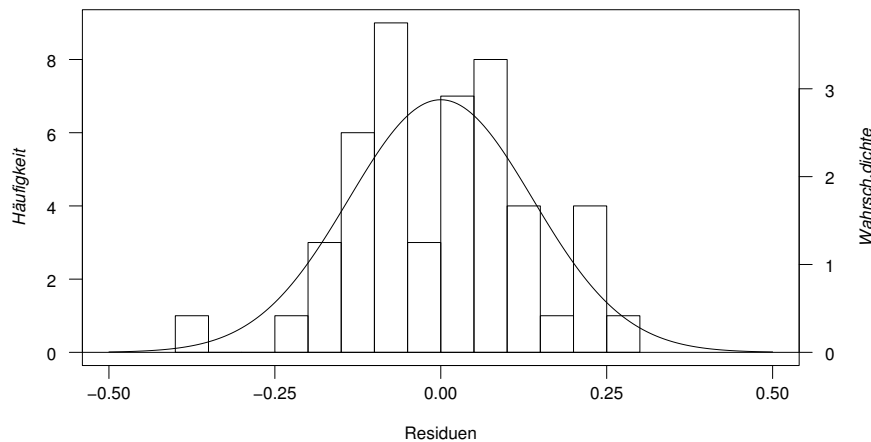


Abbildung 4.3.a: Histogramm der Residuen für das Beispiel der Sprengungen.

* Die empirische Varianz der Residuen ist nicht gleich der geschätzten Varianz $\hat{\sigma}^2$ der Fehler, sondern gleich $(\sum R_i^2)/(n-1) = \hat{\sigma}^2(n-p)/(n-1)$. Damit das Histogramm mit der Normalverteilung-Dichte vergleichbar wird, muss die Skala auf der vertikalen Achse so gewählt werden, dass die Summe der Produkte von Balkenhöhe mal Balkenbreite gleich 1 wird.

Beachten Sie, dass die Überprüfung der Normalverteilung für die Zielgröße selbst sinnlos ist, da die Y_i ja verschiedene Erwartungswerte haben.

- b Eine weitere Darstellungsart, das **Normalverteilungs-Diagramm** oder der **normal plot**, beruht auf dem Vergleich der Quantile der empirischen Verteilung der Residuen und der Quantile der Normalverteilung (Stahel (2002), 11.3).
- c Im **Beispiel der Sprengungen** zeigt sowohl das Histogramm (vergleiche Abbildung 4.3.a) als auch das Normalverteilungs-Diagramm (Abbildung 4.3.c), dass die Daten genähert normalverteilt sein könnten. Es fällt allerdings ein verdächtig extremer Wert auf, ein so genannter **Ausreisser**, den wir bereits im Tukey-Anscombe-Diagramm gesehen haben.
- d Ein Histogramm kann nie perfekt mit einer Dichtekurve übereinstimmen. Die Häufigkeitsverteilung der Residuen wird zufällig immer wieder anders herauskommen, auch wenn Beobachtungen genau nach dem Modell erzeugt werden – beispielsweise über Zufallszahlen. Welche Abweichungen können noch als „rein zufällig“ gelten? Man kann diese Frage formal mit einem statistischen Test beantworten. Dies führt zu den **Anpassungstests** (*goodness of fit tests*). Jeder dieser Tests prüft eine bestimmte Art von Abweichungen. Wir gehen hier nicht näher auf diese Methoden ein.

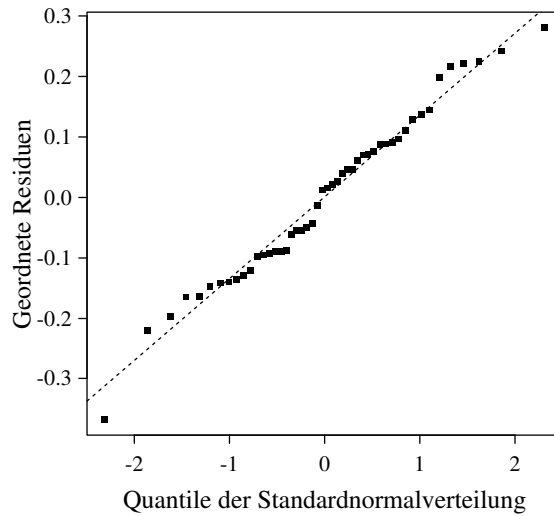


Abbildung 4.3.c: Normal plot der Residuen für das Beispiel der Sprengungen.

- e Der Vorteil einer grafischen Darstellung besteht gerade darin, dass das Auge auch Besonderheiten entdeckt, an die man vorher nicht gedacht hat. Die Entscheidung, ob ein Histogramm „nur zufällig“ von der idealen Verteilung abweicht oder nicht, braucht Übung – und diese kann man sich verschaffen, indem man durch Simulation (vergleiche 4.2.k) mit dem angepassten Modell immer neue Datensätze erzeugt. So sind die 6 **simulierten** Residuen-Histogramme in Abbildung 4.3.e (i) und die Normalverteilungs-Diagramme in Abbildung 4.3.e (ii) entstanden.

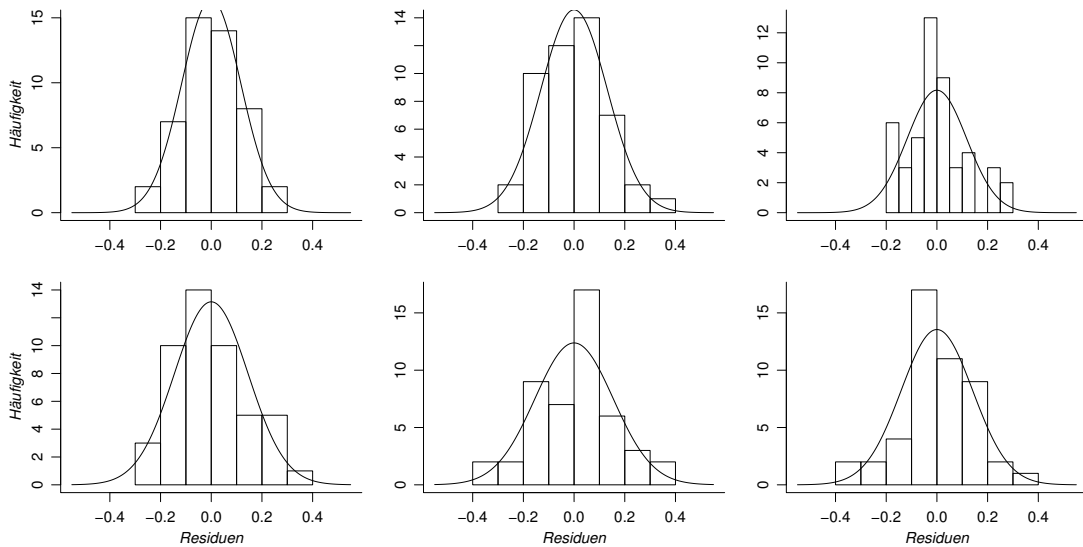


Abbildung 4.3.e (i): Histogramme von Residuen aus 6 simulierten Sätzen von Y -Werten im Beispiel der Sprengungen

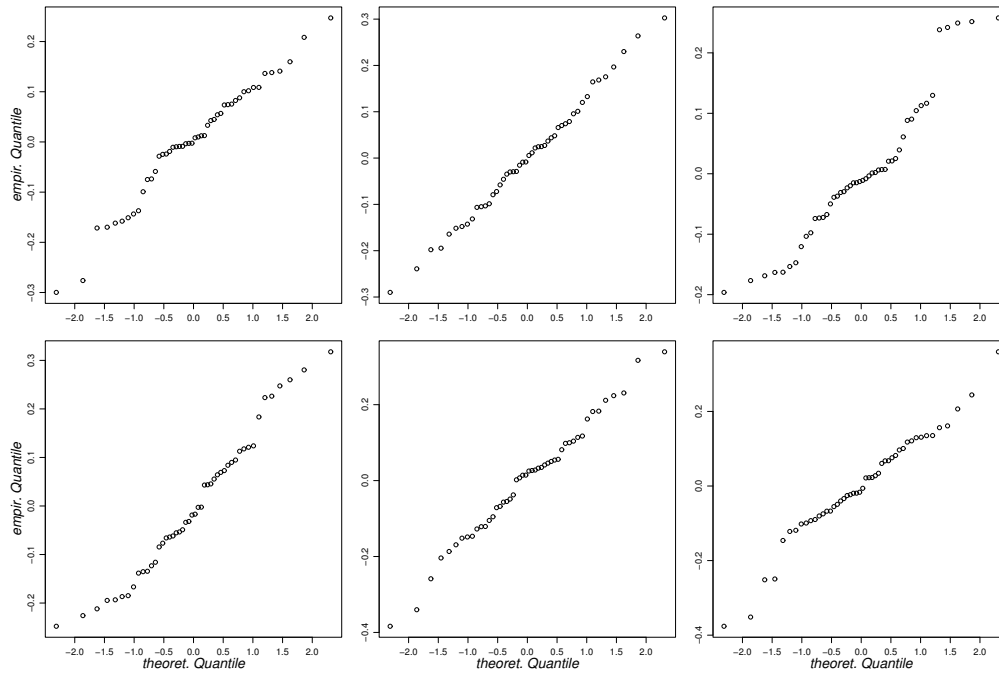


Abbildung 4.3.e (ii): Quantil-Quantil-Diagramme von Residuen aus 6 simulierten Sätzen von Y -Werten im Beispiel der Sprengungen

Nützlich ist es auch, analog zur Untersuchung der zufälligen Variation der Glättungen in 4.2.k vorzugehen und n_{rep} simulierte Normalverteilungs-Diagramme übereinander oder den daraus ermittelten „Streustreifen“ zu zeichnen.

- f Bei diesen Betrachtungen haben wir, wie eingangs angedeutet, ein wenig geschummelt. Wir wollen ja die **Verteilung der Zufallsfehler** E_i überprüfen, haben aber die Residuen R_i benützt, und das ist nicht dasselbe. Das ist mit Hilfe von Matrixalgebra nicht schwierig zu untersuchen, wie Anhang 4.A zeigt. Hier die Ergebnisse:
- g Falls die Fehler normalverteilt sind, so sind es die Residuen von einer Kleinst-Quadrate-Schätzung ebenfalls. Aber sie haben nicht die gleiche **theoretische Varianz**, auch wenn die Fehler dies erfüllen; $\text{var}\langle R_i \rangle$ hängt von $[x_i^{(1)}, x_i^{(2)}, \dots]$ ab! (Verwirrt Sie die Betrachtung der Varianz *eines* Residuums? Jedes R_i ist ja eine Zufallsvariable, die eine theoretische Varianz hat – nicht zu verwechseln mit der empirischen Varianz, die es immer nur für eine Stichprobe gibt, hier also für alle Residuen zusammen.) Es ist

$$\text{var}\langle R_i \rangle = (1 - H_{ii}) \sigma^2 .$$

Die Grösse H_{ii} ist eine Funktion aller $x_i^{(j)}$. Sie heisst englisch **leverage**, was wir mit **Hebelarm** übersetzen wollen, und wird oft als h_i notiert.

h Die Hebelarm-Werte haben einige anschauliche Bedeutungen:

- Wenn man einen Wert Y_i um Δy_i verändert, dann misst $H_{ii}\Delta y_i$ die Veränderung des zugehörigen angepassten Wertes \hat{y}_i . Wenn H_{ii} also gross ist, dann „zwingt die i te Beobachtung die Regressions-Funktion, sich an sie stark anzupassen“. Sie hat eine „grosse **Hebelwirkung**“ – daher der Name.
- Das macht auch das Ergebnis über die Varianzen qualitativ plausibel: Wenn die i te Beobachtung die Regressionfunktion stark an sich zieht, wird die Abweichung R_i tendenziell geringer, also die Varianz von R_i kleiner.
- Hebelpunkte in der Physik sind solche, die weit vom Drehpunkt entfernt sind. In unserem Zusammenhang heisst das, dass sie in gewissem Sinne weit vom „grossen Haufen“ der Punkte weg sind, was die x -Variablen betrifft.

* Die H_{ii} sind für die einfache Regression gleich $(1/n) + (x_i - \bar{x})^2 / \text{SSQ}^{(X)}$, also eine einfache Funktion des quadrierten Abstandes vom Schwerpunkt \bar{x} . In der multiplen Regression sind sie eine ebenso einfache Funktion der so genannten Mahalanobis-Distanz.

- Die leverages liegen zwischen 0 und 1. Ihr Mittelwert muss immer gleich p/n sein.

i Damit die Residuen wirklich die gleiche Verteilung haben, muss man sie also standardisieren! Man soll also für die Überprüfung der Verteilung die **standardisierten Residuen**

$$\tilde{R}_i = R_i / \left(\hat{\sigma} \sqrt{1 - H_{ii}} \right)$$

verwenden. Das Gleiche gilt für das Streuungs-Diagramm, das zeigen soll, ob die Varianzen der Fehler gleich sein können, was bedeutet, dass die Varianzen der *standardisierten* Residuen gleich sind.

Meistens sind allerdings die Unterschiede zwischen den Varianzen $\text{var}\langle R_i \rangle$ klein, so dass man auch unstandardisierte Residuen für diese Analyse verwenden kann. Wesentlich wird die Unterscheidung in der gewichteten Regression, siehe 4.7.

4.4 Zielgrösse transformieren?

a Nachdem jetzt einige Diagnose-Instrumente eingeführt sind, können wir die ersten Syndrome und Therapien besprechen. Dazu gehen wir den umgekehrten Weg von einer bekannten Krankheit zu den entsprechenden Symptomen.

- ▷ *Im Beispiel der Sprengungen wurde auf Grund von grafischen Darstellungen und theoretischen Überlegungen die Zielgrösse „Erschütterung“ logarithmiert. Wie würden die besprochenen grafischen Darstellungen aussehen, wenn die Zielgrösse nicht transformiert worden wäre? Abbildung 4.4.a zeigt es!*

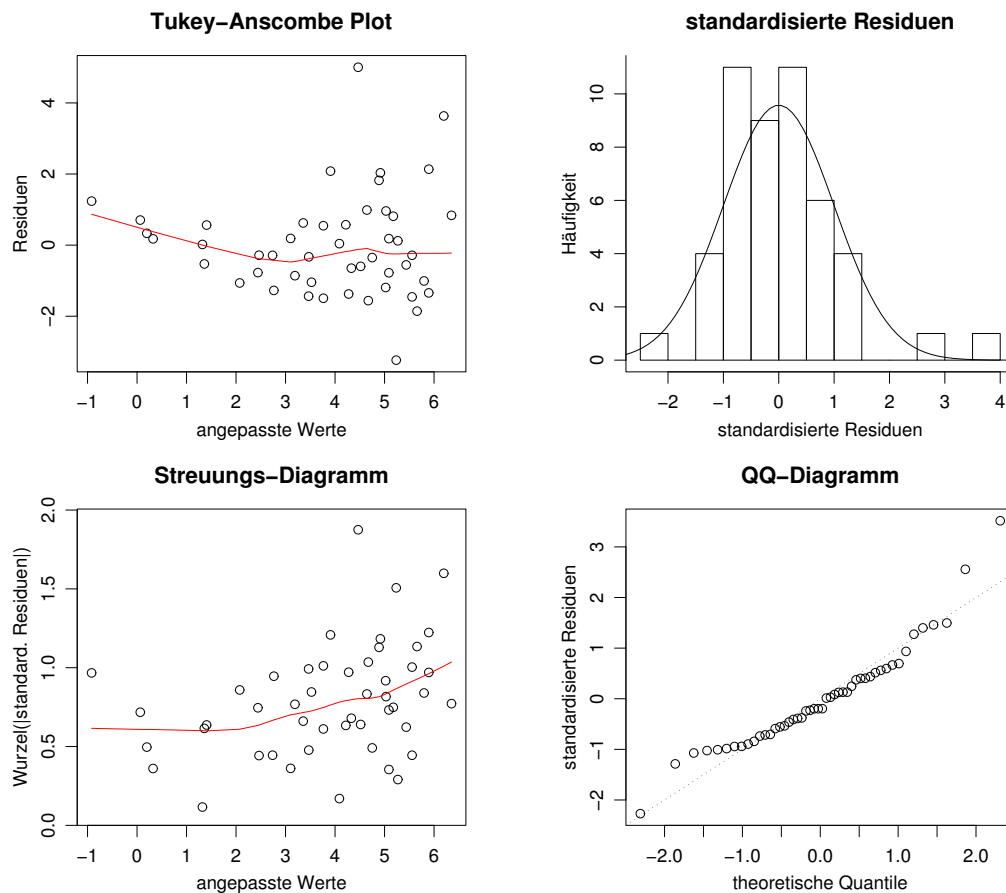


Abbildung 4.4.a: Tukey-Anscombe-Diagramm mit Streuungs-Diagramm und Histogramm und Normalverteilungs-Diagramm der standardisierten Residuen

- b Am augenfälligsten ist das Muster im Tukey-Anscombe-Diagramm: Es zeigt sich
- eine nach oben gekrümmte Glättung,
 - eine nach rechts trichterförmig zunehmende Streuung,
 - im rechten Teil eine schiefe Verteilung der Residuen – bis auf einen Ausreisser nach unten.

Im Streuungs-Diagramm wird die Zunahme der Streuung gegen rechts ebenfalls klar. Sie würde noch klarer, wenn Abweichungen von der Glättungskurve im Tukey-Anscombe-Diagramm statt der Residuen des (falschen) Modells verwendet würden.

Die Verteilung der standardisierten Residuen zeigt ebenfalls eine gewisse Schiefe. Wenn man die simulierten Bilder aus dem letzten Abschnitt ansieht (4.3.e), bleibt allerdings unklar, ob eine solche Abweichung auch zufällig zustande kommen könnte.

- c Die drei erwähnten Symptome bilden ein **Syndrom**, das nach einer **Transformation**

$$\tilde{Y} = g(Y)$$

der Zielgrösse ruft, und zwar mit einer Funktion g , die eine positive Schiefe verkleinert.

Im vorliegenden Beispiel ist die Lösung schon bekannt: Wenn die Zielgrösse logarithmiert wird, passt das Modell recht gut, wie wir bereits wissen.

Die Logarithmusfunktion ist allerdings nur eine unter vielen, die die Schiefe einer Verteilung reduzieren; alle monoton zunehmenden, nach unten gekrümmten (*konkaven*) Funktionen kommen hier in Frage. Eine weitere, oft verwendete Funktion ist die (Quadrat-) **Wurzel**, die weniger stark wirkt.

Als Transformationen der Zielgrösse kommen im vorliegenden Zusammenhang **umkehrbare** oder **monotone** Funktionen in Frage. Würde eine Funktion verwendet, die zwei verschiedenen Werten der ursprünglichen den gleichen Wert der transformierten Zielgrösse zuweist, dann würde damit die Art des untersuchten Zusammenhangs grundsätzlich verändert. Das sprengt den Rahmen der Veränderung des Modells zwecks besserer Erfüllung der Voraussetzungen. Als Grenzfall sind Funktionen zulässig, die nicht strikt, sondern nur „schwach“ monoton sind, für die also zusammenhängenden Intervallen der ursprünglichen Grösse allenfalls der gleiche transformierte Wert zugewiesen wird. Wir kommen auf mögliche Transformationen gleich zurück.

- d Im **Beispiel der basischen Böden** zeigt das Tukey-Anscombe-Diagramm (Abbildung 4.4.d) ein analoges Bild wie das Spreng-Beispiel mit untransformierter Zielgrösse – in umgekehrter Richtung und viel schwächer: Die Glättung zeigt eine leichte Krümmung nach unten, die Streuung nimmt (für $\hat{y} > 4$) gegen rechts leicht ab und die Verteilung der Residuen ist auf die unübliche Seite schief.

Hier hilft eine Transformation, die eine negative Schiefe reduziert, also eine mit einer monoton zunehmenden, *konvexen* Funktion. Erfahrung und Probieren führte in diesem Fall zu $\tilde{Y} = Y^2$. Das Tukey-Anscombe-Diagramm zeigt danach keine Abweichungen von den Modellannahmen mehr. Die Residuen sind etwa symmetrisch verteilt.

* Die Transformation $\tilde{Y} = Y^2$ ist selten nützlich. Sie ist auch nicht die einzig richtige, sondern eine einfache, die zum Ziel führt. Man kann versuchen, plausibel zu machen, weshalb eine solche Transformation in diesem Beispiel eine Bedeutung hat: Vielleicht ist die quadrierte Baumhöhe etwa proportional zur Blattfläche.

- e Ein Glücksfall, dass alle Abweichungen mit der gleichen Transformation beseitigt werden können! – Dieser Glücksfall tritt erstaunlich häufig ein. (Wenn Sie gerne philosophieren, können Sie sich nach dem Grund dieser empirischen Erscheinung fragen, die allerdings wohl kaum je mit einer empirischen Untersuchung quantitativ erfasst wurde.)
- f **Welche Transformationen** soll man in Betracht ziehen, um das beschriebene Syndrom zu kurieren? Die folgenden Empfehlungen beruhen wieder auf Erfahrungen der angewandten Statistik, auf Plausibilität, Einfachheit und ähnlichen „unexakten“ Grundlagen.

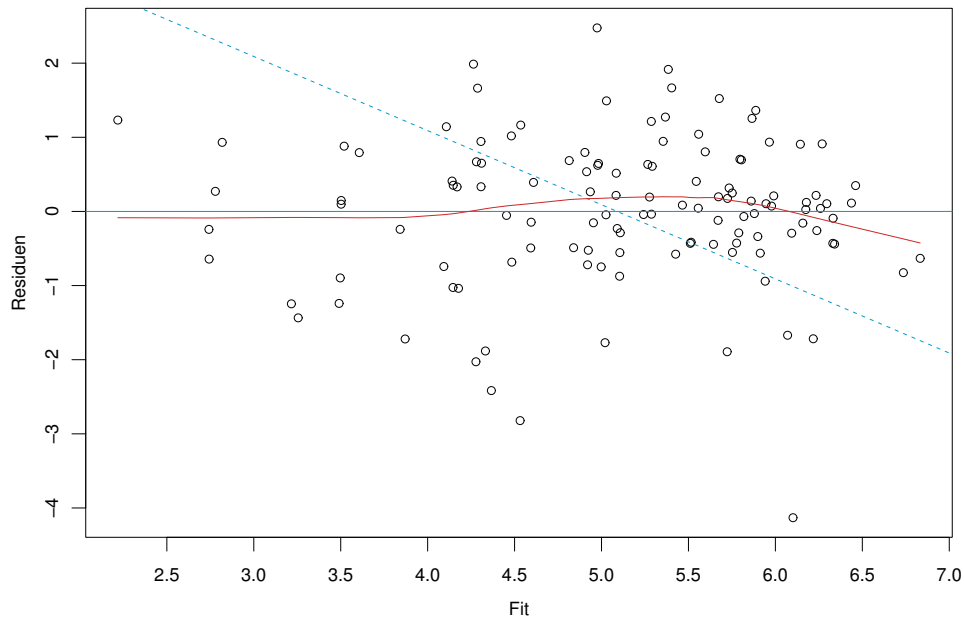


Abbildung 4.4.d: Tukey-Anscombe-Diagramm für das Beispiel der basischen Böden

g

Als nützlich erweisen sich sehr oft

- die Logarithmus-Transformation für **Konzentrationen und Beträge** – also für stetige Zufallsvariable, die nur positive Werte haben können –
- die Wurzeltransformation für **Zählraten** und
- die so genannte Arcus-Sinus-Transformation $\tilde{y} = \arcsin \sqrt{y}$ für **Anteile** (Prozentzahlen/100).

Diese Transformationen haben von J. W. Tukey den Namen **first aid transformations** erhalten und **sollten für solche Daten immer angewendet werden**, wenn es keine Gegen Gründe gibt – und zwar auch für erklärende Variable.

- h Wenn in einer einfachen Regression sowohl die erklärende Variable als auch die Zielgröße Konzentrationen sind, führt die Regel zu $\tilde{Y} = \log_{10}\langle Y \rangle$ und $\tilde{X} = \log_{10}\langle X \rangle$. Aus $\tilde{Y} = \alpha + \beta \tilde{x}_i + E_i$ wird $\log_{10}\langle Y_i \rangle = \alpha + \beta \log_{10}\langle x_i \rangle + E_i$ und

$$Y_i = 10^\alpha x_i^\beta 10^{E_i},$$

also ein **Potenzgesetz** für die ursprünglichen Größen (vergleiche 2.1.d). Falls $\beta = 1$ ist, sind die Konzentrationen proportional bis auf einen **multiplikativen zufälligen Fehler**. Wenn das lineare Modell der logarithmierten Größen weitere Terme enthält, dann wirken diese auf die untransformierte Zielgröße multiplikativ. Für eine zusätzliche kontinuierliche Ausgangsgröße kommt ein multiplikativer Potenz-Term $x_i^{(2)\beta_2}$ hinzu. Im Fall einer Indikator-Variablen, beispielsweise für eine neue Behandlung, ist die Wirkung einfacher: Die neue Behandlung bewirkt gemäss Modell eine proportional Erhöhung (oder Erniedrigung) von Y um den Faktor 10^{β_2} .

- i Die **Logarithmus-Transformation** ist also von besonderer Bedeutung. Sie ist vom datenanalytischen Gesichtspunkt her dann richtig, wenn die Standardabweichung der Residuen etwa proportional zu den angepassten Werten ist. Sie ist allerdings nur anwendbar, wenn die Zielgrösse nur positive Werte haben kann. Das allerdings gilt oft auch für Variable, für die der Wert 0 auftreten kann. Man muss dann die Logarithmus-Transformation leicht abändern, damit die **Nullen nicht wegfallen**. Beobachtungen mit $Y_i = 0$, also diejenigen mit dem kleinsten Wert der Zielgrösse, wegfällen zu lassen, müsste zu einer systematischen Verfälschung der Resultate führen!

Die einfachste Formel zur Abänderung der Logarithmus-Funktion lautet $\tilde{Y} = \log(Y + c)$ mit einer geeigneten Konstanten c . Oft sieht man, gemäss dem Prinzip der Einfachheit, die Wahl von $c = 1$. Da die Wirkung dieser Wahl stark vom Bereich der untransformierten Werte Y_i abhängt, sollte man diese Wahl eher als „einfältig“ bezeichnen. Die Wahl soll von der Verteilung der positiven Y_i abhängen. Wären diese lognormal verteilt, dann würde $c = \text{med}\langle Y_k \rangle / s^{2.9}$ mit $s = \text{med}\langle Y_k \rangle / q_{0.25}\langle Y_k \rangle$ eine Schätzung für das 2.5%-Quantil ergeben ($q_{0.25}$ ist das untere Quartil). Diese Konstante hat also die gleiche Grössenordnung wie die kleinsten positiven beobachteten Werte. Ihre Wahl ist immer noch willkürlich, aber sie macht die Wirkung der Transformation wenigstens von der Wahl der Messeinheit von Y unabhängig.

- j* **Box-Cox-Transformationen**. Damit man möglichst nicht-schiefe Fehler-Verteilungen erreichen kann, kann man eine ganze „Familie“ von Transformationen einführen. Von Box und Cox stammt der Vorschlag

$$g_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{für } \lambda \neq 0, \\ \ln(x) & \text{für } \lambda = 0 \end{cases} .$$

(für positive x). Bis auf Verschiebung um -1 und Multiplikation mit $1/\lambda$ sind dies die Potenzen x^λ . Diese Skalierung hat den Vorteil, dass im Grenzfall $\lambda \rightarrow 0$ die Logarithmus-Funktion herauskommt, was die Definition für diesen Fall begründet. Die Schiefe wird grösser für $\lambda > 1$; für $\lambda < 1$ nimmt die Schiefe ab.

Es wurde auch vorgeschlagen, die Grösse λ als zusätzlichen Parameter ins Modell aufzunehmen und nach dem Prinzip der Maximalen Likelihood zu schätzen. Für die Interpretation kann es einfacher sein, sich auf „einfache Werte“ von λ zu beschränken wie: Quadrat: $\lambda = 2$; keine Transformation (bis auf eine Verschiebung um 1): $\lambda = 1$; Quadrat-Wurzel: $\lambda = 0.5$; Logarithmus: $\lambda = 0$; Kehrwert: $\lambda = -1$.

- k Wie die Betrachtung in 4.4.h deutlich macht, **ändert sich** mit der Transformation der Zielgrösse auch die **Regressionsfunktion**. In einigen Anwendungen ist das nicht zulässig, da die (lineare) Regressionsfunktion für die untransformierte Zielgrösse theoretisch begründet ist.

Das ist beispielsweise im Beispiel der Schadstoffe im Tunnel (1.1.d) der Fall: Die gesamten Schadstoffe setzen sich nach einer offensichtlichen „physikalischen Gesetz“ additiv aus den Schadstoffen zusammen, die die beiden Fahrzeugkategorien ausstossen. In einem solchen Fall muss man zu einem allgemeineren Regressionsmodell übergehen, indem man entweder die Voraussetzungen der gleichen Varianz (b) und der Normalverteilung (c) fallen lässt oder ein **nicht-lineares Modell** verwendet.

- l Wenn keine Theorie die Transformation verbietet, kann es natürlich noch vorkommen, dass der erwähnte Glücksfall nicht eintritt, dass also eine Krümmung der Glättung, eine Abhängigkeit der Varianz vom angepassten Wert und die Form der Verteilung der Residuen nicht durch eine einzige Transformation aus der Welt zu schaffen sind.

Sind zum Beispiel die Gleichheit der Varianzen (b) und die Normalverteilung (c) in Ordnung, aber die Regressionsfunktion verbesserungsbedürftig, dann soll man zunächst prüfen, ob sie sich durch Transformationen der erklärenden Variablen oder durch Zusatzterme linearisieren lässt (siehe Abschnitt 4.6). Wenn das nicht hilft, kann man die Zielgröße trotzdem transformieren und nachher die anderen Voraussetzungen, die dann verletzt sein können, durch Gewichtung und robuste Schätzung berücksichtigen.

- m Gekrümmte Glättungen im Tukey-Anscombe-Diagramm lassen sich nicht immer mit Transformation der Zielgröße kurieren. Wenn beispielsweise die wahre Regressionsfunktion eine quadratische Funktion von $X^{(1)}$ ist, die im Bereich der Daten ein Maximum oder ein Minimum erreicht (vergleiche 3.2.v), während das gewählte Modell keinen quadratischen Term enthält, dann ergibt sich eine gekrümmte Glättung, und es ist klar, dass keine Transformation der Zielgröße diese Erscheinung zum Verschwinden bringt.

Eine monotone Transformation der Zielgröße kann einen Zusammenhang mit einer Ausgangsgröße nur dann linear machen, wenn dieser Zusammenhang selbst monoton ist. Nun sind im Tukey-Anscombe-Diagramm die Residuen in vertikaler Richtung abgetragen, nicht die Y -Werte. Man kann also entweder zum Diagramm der beobachteten Y -Werte gegen die angepassten zurückgehen (3.1.h) – oder ins Tukey-Anscombe-Diagramm eine **Referenzlinie** einzeichnen, die Punkte mit konstanten Y -Werten verbindet, wie dies in 4.2.g erwähnt wurde. Eine monotone Transformation der Zielgröße kann nur helfen, wenn die Glättung jede Parallele zur Referenzlinie (jede Gerade der Form $Y = \text{konstant}$) nur einmal schneidet.

4.5 Ausreisser und langschwänzige Verteilung

- a Im Beispiel der Sprengungen haben wir eine oder zwei Beobachtungen als **Ausreisser** bezeichnet. Der Begriff des Ausreissers ist nicht klar definiert. Es handelt sich um eine Beobachtung, die schlecht zu einem Modell passt, das für die Mehrheit der Daten angebracht ist. Im Fall einer einfachen Stichprobe ist ein Ausreisser eine Beobachtung, die, gemessen an der Streuung der Daten, weit vom Median entfernt ist. In der Regression spielt das Modell eine wesentliche Rolle. Vor allem haben Transformationen einen starken Einfluss darauf, welche Beobachtungen extreme Residuen erhalten.

* „Ausreisser“ ist damit ein „vager Begriff“. Dass diese in der Datenanalyse eine wichtige Funktion haben, auch wenn sie von Mathematikern meistens nicht geliebt werden, hat J. W. Tukey betont. Sie helfen, die nötigen Präzisierungen durch wohldefinierte Masszahlen kritisch zu hinterfragen und alternative „Operationalisierungen“ vorzuschlagen.

- b **Was soll man tun mit Ausreißern?** Zunächst sollen sie die zugehörigen Daten auf Richtigkeit überprüft werden. Es ist leicht einzusehen, dass Ausreisser im Tukey-Anscombe-Diagramm durch **grobe Fehler** sowohl in der Zielgrösse als auch in einer wichtigen erklärenden Grösse verursacht sein können.
Findet man keine genügenden Gründe, an der Richtigkeit der Werte zu zweifeln, dann wird man zunächst mit den weiteren Methoden der Residuen-Analyse nach Erklärungen für die „ungewöhnliche“ Beobachtung und Verbesserungen des Modells suchen. Ausreisser sind (wie im menschlichen Zusammenhang) etwas Besonderes, aber nichts „Schlechtes“, sondern manchmal die wertvollsten Beobachtungen im Datensatz!
Fördert auch die Suche nach Modell-Veränderungen nichts zu Tage, dann kann der Ausreisser auch durch eine ungewöhnlich grosse Zufallsabweichung zustande gekommen sein; solche werden durch langschwänzige Verteilungen mit grösserer Wahrscheinlichkeit erzeugt.
- c Schiefe Verteilungen versucht man, wie im vorherigen Abschnitt erwähnt, durch Transformationen zum Verschwinden zu bringen. Zeigt der normal plot eine einigermaßen symmetrische Verteilung, die aber **langschwänzig** ist, dann nützen Transformationen der Zielgrösse meistens nichts.
Man kann die extremsten Beobachtungen weglassen, bis die Langschwänzigkeit verschwindet oder zu viele (z. B. mehr als 5%) eliminiert werden. Resultate, die man mit den übriggebliebenen Beobachtungen erhält, sind aber mit Vorsicht zu benützen. Bei Tests und Vertrauensintervallen stimmt die Irrtums-Wahrscheinlichkeit nicht mehr. Die weggelassenen Beobachtungen soll man als Ausreisser auf ihre Richtigkeit speziell überprüfen, und auf alle Fälle sind sie im Bericht zu erwähnen.
- d* Die Kleinste-Quadrate-Methoden sind bei langschwänzigen Verteilungen der Fehler nicht optimal. **Robuste Methoden** sind in diesem Fall deutlich besser; sie liefern effizientere Schätzungen und mächtigere Tests. Gleiches gilt, wenn sich einzelne **Ausreisser** zeigen; der Fall einer Normalverteilung mit Ausreißern ist ein Spezialfall einer langschwänzigen Verteilung.

4.6 Residuen und erklärende Variable

- a Im Tukey-Anscombe-Diagramm können sich Abweichungen von der angenommenen Form der Regressionsfunktion und von der Voraussetzung der gleichen Varianzen zeigen. Ähnliches kann auch zu Tage treten, wenn als horizontale Achse statt \hat{Y} eine **Ausgangs-Variable** gewählt wird.
▷ *Abbildung 4.6.a zeigt diese Streudiagramme für die zwei kontinuierlichen Ausgangsgrössen im Beispiel der Sprengungen. Wieder wurden zur Beurteilung der Glättung 19 „zufällige Glättungen“ eingezeichnet.*
- b Wie beim Tukey-Anscombe-Diagramm erscheint auch hier eine **Referenzlinie**, die Punkte gleicher Y -Werte verbinden soll. Da Y_i aber nicht die Summe einer linearen Funktion von $x_i^{(j)}$ und dem Residuum R_i ist, ist die genaue Bedeutung der Referenzgeraden etwas komplizierter zu formulieren: sie verbindet Punkte, für die die Summe aus dem geschätzten Effekt der betrachteten erklärenden Variablen $X^{(j)}$ und den Re-

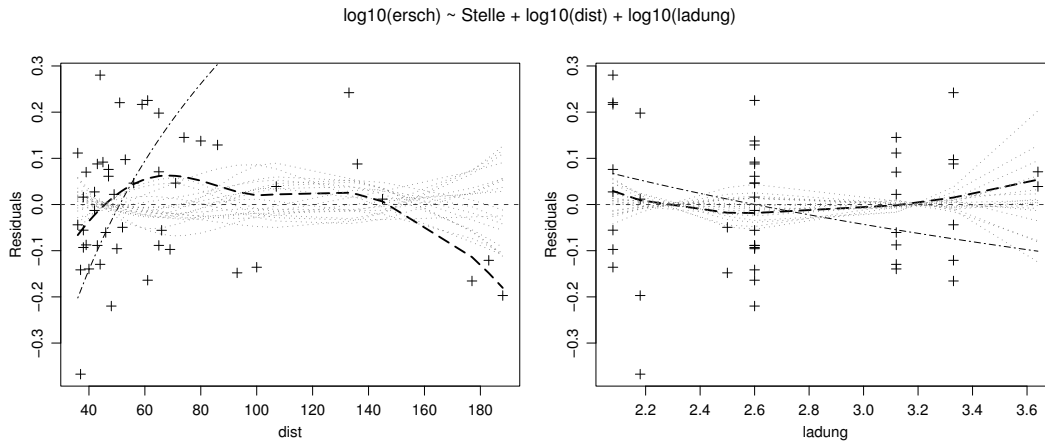


Abbildung 4.6.a: Streudiagramme der Residuen gegen zwei erklärende Variable, mit Glättung (---) und Referenzlinie $Y = \text{konstant}$ (- · - · -)

siduen, also

$$\widehat{\beta}_j x_i^{(j)} + R_i = \text{const}$$

ist. Der erste Term wird im Englischen auch *component effect* genannt. Die Summe der beiden kann auch geschrieben werden als $Y_i - \sum_{\ell \neq j} \widehat{\beta}_\ell x_i^{(\ell)}$, was als beobachteten Wert, „korrigiert für die Effekte der anderen Regressoren“, angesprochen werden kann. Wenn ein Regressor $X^{(j)}$ durch Transformation aus einer (oder mehreren) Ausgangs-Variablen $U^{(j)}$ ausgerechnet wurde, stellt sich die Frage, ob die Residuen gegen die untransformierte oder die transformierte Variable dargestellt werden sollen.

- ▷ *Im Beispiel wurden sowohl die Distanz als auch die Ladung logarithmiert. In der Abbildung wurden die untransformierten Werte benützt, was dazu führt, dass die Referenzlinie keine Geraden ist. Die Begründung für diese Wahl folgt unten (4.6.e).*
- c Eine Abweichung der Form der Regressionsfunktion, die sich im Streudiagramm der Residuen gegen $X^{(j)}$ allenfalls zeigt, kann oft durch **Transformation der erklärenden Variablen** $X^{(j)}$ zum Verschwinden gebracht werden.

Häufig wird man eine solche Abweichung bereits im Tukey-Anscombe-Diagramm gesehen haben. Vielleicht musste man aber auf eine Transformation der Zielgrösse verzichten, weil sonst die vorhandene Symmetrie und Gleichheit der Varianzen der Residuen zerstört worden wäre.

Kann eine monotone Transformation von $U^{(j)}$ helfen? Wie im Tukey-Anscombe-Diagramm hilft die **Referenzlinie**, diese Frage zu beantworten. Die Differenz zwischen der Nulllinie (der horizontalen Achse) und der Referenzlinie misst den Einfluss der Ausgangsgrösse $U^{(j)}$ auf die Zielgrösse gemäss Modell. Die Differenz zwischen der Glättung und der Referenzlinie dagegen zeigt, wie der Einfluss geschätzt wird, wenn er nicht auf die lineare Form $\beta_j X^{(j)}$ eingeschränkt wird.

- ▷ *Im Beispiel ist dieser flexibel geschätzte Einfluss für kleine Distanzen kleiner und für grosse Distanzen grösser als der Einfluss gemäss Modell. Würde die Glättung der Nulllinie folgen, dann würde der Einfluss gerade der im Modell angenommenen*

Form entsprechen. Da der flexibel geschätzte Einfluss immerhin monoton mit der erklärenden Variablen abnimmt, hat man mit einer monotonen Transformation dieser Variablen eine Chance, die Krümmung weg zu bringen.

Die Transformation müsste grosse Werte der erklärenden Variablen auseinander ziehen. Da es sich um den Logarithmus der Distanz handelt, kann man es mit ent-logarithmieren versuchen. Konsequenterweise ent-logarithmieren wir auch die erklärende Grösse Ladung. Abbildung 4.6.c zeigt die Diagramme für das entsprechend geänderte Modell. Die Transformation zeigt für die Distanz den erwünschten Erfolg. Für die Ladung ist die Wirkung gering; die Logarithmus-Transformation wirkt für die Ladung näherungsweise als lineare Funktion, da der Variationskoeffizient relativ klein ist.

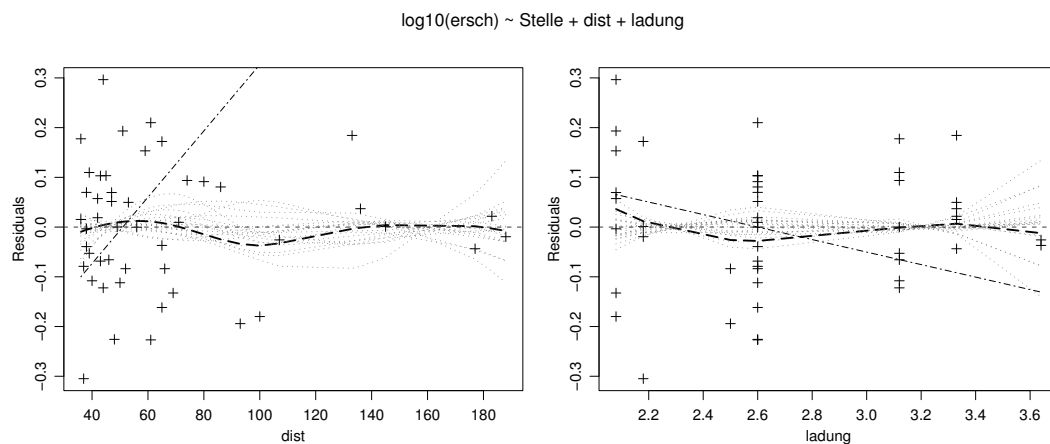


Abbildung 4.6.c: Streudiagramm der Residuen gegen die erklärenden Variablen Distanz und Ladung, die hier unlogarithmiert im Modell stehen

Im vorliegenden Fall haben die (Rück-) Transformationen den Nachteil, dass die einfache physikalische Interpretation verloren geht. Wenn wir nur an guter Vorhersage interessiert sind, können wir auf die Begründung verzichten. Allerdings ist bei der Verallgemeinerbarkeit der Studie auf andere Tunnels dann erhöhte Skepsis am Platz.

- d Wenn keine Transformation von $X^{(j)}$ zum Ziel führt, kann ein zusätzlicher, **quadratischer Term** $X^{(j)2}$ helfen. Eine einfache lineare Regression wird dann zu einer quadratischen (siehe 3.2.v).
- e* Wieso werden in den Darstellungen nicht die transformierten Variablen für die horizontale Achse verwendet? Wenn die Transformation nicht „erfolgreich“ war, dann sollte man einen neuen Versuch starten. Wurde die transformierte Variable auf der horizontalen Achse verwendet, dann kann die Abbildung nur eine Transformation der Transformierten nahelegen – das kann zu einer komplizierten, wenig sinnvollen Lösung führen. Wenn die untransformierte Variable verwendet wird, kann man mit der Abbildung direkt eine neue, einfache Transformation bestimmen. – Falls ein quadratischer Term im Modell vorkommt, ist es wenig sinnvoll, die Residuen gegen diesen Regressor aufzutragen. Es ist informativer, die untransformierte Ausgangsgrösse zu verwenden, und diese ist normalerweise sowieso ebenfalls im Modell vorhanden, weshalb für sie so oder so eine entsprechende Abbildung gezeichnet wird.

Deshalb werden von der Funktion `regr` die Residuen gegen alle in der Modellformel vorkommenden Variablen aufgetragen, nicht gegen Regressoren resp. Terme der Formel.

Wenn Wechselwirkungen im Modell sind (oder andere Regressoren, die aus mehreren Ausgangsgrößen berechnet werden), muss neu geklärt werden, wie der Effekt einer Ausgangsgröße $U^{(j)}$ gemessen werden soll. Antwort: Man setzt alle anderen Ausgangs-Variablen auf einen „typischen Wert“ u_k (Median für kontinuierliche und Modus für kategorielle Variable) und verwendet die Vorhersage $\hat{y}(u_1, \dots, u_{j-1}, U^{(j)}, u_{j+1}, \dots)$ als Funktion des variierenden $U^{(j)}$ als „component effect“ $\hat{\gamma}^{(j)}$.

- f Im Modell wird als nächstes vorausgesetzt, dass die **Effekte von zwei erklärenden Variablen sich addieren**. Diese Annahme soll ebenfalls grafisch überprüft werden. Dazu braucht es ein dreidimensionales Streudiagramm von $x_i^{(j)}, x_i^{(k)}$ und den Residuen R_i . Etliche Programme erlauben es, einen dreidimensionalen Eindruck auf einem zweidimensionalen Bildschirm durch Echtzeit-Rotation zu gewinnen.

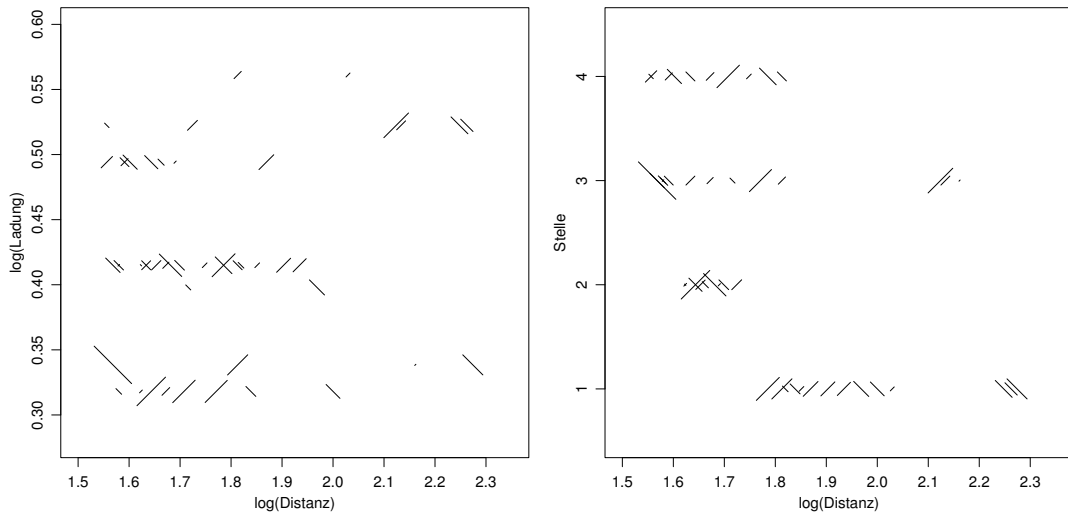


Abbildung 4.6.f (i): Residuen in Abhängigkeit von zwei erklärenden Variablen im Beispiel der Sprengungen

Auf dem Papier ist der dreidimensionale Eindruck schwieriger zu erreichen. Abbildung 4.6.f zeigt eine spezielle Art der Darstellung für das Beispiel der Sprengungen. Darin wird die Größe des i ten Residuums durch ein strichförmiges Symbol dargestellt, das am Ort $[x_i^{(1)}, x_i^{(2)}]$ platziert wird. Die Länge des Striches ist proportional zum Absolutbetrag des Residuums und die Steigung von $+1$ oder -1 gibt das Vorzeichen wieder.

- g Im linken Diagramm sind die beiden erklärenden Variablen kontinuierlich. Wenn in einem solchen Diagramm Gebiete sichtbar werden, in denen die meisten Striche in der einen Richtung verlaufen, deutet dies eine so genannte **Wechselwirkung** an. Der einfachste Fall besteht darin, dass die Residuen links unten und rechts oben vorwiegend positiv und links oben und rechts unten eher negativ sind – oder umgekehrt. Eine solche Wechselwirkung kann die durch einen zusätzlichen Term $+\beta_{m+1}x_i^{(m+1)}$ mit $x_i^{(m+1)} = x_i^{(j)}x_i^{(k)}$ im Modell berücksichtigt werden kann.

Im rechten Diagramm ist die in vertikaler Richtung gezeichnete Variable ein Faktor (die Stelle). Es zeigt sich für Stelle 1 eine Tendenz zu negativen Residuen für grosse und positiven für kleinere Distanzen; für Stelle 3 ist es gerade umgekehrt. Das deutet eine Wechselwirkung zwischen dem Faktor Stelle und der (logarithmierten) Distanz an, vergleiche 3.2.t. Eine solche Wechselwirkung lässt sich noch einfacher entdecken in einem Streudiagramm der Residuen gegen die kontinuierliche erklärende Variable, mit verschiedenen Symbolen für die verschiedenen Faktorwerte (Abbildung 4.6.g (ii)).

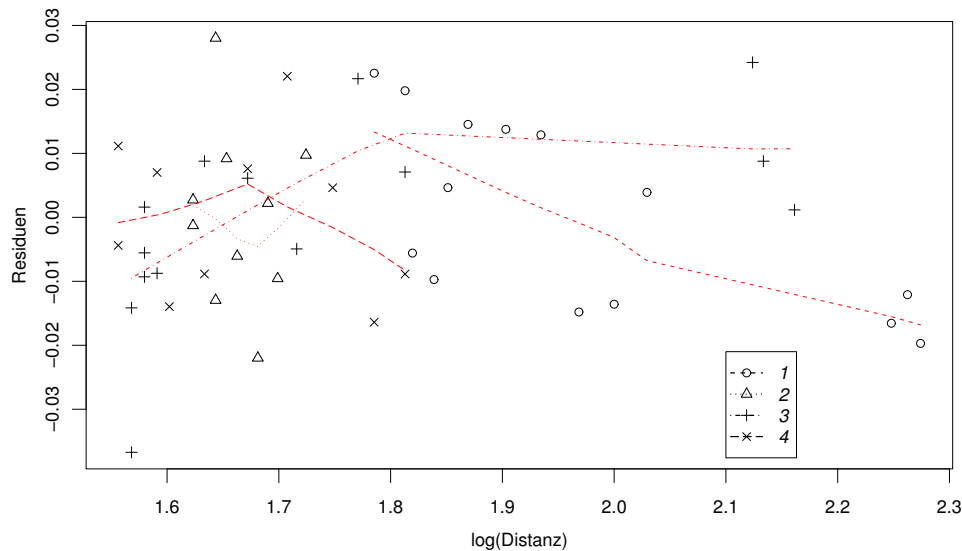


Abbildung 4.6.g (ii): Residuen gegen eine erklärende Variable, mit verschiedenen Symbolen und Glättungen für die verschiedenen Werte eines Faktors

- h In den Streudiagrammen der Residuen gegen die erklärenden Variablen kann sich auch zeigen, dass die **Streuung der Residuen** von $X^{(j)}$ abhängt. Dann gibt die **gewichtete Regression** korrekte Ergebnisse.

4.7 Gewichtete lineare Regression

- a Die **Varianzen** der einzelnen Zufallsfehler, die wir mit $\sigma_i^2 = \text{var}\langle E_i \rangle$ bezeichnen wollen, sollen nun nicht mehr als gleich ($= \sigma^2$) vorausgesetzt werden.

Wir gehen zunächst davon aus, dass die σ_i^2 bekannt seien. Dann ist es sicher sinnvoll, den Beobachtungen mit kleinerer Zufallsstreuung, also den präziseren Beobachtungen, in der Regressionsrechnung grösseres **Gewicht** zu geben. Statt der gewöhnlichen Quadratsumme $\text{SSQ}^{(E)}$ kann man eine gewichtete Version davon, $\sum_i w_i R_i^2$, minimieren. Die Gewichte w_i sollen für steigende σ_i fallen. Nach dem Prinzip der Maximalen Likelihood ist $w_i = 1/\sigma_i^2$ optimal.

* Die Wahrscheinlichkeits-Dichte für eine Beobachtung $Y_i = y_i$ ist unter dieser Annahme nämlich $1/(\sigma_i \sqrt{2\pi}) \exp\langle -(r_i^2)/(2\sigma_i^2) \rangle$ (mit $r_i = y_i - (\beta_0^* + \sum_j \beta_j^* x_i^{(j)})$). Wie in 2.A.0.a) ergibt sich durch Logarithmieren und Summieren die Quadratsumme, diesmal die gewichtete.

- b ▷ **Beispiel starke Wechselwirkung.** In Experimenten der Hochenergie-Physik wurde in den 1970er Jahren die starke Wechselwirkungskraft untersucht. In einem Versuch trifft ein Elementarteilchenstrahl auf eine Protonenquelle, und es entstehen verschiedene neue Elementarteilchen, von denen eine Sorte durch einen Detektor erfasst wird. Genaueres findet man in Weisberg (1990, Ex. 4.1).

u_i	Y_i	σ_i	u_i	Y_i	σ_i
4	367	17	15	239	6
6	311	9	20	220	6
8	295	9	30	213	6
10	268	7	70	193	5
12	253	7	150	192	5

Tabelle 4.7.b: Daten des Beispiels der starken Wechselwirkung: Energie des Teilchenstromes u_i , Anteil erfasste Teilchen Y_i und Standardabweichung σ_i der Zufallsabweichungen E_i

Die Daten in Tabelle 4.7.b enthalten die Energie u des Teilchenstromes und die Zielgrösse Y , die proportional zum Verhältnis der erfassten Teilchen zu den eingeschossenen Teilchen ist. Zudem kann man eine theoretische Standardabweichung σ_i für jedes Y_i (oder jeder Zufalls-Abweichung E_i) bestimmen; diese Grössen sind in der Tabelle ebenfalls enthalten. Für beide Grössen bildet die Logarithmus-Funktion die „first aid transformation“. Deshalb sind die beiden Variablen in Abbildung 4.7.b links mit logarithmischen Skalen gezeigt.

Gemäss einer Theorie sollte $Y \approx \beta_0 + \beta_1 u^{-1/2}$ sein. Das Streudiagramm der Zielgrösse gegen $x = u^{-1/2}$ (rechtes Diagramm) sollte gemäss Theorie einen linearen Zusammenhang zeigen. Er sieht eher quadratisch aus. Dennoch wird auch eine einfache lineare Regression angepasst. Man kann fragen (s. 4.8.a), ob die Abweichungen auch zufällig sein könnten.

- c Nun kennt man die Standardabweichung σ_i sozusagen nie. Es genügt aber, die **relativen Genauigkeiten** oder Streuungen zu kennen, also $\text{var}\langle E_i \rangle = \sigma^2 v_i$ anzunehmen, wobei man v_i kennt und nur σ aus den Daten bestimmen muss. Man minimiert dann $\sum_i R_i^2 / v_i$.

Im vorhergehenden Abschnitt wurde erwähnt, dass sich in einem Streudiagramm der Residuen gegen eine Ausgangsgrösse $U^{(j)}$ zeigen kann, dass die Streuung von $U^{(j)}$ abhängt. Dann kann man versuchen, eine Funktion v anzugeben, die diese Abhängigkeit beschreibt, für die also $\text{var}\langle E_i \rangle \approx \sigma^2 v\langle u_i^{(j)} \rangle$ angenommen werden kann. Nun wendet man gewichtete Regression an mit den Gewichten $w_i = 1/v\langle u_i^{(j)} \rangle$.

* Schwieriger wird die Überlegung, wenn die Streuung der Residuen vom angepassten Wert \hat{y}_i abhängt. Man geht dann oft so vor, dass man zuerst das Modell ohne Gewichte anpasst und die so berechneten angepassten Werte als Grundlage für eine verfeinerte, gewichtete Regressionsrechnung benützt. Ein solches Vorgehen birgt aber Tücken – vor allem, wenn man auf die Idee verfällt, es zu wiederholen: Die geschätzte Regressionsfunktion kann sich dann zu sehr an (zufälligerweise) klein ausgefallene Y -Werte anpassen.

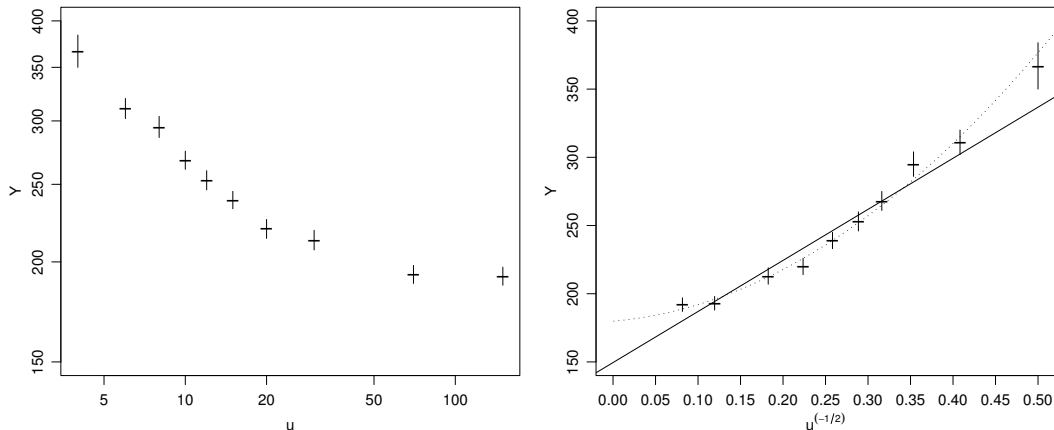


Abbildung 4.7.b: Daten des Beispiels der starken Wechselwirkung mit logarithmischen Achsen (links) und mit transformierter Energie (rechts). Im zweiten Fall sind die geschätzten Regressionsfunktionen mit linearem Modell (entsprechend der physikalischen Theorie) und quadratischem Modell eingezeichnet.

- d Es ist nicht schwierig, die **Koeffizienten**, die die gewichtete Quadratsumme minimieren, anzugeben und ihre Verteilung auszurechnen, siehe 4.e. Es sei \mathbf{W} die Diagonalmatrix mit den Diagonal-Elementen w_i . Dann wird

$$\hat{\underline{\beta}} = (\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{W} \underline{Y}.$$

Die Schätzung ist immer noch erwartungstreu und die Varianzen der $\hat{\beta}_j$ sind gleich den Diagonalelementen von $\sigma^2(\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}})^{-1}$.

Schliesslich ist die Varianz eines Residuums R_i wichtig für die Bestimmung von standardisierten Residuen. Diese werden

$$\begin{aligned} \tilde{R}_i &= R_i / \left(\hat{\sigma} \sqrt{1/w_i - (\mathbf{H}_W)_{ii}} \right) \quad \text{mit} \\ \mathbf{H}_W &= \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T. \end{aligned}$$

- e Welche **Residuen** soll man in grafischen Darstellungen verwenden? Nun ist der Unterschied zwischen standardisierten und unstandardisierten Residuen nicht mehr zu vernachlässigen. Generell gilt:
- Für die Beurteilung der Verteilung (im Normalverteilungs-Diagramm) und der Streuung der Fehler (im Streuungs-Diagramm) verwendet man standardisierte Residuen.
 - Wenn es um die Eignung der Regressionsfunktion geht (Tukey-Anscombe Diagramm und Streudiagramme der Residuen gegen die erklärenden Variablen), kommen unstandardisierte Residuen zum Zug.

In beiden Fällen ist es sinnvoll, die Gewichte w_i durch die Grösse der gezeichneten Symbole darzustellen.

- f Zur Überprüfung der Wahl der Gewichte sollen die Residuen analog zum Streuungsdiagramm gegen die Gewichte selbst aufgetragen werden.
- ▷ Für das Beispiel der starken Wechselwirkung mit quadratischem Modell zeigt Abbildung 4.7.f keine Hinweise, dass die Streuung der standardisierten Residuen von den Gewichten abhängen würden. Die Gewichtung scheint damit in Ordnung zu sein. Die eingezeichnete Glättung (die, wie im scale-location plot (4.2.o) für wurzeltransformierte Absolutwerte gerechnet und zum Zeichnen zurücktransformiert wurde) ist kaum ernst zu nehmen, da die Zahl der Beobachtungen zu klein ist.

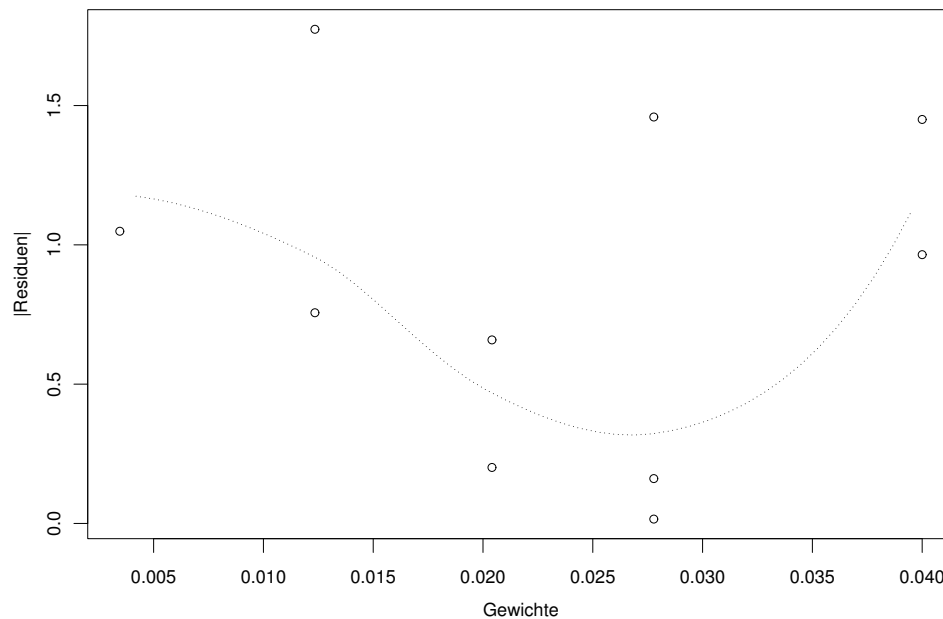


Abbildung 4.7.f: Absolute Residuen aus dem quadratischen Modell gegen Gewichte im Beispiel der starken Wechselwirkung

4.8 * Gesamthafte Überprüfung

a* Residuenanalysen können zu immer neuen Ideen führen, wie das Modell noch zu verbessern wäre. Idealerweise möchte man eine Methode haben, die sagt, wann es genügt ist.

Eine Idee zu einer solchen Methode beruht darauf, dass das Modell genügt, wenn die Residuen sich im Bereich der „natürlichen Streuung“ der Fehler bewegen. In gewissen Situationen kennt man eine solche Streuung, beispielsweise eine Mess-ungenauigkeit. In anderen Fällen gibt es Methoden, eine „natürlichen Streuung“ der Fehler zu schätzen. Die Grundidee aller Tests für die **Anpassung** oder den **lack of fit** besteht darin, die mit der Regressionsmethodik geschätzte Varianz $\hat{\sigma}^2$ der Fehler mit einer anderen Schätzung $\tilde{\sigma}^2$ zu vergleichen, die unabhängig davon gewonnen wird. Falls das Modell stimmt, sollte $\hat{\sigma}^2 \approx \tilde{\sigma}^2$ sein. Andernfalls ist $\hat{\sigma}^2$ grösser, weil die Residuen R_i zusätzlich zur zufälligen Streuung noch einen systematischen Fehler enthalten. Die Testgrösse ist jeweils das Verhältnis $T = \hat{\sigma}^2 / \tilde{\sigma}^2$. Ist diese Grösse signifikant grösser als 1, dann muss das Modell als unvollständig gelten.

b* Gegen solche Tests müssen allerdings die gleichen Bedenken wie gegen alle Anpassungstests angefügt werden: Die Anwendung von Tests ist für diese Problemstellung eigentlich nicht angebracht, denn **man möchte gerne die Nullhypothese beweisen**. Das ist bekanntlich nicht möglich; wir können eine Nullhypothese nur verwerfen oder beibehalten. Es kann gut sein, dass die Voraussetzung, die überprüft werden soll, verletzt ist, und dass trotzdem kein signifikantes Testergebnis entsteht (Fehler 2. Art).

c* Der einfachste Fall liegt vor, wenn eine Varianz für die Fehler aus einer anderen Quelle bekannt ist. Das ist der Fall, wenn Angaben zur Messgenauigkeit der Zielgrösse vorliegen. Allerdings sind diese oft vorsichtig, also die Ungenauigkeiten grösser angegeben, als sie in Wirklichkeit sind.

Sind die Ungenauigkeiten der Messfehler durch $\sigma_i^2 = \text{var}\langle E_i \rangle$ gegeben, dann lautet die Testgrösse $T = \sum_i R_i^2 / \sigma_i^2$; sie ist chiquadrat-verteilt, $\sim \chi_{n-p}^2$, falls die Varianzen stimmen und man sie bei der Schätzung mit gewichteter Regression berücksichtigt hat.

d* ▷ Im Beispiel der starken Wechselwirkung (4.7.b) waren die Standardabweichungen der E_i aus physikalischer Theorie bekannt. Für das lineare Modell erhält man als Residuen 30.3, 8.6, 13.1, 0.1, -4.6, -7.2 -13.3, -4.9, -1.3, 11.9; der Testwert $T = 19.3$ führt zum P-Wert $p = 0.013$. Das lineare Modell genügt also nicht – was dem visuellen Eindruck von Abbildung 4.7.b entspricht. Für die quadratische Regressionsfunktion erhält man dagegen die Residuen -9.67, -4.10, 11.16, 3.16, 0.97, -0.06, -5.87, 0.66, -3.00, 3.21 und daraus $T = 4.04$ und $p = 0.78$.

In diesem Beispiel – und allgemein in der einfachen linearen Regression – ist allerdings dieser Anpassungstest nicht besonders geeignet. Die naheliegenden Alternativen bestehen in einer „einfachen“ Krümmung, und gegen solche Alternativen ist es normalerweise effizienter, die Signifikanz eines quadratischen Terms zu prüfen. Im Beispiel wird der entsprechende P-Wert mit 0.0013 eine Grössenordnung kleiner als der P-Wert des lack-of-fit-Tests.

e* Wenn für die **gleichen X-Werte** $[x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}]$ **mehrere Beobachtungen** $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ gemacht werden, ergibt sich die Möglichkeit einer unabhängigen Schätzung von σ . (Normalerweise würden wir die Y-Werte durchnummerieren und hätten mehrere gleiche X-Werte-Kombinationen. Der unübliche zweite Index von Y_{ih} vereinfacht die folgende Überlegung.) Man kann dann die Varianz σ^2 der Fehler statt wie üblich auch nur aus der Streuung innerhalb dieser Gruppen schätzen, nämlich durch

$$\tilde{\sigma}^2 = \frac{1}{n-g} \sum_{i=1}^g \sum_{h=1}^{n_i} (Y_{ih} - \bar{Y}_{i.})^2 = \frac{1}{n-g} \text{SSQ}^{(rep)},$$

wobei $\bar{Y}_{i.}$ das Mittel über die n_i Beobachtungen zu den X-Werten $[x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}]$ und g die Anzahl solcher Beobachtungs-Gruppen ist, während $\text{SSQ}^{(rep)}$ die „Quadratsumme der Replikate“ bezeichnet.

Die Testgrösse

$$T = \frac{(\text{SSQ}^{(E)} - \text{SSQ}^{(rep)}) / (g-p)}{\text{SSQ}^{(rep)} / (n-g)}$$

hat unter der Nullhypothese eine F-Verteilung mit $g-p$ und $n-g$ Freiheitsgraden. (Falls $g < p$ ist, sind die Parameter nicht schätzbar; für $g = p$ ist T ebenfalls nicht definiert.)

Als Begründung denke man sich das betrachtete Modell erweitert durch je eine Indikatorvariable für jede der g Gruppen. Der Test ist ein F-Test zum Vergleich des betrachteten mit dem so erweiterten Regressionsmodell.

f* Wenn keine Gruppen von Beobachtungen mit gleichen X -Werten vorhanden sind, können Paare von „benachbarten“ X -Kombinationen $[x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}]$ und $[x_h^{(1)}, x_h^{(2)}, \dots, x_h^{(m)}]$ gesucht werden. Die quadrierten Differenzen $(R_i - R_h)^2$ der entsprechenden Residuen sollte im Mittel etwa $2\hat{\sigma}^2$ betragen. Man kann dies grafisch überprüfen, indem man $(R_i - R_h)^2$ gegenüber einem geeigneten Distanzmass $d\langle x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}; x_h^{(1)}, x_h^{(2)}, \dots, x_h^{(m)} \rangle$ in einem Streudiagramm aufträgt. Der Vorschlag stammt von Daniel and Wood (1980, Abschnitt 7.10), die

$$d\langle x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}; x_h^{(1)}, x_h^{(2)}, \dots, x_h^{(m)} \rangle = \sum_j (\hat{\beta}_j(x_i^{(j)} - x_h^{(j)}))^2 / \hat{\sigma}^2$$

benützen.

4.9 Unabhängigkeit

- a Die letzte Voraussetzung, die zu überprüfen bleibt, ist die **Unabhängigkeit** der zufälligen Fehler. Wenn die Beobachtungen eine natürliche, insbesondere eine **zeitliche Reihenfolge** einhalten, soll man die Residuen R_i in dieser Reihenfolge auftragen.
- ▷ *Im Beispiel der Sprengungen (Abbildung 4.9.a) sieht man allenfalls am Schluss einen Abfall; dies dürfte jedoch im Bereich eines Zufalls-Phänomens liegen.*

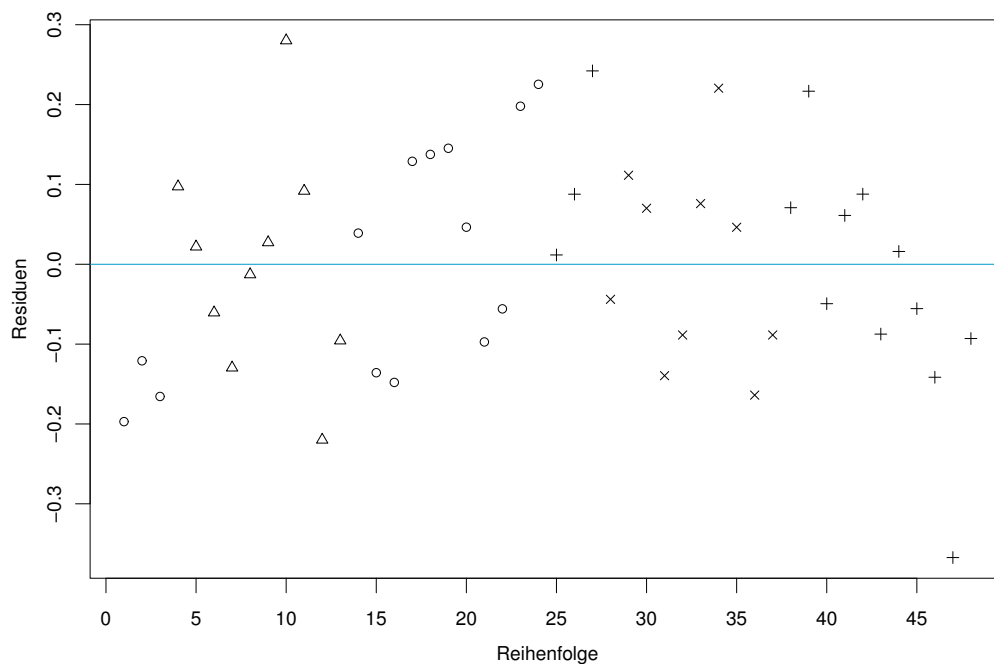


Abbildung 4.9.a: Residuen gegen Reihenfolge im Beispiel der Sprengungen. Die verschiedenen Stellen sind mit verschiedenen Symbolen dargestellt.

- b* Die Programme liefern häufig Tests, die die Unabhängigkeit überprüfen. Am bekanntesten ist der **Durbin-Watson-Test**. Wenn die Zufallsfehler positiv korreliert sind, dann unterscheiden sich aufeinanderfolgende Residuen weniger, als wenn sie unabhängig sind. Deshalb sollte die Teststatistik

$$T = \sum_{i=2}^n (R_i - R_{i-1})^2 / \sum_{i=1}^n R_i^2$$

in diesem Fall klein ausfallen. Leider ist die Verteilung der Teststatistik unter der Nullhypothese der Unabhängigkeit der E_i von der Design-Matrix \mathbf{X} abhängig (da ja die R_i trotzdem korreliert sind, siehe 4.d). Durbin und Watson ist es immerhin gelungen, ein Intervall anzugeben, in dem die wahre kritische Grenze für den Test liegen muss. Deshalb ist die Schlussweise im Durbin-Watson-Test unüblich: Man erhält aus Tabellen (die der Computer hoffentlich kennt) zwei Grenzen c' und c'' mit $c' < c''$ und schliesst

- auf Verwerfung der Unabhängigkeit, falls $T < c'$,
- auf Beibehaltung der Unabhängigkeit, falls $T > c''$,
- gar nichts (unentscheidbar), falls T dazwischen liegt.

(Vielleicht entschliesst sich jemand gelegentlich, dieses Problem mit den heutigen Rechenmöglichkeiten befriedigender zu lösen!)

- c Oft ist jede Beobachtung mit einem **Ort** verbunden, und es ist plausibel, dass die Beobachtungen an benachbarten Orten ähnlicher sind als für weit entfernte Orte. Solche räumliche Korrelationen zeigen sich im **Beispiel der basischen Böden**. Die Bäume wurden in einem regelmässigen Gitter gepflanzt. Für die Gitterpunkte sind in Abbildung 4.9.c die Residuen auf gleiche Weise dargestellt wie in Abbildung 4.6.f.

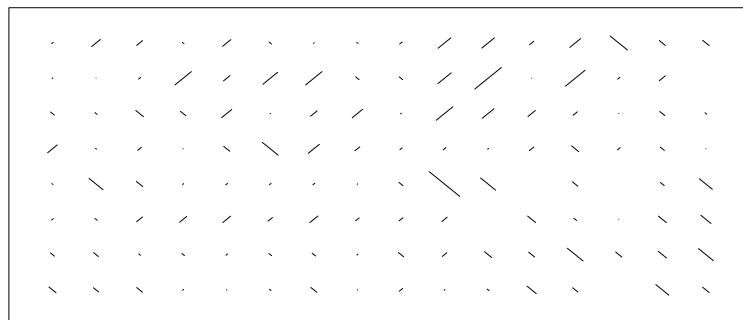


Abbildung 4.9.c: Residuen und räumliche Anordnung der Beobachtungen im Beispiel der basischen Böden

Benachbarte Punkte scheinen in der Tat ähnliche Residuen aufzuweisen. In der rechten unteren Ecke sind alle Residuen negativ. Es ist eine Abhängigkeit zwischen den Fehlern vorhanden, die sich geografisch zeigt.

- d Wenn Korrelationen – zeitliche, räumliche oder andere – vorliegen, dann sind die P-Werte der üblichen Tests häufig grob falsch. Methoden, die Korrelationen berücksichtigen, laufen unter der Bezeichnung **Verallgemeinerte Kleinste Quadrate**. Wir kommen im Block Regression von Zeitreihen auf das Problem zurück.

4.10 Einflussreiche Beobachtungen

- a **Ausreisser** wurden schon in 4.5.a diskutiert. Manchmal verschwinden sie durch Verbesserungen des Modells. Soweit sie stehen bleiben, stellt sich die Frage, wie stark sie die Analyse beeinflussen. Weshalb ist das wichtig? Wenn es sich um fehlerhafte Beobachtungen handelt, wird die Analyse verfälscht. Wenn es korrekte Beobachtungen sind und sie die Ergebnisse stark prägen, ist es nützlich, dies zu wissen. Man wird dann als Interpretation die Möglichkeit bedenken, dass die Ausreisser aus irgendeinem Grund nicht zur gleichen Grundgesamtheit gehören, und dass das an die übrigen Beobachtungen angepasste Modell die „typischen“ Zusammenhänge in sinnvoller Weise wiedergibt.
- b Der **Effekt eines Ausreissers** auf die Resultate kann untersucht werden, indem die Analyse ohne die fragliche Beobachtung wiederholt wird. Auf dieser Idee beruhen die „(influence) **diagnostics**“, die von etlichen Programmen als grosse Tabellen geliefert werden: Die Veränderung aller möglichen Resultatgrössen (Schätzwerte, Teststatistiken) beim Weglassen der i ten Beobachtung werden für alle i angegeben. (Dazu muss nicht etwa die Analyse n mal wiederholt werden; es sind starke rechnerische Vereinfachungen möglich, so dass der zusätzliche Rechenaufwand unbedeutend wird.) Es ist nützlich, diese diagnostics zu studieren. Leider zeigen sie aber oft nicht, was passieren würde, wenn man zwei oder mehrere Ausreisser gleichzeitig weglässt – die Effekte müssen sich nicht einfach addieren.
- c Ein wesentlicher Teil dieser Tabellen kann glücklicherweise mit einer einzigen grafischen Darstellung erfasst werden, die wir **Hebelarm-Diagramm** (*leverage plot*) nennen wollen. Etliche influence diagnostics sind nämlich Funktionen des i ten Residuum R_i , der leverage H_{ii} (4.3.h) und der geschätzten Standardabweichung $\hat{\sigma}$.

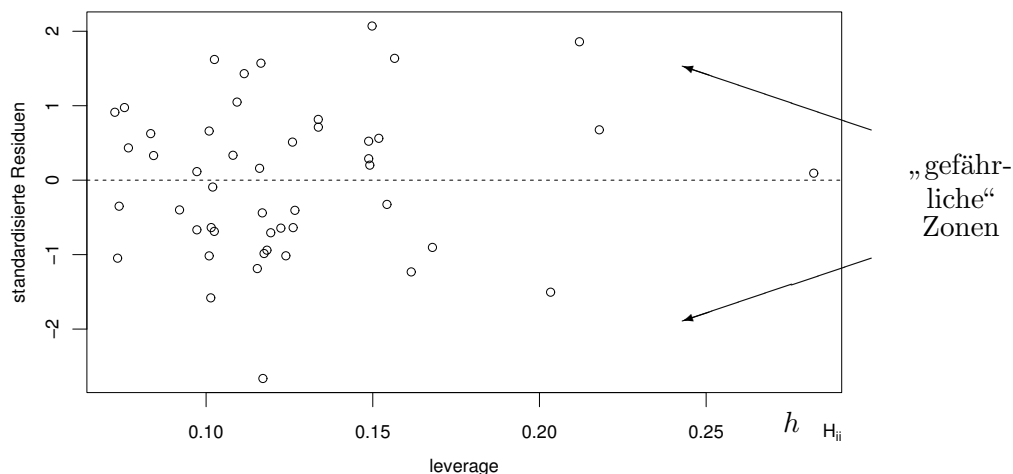


Abbildung 4.10.c: Hebelarm-Diagramm für das Beispiel der Sprengungen

Die (Beträge der) Einfluss-Indikatoren sind jeweils grösser für grössere $|R_i|$ und grössere H_{ii} . Für die grafische Darstellung verwendet man aber besser die standardisierten Residuen \tilde{R}_i , die ja selbst aus R_i , H_{ii} und $\hat{\sigma}$ berechnet werden (4.3.i). In einem Streu-

diagramm der \tilde{R}_i gegen die H_{ii} sind die „gefährlichen“ Beobachtungen rechts, oben und unten, zu finden (Abbildung 4.10.c). Es gibt allerdings keine eindeutigen Grenzen, die festlegen, wo die „Gefährlichkeit“ beginnt.

Im Beispiel ist die grösste leverage bedenklich gross und die beiden extremeren Residuen der Beobachtungen mit $H_{ii} > 0.2$ sind ebenfalls beachtenswert. Es könnte sich lohnen, die Analyse versuchsweise ohne diese Beobachtungen zu wiederholen.

- d Neben den standardisierten Residuen gibt es auch so genannte **studentisierte Residuen**. Das i te studentisierte Residuum misst die Differenz zwischen Y_i und dem angepassten Wert, der sich ergäbe, wenn man die i te Beobachtung zum Anpassen des Modells nicht verwenden würde. Diese Differenz wird noch geeignet standardisiert. Man würde erwarten, dass man zur Berechnung dieser Grössen für jede Beobachtung das Modell neu anpassen müsse. Es zeigt sich aber, dass sie sich als relativ einfache Funktion aus R_i , H_{ii} und $\hat{\sigma}$ ergeben.
- e Die **Distanz von Cook** fasst die Veränderungen aller angepassten Werte \hat{y}_i beim Weglassen der i ten Beobachtung zu einer Zahl zusammen (nämlich zu ihrer Quadratsumme $(\hat{y}_{(-i)} - \hat{y})^T (\hat{y}_{(-i)} - \hat{y})$, dividiert durch $p\hat{\sigma}^2$). Sie lässt sich schreiben als

$$d_i^{(C)} = \frac{R_i^2 H_{ii}}{p\hat{\sigma}^2 (1 - H_{ii})^2} = (1/p) \tilde{R}_i^2 H_{ii}/(1 - H_{ii}),$$

ist also ebenfalls eine Funktion der drei erwähnten Grössen.

Im Programmsystem R werden die $d_i^{(C)}$ in der Reihenfolge der Beobachtungen im Datensatz routinemässig grafisch dargestellt.

- f Der Einfluss einzelner Beobachtungen auf einen **einzelnen Regressionskoeffizienten** β_j zeigt sich in einem speziellen Streudiagramm, das **added variable plot** oder **partial regression leverage plot** genannt wird. (Das erste könnte man als „Diagramm für zusätzliche Variable“ übersetzen.) Es zeigt die Residuen einer Regressionsanalyse ohne die entsprechende erklärende Variable $X^{(j)}$, aufgetragen gegen „korrigierte“ Werte von $X^{(j)}$. Diese Werte erhält man als Residuen in einer Regression von $X^{(j)}$ (als „Zielvariable“) auf die übrigen erklärenden Variablen – mit der Bildung solcher Residuen schaltet man die „indirekten Einflüsse“ von $X^{(j)}$ auf Y aus.

Wenn man in diesem Streudiagramm eine Gerade (mit Kleinsten Quadraten) anpasst, so hat sie genau die Steigung $\hat{\beta}_j$, die auch bei der Schätzung aller Koeffizienten im gesamten Modell herauskommt. Das Diagramm zeigt, wie diese „Steigung“ zustandekommt, also insbesondere, welche Beobachtungen einen starken Einfluss auf sie ausüben.

In Abbildung 4.10.f fällt ein Punkt im linken Teil auf, der einen starken Einfluss auf den geschätzten Koeffizienten der Distanz hat. Es handelt sich um unseren altbekannten Ausreisser.

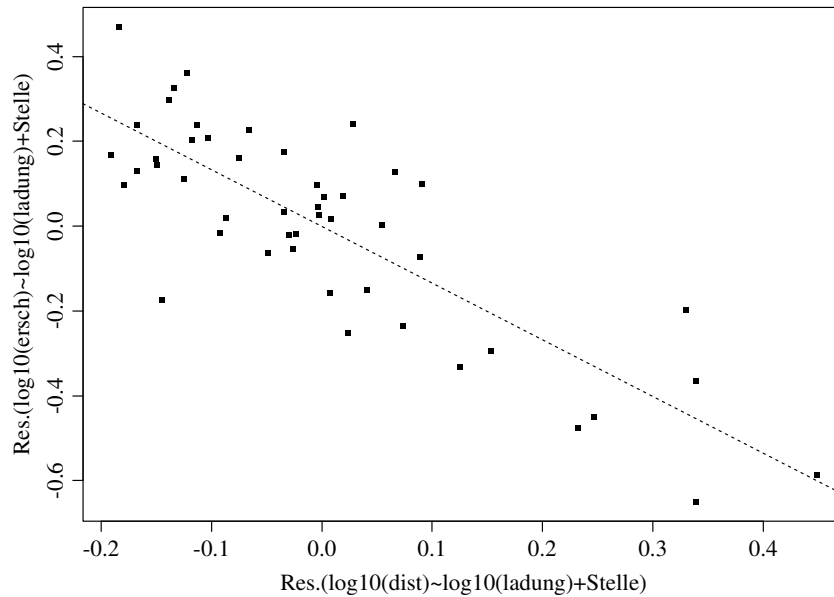


Abbildung 4.10.f: Added variable plot für die logarithmierte Distanz im Beispiel der Sprengungen

4.A Theoretische Verteilung der Residuen

- a Die **angepassten Werte** kann man mit Hilfe der in 3.4.g hergeleiteten Matrix-Formel einfach schreiben,

$$\begin{aligned}\underline{\hat{y}} = \widetilde{\mathbf{X}}\underline{\hat{\beta}} &= \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\underline{Y} \\ &=: \mathbf{H}\underline{Y}.\end{aligned}$$

Die Matrix \mathbf{H} heisst **Projektionsmatrix** (von \underline{Y} auf den Raum, der durch die erklärenden Variablen $\underline{X}^{(j)}$ aufgespannt wird) oder **Hut-Matrix (hat matrix)** – „sie setzt dem Y den Hut auf!“

Die **Diagonal-Elemente** H_{ii} von \mathbf{H} haben eine besondere Bedeutung: Wenn man einen Wert Y_i um Δy_i verändert, dann misst, wie die Gleichung zeigt, $H_{ii}\Delta y_i$ die Veränderung des zugehörigen angepassten Wertes \hat{y}_i .

- b Nun zur **Verteilung der Residuen**. !!! Hier werden noch Voraussetzungen an die Kenntnisse gemacht, die nicht erfüllt sind.

Zunächst ist einfach festzustellen, dass jedes Residuum den Erwartungswert 0 hat,

$$\mathcal{E}(\underline{R}) = \mathcal{E}(\underline{Y}) - \widetilde{\mathbf{X}}\mathcal{E}(\underline{\hat{\beta}}) = \widetilde{\mathbf{X}}\underline{\beta} - \widetilde{\mathbf{X}}\underline{\beta} = \underline{0}.$$

Für die Berechnung der **Varianz** schreiben wir zuerst

$$\underline{R} = \underline{Y} - \underline{\hat{y}} = \mathbf{I}\underline{Y} - \mathbf{H}\underline{Y} = (\mathbf{I} - \mathbf{H})\underline{Y}$$

und erhalten daraus

$$\begin{aligned}\text{var}\langle \underline{R} \rangle &= (\mathbf{I} - \mathbf{H}) \text{var}\langle \underline{Y} \rangle (\mathbf{I} - \mathbf{H})^T = \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2 (\mathbf{I} - \mathbf{H} - \mathbf{H}^T + \mathbf{H}\mathbf{H}^T) .\end{aligned}$$

Es ist $\mathbf{H} = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T$ und deshalb $\mathbf{H}^T = \mathbf{H}$ und

$$\begin{aligned}\mathbf{H}\mathbf{H}^T &= \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \\ &= \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T = \mathbf{H} .\end{aligned}$$

Also gilt

$$\text{var}\langle \underline{R} \rangle = \sigma^2 (\mathbf{I} - \mathbf{H}) .$$

Die Varianzen der einzelnen Residuen stehen in der Diagonalen dieser Matrix, $\text{var}\langle R_i \rangle = (1 - H_{ii}) \sigma^2$.

- c Die Gleichung $\underline{R} = (\mathbf{I} - \mathbf{H}) \underline{Y}$ zeigt, dass die R_i und damit auch die „halb-standardisierten“ Residuen $R_i / \sqrt{1 - H_{ii}}$ Linearkombinationen der normalverteilten Y_i sind. Sie sind deshalb selbst normalverteilt; es gilt $R_i / \sqrt{1 - H_{ii}} \sim \mathcal{N}\langle 0, \sigma^2 \rangle$.
- d* Gemäss der Formel $\text{var}\langle \underline{R} \rangle = \sigma^2 (\mathbf{I} - \mathbf{H})$ sind die Residuen korreliert,

$$\text{cov}\langle R_i, R_k \rangle = -\sigma^2 H_{ik} .$$

- e **Gewichtete Regression.** Es sei \mathbf{W} die Diagonalmatrix mit den Diagonal-Elementen w_i . Dann ist

$$Q\langle \underline{\beta}^* \rangle = \sum_i w_i R_i^2 = \underline{R}^T \mathbf{W} \underline{R}$$

zu minimieren. Es ergeben sich die Normalgleichungen

$$\widetilde{\mathbf{X}}^T \mathbf{W} \underline{R} = \underline{0} \quad \text{oder} \quad \widetilde{\mathbf{X}}^T \mathbf{W} (\underline{Y} - \widetilde{\mathbf{X}} \underline{\hat{\beta}}) = \underline{0} \quad \Rightarrow \quad \widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}} \underline{\hat{\beta}} = \widetilde{\mathbf{X}}^T \mathbf{W} \underline{Y}$$

und daraus, mit $\mathbf{C}_W = \widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}}$,

$$\underline{\hat{\beta}} = \mathbf{C}_W^{-1} \widetilde{\mathbf{X}}^T \mathbf{W} \underline{Y} .$$

Die Erwartungstreue ist einfach nachzurechnen. Da $\text{var}\langle Y_i \rangle = \sigma^2 / w_i$ und deshalb $\text{var}\langle \underline{Y} \rangle = \sigma^2 \mathbf{W}^{-1}$ gilt, wird

$$\begin{aligned}\text{var}\langle \underline{\hat{\beta}} \rangle &= \mathbf{C}_W^{-1} \widetilde{\mathbf{X}}^T \mathbf{W} \cdot \sigma^2 \mathbf{W}^{-1} \cdot \mathbf{W} (\mathbf{C}_W^{-1} \widetilde{\mathbf{X}}^T)^T = \sigma^2 \mathbf{C}_W^{-1} \widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}} (\mathbf{C}_W^{-1})^T \\ &= \sigma^2 (\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}})^{-1} .\end{aligned}$$

f Die Residuen sind jetzt gleich

$$\underline{R} = (\mathbf{I} - \widetilde{\mathbf{X}} \mathbf{C}_W^{-1} \widetilde{\mathbf{X}}^T \mathbf{W}) \underline{Y} = (\mathbf{I} - \mathbf{H}_W \mathbf{W}) \underline{Y},$$

wenn wir $\mathbf{H}_W = \widetilde{\mathbf{X}} \mathbf{C}_W^{-1} \widetilde{\mathbf{X}}^T$ setzen. Ihre Kovarianzmatrix wird

$$\begin{aligned} \text{var}(\underline{R}) &= (\mathbf{I} - \mathbf{H}_W \mathbf{W}) \cdot \sigma^2 \mathbf{W}^{-1} \cdot (\mathbf{I} - \mathbf{H}_W \mathbf{W})^T \\ &= \sigma^2 (\mathbf{W}^{-1} - \mathbf{H}_W \mathbf{W} \mathbf{W}^{-1} - \mathbf{W}^{-1} \mathbf{W} \mathbf{H}_W + \mathbf{H}_W \mathbf{W} \mathbf{W}^{-1} \mathbf{W} \mathbf{H}_W) \\ &= \sigma^2 (\mathbf{W}^{-1} - \mathbf{H}_W). \end{aligned}$$

Die standardisierten Residuen sind also

$$\widetilde{R}_i = R_i / \left(\widehat{\sigma} \sqrt{1/w_i - (\mathbf{H}_W)_{ii}} \right).$$

4.S S-Funktionen

- a Die Funktion `plot` zeigt, wenn man sie auf das Resultat einer Regressions-Anpassung anwendet, Diagramme, die der Residuen-Analyse dienen. Grundlegend ist dabei der Tukey-Anscombe plot (Residuen gegen angepasste Werte), und zudem wird normalerweise ein QQ-plot (Normalverteilungs-Diagramm) der Residuen und der scale-location plot (Absolutbeträge der Residuen gegen angepasste Werte) zur Überprüfung der Homogenität der Varianzen dargestellt.
- b Wenn die Regression mit `lm` angepasst wurde, werden zudem die Werte der Cook-Distanz in der Reihenfolge der Beobachtungen gezeichnet.

Für ein Objekt, das mit `regr` erzeugt wurde, wird von der entsprechenden `plot`-Methode stattdessen der leverage plot (Residuen gegen Hebelwerte H_{ii}) gezeigt. Einflussreiche Beobachtungen befinden sich rechts oben und unten. Zudem werden die Residuen gegen die Reihenfolge der Beobachtungen aufgetragen. Schliesslich wird die unten beschriebene Funktion `plresx` für alle Variablen, die in der Modellformel vorkommen, aufgerufen. Als Alternative (oder zusätzlich) zum Tukey-Anscombe-Diagramm kann die Zielgrösse statt der Residuen gegen die angepassten Werte aufgetragen werden.

Das Ziel der `plot`-Methode für die Ergebnisse von `regr` ist es, für den „Normalfall“ eine möglichst vollständige Residuen-Analyse zu präsentieren. Erfahrungsgemäss beschränkt sich die Residuen-Analyse der meisten Benutzer nämlich darauf, anzusehen, was die Funktion `plot` automatisch liefert.

- c Die **Argumente** `smooth` und `smooth.sim` von `plot` für `regr`-Objekte. In allen geeigneten Grafiken wird eine glatte Kurve eingezeichnet, ausser wenn `smooth=FALSE` gesetzt wird. Wenn `smooth` nicht selbst eine Funktion ist, wird `lowess` verwendet. Es werden `smooth.sim=19` Datensätze der Zielgrösse entsprechend dem angepassten Modell erzeugt und angepasst und die Ergebnisse der Glättungsmethode jeweils mit eingezeichnet (in schwächerer Farbe), damit die „Zufälligkeit“ der Glättung beurteilt werden kann. Wie man damit sehen kann, passt sich eine Glättung an den Rändern meist zu stark den Beobachtungen an. Die Glättung im scale-location plot beruht auf den Wurzeln der Absolutbeträge der Residuen, auch wenn die Absolutbeträge (und die zurücktransformierte Glättung) gezeigt werden (im Gegensatz zur Methode für `lm`).
- d **Funktion** `termpplot`. (nur R ? sonst `plot.gam` ?) Residuen, genauer partial residuals, werden gegen die Ausgangsgrössen aufgetragen.
- e **Funktion** `plresx`. (Zusatzfunktion zu `regr`.) Diese Funktion leistet Ähnliches wie `termpplot`: Die Residuen werden gegen die erklärenden Variablen aufgetragen. Im Normalfall werden die Residuen (ohne „component effect“) verwendet; dafür wird die Referenzlinie, die konstanten Y -Werten entspricht (und gleich den negativen component effects ist), eingezeichnet. Die Argumente `smooth` und `smooth.sim` funktionieren wie oben.
- f Die Funktionen rufen für jede grafische Darstellung die Funktion `g.stamp` auf, die zur Dokumentation des grafischen Outputs dient.

Literaturverzeichnis

- Agresti, A. (1990). *Categorical Data Analysis*, Wiley, N.Y.
- Agresti, A. (1996). *Introduction to categorical data analysis*, Wiley Series in Probability & Math. Statistics, Wiley, New York.
- Christensen, R. (1990). *Log-linear models*, Springer, N.Y.
- Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.
- Clogg, C. C. and Shihadeh, E. S. (1994). *Statistical models for ordinal variables*, Sage, Thousand Oaks, CA.
- Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table, *Communications in Statistics – Theory and Methods* **A9**: 1025–1041.
- Collet, D. (1991, 1999). *Modelling binary data*, Chapman & Hall/CRC Press LLC, Boca Raton, Florida.
- Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*, Wiley, N.Y.
- Cox, D. R. (1989). *Analysis of Binary Data*, 2nd edn, Chapman and Hall, London.
- Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*, Chapman and Hall, London.
- Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N.Y.
- Davies, P. (1995). Data features, *Statistica Neerlandica* **49**: 185–245.
- Devore, J. L. (1991). *Probability and Statistics for Engineering and the Sciences*, 3rd edn, Duxbury Press, Belmont, California.
- Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd edn, Chapman and Hall, London.
- Draper, N. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edn, Wiley, N.Y.
- Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer-Verlag, New York.
- Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics, *Journal of the American Statistical Association* **87**: 178–183.
- Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N.Y.
- Haaland, P. D. (1989). *Experimental Design in Biotechnology*, Marcel Dekker, N.Y.
- Hartung, J., Elpelt, B. und Klösener, K. (1998). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 11. Aufl., Oldenbourg, München.

- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*, Wiley, N.Y.
- Linder, A. und Berchtold, W. (1982). *Statistische Methoden II: Varianzanalyse und Regressionsrechnung*, Birkhäuser, Basel.
- Lindsey, J. K. (1995). *Modelling Frequency and Count Data*, number 15 in *Oxford Statistical Science Series*, Clarendon Press, Oxford.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.
- Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, Massachusetts.
- Myers, R. H., Montgomery, D. C. and Vining, G. G. (2001). *Generalized Linear Models. With Applications in Engineering and the Sciences*, Wiley Series in Probability and Statistics, Wiley, NY.
- Ryan, T. P. (1997). *Modern Regression Methods*, Series in Probability and Statistics, Wiley, N.Y. includes disk
- Sachs, L. (1997). *Angewandte Statistik*, 8. Aufl., Springer, Berlin.
- Sen, A. and Srivastava, M. (1990). *Regression Analysis; Theory, Methods, and Applications*, Springer-Verlag, N.Y.
- Stahel, W. A. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3. Aufl., Vieweg, Wiesbaden.
- Stahel, W. A. (2002). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 4. Aufl., Vieweg, Wiesbaden.
- van der Waerden, B. L. (1971). *Mathematische Statistik*, 3. Aufl., Springer, Berlin.
- Vincze, I. (1984). *Mathematische Statistik mit industriellen Anwendungen*, Band 1, 2, 2. Aufl., Bibliographisches Institut, Mannheim.
- Weisberg, S. (1990). *Applied Linear Regression*, 2nd edn, Wiley, N.Y.
- Wetherill, G. (1986). *Regression Analysis with Applications*, number 27 in *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.