Lineare Regression

Werner Stahel
Seminar für Statistik, ETH Zürich

Mai 2008 / Sept. 2013

Unterlagen zum Teil 1 der Vorlesung / des Kurses in Angewandter Regression

1 Einführung in die statistische Regressionsrechnung

1.1 Beispiele zur linearen Regression

a In der Wissenschaft, in der Technik und im Alltag fragen wir immer wieder danach, wie eine Grösse, die uns speziell interessiert, von anderen Grössen abhängt. Diese grundlegende Frage behandelt die statistische Regression, die deshalb wohl (neben einfachen grafischen Darstellungen) die am meisten verwendete Methodik der Statistik darstellt.

In diesem Abschnitt soll mittels Beispielen zur „gewöhnlichen“ linearen Regression in die Problemstellung eingeführt werden, bevor ein Überblick über die verschiedenen, allgemeineren Regressions-Modelle geboten wird.

b ▷ **Beispiel Sprengungen**. Beim Bau eines Strassentunnels zur Unterfahrung einer Ortschaft muss gesprengt werden. Die Erschütterung der Häuser darf dabei einen bestimmten Wert nicht überschreiten. In der Nähe der Häuser muss daher vorsichtig gesprengt werden, was natürlich zu erhöhten Kosten führt. Es lohnt sich, eine Regel zu entwickeln, die angibt, wie stark in welcher Situation gesprengt werden darf.

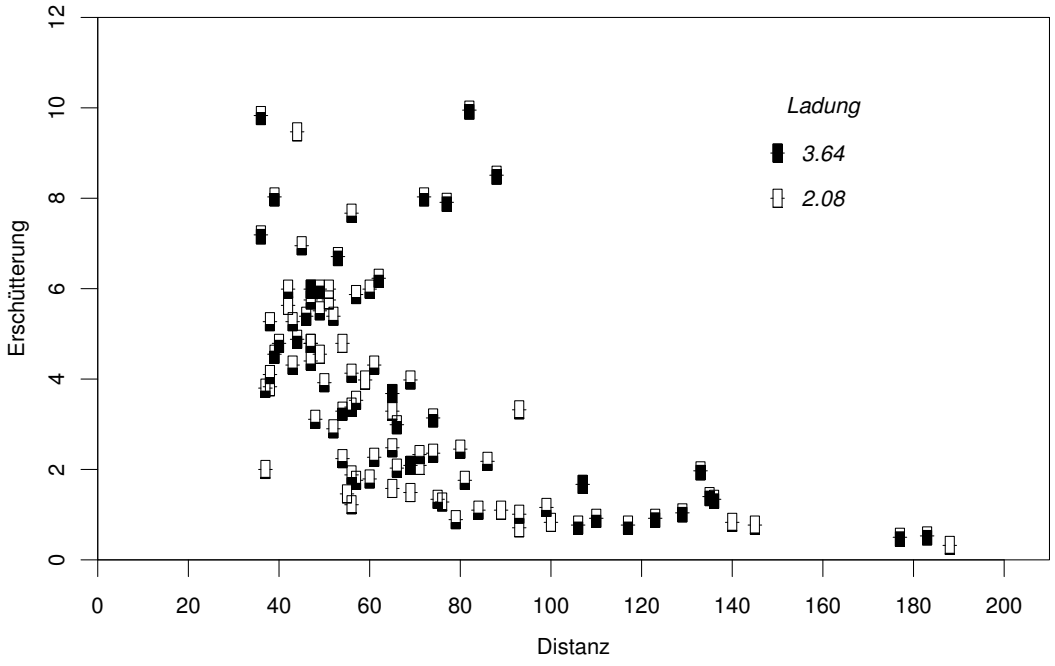

Abbildung 1.1.c: Erschütterung in Abhängigkeit von der Distanz für verschiedene Ladungen

Die Erschütterung ist abhängig von der Sprengladung, von der Distanz zwischen dem Spreng- und dem Messort, von der Art des Untergrund-Materials zwischen diesen Punkten, vom Ort der Sprengung im Tunnelprofil und möglicherweise von weiteren Grössen. Wäre die Erschütterung eine exakte, bekannte Funktion dieser Grössen und könnte man sie bei einer geplanten Sprengung alle genau erfassen, dann könnte man die Sprengladung ausrechnen, die zu einer gerade noch tolerierbaren Erschütterung führt. ◁

c Beginnen wir, mathematische Symbole und Sprachregelungen einzuführen!

Die **Zielgrösse** y (englisch *target variable)* – die Erschütterung – hängt über eine Funktion h von den **Eingangsgrössen** oder **erklärenden Variablen** $x^{(1)}$, $x^{(2)}$, …, $x^{(m)}$ *(explanatory variables)* – Ladung, Distanz, Spreng-Situation, Untergrundart – ab.

Bemerkungen zur Wortwahl. Der Ausdruck „erklärende Variable“ ist geeignet, wenn diese die Ursachen für die Zielgrösse darstellen. Da dies in vielen Anwendungen nicht gewährleistet ist – eine Regression kann dazu dienen, aus der Grösse der Wirkung auf den Wert der verursachenden Variablen zu schliessen – bevorzugen wir hier den Ausdruck Eingangsgrösse, der diesbezüglich etwas neutraler tönt.

Die ebenfalls gebräuchlichen Ausdrücke **„unabhängige Variable“** für die $x^{(j)}$ und **„abhängige Variable“** für y sind irreführend, da sie mit stochastischer Unabhängigkeit nichts zu tun haben.

* Der Ausdruck Ausgangsgrösse – Grösse, von der man ausgeht – wäre vom umgangssprachlichen Gebrauch ebenfalls naheliegend, aber im Zusammenhang mit Systemen, die Eingangs- und Ausgangsgrössen haben, bezeichnet er das genaue Gegenteil.

d Im Idealfall sollte also

$$y_i = h\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\rangle$$

für jede **Beobachtung** i (jede Sprengung) gelten.

▷ Leider existiert eine solche Formel nicht, und das Untergrundmaterial ist sowieso nicht genau genug erfassbar. Abbildung 1.1.d zeigt die Erschütterung in Abhängigkeit von der Distanz für verschiedene Ladungen. (Die Daten stammen vom Bau der Unterfahrung von Schaffhausen. Sie wurden freundlicherweise vom Ingenieurbüro Basler und Hoffmann, Zürich, zur Verfügung gestellt.) ◁

e Die statistische Regressionsrechnung geht davon aus, dass eine Formel wenigstens „ungefähr“ gilt – bis auf Abweichungen, die „zufällig“ genannt werden. Wir schreiben

$$Y_i = h\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\rangle + E_i$$

und nennen die E_i die **Zufallsfehler**. Die Vorstellungen, wie gross solche Abweichungen sind, werden mit einer Wahrscheinlichkeits-Verteilung formuliert. Oft wird dafür die Normalverteilung verwendet.

Man wird mit Hilfe dieses Modells trotz der Unsicherheit eine Regel für die zu wählende Grösse der Sprengladung herleiten können. Allerdings muss man zulassen, dass gemäss Modell auch eine zu grosse Erschütterung mit einer gewissen Wahrscheinlichkeit auftreten kann. Will man diese Wahrscheinlichkeit klein halten, so muss man entsprechend vorsichtig sprengen. Die statistische Regressionsrechnung gibt einen Zusammenhang zwischen der Ladung und der Wahrscheinlichkeit einer zu grossen Erschütterung bei einer bestimmten Distanz an.

Dieses Beispiel wird uns in den kommenden Abschnitten begleiten. Auf die Antworten müssen Sie deshalb noch eine Weile warten.

f ▷ **Beispiel Schadstoffe im Tunnel**. Die Schadstoffe, die vom motorisierten Verkehr ausgestossen werden, bilden einen wesentlichen Bestandteil der Belastung der Luft. Um die Grösse dieser Belastung zu schätzen, werden für die Fahrzeuge so genannte **Emissionsfaktoren** bestimmt. Dies kann einerseits auf dem Prüfstand geschehen, auf dem die Strasse mit Rollen simuliert wird. Der Widerstand der Rollen wird dabei variiert, so dass ein typischer „Fahrzyklus“ durchgespielt werden kann. – Andererseits eignen sich Strassentunnels mit Ein-Richtungs-Verkehr für Messungen unter realen Bedingungen. Misst man Schadstoff-Konzentrationen am Anfang und am Schluss des Tunnels und zählt, wie viele Fahrzeuge durch den Tunnel fahren, so kann man ebenfalls Emissionsfaktoren ausrechnen. Allerdings erhält man zunächst nur einen gemittelten Faktor für jeden gemessenen Schadstoff, und dieser lässt sich nicht ohne zusätzliche

Erkenntnisse auf andere Strassenabschnitte übertragen. Wenn man die Anzahl der Fahrzeuge nach Fahrzeug-Kategorien aufteilen kann, dann kann man immerhin mit Regressionsrechnung zu einem Emissionsfaktor für jede Fahrzeug-Kategorie kommen.

Während einer Woche im September 1993 wurden in der Südröhre des Gubrist-Tunnels nördlich von Zürich solche Messungen durchgeführt. Die Schadstoff-Konzentrationen am Anfang und am Ende wurden gemessen und die Luftströmung erfasst. Daraus lässt sich die Schadstoff-Emission Y pro Kilometer für alle durchgefahrenen Fahrzeuge zusammen berechnen. Von einem Schlaufen-Detektor im Strassenbelag wurden die Fahrzeuge in zwei Kategorien gezählt: Auf Grund des Abstands von Vorder- und Hinterachse wurden die Lastwagen von den übrigen Fahrzeugen getrennt. Es bezeichne $x^{(1)}$ die Anzahl „Nicht-Lastwagen“ und $x^{(2)}$ die Anzahl Lastwagen. Die gesamten Emissionen in der Zeitperiode i setzen sich zusammen gemäss

$$Y_i = \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + E_i \ ,$$

wobei θ_1 die durchschnittliche Emission pro Nicht-Lastwagen und θ_2 diejenige pro Lastwagen bedeutet – also die Grössen, an denen wir in der Studie primär interessiert sind. Die „Zufallsfehler“ E_i entstehen durch Variationen in Bauart und Zustand der Fahrzeuge, durch zeitliche Abgrenzungs-Schwierigkeiten und durch Mess-Ungenauigkeiten.

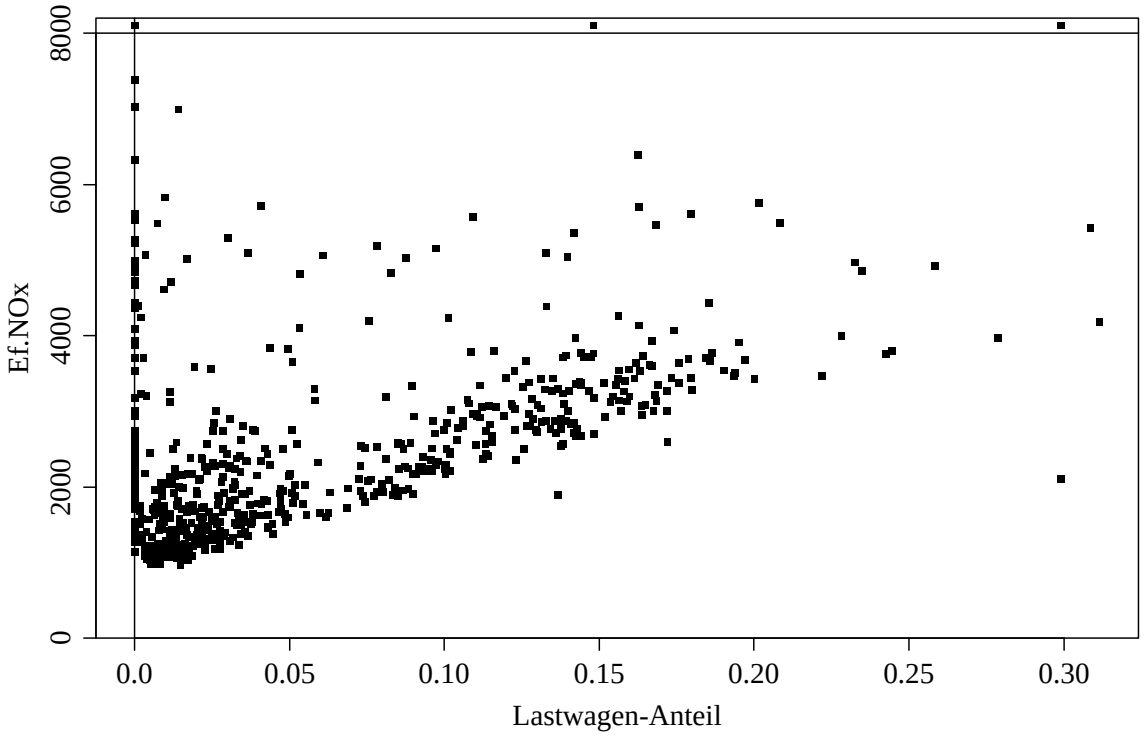

Abbildung 1.1.f: Emissionsfaktor für NO_X und Lastwagen-Anteil, gemittelt über jeweils 15 Minuten, im Beispiel der Schadstoffe im Tunnel. Drei extrem hohe Y-Werte sind im Bildrand dargestellt.

g ▷ Die Formel lässt sich in eine üblichere und vielleicht noch einfachere Form bringen: Wir dividieren Y_i, $x_i^{(1)}$ und $x_i^{(2)}$ durch die gesamte Anzahl Fahrzeuge $x_i^{(1)} + x_i^{(2)}$ und erhalten $\widetilde{Y}_i = \theta_1 \widetilde{x}_i^{(1)} + \theta_2 \widetilde{x}_i^{(2)} + \widetilde{E}_i$, wobei $\widetilde{Y}_i$ der „mittlere Emissionsfaktor“ für die Zeitperiode i und $\widetilde{x}_i^{(1)}$ und $\widetilde{x}_i^{(2)}$ die Anteile der Nicht-Lastwagen und der Lastwagen bedeuten. Da $\widetilde{x}_i^{(1)} = 1 - \widetilde{x}_i^{(2)}$ ist, gilt

$$\widetilde{Y}_i = \theta_1 + (\theta_2 - \theta_1)\widetilde{x}_i^{(2)} + \widetilde{E}_i \ .$$

Mit weniger komplizierten Symbolen geschrieben sieht das so aus:

$$Y_i = \alpha + \beta x_i + E_i \ .$$

Dies ist das Modell einer so genannten **einfachen linearen Regression**. Die Konstanten α und β nennen wir **Koeffizienten** oder **Parameter** des Modells. Wir wollen sie aus den Daten der Studie bestimmen, also **schätzen**.

In Abbildung 1.1.f zeigt sich als Tendenz eine lineare Zunahme des mittleren Emissionsfaktors für NO_x mit zunehmendem Lastwagen-Anteil, wie es dem besprochenen Modell entspricht. ◁

h ▷ **Beispiel Lastwagen-Anteil**. Der Schlaufen-Detektor zählt zwar die gesamte Zahl der Fahrzeuge zuverlässig, kann aber den Anteil der Lastwagen nur ungenau erfassen. Deshalb (unter anderem) wurde der Verkehr zeitweise mit Video aufgenommen und der Lastwagen-Anteil auf diesen Aufnahmen genau ausgezählt. Da dies teurer war, konnte nicht der ganze Zeitraum abgedeckt werden. Abbildung 1.1.h zeigt, dass die Schlaufen-Zählung systematische und zufällige Abweichungen von der Video-Zählung aufweist. Die zufälligen Abweichungen kommen teilweise zustande, weil die Schlaufe am Anfang, die Kamera aber am Ende des Tunnels installiert war, und die Abgrenzung der Mess-Intervalle nicht entsprechend korrigiert wurde. (Die Fahrzeit beträgt etwa 3 Minuten, die Intervalle dauerten 15 Minuten.)

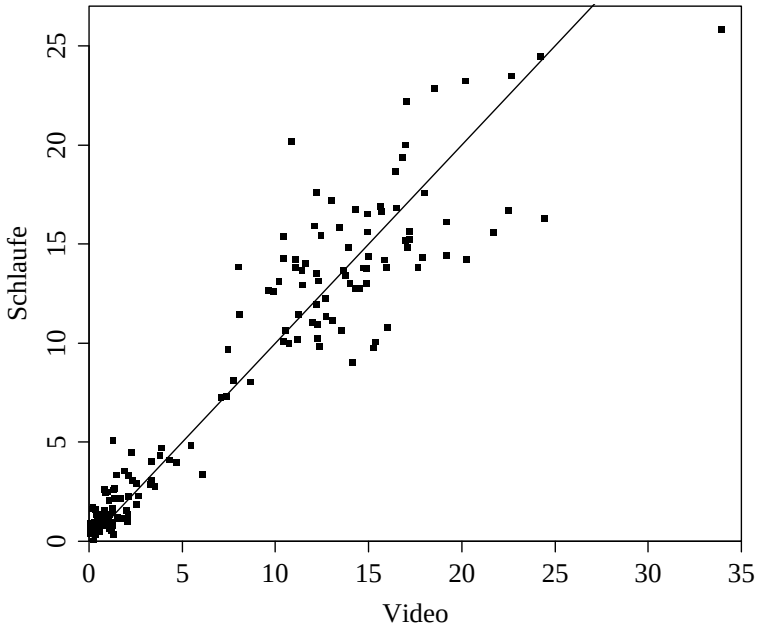

Abbildung 1.1.h: Lastwagen-Anteil (in Prozenten) gemäss Schlaufen- und Videozählung. Die Gerade stellt die Gleichheit ($y = x$) dar.

Es ergibt sich die weit verbreitete Situation, dass der Wert einer interessierenden Grösse auf Grund der Messung einer mit ihr zusammenhängenden anderen Grösse mittels einer Umrechnungsformel ermittelt werden soll. Dabei kann die Messung auf einer ganz anderen Skala erfolgen; beispielsweise wird eine Konzentration mittels einer optischen Durchlässigkeit erfasst.

Man geht zunächst davon aus, dass für einen gegebenen exakten Wert x_i die Messung Y_i sich aus einem „Idealwert“ $h\langle x_i\rangle$ und einem Messfehler E_i zusammensetzt. Das entspricht einem Regressionsmodell. Man bestimmt die Funktion h mittels Messungen Y_i, für die der zugehörige Wert x_i bekannt ist. In der Anwendung wird aber nicht von x auf Y, sondern von einem Messwert Y auf den gesuchten Wert x geschlossen. Aus dieser Umkehrung ergeben sich gewisse zusätzliche Probleme.

Dieses Vorgehen entspricht der **Eichung** eines Messgeräts. Man misst Proben mit bekanntem exaktem Wert (z. B. bekannter Konzentration) und liest die Messung ab. Dann wird die Ablese-Skala ajustiert, was der Schätzung und Verwendung der Funktion h in unserem allgemeineren Zusammenhang entspricht. ◁

i ▷ **Beispiel basische Böden.** In Indien behindern basische Böden, also tiefe Säurewerte oder hohe pH-Werte, Pflanzen beim Wachstum. Es werden daher Baumarten gesucht, die eine hohe Toleranz gegen solche Umweltbedingungen haben. In einem Freilandversuch wurden auf einem Feld mit grossen lokalen Schwankungen des pH-Wertes 120 Bäume einer Art gepflanzt und ihre Höhe Y_i nach 3 Jahren gemessen. Abbildung 1.1.i zeigt die Ergebnisse mit den zugehörigen pH-Werten $x_i^{(1)}$ des Bodens zu Beginn des Versuchs. Zusätzlich wurde eine Variable $x_i^{(2)}$ gemessen, die einen etwas anderen Aspekt der „Basizität" erfasst (der Logarithmus der so genannten sodium absorption ratio, SAR). Dieses Beispiel hat also zwei Eingangsgrössen.

Ein Hauptziel der Untersuchung besteht darin, für gegebene Werte der beiden Eingangsgrössen an einem möglichen Pflanzort bestimmen zu können, wie gut ein solcher Baum dort wohl wachsen wird. Es stellt sich zusätzlich die Frage, ob die Messung der zweiten Grösse $x^{(2)}$ dazu überhaupt etwas beiträgt, oder ob der pH ($x^{(1)}$) allein auch genügt. ◁

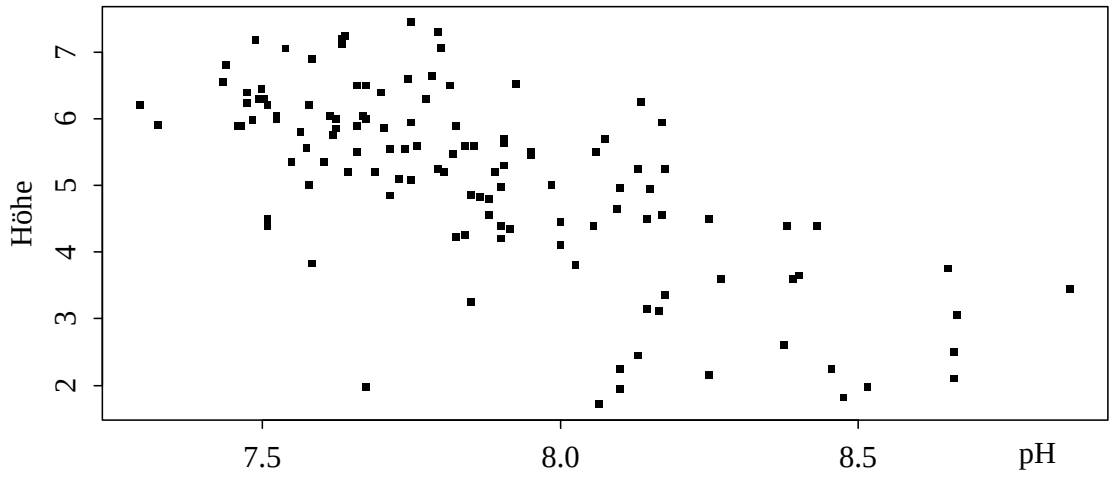

Abbildung 1.1.i: Baumhöhe in Abhängigkeit vom pH für das Beispiel der basischen Böden

j ▷ **Beispiel Antikörper-Produktion**. Grössere Mengen von Antikörpern werden in biotechnologischen Prozessen gewonnen. Dazu werden biotechnologisch veränderte Zellen, die den entsprechenden Antikörper produzieren können, Wirtstieren (z. B. Mäusen) injiziert. Nach einer gewissen Zeit beginnen diese Zellen Antikörper zu produzieren und auszuscheiden. Die ausgeschiedene Flüssigkeit wird dann eingesammelt und weiter verarbeitet. Dieses Beispiel wird ausführlich in Haaland (1989) dargestellt und analysiert. Es dient uns hier nur zur Illustration der Fragestellung.

Die Zellen können erfahrungsgemäss nur Antikörper produzieren, wenn das Immunsystem der Wirtstiere geschwächt wird. Dies kann durch 4 Faktoren geschehen. Es wird zudem vermutet, dass die Menge der injizierten Zellen und deren Entwicklungsstand die Antikörper-Produktion beeinflusst.

Da es für so komplexe biologische Prozesse keine theoretischen Modelle gibt, werden die relevanten Prozessfaktoren durch ein Experiment ermittelt. Ein solches Experiment braucht viele Mäuse, ist zeitaufwändig und kostet Geld. Mit einer geschickten Versuchsanordnung können unter geringstmöglichem Aufwand die wichtigen Prozessfaktoren ermittelt werden. Hier hilft die **statistische Versuchsplanung**. ◁

k ▷ Als relevante Prozessfaktoren wurden in dieser Studie zwei Prozessfaktoren identifiziert, nämlich die Dosis von Co^{60} Gamma-Strahlen und die Anzahl Tage zwischen der Bestrahlung und der Injektion eines reinen Öls (englische Bezeichnung *pristane*). Diese beiden Prozessfaktoren sollen nun so eingestellt werden, dass eine möglichst optimale Menge von Antikörpern durch die veränderten Zellen produziert wird.

Dazu wollen wir ein empirisches Modell $Y_i = h\langle x_i^{(1)}, x_i^{(2)}\rangle + E_i$ finden, das die Ausbeute Y von Antikörpern möglichst gut aus den beiden Prozessfaktoren $x^{(1)}$ und $x^{(2)}$ vorhersagt. Als Funktion h wird oft ein quadratisches Polynom in den Variablen $x^{(1)}$ und $x^{(2)}$ verwendet. Mit dem aus den Daten bestimmten Modell lässt sich dann die optimale Einstellung $[x_o^{(1)}, x_o^{(2)}]$ der Prozessfaktoren bestimmen. ◁

1.2 Fragestellungen

a Von der Problemstellung her können die Anwendungen der Regression in Gruppen eingeteilt werden:
• **Vorhersage, Prognose, Interpolation**. Im Beispiel der Sprengungen soll eine Formel helfen, für gegebene Distanz und Ladung die Erschütterung „vorherzusagen“. Es interessiert nicht nur der mittlere zu erwartende Wert, sondern auch eine obere Grenze, über der die Erschütterung nur mit kleiner Wahrscheinlichkeit liegen wird. (Die Begriffe Vorhersage und Prognose werden meistens für eine zeitliche Extrapolation in die Zukunft verwendet. Hier spielt die Zeit keine Rolle – ausser dass die Problemstellung nur wesentlich ist, wenn die Sprengung noch nicht erfolgt ist.)

b • **Schätzung von Parametern**. Im Beispiel des Gubrist-Tunnels sollen zwei Konstanten, die Emissionsfaktoren für Lastwagen und für übrige Fahrzeuge, bestimmt werden.

c • **Bestimmung von Einflussgrössen**. Im Beispiel der Antikörper-Produktion müssen zunächst aus mehreren in Frage kommenden Eingangsgrössen diejenigen herausgefunden werden, die die Zielvariable wesentlich beeinflussen. In vielen Forschungs-Projekten steht diese Frage ebenfalls im Vordergrund: Von welchen Grössen wird eine Zielgrösse eigentlich beeinflusst?

d • **Optimierung**. Im Beispiel der Antikörper-Produktion sollten optimale Produktionsbedingungen gefunden werden. In allen Bereichen der Produktion ist diese Frage offensichtlich von grundlegender Bedeutung.

e • **Eichung**. Auf Grund der ungenauen und systematisch verfälschten Angabe des Schlaufen-Detektors soll der Anteil der Lastwagen bestimmt werden. Diese Problemstellung kombiniert Elemente der Vorhersage und der Schätzung von Parametern.

f Der Block Regression 1 wird sich vor allem mit den ersten drei Fragen befassen.

1.3 Ausblick

a In der **linearen Regression**, die im Folgenden behandelt wird, setzt man voraus,
• dass die Zielgrösse eine kontinuierliche Variable ist,
• dass die zufälligen Abweichungen E_i einer Normalverteilung folgen und von einander statistisch unabhängig sind
• und dass die Funktion h von einer einfachen Form ist, nämlich in einem gewissen Sinne linear (siehe 3.2.w). Die gleichen Fragestellungen werden auch in der Varianzanalyse 1 behandelt, mit anderen Schwerpunkten bezüglich der Art der Eingangsgrössen.

b Am Ende dieses Blockes und in späteren Blöcken wird dieser Ansatz in vielen Richtungen erweitert:

• Wenn die Funktion h nicht im erwähnten Sinne linear ist, kommt die **nichtlineare Regression** zum Zug.

c • Wenn die Beobachtungen der Zielgrösse und der erklärenden Grössen in einer zeitlichen Abfolge auftreten, entstehen normalerweise besondere Probleme durch entsprechende Korrelationen. Diese Besonderheiten werden in der Theorie der **Zeitreihen** behandelt.

d • Man kann an mehreren Zielgrössen interessiert sein. Eine einfache Art, damit umzugehen, besteht darin, für jede von ihnen eine separate Regressionsrechnung durchzuführen. Die multivariate Statistik zeigt, wie man bei gemeinsamer Betrachtung mit **multivariater Regression und Varianzanalyse** noch etwas darüber hinaus gewinnen kann.

e • Die Annahme der Normalverteilung für die E_i ist oft nur näherungsweise erfüllt. Die Methoden, die wir im Folgenden kennen lernen, sind dann nicht mehr gut geeignet. Besser fährt man mit den Methoden der **robusten Regression**.

f • Die interessierende Zielgrösse kann eine zweiwertige Variable (Ja/Nein) sein. Das führt zur **logistischen Regression**. Ist die Zielvariable eine Zählgrösse, eine diskrete geordnete oder eine nominale Variable, so sind die **verallgemeinerten linearen Modelle** anzuwenden, zu denen auch das gewöhnliche und das logistische Regressionmodell gehören.

g • Zeiten bis zum Ausfall eines Gerätes oder bis zum Eintreffen eines anderen Ereignisses folgen meist anderen Verteilungen als der üblicherweise verwendeten Normalverteilung. Ausserdem werden solche Ereignisse oft nicht für alle Beobachtungseinheiten abgewartet, was zu so genannt zensierten Daten führt. Es gibt auch für solche Daten geeignete Regressionsmethoden, die im Gebiet der **Überlebenszeiten** (*survival* oder *failure time data*) behandelt werden.

h • In der linearen Regression werden nur die Abweichungen E_i als Zufallsvariable modelliert. Manchmal kann es auch sinnvoll sein, die **Parameter** selbst durch **Zufallsgrössen** zu ersetzen. Dies kommt vor allem in einem weiterführenden Gebiet der Varianzanalyse (repeated measures und „Spaltanlagen“, *split plot designs)* zum Zug, wo man von **zufälligen Effekten** spricht.

i • In all diesen Modellen ist die Regressionsfunktion ein Mitglied einer Schar von vorgegebenen Funktionen, die durch einen oder mehrere Parameter charakterisiert ist. Es geht dann darum, diese(n) Parameter zu bestimmen. Was wir intuitiv oft wollen, ist kein in solcher Weise vorgegebener Funktionstyp, sondern einfach eine „glatte Funktion“. Man spricht von „**Glättung**“ der Daten. Wie man eine solche Idee mathematisch formuliert und die entsprechende Funktion schätzt, untersucht die **nichtparametrische Regression**.

j In all diesen Verallgemeinerungen erscheinen immer wieder die gleichen Grundideen, die wir nun an Hand der linearen Regression – zunächst mit einer einzigen erklärenden Variablen, nachher mit mehreren – einführen wollen.

Die folgenden Unterlagen für die einfache Regression enthalten Repetitions-Abschnitte zu den Begriffen der Schliessenden Statistik. Sie sollen den Einstieg vor allem jenen erleichtern, die nicht gerade den entsprechenden Block des Nachdiplomkurses hinter sich haben.

2 Einfache lineare Regression

2.1 Das Modell

a ▷ **Beispiel Sprengungen** (1.1.b). Wir untersuchen zunächst die Abhängigkeit der Erschütterung von der Distanz bei konstanter Ladung. Im Streudiagramm Abbildung 2.1.a sind beide Achsen logarithmisch dargestellt. Die logarithmierte Erschütterung hängt gemäss der Figur ungefähr linear von der logarithmierten Distanz ab; einfacher gesagt, die Punkte in der Figur streuen um eine Gerade. ◁

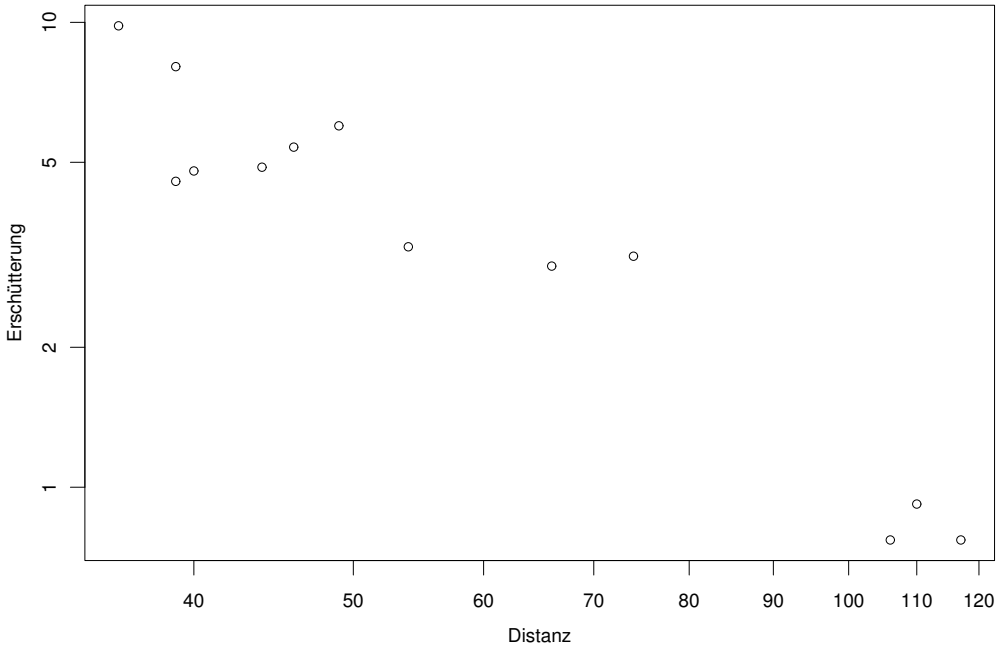

Abbildung 2.1.a: Distanz und Erschütterung bei Sprengungen mit Ladung 3.12. Die Achsen sind logarithmisch dargestellt

b Eine **Gerade** ist wohl die einfachste Funktion, die eine Abhängigkeit ausdrücken kann. Alle Punkte $[x_i, y_i]$ auf einer Geraden folgen der Geradengleichung

$$y_i = \alpha + \beta x_i$$

mit geeigneten Zahlen α und β. Die erste, α, ist der „**Achsenabschnitt**“ und β misst die **Steigung** der Geraden. Da β als Faktor vor der Eingangs-Variablen auftritt, wird es als **(Regressions-) Koeffizient** von X bezeichnet. Wenn $\alpha = 0$ ist, geht die Gerade durch den Nullpunkt.

c Im Beispiel scheinen die *logarithmierten* Daten ungefähr einer Beziehung zu folgen, die sich durch eine Gerade darstellen lässt. Immer wieder wird gefragt, ob denn eine **Transformation** nicht eine unerlaubte „**Daten-Manipulation**“ sei. Hier wird folgende These vertreten:

Daten verlangen keine Gerechtigkeit. Unser Ziel ist es, Zusammenhänge und Strukturen zu erkennen und wenn möglich zu verstehen. Dazu bauen wir Modelle auf, die deterministische, gut interpretierbare Zusammenhänge mit zufälligen Grössen verbinden. Es ist wichtig, dass wir sorgfältig prüfen, wie eng die „Übereinstimmung“ der Modelle mit den Daten ist. Ob die Modelle aber für Rohdaten oder für daraus abgeleitete Grössen formuliert sind, ist keine Frage der wissenschaftlichen Redlichkeit, sondern höchstens eine der einfachen **Interpretierbarkeit**.

Im Beispiel werden wohl wenige dagegen Einspruch erheben, dass für die grafische Darstellung logarithmisch geteilte Achsen verwendet werden. Dem entspricht, wie erwähnt, das Rechnen und Modellieren mit logarithmisch transformierten Daten und Zufallsgrössen.

d In vielen Anwendungen gibt es fachliche Theorien, die einen linearen Zusammenhang zwischen logarithmierten Grössen beinhalten. Im Beispiel ist anzunehmen, dass die Erschütterung proportional zur Ladung und umgekehrt proportional zur quadrierten Distanz sein sollten, also

$$\begin{aligned} \text{Erschütterung} &\approx \text{const} \cdot \text{Ladung}/(\text{Distanz})^2 \qquad \text{oder} \\ \log(\text{Erschütterung}) &\approx \log(\text{const}) + \log(\text{Ladung}) - 2 \cdot \log(\text{Distanz}) \; . \end{aligned}$$

Für die logarithmierten Grössen lässt sich also ein linearer Zusammenhang herleiten. Da die Ladung hier konstant gehalten wurde, müssten die Punkte [log(Distanz), log(Erschütterung)] idealerweise auf einer Geraden liegen.

Gemäss Modell wäre die Steigung schon bekannt – ein seltener Fall. Wir wollen davon ausgehen, dass die logarithmierten Grössen etwa linear zusammenhängen, aber die Steigung der Geraden zunächst nicht festlegen.

e Als nächstes werden Sie wohl eine Gerade in das Streudiagramm legen wollen. Das ist eine Aufgabe der zusammenfassenden Beschreibung, also der Beschreibenden Statistik. Die bekannteste Regel, wie die zu den Daten passende Gerade zu bestimmen sei, heisst „Kleinste Quadrate“. Wir werden sie bald einführen (2.2.c); das Resultat für das Beispiel zeigt Abbildung 2.2.a.

Wenn die Daten als „die Wahrheit“ gelten, dann ist dies „die richtige“ Gerade. Allen ist aber klar, dass die Daten auch anders hätten herauskommen können – dass der Zufall mitgespielt hat. Mit anderen Daten wäre auch die Gerade nicht die selbe. Die erhaltene Gerade ist also zufällig, ungenau. Wie sollen wir den Zufall, die Ungenauigkeit erfassen?

Die Antwort auf diese Frage gibt die Schliessende oder Analytische Statistik, die auf der Wahrscheinlichkeitsrechnung beruht. Um sie zu verstehen, müssen wir zunächst eine Modellvorstellung entwickeln, die sagt, welche anderen Datensätze „ebenso gut“ möglich gewesen wären wie der in Abbildung 2.1.a festgehaltene. Wir vergessen dazu zunächst diese Daten und überlegen uns ein **Wahrscheinlichkeitsmodell**, das die gegebene Situation beschreibt.

f Zunächst überlegen wir, wie ein Wert Y_i der Zielgrösse aussehen wird, der zur Eingangsgrösse x_i gemessen wird – im Beispiel, wie gross wohl die logarithmierte Erschütterung ist, wenn die logarithmierte Distanz zum Sprengort $x_i = \log_{10}\langle 50 \rangle$ beträgt. Gemäss dem bisher Gesagten ist dies gleich dem Funktionswert $\alpha + \beta x_i$, bis auf eine Abweichung E_i, die wir jetzt als Zufallsvariable betrachten,

$$Y_i = \alpha + \beta x_i + E_i \; .$$

Wir nehmen an, dass die Abweichungen E_i, $i = 1, \ldots, n$, eine bestimmte Verteilung haben – alle die gleiche – und stochastisch unabhängig (insbesondere unkorreliert) seien. Sie bilden also eine Zufalls-Stichprobe. Es zeigt sich, dass die Annahme einer Normalverteilung zu den mathematisch einfachsten Resultaten führt. Die Normalverteilung soll Erwartungswert 0 und Varianz σ^2 haben. Wir notieren das als $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$.

g Das Modell wird erst dann konkret, wenn wir die drei Zahlen α, β und σ festlegen. Diese Situation ist in der Wahrscheinlichkeitsrechnung und in der Statistik üblich: Es wird ein Modell zunächst nur bis auf ein paar Konstante festgelegt. Diese Konstanten nennt man **Parameter** der Verteilung. Die „Normalverteilung“ ist eigentlich keine Verteilung, sondern eine **Verteilungs-Familie**; erst wenn Erwartungswert und Varianz festgelegt sind, entsteht daraus *eine* Verteilung.

In vielen Anwendungsgebieten wird das Wort Parameter für eine gemessene Grösse verwendet – was in der Statistik als Variable bezeichnet wird. Ein anderes Wort dafür ist Merkmal. Wir hoffen auf Ihr Verständnis für diese Sprachkonfusion.

h Eine Modell-Vorstellung entsteht in unseren Köpfen. Wir wollen auch gleich noch die Parameter „erfinden“. Abbildung 2.1.h veranschaulicht das Modell der linearen Regression mit den Parameter-Werten $\alpha = 4$, $\beta = -2$ und $\sigma = 0.1$. Die Wahrscheinlichkeiten, mit denen bestimmte Werte für die Y-Variable erwartet werden, sind mit den Wahrscheinlichkeitsdichten dargestellt.

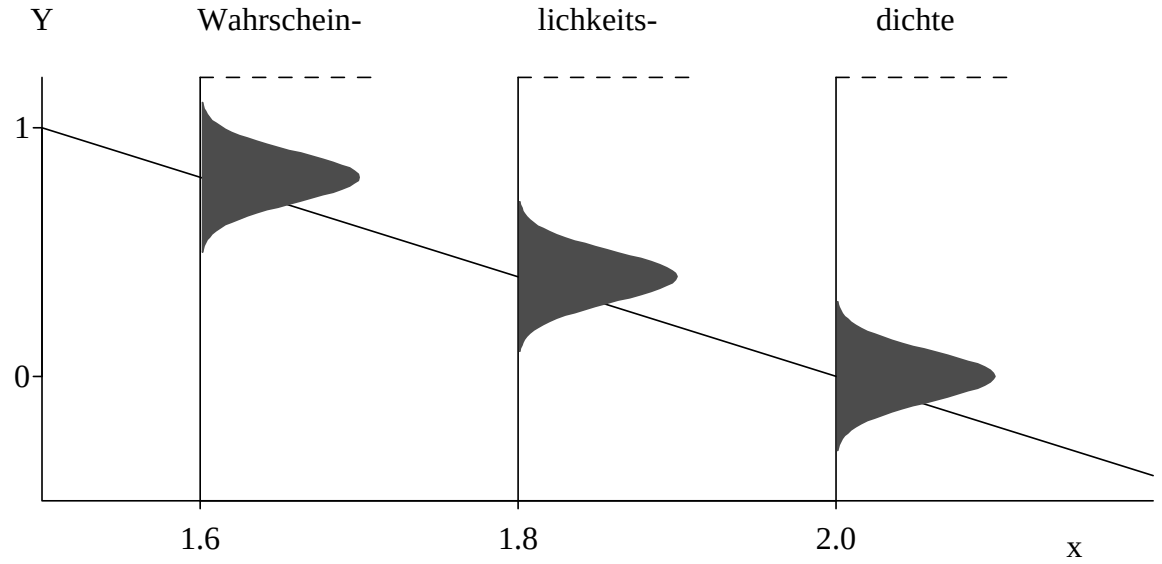

Abbildung 2.1.h: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ für drei Beobachtungen Y_1, Y_2 und Y_3 zu den x-Werten $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$

i Als zweite Veranschaulichung wollen wir **Zufallszahlen** gemäss unserm Modell ziehen und darstellen, also Beobachtungen, die dem Modell entsprechen, **simulieren**. Drei standard-normalverteilte Zufallszahlen, die mit $\sigma = 0.1$ multipliziert werden, bilden ein mögliches Ergebnis für die drei zufälligen Abweichungen E_1, E_2 und E_3. Ein Zufallszahl-Generator lieferte die vier Dreiergruppen

$$\begin{aligned} -0.419,\ -1.536,\ -0.671\ ; \qquad & 0.253,\ -0.587,\ -0.065\ ; \\ 1.287,\ 1.623,\ -1.442\ ; \qquad & -0.417,\ 1.427,\ 0.897\ . \end{aligned}$$

Wenn $4 - 2x_i$ mit $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$ dazugezählt werden, erhält man je die entsprechenden Werte für Y_1, Y_2 und Y_3. In Abbildung 2.1.i sind die so „simulierten“ Ergebnisse dargestellt.

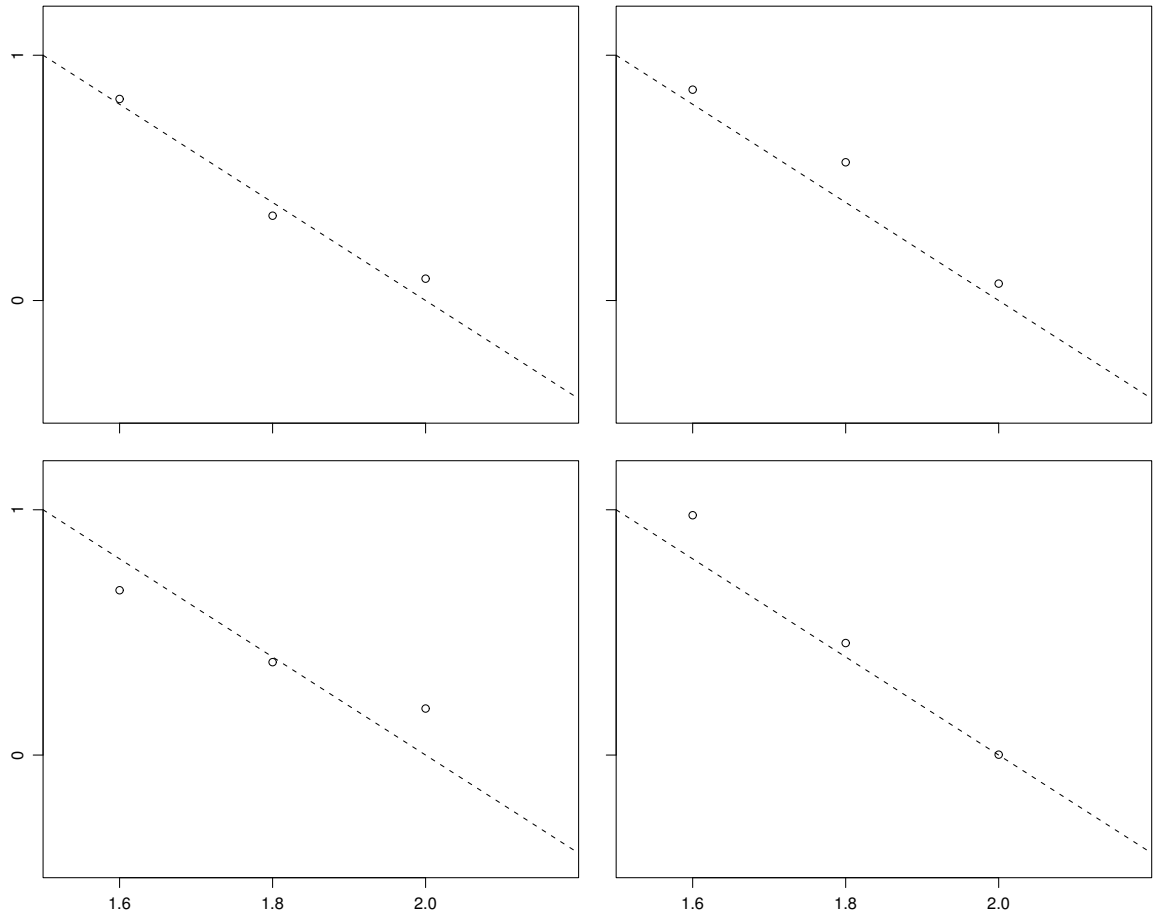

Abbildung 2.1.i: Vier simulierte Ergebnisse für drei Messungen gemäss dem Modell $Y_i = 4 - 2x_i + E_i$. Die gestrichelten Geraden stellen den hier bekannten „wahren“ Zusammenhang $y = 4 - 2x$ dar.

2.2 Schätzung der Parameter

a ▷ Kehren wir zu konkreten Daten zurück! Abbildung 2.2.a zeigt die Daten des **Beispiels der Sprengungen** mit einer Geraden, die zu den Daten passt. Sie legt die Parameter α und β des Regressionsmodells fest. ◁

b Um allgemein den Daten ein best-passendes Modell zuzuordnen, müssen die Parameter mit geeigneten Regeln festgelegt werden. Die Funktionen, die den Daten die best-passenden Werte zuordnen, heissen **Schätzfunktionen** oder **Schätzungen**.

c Es gibt einige allgemeine Prinzipien, nach denen solche Regeln aufgestellt werden können. Das berühmteste für unseren Fall ist das Prinzip der **Kleinsten Quadrate**. Darin werden die Parameter so bestimmt, dass die Summe der quadrierten Abweichungen

$$\sum_{i=1}^{n} r_i^2 \,, \qquad r_i = y_i - (\alpha + \beta x_i)$$

minimal wird. Wenn die Fehler E_i normalverteilt sind, dann kann dieses Kriterium aus dem Prinzip der Maximalen Likelihood hergeleitet werden.

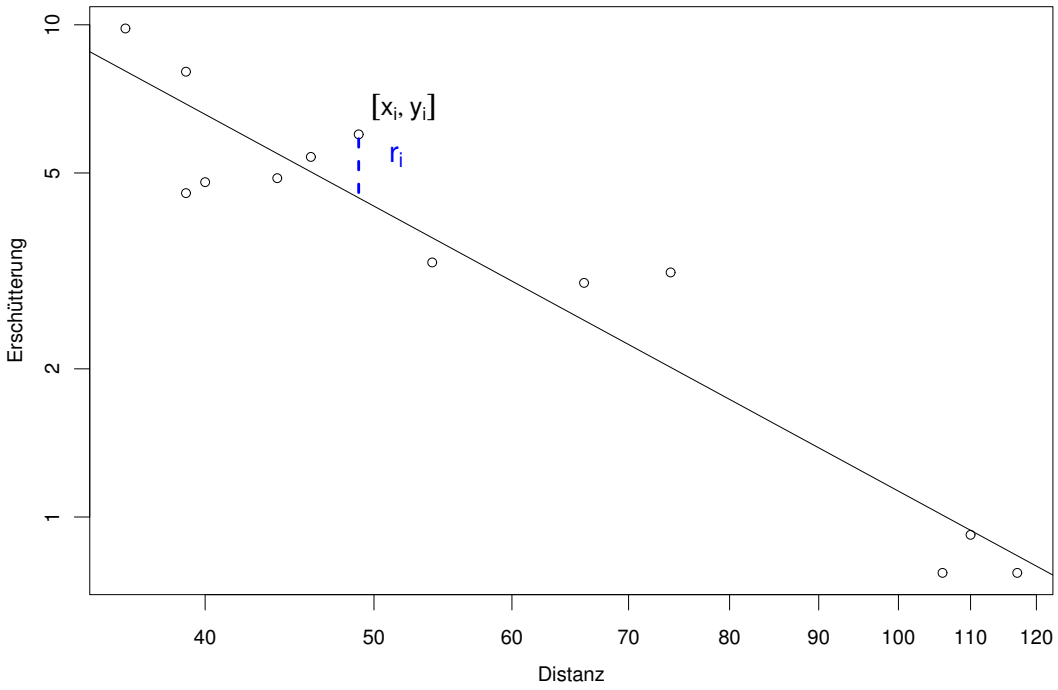

Abbildung 2.2.a: Geschätzte Gerade für das Beispiel der Sprengungen

Die Schätzfunktionen lauten dann

$$\begin{aligned} \widehat{\beta} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \widehat{\alpha} &= \overline{Y} - \widehat{\beta}\bar{x}\,. \end{aligned}$$

Weitere Details sind im Anhang 2.A beschrieben.

Es gibt in unserem Modell einen weiteren Parameter, die Varianz σ^2 der zufälligen Abweichungen. Diese Grösse muss ebenfalls aus den Daten geschätzt werden. Man braucht sie allerdings nicht, um die best-passende Gerade zu bestimmen. Wir stellen das Thema deshalb zurück (2.2.n).

d* Eine best-passende Gerade würde anschaulich eher so bestimmt, dass die Abstände der Punkte von der Geraden, senkrecht zur Geraden gemessen, möglichst klein würden. Man nennt die Methode, die die Quadratsumme dieser Abstände minimiert, **orthogonale Regression**. Das Modell, das wir in 2.1.f formuliert haben, sagt aber, der „Idealpunkt“ $[x_i, \alpha + \beta x_i]$ auf der Geraden werde durch die zufälligen Abweichungen E_i in Y-Richtung verschoben, nicht senkrecht zur Geraden. – Im Zusammenhang mit einem anderen Modell für die Wirkung des Zufalls ist die orthogonale Regression in der Tat die angebrachte Methode, vergleiche 6.1.j.

e Eine Schätzung ist eine Funktion, die den n Beobachtungen *eine* Zahl und damit den n Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$, die wir als Modell für die Daten benützen, *eine* Zufallsvariable zuordnet. Also sind **Schätzungen** selbst auch **Zufallsvariable**. Üblicherweise werden sie mit einem Hut über dem zu schätzenden Parameter bezeichnet, z. B. $\widehat{\alpha}$, $\widehat{\beta}$.

Zufallsvariable streuen. Dies kann in Abbildung 2.2.e beobachtet werden. In dieser Abbildung wurden jeweils die zu den Punkten aus Abbildung 2.1.i am besten passenden Geraden eingezeichnet. Die geschätzten Geraden und damit die entsprechenden geschätzten Parameter streuen um die „wahre“ Gerade respektive um die „wahren“ Parameter.

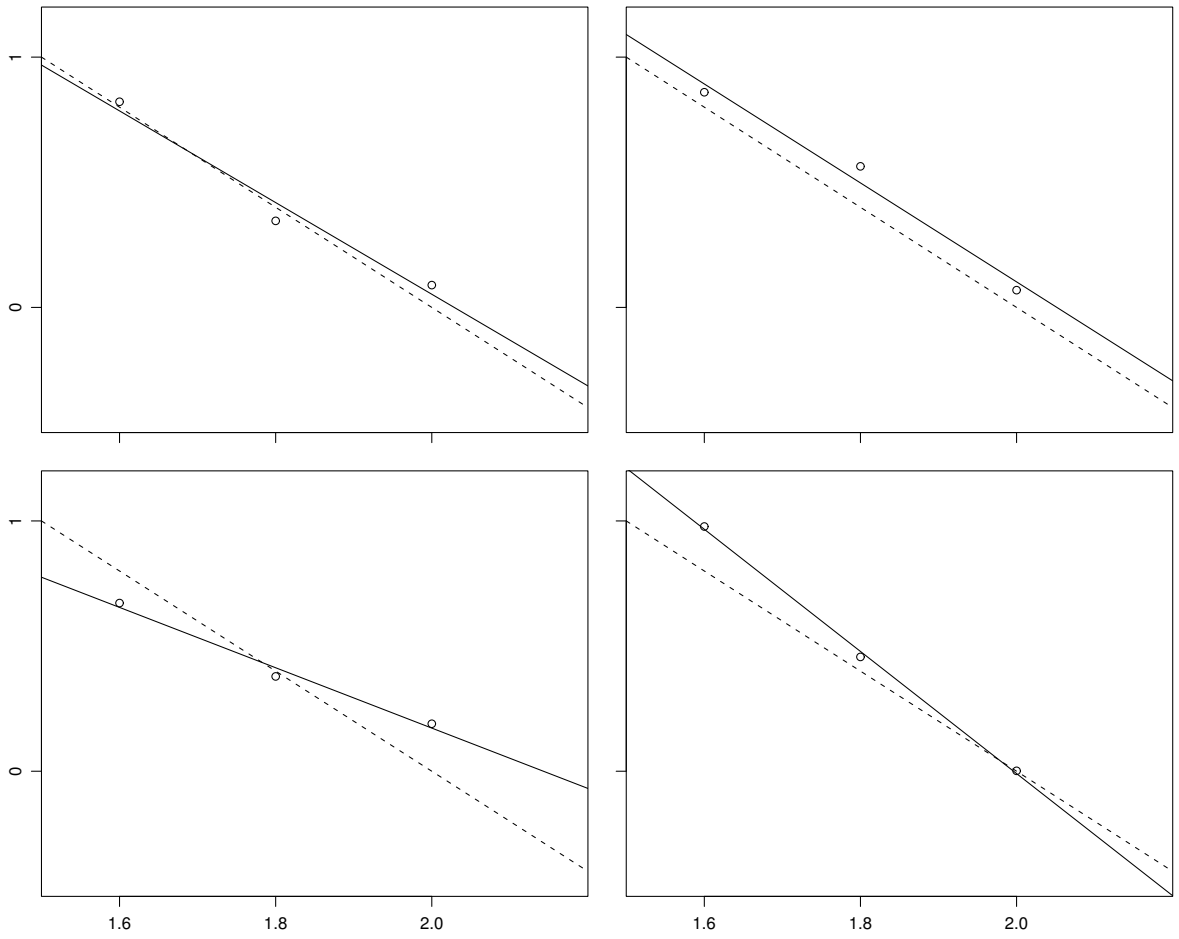

Abbildung 2.2.e: Vier simulierte Ergebnisse für drei Messungen mit den geschätzten (ausgezogenen) Geraden

f Da Schätzungen Zufallsvariable sind, können wir **Eigenschaften von Schätzungen** mit Hilfe des Wahrscheinlichkeitsmodells studieren. Dazu vergessen wir wieder für einen Moment die konkreten Daten. Wir nehmen jetzt an, wir kennen das Modell für die Beobachtungen genau, die Werte der Parameter eingeschlossen. Überlegen wir uns, was ein armer Forscher, der die Parameter α und β nicht kennt, als Schätzwerte erhalten könnte und welche Wahrscheinlichkeiten diese Werte haben würden – kurz, wie die **Verteilung der Schätzfunktion** aussieht.

g Diese Verteilung kann mit Hilfe der Wahrscheinlichkeitstheorie bestimmt werden. Anschaulicher ist es, wenn wir **Modell-Experimente** betrachten. Dazu werden Zufallszahlen gemäss dem Modell gezogen analog dem Beispiel in Abbildung 2.2.e. Dann werden die Parameter für diese **simulierten Beobachtungen** geschätzt. Dieses Vorgehen wird nun m mal wiederholt, und wir erhalten daraus m Schätzwerte für die Parameter α und β. In Abbildung 2.2.g sind 1000 Schätzwerte der Steigung β in einem Histogramm zusammengefasst.

h Wie gesagt, die Verteilungen der Schätzungen lassen sich mit Hilfe der Wahrscheinlichkeitsrechnung direkt aus den Annahmen über die Verteilung der Messfehler bestimmen. Wir haben angenommen, dass diese unabhängig und normalverteilt sind. Daraus folgt nun, dass die Kleinste-Quadrate-Schätzungen $\widehat{\alpha}$ und $\widehat{\beta}$ ebenfalls normalverteilt sind, nämlich

$$\widehat{\beta} \sim \mathcal{N}\left\langle \beta, \sigma^{(\beta)2} \right\rangle \text{ und } \widehat{\alpha} \sim \mathcal{N}\left\langle \alpha, \sigma^{(\alpha)2} \right\rangle ,$$

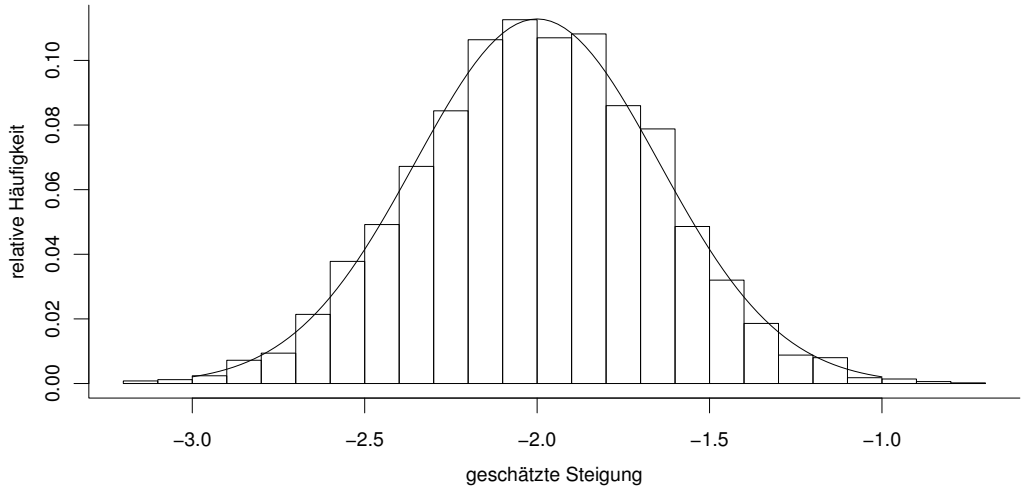

Abbildung 2.2.g: Simulierte und theoretische Verteilung der Schätzung $\widehat{\beta}$ der Steigung

wobei $\sigma^{(\beta)}$, $\sigma^{(\alpha)}$ und die so genannte Quadratsumme $\mathrm{SSQ}^{(X)}$ der x-Werte definiert sind als

$$\sigma^{(\beta)2} = \sigma^2/\mathrm{SSQ}^{(X)} \qquad \sigma^{(\alpha)2} = \sigma^2\left(\tfrac{1}{n} + \bar{x}^2/\mathrm{SSQ}^{(X)}\right)$$

$$\mathrm{SSQ}^{(X)} \;=\; \sum_{i=1}^{n}(x_i - \bar{x})^2 \,.$$

Für mathematisch Interessierte ist die Herleitung im Anhang 2.B beschrieben.

i* Die Methode der Kleinsten Quadrate ist zwar die bekannteste Schätzmethode für die Parameter, aber nicht die einzige. Man könnte auch den Punkt mit dem kleinsten und den mit dem grössten x-Wert miteinander verbinden und erhielte auch eine Gerade – meist gar nicht eine allzu schlechte. Es würde wohl kaum jemand diese Regel, eine Gerade an Daten anzupassen, ernsthaft zum allgemeinen Gebrauch empfehlen. Wieso nicht? Diese Frage kann solide beantwortet werden, wenn man die Verteilung von verschiedenen Schätzfunktionen für den gleichen Parameter miteinander vergleicht.

j* Die oben genannten Ergebnisse sagen unter anderem, dass der Erwartungswert der Schätzung $\widehat{\beta}$ der Steigung gleich dem „wahren“ Wert der Steigung β sei, und Analoges gilt für den Achsenabschnitt. Man nennt diese Eigenschaft **Erwartungstreue**. Das ist sicher eine nützliche Eigenschaft: Wenn die Schätzung schon notwendigerweise streuen muss, dann hoffentlich wenigstens um den Wert, den sie schätzen sollte.

(Wenn dies für eine Schätzung nicht gilt, so spricht man von einem **Bias**, definiert als Differenz zwischen dem Erwartungswert der Schätzung $\widehat{\theta}$ und dem vorgegebenen Parameterwert θ.)

k* Eine Schätzung streut, wie gesagt, notwendigerweise. Es ist natürlich anzustreben, dass sie möglichst wenig streut. Das kann man mit der **Varianz der Schätzung** messen – für $\widehat{\beta}$ haben wir $\mathrm{var}\langle\widehat{\beta}\rangle = \sigma^2/\mathrm{SSQ}^{(X)}$ angegeben. (Wenn eine Schätzung $\widehat{\theta}$ nicht erwartungstreu ist, ist der **Mittlere Quadratische Fehler**, englisch *mean squared error*, $\mathrm{MSE} = \mathcal{E}\langle(\widehat{\theta} - \theta)^2\rangle$ ein geeigneteres Mass.)

Je grösser die Varianz (oder der MSE), desto schlechter die Schätzung. Um zwei Schätzungen zu vergleichen, wählt man das umgekehrte Verhältnis der Varianzen und definiert es als die **relative Effizienz** der Schätzungen. Die (absolute) Effizienz einer Schätzung ist ihre relative Effizienz verglichen mit der „besten“ Schätzung, also mit jener mit der kleinsten Varianz. Es zeigt sich, dass die Kleinsten Quadrate unter den hier gemachten Voraussetzungen zu solchen besten Schätzungen führen.

l* Wieso denn so viele Begriffe? Wenn doch die besten Schätzungen so einfach zu bestimmen sind, kann man doch alle anderen sowieso vergessen! Das werden wir auch ziemlich lange tun. Später werden wir uns daran erinnern, dass all diese Theorie auf der Annahme beruht, dass die Zufallsfehler normalverteilt seien. Wenn dies nicht stimmt, dann sind die genannten Schätzungen nicht mehr die besten – so genannte **robuste** Schätzungen sind dann besser. Vorläufig aber gilt:

m Die **Kleinste-Quadrate-Schätzungen** $\widehat{\alpha}$ und $\widehat{\beta}$ sind

- erwartungstreu und normalverteilt mit den oben angegebenen Varianzen und
- die besten Schätzungen,

sofern die Zufallsfehler unabhängig sind und alle die gleiche Normalverteilung $\mathcal{N}\langle 0, \sigma^2 \rangle$ haben.

n Bis jetzt haben wir uns ausschliesslich mit den beiden Parametern, welche die Gerade bestimmen, beschäftigt. Nun kümmern wir uns noch um den Parameter $\sigma^2 = \mathrm{var}\langle E_i \rangle$, der die **Varianz der Fehlerverteilung** festlegt. Die „zufälligen Fehler" E_i können weder direkt beobachtet noch aus $E_i = Y_i - (\alpha + \beta x_i)$ hergeleitet werden, da α und β unbekannt sind; sonst könnte man deren empirische Varianz berechnen. Bekannt sind wenigstens, als „Näherungswerte" für die E_i, die so genannten **Residuen**

$$R_i = Y_i - (\widehat{\alpha} + \widehat{\beta} x_i) \,,$$

die Differenzen zwischen den Beobachtungen Y_i und den **angepassten Werten** $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$ (englisch *fitted values*). Deren empirische Varianz ist $\frac{1}{n-1} \sum_{i=1}^{n} (R_i - \overline{R})^2$. Der Nenner $n-1$ in der Definition der empirischen Varianz wurde eingeführt, um sie im Falle einer einfachen Stichprobe erwartungstreu zu machen. Rechnungen zeigen, dass wir im vorliegenden Fall der einfachen Regression durch $n-2$ teilen müssen, um dies zu erreichen. Da immer $\overline{R} = 0$ gilt, ist

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} R_i^2$$

die gebräuchliche, erwartungstreue Schätzung von σ^2.

o* Ein Vielfaches der geschätzten Varianz, $(n-2)\widehat{\sigma}^2/\sigma^2$, ist chi-quadrat-verteilt mit $n-2$ Freiheitsgraden und unabhängig von $\widehat{\alpha}$ und $\widehat{\beta}$. Auf eine Herleitung wollen wir verzichten.

2.3 Tests und Vertrauensintervalle

a Im letzten Abschnitt haben wir uns damit beschäftigt, wie man die Parameter des Modells aus den Daten bestimmen kann. Eine nahe liegende Frage kann nun sein, ob die Daten mit einem Modell mit (teilweise) vorgegebenen Parametern verträglich ist – im Beispiel, ob die Steigung der Geraden wirklich gleich -2 sein kann (vergleiche 2.1.d).

Obwohl die geschätzte Steigung $\widehat{\beta} = -1.92$ ist, könnte dies zutreffen, da ja die Schätzung eine Zufallsvariable ist und demnach vom „wahren Wert" $\beta = -2$ abweichen wird. Wir können also nicht zwingend schliessen, dass die beobachteten Werte dem vorgegebenen Modell widersprechen. Die Frage ist, ob der geschätzte Wert $\widehat{\beta} = -1.92$ bloss auf Grund des Zufalls vom postulierten Wert $\beta_0 = -2$ verschieden ist, oder ob die Abweichung so gross ist, dass wir *das Modell mit* $\beta_0 = -2$ *als nicht zutreffend ablehnen müssen.* Diese Frage wird mit einem **statistischen Test** beantwortet.

Allgemeiner kann man fragen, welche Parameterwerte auf Grund der Daten als plausibel erscheinen. Diese Frage führt auf die so genannten **Vertrauensintervalle**.

Hier geben wir stichwortartig das Vorgehen zur Beantwortung dieser Fragen an.

b Der statistische **Test** soll die Nullhypothese

$$H_0: \quad \beta = \beta_0 = -2$$

prüfen. Die vollständige Nullhypothese lautet: Die Beobachtungen folgen dem Modell der einfachen linearen Regression mit $\beta = -2$ und beliebigem α und σ.

Als **Alternative** H_A zieht man in Betracht, dass $\beta \neq -2$ sei, während die anderen Annahmen (Fehlerverteilung, Unabhängigkeit) der Nullhypothese weiterhin gelten. Die Alternative $\beta \neq -2$ umfasst also die Modelle mit allen Parameterwerten ausser dem Wert β_0, der durch die Nullhypothese festgelegt ist; es sind die Parameterwerte auf beiden Seiten des Wertes β_0 durch die Alternative abgedeckt. Diese heisst daher **zweiseitige Alternative**.

In gewissen Anwendungen ist man bloss an Alternativen auf einer Seite interessiert – beispielsweise, wenn Abweichungen auf die eine Seite sowieso nicht auftreten können. Dann zieht man nur die entsprechende **einseitige Alternative** – hier $\beta > -2$ (oder $\beta < -2$) – in Betracht. Als Nullhypothese prüft man dann nicht nur den Grenzfall, sondern auch die andere Seite – hier $\beta \leq -2$ (oder $\beta \geq -2$).

Als **Teststatistik** eignet sich (wie üblich) eine standardisierte Form der Differenz zwischen Schätzung und postuliertem Wert des Parameters,

$$T = \frac{\widehat{\beta} - \beta_0}{\mathrm{se}^{(\beta)}}, \quad \mathrm{se}^{(\beta)} = \sqrt{\widehat{\sigma}^2 / \mathrm{SSQ}^{(X)}}\,.$$

Die Grösse $\mathrm{se}^{(\beta)}$ entspricht $\sigma^{(\beta)}$ von 2.2.h; da der Parameter σ in jener Formel nicht als bekannt angenommen werden kann, wird er durch seine Schätzung $\widehat{\sigma}$ ersetzt. $\mathrm{se}^{(\beta)}$ (manchmal auch $\sigma^{(\beta)}$) wird **Standardfehler** genannt.

Die Teststatistik T hat, falls das Modell der Nullhypothese gilt, eine so genannte t-Verteilung mit $n-2$ Freiheitsgraden. Dies ist der „**t-Test**“ für den Koeffizienten β.

c **P-Wert**. Der P-Wert ist ein standardisiertes Mass dafür, „wie typisch“ ein Wert der Teststatistik ist oder wie gut die Daten mit dem Modell der Nullhypothese übereinstimmen. Man braucht dazu die kumulative Verteilungsfunktion $F^{(T)}$ der Teststatistik, die der Nullhypothese entspricht. Abbildung 2.3.c veranschaulicht die Rechnung für den Fall eines zweiseitigen Tests. (Der Anschaulichkeit halber wurde $\widehat{\beta}$ als Teststatistik verwendet. Das wäre sinnvoll, wenn man σ kennen würde.)

Der P-Wert ist, anschaulich gesprochen, die Fläche unter der Dichtekurve für den Bereich von Werten der Teststatistik, die „extremer“ sind als der beobachtete Wert. Er misst also die Wahrscheinlichkeit, extremere Werte der Teststatistik als den beobachteten zu erhalten, falls die Nullhypothese stimmt. (Im Falle von diskreten Teststatistiken muss „extremer“ durch „mindestens so extrem“ ersetzt werden.) Wenn er klein genug ist, dann sagt man, „die Daten weichen signifikant von der Nullhypothese ab“, oder, falls $\beta_0 = 0$ getestet wird, der Einfluss der Eingangsgrösse auf die Zielgrösse ist „statistisch gesichert“ oder Ähnliches. „Klein genug“ heisst nach üblicher *Konvention* kleiner als 0.05.

Die gewählte Grenze von 0.05=5% wird **Niveau** des Tests genannt. Sie ist gleich der Wahrscheinlichkeit eines Fehlers „erster Art“, der darin besteht, die Nullhypothese zu verwerfen, falls sie gilt. Falls Sie diesen Begriff noch nicht kennen, ist wohl eine Erklärung nützlich: Wahrscheinlichkeiten gibt es nur unter der Annahme eines bestimmten Modells für die Beobachtungen. Wir setzen dafür die Annahmen der Nullhypothese ein und berechnen dann die Wahrscheinlichkeit, dass die Test-Entscheidung „signifikante Abweichung von der Nullhypothese“ lautet, was unter der gemachten Annahme eine Fehlentscheidung ist. Das ist der Fall, wenn der P-Wert unter 5%

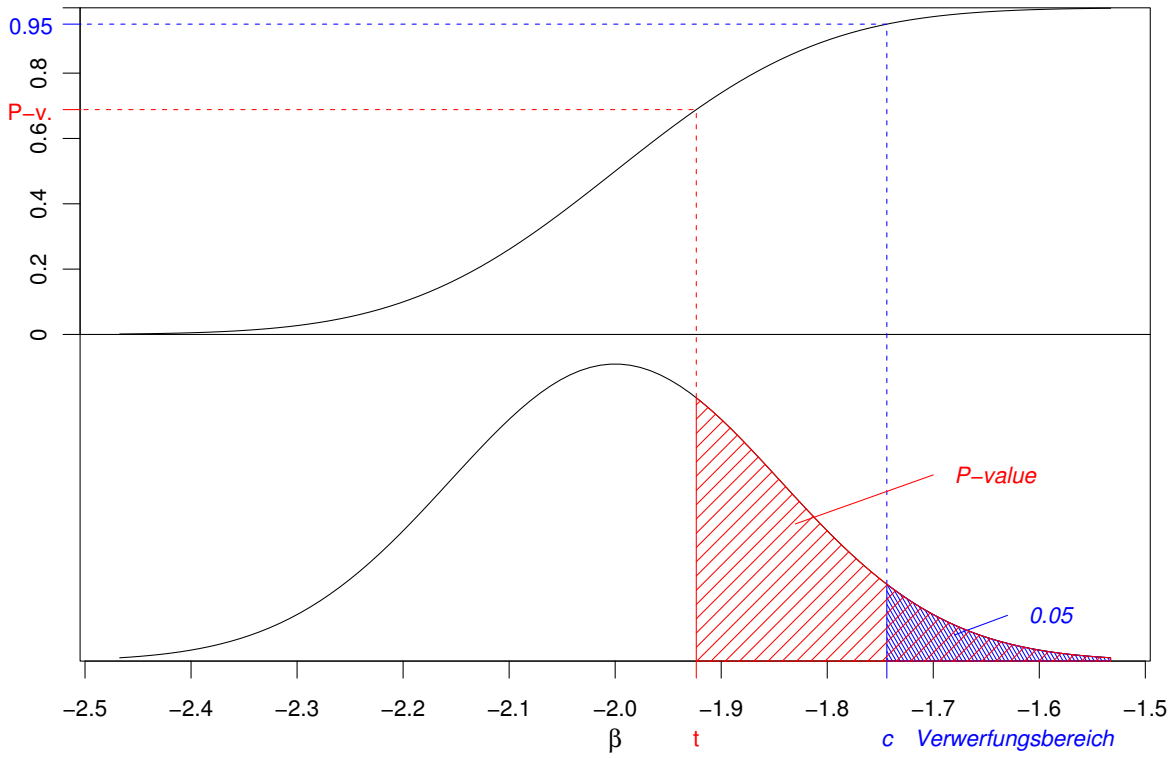

Abbildung 2.3.c: Veranschaulichung des P-Wertes und des Verwerfungsbereiches für einen zweiseitigen Test. Die obere Kurve stellt die kumulative Verteilungsfunktion, die untere die Dichte der Verteilung der Teststatistik dar.

liegt. Die Grösse „P-Wert“ ist gerade so konstruiert, dass für die Entscheidungsregel „signifikant falls P-Wert ≤ 0.05“ die obige Wahrscheinlichkeit 5% beträgt. Gleiches gilt natürlich auch für andere Niveaus; der P-Wert erlaubt es, für beliebige Niveaus die Entscheidung über signifikante Abweichung von der Nullhypothese sofort abzulesen. (Genaueres zum Thema siehe Stahel, 2000, Kap. 8.7).

d Statt einer Schranke für den P-Wert kann man eine entsprechenden Schranke c für die Teststatistik angeben. Das erspart die Umrechnung der Teststatistik in den P-Wert und war deshalb früher üblich. Die Schranke erhält man aus Tabellen. Für die t-Verteilung wie für die F-Verteilung, die wir später noch antreffen werden, sind solche Tabellen verbreitet und entsprechende Funktionen sind in Computer-Umgebungen verfügbar. Der P-Wert, der von Statistik-Programmen ebenfalls angegeben wird, kann aber, wie gesagt, ohne Tabellen beurteilt werden und ist deshalb handlicher.

e ▷ Einen **Computer-Output** für das Beispiel der Sprengungen zeigt Tabelle 2.3.e. Für den Test der Nullhypothese $\beta = 0$ (und für $\alpha = 0$) sind der Wert der Teststatistik $T = T^{(\beta)}$ (und die analog gebildete Teststatistik $T^{(\alpha)}$) und der zugehörige P-Wert angegeben. Die Teststatistiken sind unter der Nullhypothese t-verteilt; wir prüfen also die Steigung und den Achsenabschnitt mit einem **t-Test**.

Regression Analysis - Linear model: Y = a+bX

Dependent variable: log10(ersch)			Independent variable: log10(dist)	
Parameter	Estimate	Standard Error	T Value	(*P-Wert*) Prob. Level
Intercept	$\widehat{\alpha} = 3.8996$	$\mathrm{se}^{(\alpha)} = 0.3156$	$T^{(\alpha)} = 12.36$	0
Slope	$\widehat{\beta} = -1.9235$	$\mathrm{se}^{(\beta)} = 0.1783$	$T^{(\beta)} = -10.79$	0

R-squared = $0.9136 = r^2_{XY}$
Std.dev. of Error = $\widehat{\sigma} = 0.1145$ on $n - 2 = 11$ degrees of freedom
F-statistic: 116.4 on 1 and 11 degrees of freedom, the p-value is 3.448e-07

Tabelle 2.3.e: Computer-Output für das Beispiel der Sprengungen

f ▷ Für die Nullhypothese $\beta = \beta_0 = -2$ erhält man $T = (\widehat{\beta} - \beta_0)/\,\mathrm{se}^{(\beta)} = (-1.92-(-2))/0.1783 = 0.429$. Die kritische Grenze c für die t-Verteilung mit 11 Freiheitsgraden ist gemäss einer Tabelle 2.201. Also ist die Abweichung bei weitem nicht signifikant. Das kann man auch feststellen, wenn man den Rechner den P-Wert bestimmen lässt. Er beträgt 0.676, ist also viel höher als 0.05. ◁

g Nun zur Frage, welche Parameterwerte auf Grund der Daten plausibel erscheinen.

Das Vertrauensintervall umfasst alle Parameterwerte, die auf Grund eines bestimmten statistischen Tests nicht abgelehnt werden. Jedes Vertrauensintervall entspricht also einer bestimmten Test-Regel.
Für die Steigung in der einfachen linearen Regression ergibt sich das Intervall

$$\widehat{\beta} - q\ \mathrm{se}^{(\beta)} \le \beta \le \widehat{\beta} + q\ \mathrm{se}^{(\beta)}$$

wobei $q = q^{t_{n-2}}_{0.975}$ das 0.975-Quantil der genannten t-Verteilung ist. Man schreibt dies oft als

$$\widehat{\beta} \pm q\ \mathrm{se}^{(\beta)}\,, \quad \mathrm{se}^{(\beta)} = \widehat{\sigma} \Big/ \sqrt{\mathrm{SSQ}^{(X)}}\ .$$

h ▷ Im Output (Tabelle 2.3.e) findet man die nötigen Angaben für das Vertrauensintervall von β: Man erhält $-1.9235 \pm 2.201 \cdot 0.1783 = -1.9235 \pm 0.3924$, also das Intervall von -2.32 bis -1.53. (Gute Programme liefern das Vertrauensintervall direkt.) Der Wert -2 liegt klar in diesem Intervall, was nochmals zeigt, dass das Modell mit Steigung -2 sehr gut mit den Daten verträglich ist. ◁

i Damit haben wir die **drei Grundfragen** der parametrischen Statistik behandelt:

1. Welcher Wert ist für den (respektive jeden) Parameter am plausibelsten? Die Antwort wird durch eine **Schätzung** gegeben.
2. Ist ein bestimmter Wert plausibel? Die Entscheidung trifft man mit einem **Test**.
3. Welche Werte sind insgesamt plausibel? Als Antwort erhält man eine ganze Menge plausibler Werte, die meistens ein Intervall bilden – das **Vertrauensintervall** oder **Konfidenzintervall**.

2.4 Vertrauens- und Vorhersage-Bereiche

a Im **Beispiel der Sprengungen** kann man fragen, wie gross die Erschütterung sein wird, wenn die Distanz zur Sprengstelle 50 m beträgt. Zunächst fragen wir nach dem Erwartungswert der Erschütterung bei 50 m Distanz. Allgemein interessiert man sich oft für den **Funktionswert** $h\langle x_0\rangle$ an einer bestimmten Stelle x_0. Kann man dafür ein **Vertrauensintervall** erhalten?

Laut Modell ist $h\langle x_0\rangle = \alpha + \beta x_0$. Wir wollen die Hypothese $h\langle x_0\rangle = \eta_0$ („eta“) testen. Üblicherweise legt eine Hypothese einen bestimmten Wert für einen *Parameter* des Modells fest. Das „Rezept“ lässt sich aber ohne weiteres auf eine aus den ursprünglichen Parametern abgeleitete Grösse übertragen, wie es $\eta = \alpha + \beta x$ ist.

b Als Testgrösse für die genannte Hypothese verwenden wir wie üblich die Schätzung

$$\widehat{\eta} = \widehat{\alpha} + \widehat{\beta} x_0 \ .$$

Erwartungswert und Varianz von $\widehat{\eta}$ sind nicht schwierig zu bestimmen.

* Es ist $\mathcal{E}\langle\widehat{\eta}\rangle = \mathcal{E}\langle\widehat{\alpha}\rangle + \mathcal{E}\left\langle\widehat{\beta}\right\rangle x_0 = \alpha + \beta x_0 = \eta_0$. Um die Varianz zu bestimmen, schreiben wir $\widehat{\eta} = \widehat{\gamma} + \widehat{\beta}(x_0 - \bar{x})$ mit $\widehat{\gamma} = \widehat{\alpha} + \widehat{\beta}\bar{x} = \overline{Y}$ und erhalten, da $\operatorname{cov}\left\langle\overline{Y}, \widehat{\beta}\right\rangle = 0$ ist,

$$\operatorname{var}\langle\widehat{\eta}\rangle = \operatorname{var}\langle\widehat{\gamma}\rangle + \operatorname{var}\left\langle\widehat{\beta}\right\rangle (x_0 - \bar{x})^2 = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}} \right) \ .$$

Wenn, wie üblich, σ^2 unbekannt ist, bildet man die Testgrösse

$$T = \frac{\widehat{\eta} - \eta_0}{\mathrm{se}^{(\eta)}} \ , \qquad \mathrm{se}^{(\eta)} = \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSQ}^{(X)}}} \ ,$$

die unter der Nullhypothese eine t-Verteilung mit $n - 2$ Freiheitsgraden hat.
Das Vertrauensintervall für $\eta = h\langle x_0\rangle$ wird dann

$$(\widehat{\alpha} + \widehat{\beta} x_0) \pm q \ \mathrm{se}^{(\eta)} \ ,$$

wobei $q = q_{0.975}^{t_{n-2}}$ wieder das 0.975-Quantil der t-Verteilung mit $n - 2$ Freiheitsgraden ist.

c Der Ausdruck für das Vertrauensintervall gilt für beliebiges x_0, und es ist nahe liegend, die Grenzen des Intervalls als Funktionen von x_0 aufzuzeichnen (Abbildung 2.4.c, innere Kurven). Das ergibt ein „Band“, das für $x_0 = \bar{x}$ am schmalsten ist und gegen beide Seiten langsam breiter wird. In der Mitte des Bandes liegt die geschätzte Gerade (fitted line) $\widehat{\alpha} + \widehat{\beta} x$. Aus diesem Bild lässt sich für einen beliebigen x-Wert x_0 das **Vertrauensintervall für den Funktionswert** $h\langle x_0\rangle$ ablesen.

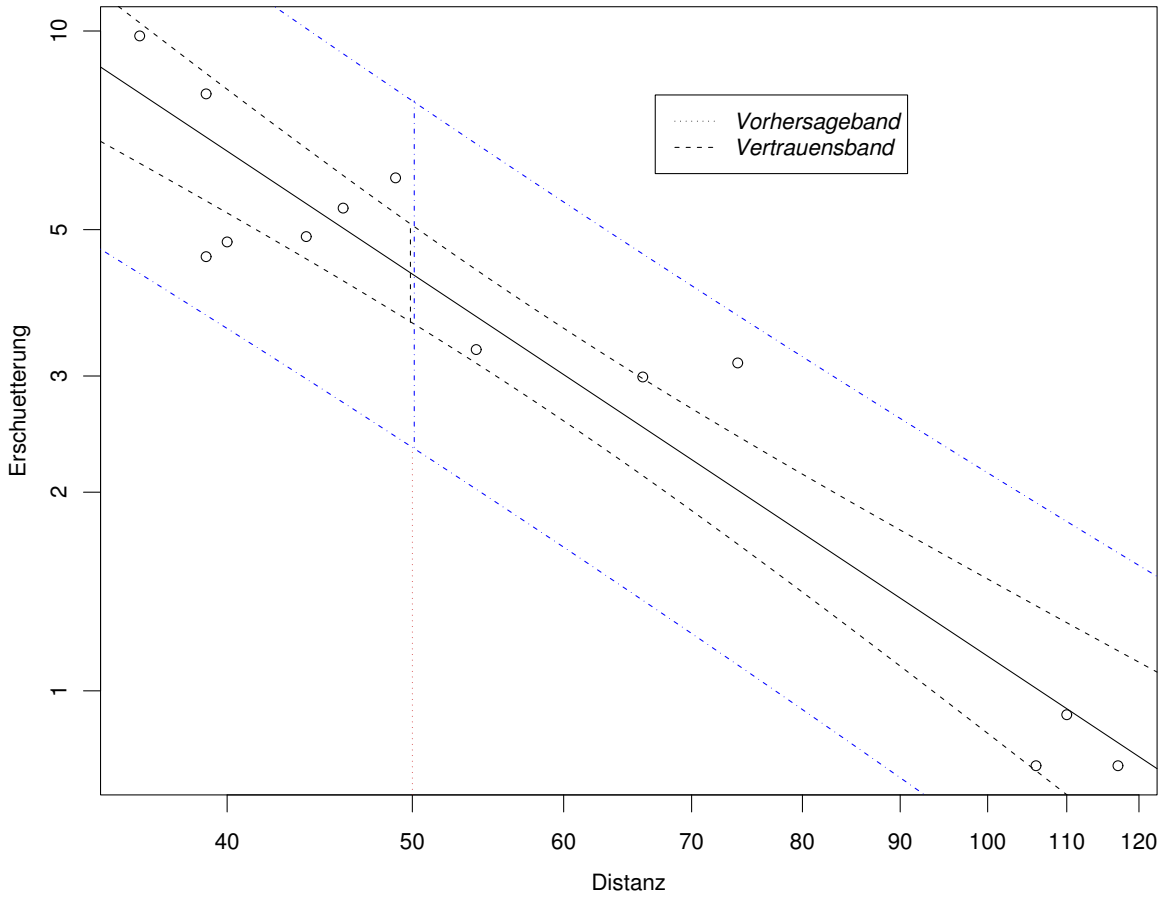

Abbildung 2.4.c: Vertrauensband für den Funktionswert $h\langle x\rangle$ und Vorhersage-Band für eine weitere Beobachtung im Beispiel der Sprengungen

d Das betrachtete „Vertrauensband“ gibt an, wo die *idealen Funktionswerte* $h\langle x\rangle$, also die Erwartungswerte von Y bei gegebenen x, liegen. Die Frage, in welchem Bereich eine **künftige Beobachtung** zu liegen kommen, ist damit nicht beantwortet. Sie ist aber oft interessanter als die Frage nach dem idealen Funktionswert; man möchte beispielsweise wissen, in welchem Bereich der zu messende Wert der Erschütterung bei 50 m Distanz liegen wird. Dieser muss schliesslich unter dem festgelegten Grenzwert bleiben!
Eine solche Angabe ist eine Aussage über eine *Zufallsvariable* und ist prinzipiell zu unterscheiden von einem Vertrauensintervall, das über einen *Parameter*, also eine feste, aber unbekannte Zahl, etwas aussagt. Entsprechend der Fragestellung nennen wir den jetzt gesuchten Bereich **Vorhersage-Intervall** oder **Prognose-Intervall**.

Es ist klar, dass dieses Intervall breiter ist als das Vertrauensintervall für den Erwartungswert, da ja noch die Zufallsabweichung der zukünftigen Beobachtung berücksichtigt werden muss. Das Ergebnis ist in Abbildung 2.4.c auch eingezeichnet.

e* Herleitung: Die Zufallsvariable Y_0 sei also der Wert der Zielgrösse bei einer Beobachtung mit Eingangsgrösse x_0. Da wir die wahre Gerade nicht kennen, bleibt uns nichts anderes übrig, als die Abweichung der Beobachtung von der geschätzten Geraden zu untersuchen,

$$R_0 = Y_0 - (\widehat{\alpha} + \widehat{\beta}x_0) = \big(Y_0 - (\alpha + \beta x_0)\big) - \big((\widehat{\alpha} + \widehat{\beta}x_0) - (\alpha + \beta x_0)\big) \,.$$

Auch wenn α und β unbekannt sind, kennen wir die Verteilungen der Ausdrücke in den grossen Klammern: Beides sind normalverteilte Zufallsvariable, und sie sind unabhängig, weil die erste nur von der „zukünftigen“ Beobachtung Y_0, die zweite nur von den Beobachtungen $Y_1, \ldots, Y_n$ abhängt, die zur ge-

schätzten Geraden führten. Beide haben Erwartungswert 0; die Varianzen addieren sich zu

$$\operatorname{var}\langle R_0\rangle = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\mathrm{SSQ}^{(X)}}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\mathrm{SSQ}^{(X)}}\right) .$$

Daraus ergibt sich das Vorhersage-Intervall

$$\widehat{\alpha} \;+\; \widehat{\beta}x_0 \pm q\widehat{\sigma}\,\sqrt{1+\tfrac{1}{n}+(x_0-\bar{x})^2/\mathrm{SSQ}^{(X)}} = \widehat{\alpha} \;+\; \widehat{\beta}x_0 \pm q\sqrt{\widehat{\sigma}^2+(\mathrm{se}^{(\eta)})^2}\,,$$

wobei wieder $q = q^{t_{n-2}}_{0.975}$ bedeutet. (Der zweite Ausdruck gilt auch für die multiple Regression.)

f Die Interpretation dieses „Vorhersage-Bandes“ ist nicht ganz einfach: Es gilt nach der Herleitung, dass

$$P\left\langle\, V_0^*\langle x_0\rangle \le Y_0 \le V_1^*\langle x_0\rangle \,\right\rangle = 0.95$$

ist, wobei $V_0^*\langle x_0\rangle$ die untere und $V_1^*\langle x_0\rangle$ die obere Grenze des Vorhersage-Intervalls ist. Wenn wir aber eine Aussage für mehr als eine zukünftige Beobachtung machen wollen, dann ist die Anzahl der Beobachtungen im Vorhersage-Band *nicht* etwa binomialverteilt mit $\pi = 0.95$. Die Ereignisse, dass die einzelnen zukünftigen Beobachtungen ins Band fallen, sind nämlich nicht unabhängig; sie hängen über die zufälligen Grenzen V_0^* und V_1^* voneinander ab. Wenn beispielsweise die Schätzung $\widehat{\sigma}$ zufälligerweise merklich zu klein herauskam, bleibt für alle zukünftigen Beobachtungen das Band zu schmal, und es werden zu viele Beobachtungen ausserhalb des Bandes liegen.

Um sicher zu gehen, dass mindestens 95% aller zukünftigen Beobachtungen im Intervall liegen, muss dieses nochmals vergrössert werden. Genaueres ist unter dem Stichwort **Toleranz-Intervall** beispielsweise in Hartung, Elpelt und Klösener (2002, §IV.1.3.3) nachzulesen.

g* Der Vollständigkeit halber sei noch ein weiteres Band mit der gleichen, hyperbolischen Form erwähnt, das in der einfachen Regression manchmal angegeben wird. Man kann zunächst einen Test für eine gemeinsame Hypothese über α und β, $H_0 : \alpha = \alpha_0$ und $\beta = \beta_0$, angeben und daraus einen Vertrauensbereich für das Wertepaar $[\alpha, \beta]$ erhalten. Es ergibt sich eine Ellipse in der $[\alpha, \beta]$-Ebene. Jedem Punkt in dieser Ellipse entspricht eine Gerade in der $[x, y]$-Ebene. Wenn man sich alle plausiblen Geraden eingezeichnet denkt, verlaufen sie in einem Band mit hyperbolischen Begrenzungslinien, den so genannten **Enveloppen der plausiblen Geraden** (im Sinne eines Vertrauensbereichs).

2.A Kleinste Quadrate

a Eine klare Begründung für die Forderung nach „Kleinsten Quadraten“ liefert das Prinzip der **Maximalen Likelihood**. Wir nehmen ja $E_i \sim \mathcal{N}(0, \sigma^2)$ an. Daraus folgt, dass die Wahrscheinlichkeitsdichte für eine einzelne Beobachtung, wenn $[\alpha^*, \beta^*]$ die wahren Parameter sind, gleich

$$f\langle y_i\rangle = c \cdot \exp\left\langle -\frac{\big(y_i - (\alpha^* + \beta^* x_i)\big)^2}{2\sigma^2} \right\rangle = c \cdot \exp\left\langle \frac{-r_i\langle\alpha^*, \beta^*\rangle^2}{2\sigma^2} \right\rangle$$

ist; dabei ist $r_i\langle\alpha^*, \beta^*\rangle = y_i - (\alpha^* + \beta^* x_i)$, analog zu 2.2.n, und c ist eine Konstante, die wir nicht genau aufzuschreiben brauchen. Die gemeinsame Dichte für alle Beobachtungen ist das Produkt all dieser Ausdrücke, für $i = 1, 2, \ldots, n$.

Das Prinzip der Maximalen Likelihood besteht darin, die Parameter so zu wählen, dass diese Dichte möglichst gross wird.

Die Rechnungen werden einfacher, wenn man logarithmiert. Das ergibt

$$\textstyle\sum_{i=1}^n \big(\log\langle c\rangle - r_i\langle\alpha^*, \beta^*\rangle^2/(2\sigma^2)\big) = n\log\langle c\rangle - \frac{1}{2\sigma^2}\sum_{i=1}^n r_i^2\langle\alpha^*, \beta^*\rangle\,.$$

Die Parameter, die die Dichte maximieren, tun dies auch für die logarithmierte Dichte. Da $n\log\langle c\rangle$ und σ^2 nicht von α^* oder β^* abhängen, kann man sie zur Maximierung weglassen. Maximierung von $-\sum_i r_i^2\langle\alpha^*, \beta^*\rangle$ bedeutet die Suche nach „Kleinsten Quadraten“.

b Lässt man Konstante, die nicht von α und β abhängen, weg, dann muss man also $\sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$ als Funktion von α und β minimieren. Wir leiten also ab

$$\begin{aligned} \frac{\partial}{\partial\alpha}\sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 &= \sum_{i=1}^n 2(y_i - (\alpha + \beta x_i))(-1) \\ \frac{\partial}{\partial\beta}\sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 &= \sum_{i=1}^n 2(y_i - (\alpha + \beta x_i))(-x_i) \end{aligned}$$

und setzen die Ableitung null; wir erhalten

$$\begin{aligned} n\widehat{\alpha} &= \sum_{i=1}^n y_i - \widehat{\beta}\sum_{i=1}^n x_i \\ \widehat{\beta}\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i - \widehat{\alpha}\sum_{i=1}^n x_i \end{aligned}$$

Das kann man umformen zu

$$\begin{aligned} \widehat{\beta}\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i - \bar{y}\sum_{i=1}^n x_i + \widehat{\beta}\bar{x}\sum_{i=1}^n x_i \\ \widehat{\alpha} &= \bar{y} - \widehat{\beta}\bar{x} \\ \widehat{\beta}\sum_{i=1}^n x_i(x_i - \bar{x}) &= \sum_{i=1}^n (y_i - \bar{y})x_i \\ \widehat{\beta} &= \frac{\sum_{i=1}^n (y_i - \bar{y})x_i}{\sum_{i=1}^n x_i(x_i - \bar{x})} \end{aligned}$$

Der Ausdruck für $\widehat{\beta}$ kann nochmals umgeformt werden: Da $\sum_{i=1}^n (x_i - \bar{x}) = 0$ und $\sum_{i=1}^n (y_i - \bar{y}) = 0$ gilt, können wir vom Zähler $\sum_{i=1}^n (y_i - \bar{y})\,\bar{x} = 0$ und vom Nenner $\sum_{i=1}^n (x_i - \bar{x})\,\bar{x} = 0$ abzählen. Dann erhalten wir den üblichen Ausdruck

$$\widehat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

für die geschätzte Steigung. So weit die Herleitung der Kleinste-Quadrate-Schätzungen von α und β.

2.B Verteilung der geschätzten Parameter

a In einem ersten Schritt wollen wir den **Erwartungswert** der Schätzung $\widehat{\beta}$ bestimmen.

Zur Abkürzung schreiben wir für die so genannte Quadratsumme der x-Werte $\mathrm{SSQ}^{(X)} = \sum_{i=1}^n (x_i - \bar{x})^2$ und $\widetilde{x}_i = (x_i - \bar{x})/\mathrm{SSQ}^{(X)}$. Es gilt $\sum_i \widetilde{x}_i = 0$ und deshalb

$$\widehat{\beta} = \sum_{i=1}^n \widetilde{x}_i(Y_i - \overline{Y}) = \sum_{i=1}^n \widetilde{x}_i Y_i - \overline{Y}\sum_{i=1}^n \widetilde{x}_i = \sum_{i=1}^n \widetilde{x}_i Y_i \,.$$

Mit Hilfe der allgemeinen Regeln $\mathcal{E}\langle a + bX\rangle = a + b\,\mathcal{E}\langle X\rangle$ und $\mathcal{E}\langle X + Y\rangle = \mathcal{E}\langle X\rangle + \mathcal{E}\langle Y\rangle$ ergibt sich

$$\mathcal{E}\langle\widehat{\beta}\rangle = \sum_{i=1}^n \widetilde{x}_i\,\mathcal{E}\langle Y_i\rangle = \sum_{i=1}^n \widetilde{x}_i(\alpha + \beta x_i) = \alpha\sum_{i=1}^n \widetilde{x}_i + \beta\sum_{i=1}^n \widetilde{x}_i x_i \,.$$

Wegen $\sum_{i=1}^n \widetilde{x}_i = 0$ fällt der erste Term weg, und

$$\sum_{i=1}^n \widetilde{x}_i x_i = \sum_{i=1}^n \widetilde{x}_i(x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2/\mathrm{SSQ}^{(X)} = 1 \,.$$

Daraus folgt die Erwartungstreue von $\widehat{\beta}$, $\mathcal{E}\langle\widehat{\beta}\rangle = \beta$.

b Die **Varianz von** $\widehat{\beta}$ ergibt sich ebenfalls aus den entsprechenden allgemeinen Regeln für die lineare Transformation, $\mathrm{var}\langle a+bX\rangle = b^2\,\mathrm{var}\langle X\rangle$, und für die Summe von unabhängigen Zufallsvariablen, $\mathrm{var}\langle X+Y\rangle = \mathrm{var}\langle X\rangle + \mathrm{var}\langle Y\rangle$,

$$\begin{aligned}\mathrm{var}\langle\widehat{\beta}\rangle &= \mathrm{var}\,\langle\textstyle\sum_{i=1}^n \widetilde{x}_i Y_i\rangle = \sum_{i=1}^n \widetilde{x}_i^2\ \mathrm{var}\langle Y_i\rangle \\ &= \sigma^2 \textstyle\sum_{i=1}^n (x_i-\bar{x})^2 \Big/ \Big(\mathrm{SSQ}^{(X)}\Big)^2 = \sigma^2/\mathrm{SSQ}^{(X)}\ .\end{aligned}$$

Nun sind Erwartungswert und Varianz von $\widehat{\beta}$ bekannt. Wir können auch genauer nach der Verteilung von $\widehat{\beta}$ fragen. Da $\widehat{\beta} = \sum_i \widetilde{x}_i Y_i$ eine Summe von Vielfachen (eine Linearkombination) von normalverteilten Zufallsvariablen Y_i ist, ist es selbst normalverteilt. Gesamthaft ergibt sich also $\widehat{\beta} \sim \mathcal{N}\left\langle \beta, \sigma^2/\mathrm{SSQ}^{(X)}\right\rangle$.

c Der Parameter α ist meistens weniger von Interesse. Um seine Verteilung herzuleiten, verwenden wir einen Trick, der auch später nützlich sein wird: Wir schreiben das Regressionsmodell etwas anders,

$$Y_i = \gamma + \beta(x_i-\bar{x}) + E_i = (\gamma - \beta\bar{x}) + \beta x_i + E_i\ .$$

Diese Schreibweise ändert das Modell nicht – es besteht immer noch aus einer allgemeinen Geradengleichung und einem „Fehlerterm"– nur die „Parametrisierung" ist jetzt anders. Aus $[\gamma, \beta]$ lässt sich das frühere Parameterpaar sofort ausrechnen: Der Vergleich der letzten Gleichung mit dem ursprünglichen Modell zeigt $\gamma = \alpha + \beta\bar{x}$; β ist als Parameter beibehalten worden. Ebenso hängen natürlich die Schätzungen zusammen,

$$\widehat{\gamma} = \widehat{\alpha} + \widehat{\beta}\,\bar{x} = \overline{Y}\ ;$$

die zweite Gleichheit erhält man aus 2.2.c.

d Die Verteilung von $\widehat{\gamma}$ ist einfach zu bestimmen. Es ist eine Normalverteilung mit

$$\begin{aligned}\mathcal{E}\langle\widehat{\gamma}\rangle &= \frac{1}{n}\textstyle\sum_{i=1}^n \mathcal{E}\langle Y_i\rangle = \gamma + \beta\,\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x}) = \gamma, \\ \mathrm{var}\langle\widehat{\gamma}\rangle &= \mathrm{var}\left\langle \frac{1}{n}\textstyle\sum_{i=1}^n Y_i\right\rangle = \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}\langle Y_i\rangle = \frac{\sigma^2}{n}\ ,\end{aligned}$$

da $\mathrm{var}\langle Y_i\rangle = \mathrm{var}\langle\alpha+\beta x_i + E_i\rangle = \mathrm{var}\langle E_i\rangle$ ist. Also ist $\widehat{\gamma} \sim \mathcal{N}\langle\gamma, \sigma^2/n\rangle$.

e Wie sieht die gemeinsame Verteilung von $\widehat{\gamma}$ und $\widehat{\beta}$ aus? Man kann zeigen, dass $\mathrm{cov}\langle\widehat{\gamma},\widehat{\beta}\rangle = 0$ ist. Zum Beweis formen wir zunächst $\widehat{\beta}$ und $\widehat{\gamma}$ um. Ausgehend von 2.B.0.a wird

$$\begin{aligned}\widehat{\beta} &= \textstyle\sum_{i=1}^n \widetilde{x}_i Y_i = \alpha\sum_{i=1}^n \widetilde{x}_i + \beta\sum_{i=1}^n \widetilde{x}_i x_i + \sum_{i=1}^n \widetilde{x}_i E_i = \alpha\cdot 0 + \beta\cdot 1 + \sum_{i=1}^n \widetilde{x}_i E_i \\ \widehat{\gamma} &= \overline{Y} = \gamma + \tfrac{1}{n}\beta\textstyle\sum_{i=1}^n (x_i-\bar{x}) + \frac{1}{n}\sum_{i=1}^n E_i = \gamma + \frac{1}{n}\sum_{i=1}^n E_i\ .\end{aligned}$$

Daraus ergibt sich

$$\begin{aligned}\mathrm{cov}\Big\langle\widehat{\beta},\widehat{\gamma}\Big\rangle &= \mathcal{E}\Big\langle(\widehat{\beta}-\beta)(\widehat{\gamma}-\gamma)\Big\rangle = \mathcal{E}\Big\langle\Big(\textstyle\sum_{i=1}^n \widetilde{x}_i E_i\Big)\Big(\frac{1}{n}\sum_{i=1}^n E_i\Big)\Big\rangle \\ &= \frac{1}{n}\Big(\textstyle\sum_{i=1}^n \widetilde{x}_i\,\mathcal{E}\langle E_i^2\rangle + \sum_{i=1}^n \widetilde{x}_i \sum_{j\neq i}\mathcal{E}\langle E_i E_j\rangle\Big)\ ,\end{aligned}$$

und dies ist $= 0$, da $\sum_{i=1}^n \widetilde{x}_i = 0$ und $\mathcal{E}\langle E_i E_j\rangle = 0$ für $j \neq i$.

f Jetzt ist auch die Verteilung von $\widehat{\alpha} = \widehat{\gamma} - \widehat{\beta}\,\bar{x}$ einfach zu bestimmen: Es ist die Normalverteilung mit $\mathcal{E}\langle\widehat{\alpha}\rangle = \mathcal{E}\langle\widehat{\gamma}\rangle - \bar{x}\,\mathcal{E}\langle\widehat{\beta}\rangle = \gamma - \bar{x}\beta = \alpha$ und

$$\operatorname{var}\langle\widehat{\alpha}\rangle = \operatorname{var}\left\langle(\widehat{\gamma} - \widehat{\beta}\bar{x})\right\rangle = \operatorname{var}\langle\widehat{\gamma}\rangle - 2\bar{x}\operatorname{cov}\langle\widehat{\gamma},\widehat{\beta}\rangle + \bar{x}^2\operatorname{var}\langle\widehat{\beta}\rangle = \sigma^2\left(\tfrac{1}{n} + \bar{x}^2\Big/\mathrm{SSQ}^{(X)}\right) .$$

Die Parameter $\widehat{\alpha}$ und $\widehat{\beta}$ sind im Allgemeinen korreliert: Es gilt

$$\operatorname{cov}\left\langle\widehat{\alpha},\widehat{\beta}\right\rangle = \operatorname{cov}\left\langle\widehat{\gamma} - \bar{x}\widehat{\beta},\widehat{\beta}\right\rangle = \operatorname{cov}\left\langle\widehat{\gamma},\widehat{\beta}\right\rangle - \bar{x}\,\operatorname{cov}\left\langle\widehat{\beta},\widehat{\beta}\right\rangle = -\bar{x}\operatorname{var}\left\langle\widehat{\beta}\right\rangle .$$

2.S S-Funktionen

a Am Ende jedes Kapitels wird ein solcher Anhang stehen, in dem die nützlichen S-Funktionen beschrieben sind. Sofern nichts anderes steht, sind die Angaben für die freie Software R und das kommerzielle Produkt S-Plus gültig. (Letzteres ist aber zurzeit nicht durchgehend überprüft.)

b **Funktion `lm`.** In S ist `lm` die grundlegende Funktion zur Anpassung von linearen Regressionsmodellen. Sie erzeugt als Resultat ein Objekt der Klasse `lm`, für die die zentralen generischen Funktionen spezielle Methoden kennen.

```
> r.lm <- lm(log10(ersch) ~ log10(dist), data = d.spreng)
```

c **Modell-Formeln.** Das erste Argument ist eine „Modell-Formel“. Solche Formeln enthalten Namen von Variablen, allenfalls (wie im Beispiel) Funktionsnamen und immer das Zeichen $\sim$, das die Zielgrösse auf der linken Seite mit der oder den X-Variablen (Regressoren) auf der rechten Seite verbindet. Die Variablen müssen entweder im `data.frame` enthalten sein, der als Argument `data=` angegeben wird (siehe unten) oder sie müssen als Objekte vorhanden sein.

Die Modell-Formeln werden im nächsten Abschnitt (3.S.0.a) im allgemeineren Zusammenhang behandelt.

d **Argument `data`.** Die Variablen, die in der Modell-Formel benützt werden, werden im `data.frame` gesucht, das als Argument `data` angegeben wird. Falls das Argument fehlt oder Variable nicht gefunden werden, werden sie im „global environment“ gesucht – also da, wo Sie Ihre Objekte speichern.

S ermöglicht auch, die Variablen eines `data.frame`s über die Funktion `attach` generell verfügbar zu machen, und dann muss das Argument `data` nicht gesetzt werden. Dieses Vorgehen wird aber nicht empfohlen (da Änderungen an den Variablen dann nicht in der erhofften Art wirksam werden).

e **Fehlende Werte.** Die einfachste Art, Datensätze mit fehlenden Werten zu behandeln, besteht darin, die entsprechenden ganzen Beobachtungen wegzulassen, und das wird mit dem Argument `na.action` in der Form `lm(..., na.action=na.omit, ...)` erreicht. Wenn viele Werte fehlen, kann das dazu führen dass sehr wenige oder keine Beobachtungen übrig bleiben. Methoden, die in solchen Fällen weiter helfen, sind anspruchsvoll.

f **Argument `subset`.** Mit dem Argument `subset` kann man die Analyse auf einen Teil des Datensatzes beschränken.

g **Funktion `summary`.** Die generische Funktion `summary` zeigt generell „die nützlichen“ Informationen aus einem Objekt. Wendet man sie auf das Resultat eines `lm`-Aufrufs an (also auf ein Objekt der Klasse `lm`), dann erhält man im Wesentlichen den in 2.3.e gezeigten Output (allerdings mit einer Bezeichnung von $\widehat{\sigma}$ als „Residual standard error“, die der Autor nicht versteht; ein korrekter Ausdruck wäre „estimated error standard deviation“).

h **Funktion** `predict.` Vorhersagewerte für gegebene Eingangsgrössen liefert die Funktion `predict`, wenn gewünscht auch mit Vertrauens- und Vorhersage-Intervallen. Will man nur die Vorhersagewerte für die x-Variablen des vorliegenden Datensatzes, dann genügt `fitted`. Wenn Vorhersagewerte und Intervalle für neue Werte der Eingangsgrössen berechnet werden sollen, müssen diese in Form eines `data.frame`s vorliegen – auch wenn es nur um eine Variable geht,

```
> t.pred <- predict(t.r, newdata=data.frame(x=seq(5,15,0.1)),
                    interval="prediction")
```

3 Multiple lineare Regression

3.1 Modell und Statistik

a Die Abhängigkeit einer Zielgrösse von einer Eingangsgrösse kann in einem einfachen Streudiagramm dargestellt werden. Oft wird dadurch das Wesentliche des Zusammenhangs sofort sichtbar. Die ganze Methodik der einfachen Regression wird dann nur noch zur Erfassung der Genauigkeit von Schätzungen und Vorhersagen gebraucht – in Grenzfällen auch zur Beurteilung, ob der Einfluss von X auf Y „signifikant“ sei.

Wenn der Zusammenhang zwischen einer Zielgrösse und **mehreren Ausgangsgrössen** $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ erfasst werden soll, reichen grafische Mittel nicht mehr aus. Das Modell der Regression lässt sich aber ohne Weiteres verallgemeinern zu

$$Y_i = h\left\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \right\rangle + E_i \; .$$

Über die zufälligen Fehler E_i macht man die gleichen Annahmen wie früher. Für h ist die einfachste Form wieder die lineare,

$$h\left\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)} \right\rangle = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} \; .$$

Sie führt zum Modell der **multiplen linearen Regression**. Die Parameter sind die so genannten **Koeffizienten** β_0, $\beta_1, \ldots, \beta_m$ der Eingangs-Variablen und die Varianz σ^2 der zufälligen Abweichungen E_i. Die Koeffizienten β_1, $\beta_2, \ldots, \beta_m$ sind die „Steigungen in Richtung der x-Achsen“. Den „Achsenabschnitt“ (für die Y-Achse) bezeichnen wir mit β_0 statt mit α wie in der einfachen Regression; das wird später die Notation vereinfachen.

b ▷ Im **Beispiel der Sprengungen** wurde nicht nur in unterschiedlicher Distanz vom Messort gesprengt, sondern es wurden auch verschiedene Ladungen verwendet (siehe Abbildung 1.1.b). Das multiple lineare Regressionsmodell mit $m = 2$ Eingangs-Variablen lautet

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i \; .$$

Wieder ist eine lineare Beziehung nicht für die ursprünglichen Variablen, sondern – wenn schon – für die logarithmierten Werte plausibel. Wir verwenden also $Y = \log_{10}\langle\text{Erschütterung}\rangle$, $X^{(1)} = \log_{10}\langle\text{Distanz}\rangle$ und $X^{(2)} = \log_{10}\langle\text{Ladung}\rangle$. Eine Formulierung des Modells, die der Programmeingabe näher steht, lautet

$$\texttt{log10(ersch)}_i = \beta_0 + \beta_1 \texttt{log10(dist)}_i + \beta_2 \texttt{log10(ladung)}_i + E_i \; . \quad \triangleleft$$

c Die übliche **Schätzung** der Koeffizienten β_j erfolgt wie in der einfachen Regression über die **Methode der Kleinsten Quadrate**. Ihre Verteilung ist mit Hilfe von Linearer Algebra nicht schwierig zu bestimmen(Anhänge 3.4 und 3.5), und darauf werden wieder Tests und Vertrauensintervalle aufgebaut. Auch die Streuung σ^2 wird auf die gleiche Weise wie vorher behandelt (siehe 2.2.n). Hier wollen wir sofort die Interpretation der Ergebnisse diskutieren.

d ▷ Eine **Computer-Ausgabe** für das **Beispiel der Sprengungen** zeigt Tabelle 3.1.d. (Es wurden zunächst von den sechs Messorten nur die ersten vier berücksichtigt, die gut zueinander passen.) Die Tabelle enthält die Schätzungen der Koeffizienten in der Kolonne „Value“, die geschätzte Standardabweichung des Fehlers und die nötigen Angaben für Tests, auf die wir gleich zurückkommen. ◁

Coefficients:	Value	Std. Error	t value	Pr(> |t|)	
(Intercept)	2.8323	0.2229	12.71	0.000	***
log10(dist)	-1.5107	0.1111	-13.59	0.000	***
log10(ladung)	0.8083	0.3042	2.66	0.011	*

St.dev. of Error = 0.1529 on 45 degrees of freedom
Multiple R-Squared: 0.8048
F-statistic: 92.79 on 2 and 45 degrees of freedom
p-value 1.11e-16

Tabelle 3.1.d: Computer-Output für das Beispiel der Sprengungen

e Bevor wir P-Werte interpretieren können, sollten wir überlegen, **welche Fragen** zu stellen sind. In den Beispielen könnten wir fragen (wenn es nicht so eindeutig wäre), ob die Distanz und die Ladung die Erschütterung, respektive die Basizität das Wachstum, überhaupt beeinflussen. Allgemeiner: Beeinflusst die **Gesamtheit der Eingangsgrössen** die Zielgrösse? Die Nullhypothese lautet: „Alle β_j (ausser β_0) sind $= 0$.“ Den entsprechenden Test findet man in den beiden letzten Zeilen der Tabelle 3.1.d. Es wird eine Testgrösse gebildet, die eine F-Verteilung hat; man spricht vom F-Test.

Bei einer einzigen Eingangsgrösse ist die Frage, ob sie einen Einfluss auf die Zielgrösse hat, mit dem Test der Nullhypothese $\beta = 0$ zu prüfen. Der „F-Test“, der in Tabelle 2.3.e auch aufgeführt wird, gibt in diesem Fall immer die gleiche Antwort – ist äquivalent – zum t-Test, der dort besprochen wurde.

f* Die Testgrösse ist $T = \big(\mathrm{SSQ}^{(R)}/m\big)/\big(\mathrm{SSQ}^{(E)}/(n-p)\big)$. Dabei ist die „Quadratsumme der Regression“ $\mathrm{SSQ}^{(R)} = \mathrm{SSQ}^{(Y)} - \mathrm{SSQ}^{(E)}$ die Differenz zwischen der „Quadratsumme der Zielgrösse“ oder „totalen Quadratsumme“$\mathrm{SSQ}^{(Y)} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ und der „Quadratsumme der Fehler“ $\mathrm{SSQ}^{(E)} = \sum_{i=1}^{n} R_i^2$. Ferner ist $p = m + 1$ die Zahl der Koeffizienten. Falls kein Achsenabschnitt β_0 im Modell erscheint, ist $p = m$ und $\mathrm{SSQ}^{(Y)} = \sum_{i=1}^{n} Y_i^2$. Die Freiheitsgrade der F-Verteilung sind m und $n - p$.

g ▷ Etliche Programme liefern auch eine so genannte Varianzanalyse-Tabelle. Tabelle 3.1.g zeigt entsprechend ausführlichere Angaben für das **Beispiel der basischen Böden** (1.1.i). In dieser Tabelle wird der genannte F-Test in der Zeile „Regression“ ausgewiesen; der P-Wert in dieser Zeile gibt Auskunft über die Signifikanz. ◁

Coefficients:

	Value	Std. Error	t value	Pr(> |t|)
(Intercept)	19.7645	2.6339	7.5039	0.0000
pH	-1.7530	0.3484	-5.0309	0.0000
lSAR	-1.2905	0.2429	-5.3128	0.0000

Residual standard error: $\widehat{\sigma} = 0.9108$ on $n - p = 120$ degrees of freedom
Multiple R-Squared: $R^2 = 0.5787$

Analysis of variance

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Regression	$m = 2$	$\text{SSQ}^{(R)} = 136.772$	68.386	$T = 82.43$	0.0000
Residuals	$n - p = 120$	$\text{SSQ}^{(E)} = 99.554$	$\widehat{\sigma}^2 = 0.830$		*P-Wert*
Total	122	$\text{SSQ}^{(Y)} = 236.326$			

Tabelle 3.1.g: Computer-Output für das Beispiel der basischen Böden mit Varianzanalyse-Tabelle und der im folgenden verwendeten Notation

h Die Grösse „Multiple R-Squared“ ist das Quadrat der so genannten **multiplen Korrelation**, der Korrelation zwischen den Beobachtungen Y_i und den **angepassten Werten** *(fitted values)*

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i^{(1)} + \widehat{\beta}_2 x_i^{(2)} + \ldots + \widehat{\beta}_m x_i^{(m)} \ .$$

Man kann zeigen, dass die nach Kleinsten Quadraten geschätzten Koeffizienten nicht nur die Quadratsumme der Residuen minimieren, sondern auch die Korrelation zwischen den angepassten Werten und den Beobachtungen der Zielgrösse maximieren; der maximale Wert ist die multiple Korrelation. Das Streudiagramm in Abbildung 3.1.h soll diese Korrelation veranschaulichen.

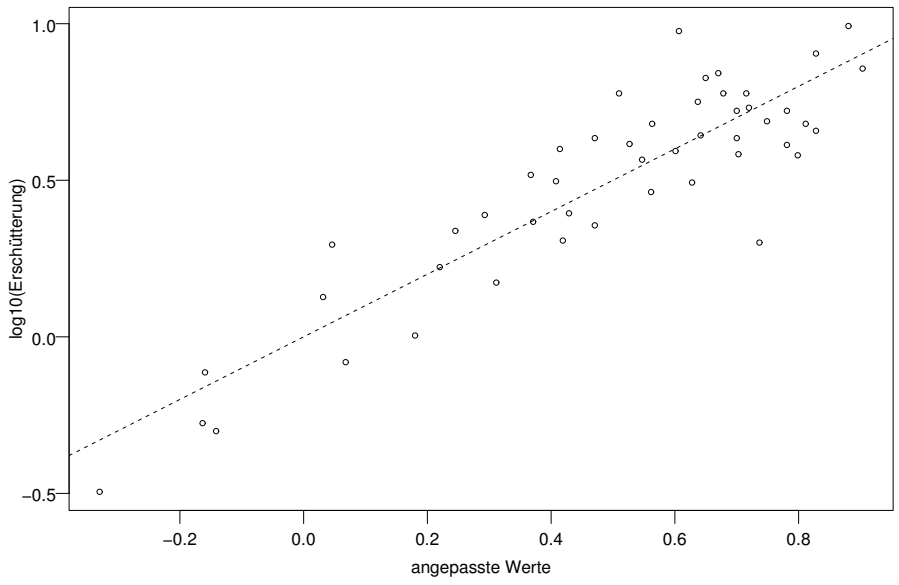

Abbildung 3.1.h: Streudiagramm der beobachteten und der angepassten Werte im Beispiel der Sprengungen

Die quadrierte multiple Korrelation wird auch **Bestimmtheitsmass** genannt, da sie den „durch die Regression bestimmten“ Anteil der Streuung der Y-Werte misst,

$$R^2 = \mathrm{SSQ}^{(R)}/\mathrm{SSQ}^{(Y)} = 1 - \mathrm{SSQ}^{(E)}/\mathrm{SSQ}^{(Y)} \ .$$

i Die Frage nach dem **Einfluss der einzelnen Variablen** $X^{(j)}$ muss man genau stellen. Der t-Wert und der P-Wert in derjenigen Zeile der Tabelle 3.1.d (oder des ersten Teils von 3.1.g), die $X^{(j)}$ entspricht, prüft, ob diese Variable aus dem Modell weggelassen werden kann, also ob die Nullhypothese $\beta_j = 0$ mit den Daten verträglich ist.

Die letzte Spalte der Tabelle enthält die übliche symbolische Darstellung der **Signifikanz**: Drei Sternchen *** für hoch signifikante Testergebnisse (P-Wert unter 0.1%), zwei Sternchen für P-Werte zwischen 0.1% und 1%, ein Sternchen für gerade noch signifikante Ergebnisse (1% bis 5 %), einen Punkt für nicht ganz signifikante Fälle (P-Wert unter 10%) und gar nichts für Zeilen mit P-Wert über 10%. Das erleichtert in grossen Tabellen das Auffinden von signifikanten Resultaten.

Im **Beispiel der basischen Böden** zeigt sich unter anderem, dass die zweite Art der Erfassung der Basizität, also $X^{(2)}$, einen Teil der Variabilität von Y erfasst, der durch den pH-Wert $X^{(1)}$ nicht „erklärt“ wird.

Die Frage, wie stark $X^{(2)}$ für sich allein, ohne Konkurrenz von $X^{(1)}$, mit Y zusammenhängt, lässt sich mit einer einfachen Regression beantworten und wird im Computer-Output der multiplen Regressionsrechnung nicht geprüft.

j Mit den Angaben der Tabelle lässt sich auch ein **Vertrauensintervall** für einen Koeffizienten β_j angeben. Es hat wie üblich die Form $\widehat{\beta}_j \pm q \ \mathrm{se}^{(\beta_j)}$, wobei $\widehat{\beta}_j$ und $\mathrm{se}^{(\beta_j)}$ in Tabelle 3.1.d unter „Value“ und „Std. Error“ zu finden sind, während der kritische Wert $q = q_{0.975}^{t_{n-2}}$ in einer Tabelle der t-Verteilung zu finden ist.

Einige Programme geben die Vertrauensintervalle direkt an.

k ▷ Im **Beispiel der Sprengungen** erhält man für den Koeffizienten von `log10(dist)` das Vertrauensintervall $-1.5107 \pm 2.014 \cdot 0.1111 = -1.5107 \pm 0.2237 = [1.2869, 1.7345]$. Nun ist der Wert -2, den wir bisher als von der Theorie vorgegeben dargestellt haben, nicht mehr im Vertrauensintervall enthalten. Der Wert -2 entspricht der ungehinderten Ausbreitung der Energie in drei Dimensionen – die Energie ist dann umgekehrt proportional zur Kugeloberfläche und damit zum quadriereten Radius. Wenn die Energie an gewissen Schichten reflektiert wird, dann ist eine weniger starke Abnahme mit der Distanz plausibel. ◁

l In diesem Skript wird eine neue Grösse eingeführt, die einerseits die Spalte „t value“ ersetzt und andererseits die Berechnung der Vertrauensintervalle erleichtert. Die t-Werte werden eigentlich nicht mehr gebraucht, um den Test auf $\beta_j = 0$ durchzuführen, da ja die p-Werte angegeben werden. Immerhin geben sie eine andere Art der „Stärke der Signifikanz“ an: Wenn sie wesentlich grösser als etwa 2 sind, dann ist der Effekt entsprechend stark gesichert, denn das 95 %-Quantil einer t-Verteilung mit nicht allzu wenigen Freiheitsgraden ist ungefähr 2. Vor allem für klar signifikante Effekte kann das eine quantitative Beurteilung erleichtern, da der p-Wert dann einfach „sehr klein“ wird.

Machen wir das exakt und führen als Mass für die **Signifikanz** den „**t-Quotienten**“ *(t ratio)* ein,

$$\widetilde{T}_j = \frac{\widehat{\beta}_j}{\mathrm{se}^{(\beta_j)} \cdot q_{0.975}^{(t_k)}} = T / q_{0.975}^{(t_k)} \ .$$

Die Stärke der Signifikanz wird jetzt nicht mehr durch Vergleich mit „ungefähr 2“, sondern mit exakt 1 beurteilt; wenn $\widetilde{T}_j$ betragsmässig grösser als 1 ist, ist der Koeffizient signifikant. $\widetilde{T}_j$ sagt direkt, wie weit innerhalb oder ausserhalb des Vertrauensintervalls der Wert 0 liegt

– im Verhältnis zur halben Länge des Intervalls. Ist der Wert 0.8, so liegt 0 innerhalb des Vertrauensintervalls, und zwar um 20% seiner halben Länge. Ist $\widetilde{T}_j = 1.2$, so liegt 0 um gleich viel ausserhalb des Intervalls. Anders ausgedrückt, ermöglicht $\widetilde{T}_j$, das Vertrauensintervall zu berechnen: Die halbe Breite des Intervalls ist $\widehat{\beta}_j/\widetilde{T}_j$ und deshalb das Vertrauensintervall selbst

$$\widehat{\beta}_j \cdot (1 \pm 1/\widetilde{T}_j) \ .$$

Tabelle 3.1.l zeigt eine Tabelle mit dieser Grösse, bezeichnet als „signif“ und wir erhalten das Vertrauensintervall für den Koeffizienten von `log10(dist)` aus $-1.511(1\pm1/6.75) = -1.511\pm0.224$, ohne das Quantil der t-Verteilung nachsehen oder abrufen zu müssen. Die Tabelle enthält ausserdem eine Spalt mit den „Freiheitsgraden“ (df), die im gegenwärtigen Zusammenhang immer gleich 1 sind, und zwei weiteren Grössen, die gleich noch erklärt werden.

Coefficients:

	coef	stcoef	signif	R2.x	df	p.value
(Intercept)	2.832	0.000	6.31	NA	1	0.000
log10(dist)	-1.511	-0.903	-6.75	0.01659	1	0.000
log10(ladung)	0.808	0.176	1.32	0.01659	1	0.011

St.dev. of Error = 0.1529 on 45 degrees of freedom
Multiple R-Squared: 0.8048
F-statistic: 92.79 on 2 and 45 degrees of freedom
p-value 1.11e-16

Tabelle 3.1.l: Resultat der S-Funktion `regr` für das Beispiel der Sprengungen

* Man könnte auch $1/\widetilde{T}_j$ als neue Grösse einführen und würde damit die Bildung des Kehrwertes bei der Berechnung des Vertrauensintervalls vermeiden. Das wäre aber als Mass für die Signifikanz ungeeignet, da ein schwacher Effekt zu einer unbegrenzten Zahl führen würde, während ein sehr stark gesicherter Effekt zu einer sehr kleinen Zahl führt.

m Eine weitere nützliche Grösse für jede X-Variable, die von einigen Programmen angegeben wird, ist der **standardisierte Regressions-Koeffizient** („stcoef“ in der Tabelle)

$$\widehat{\beta}^*_j = \widehat{\beta}_j \cdot \mathrm{sd}\left\langle X^{(j)} \right\rangle / \mathrm{sd}\langle Y \rangle \ .$$

(sd steht für die Standardabweichung.) Es ist der Koeffizient, den man erhält, wenn man alle X-Variablen und die Zielgrösse auf Mittelwert 0 und Varianz 1 standardisiert und das Modell mit den neuen Grössen anpasst. In einer einfachen Regression ist die so standardisierte Steigung gleich der Korrelation. In der multiplen Regression messen die standardisierten Koeffizienten ebenfalls die Stärke des Einflusses der einzelnen Eingangs-Variablen auf die Zielgrösse, unabhängig von den Masseinheiten oder Streuungen der Variablen. Ändert man $X^{(j)}$ um eine Standardabweichung $\mathrm{sd}\langle X^{(j)} \rangle$, dann ändert sich der geschätzte Wert der Zielgrösse um $\widehat{\beta}^*_j$ Standardabweichungen $\mathrm{sd}\langle Y \rangle$.

n* Schliesslich erscheint in der Tabelle unter der Spalte „R2.x“ ein Mass für die so genannte Kollinearität zwischen den X-Variablen. Wenn eine X-Variable stark mit den anderen zusammenhängt, führt das zu Schwierigkeiten bei der Interpretation und zu grossen Ungenauigkeiten bei der Schätzung der betroffenen Koeffizienten. Genaueres folgt in 5.3.m und 5.4.

Das hier verwendete Mass für diese Schwierigkeit wird bestimmt, indem man die Regression jeder X-Variablen $X^{(j)}$ gegen alle anderen X-Variablen durchführt und das entsprechende Bestimmtheitsmass R^2_j notiert. Auch wenn eine X-Variable, als Zielgrösse verwendet, allen Annahmen des entsprechenden Regressionsmodells widersprechen sollte, gibt das Bestimmtheitsmass einen brauchbaren Hinweis auf das Problem der Kollinearität. Der Minimalwert 0 sagt, dass $X^{(j)}$ mit den anderen Eingangsgrössen nicht (linear) zusammenhängt. Das Maximum 1 tritt auf, wenn $X^{(j)}$ von den anderen X-Variablen vollständig linear abhängt. In diesem Fall tritt sogar ein numerisches Problem auf, da die Koeffizienten nicht mehr eindeutig schätzbar sind (wie in 3.2.f).

Ein häufig verwendetes Mass für die Kollinearität ist der „Variance Inflation Factor“(VIF), der gleich $1/(1-R_j^2)$ ist. Sein Minimum ist 1; er kann beliebig gross werden.

3.2 Vielfalt der Fragestellungen

a Die Eingangs-Variablen $X^{(1)}$ und $X^{(2)}$ sind in den Beispielen kontinuierliche Messgrössen wie die Zielvariable. Das braucht allgemein nicht so zu sein.

Im Modell der multiplen Regression werden keine einschränkenden Annahmen über die X-Variablen getroffen. Sie müssen von keinem bestimmten Datentyp sein und schon gar nicht einer bestimmten Verteilung folgen. Sie sind ja nicht einmal als Zufallsvariable eingesetzt.

b* Im Beispiel der basischen Böden sind die Bodenwerte wohl ebenso zufällig wie die Baumhöhen. Für die Analyse können wir trotzdem so tun, als ob die Basizität vorgegeben wäre. Eine formale Begründung besteht darin, dass die Verteilungen gemäss Modell als bedingte Verteilungen, gegeben die $x_i^{(j)}$-Werte, aufgefasst werden.

c Eine Eingangs-Variable kann beispielsweise **binär**, also auf die Werte 0 und 1 beschränkt sein. Ist sie die einzige X-Variable, dann wird das Modell zu $Y_i = \beta_0 + E_i$ für $x_i = 0$ und $Y_i = \beta_0 + \beta_1 + E_i$ für $x_i = 1$. Das Regressionsmodell ist dann äquivalent zum Modell von zwei unabhängigen Stichproben, von denen ein allfälliger Unterschied der Lage interessiert – eine sehr übliche, einfache Fragestellung in der Statistik.

Das sieht man folgendermassen: Oft werden bei zwei Stichproben die Beobachtungen mit zwei Indices versehen: Y_{ki} ist die $i$te Beobachtung der $k$ten Gruppe ($k = 1$ oder 2) und $Y_{ki} \sim \mathcal{N}\langle \mu_k, \sigma^2 \rangle$. Es sei nun $x_{ki} = 0$, falls $k = 1$ ist, und $x_{ki} = 1$ für $k = 2$. Dann ist $Y_{ki} \sim \mathcal{N}\langle \beta_0 + \beta_1 x_{ki}, \sigma^2 \rangle$, mit $\beta_0 = \mu_1$ und $\beta_1 = \mu_2 - \mu_1$. Wenn man die Beobachtungen wieder mit einem einzigen Index durchnummeriert, ergibt sich das Regressionsmodell mit der binären x-Variablen.

d ▷ Im **Beispiel der Sprengungen** wurde die Messstelle je nach Arbeitsfortschritt verändert. Es ist plausibel, dass die örtlichen Gegebenheiten bei den Messstellen einen Einfluss auf die Erschütterung haben.

Betrachten wir zunächst den Fall von nur zwei Messstellen! Ein einfaches Modell lautet wie in 3.1.b

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i \ ,$$

wobei $X^{(1)}$ die logarithmierte Distanz sei und $X^{(2)}$ die binäre Variable, die die Messstelle bezeichnet, beispielsweise durch die Werte 0 für die erste und 1 für die zweite Messstelle. Das Modell beschreibt zwei Geraden $y = \beta_0 + \beta_1 x^{(1)}$ für die erste und $y = (\beta_0 + \beta_2) + \beta_1 x^{(1)}$ für die zweite Messstelle. Für beide Messstellen ist die gleiche Steigung β_1 wirksam; deshalb sind die beiden Geraden **parallel**. Dass die Geraden parallel sein sollen, ist eine Annahme, die in unserem Beispiel recht plausibel erscheint. Auf den allgemeineren Fall kommen wir zurück (3.2.u).

e ▷ Nun waren es aber vier Stellen, die wie üblich in einer willkürlichen Reihenfolge durchnummeriert wurden. Es ist sinnlos, die so entstehende Variable „Stellennummer“ als Eingangs-Variable $X^{(j)}$ ins Modell aufzunehmen, da eine *lineare* Abhängigkeit der Erschütterung von der Stellen-Nummer kaum plausibel ist. ◁

Eine solche Eingangs-Variable mit **nominalem** oder **kategoriellem Wertebereich** wird auch **Faktor** genannt. Um sie in ein Regressionsmodell einzubeziehen, führt man für jeden möglichen Wert (jede Stelle) eine **„Indikatorvariable“** ein,

$$x_i^{(j)} = \begin{cases} 1 & \text{falls } i\,\text{te Beobachtung aus der } j\,\text{ten Gruppe,} \\ 0 & \text{sonst.} \end{cases} .$$

Ein Modell für mehrere Gruppen j von Beobachtungen mit verschiedenen Erwartungswerten μ_j (aber sonst gleicher Verteilung) kann man schreiben als

$$Y_i = \mu_1 x_i^{(1)} + \mu_2 x_i^{(2)} + \ldots + E_i$$

mit unabhängigen, gleich verteilten E_i. Setzt man $\mu_j = \beta_j$, so steht das multiple Regressionsmodell da, allerdings ohne Achsenabschnitt β_0.

Eine binäre Variable, die eine Gruppenzugehörigkeit ausdrückt, wird als **dummy variable** bezeichnet. Eine nominale Eingangs-Variable führt so zu einem **„Block“ von dummy Variablen**.

f ▷ Im Beispiel kommt dieser Block zu den beiden andern Eingangs-Variablen hinzu (und die Nummerierung j der $X^{(j)}$ mag sich dadurch verändern). Das Modell kann man so schreiben:

$$\begin{aligned} \texttt{log10(ersch)}_i \quad = \quad & \beta_0 + \beta_1 \texttt{log10(dist)}_i + \beta_2 \texttt{log10(ladung)}_i \\ & + \gamma_1 \texttt{St1}_i + \gamma_2 \texttt{St2}_i + \gamma_3 \texttt{St3}_i + \gamma_4 \texttt{St4}_i + E_i \end{aligned} \quad \triangleleft$$

g Ein technischer Punkt: In diesem Modell lassen sich die Koeffizienten prinzipiell **nicht eindeutig** bestimmen (vergleiche 3.4.h). Es verändern sich nämlich die „Modellwerte“ $h\left\langle x_i^{(1)}, \ldots x_i^{(m)} \right\rangle$ nicht, wenn man zu allen γ_k eine Konstante dazuzählt und sie von β_0 abzählt. Eine so gebildete Kombination von Koeffizienten passt also sicher genau gleich gut zu den Beobachtungen. Man sagt deshalb, die Parameter seien **nicht identifizierbar**.

Um die Sache eindeutig zu machen, braucht man entweder **Nebenbedingungen** oder man lässt eine dummy Variable weg. Eine einfache Lösung besteht darin, $\gamma_1 = 0$ zu setzen oder, anders gesagt, die Variable `St1` nicht ins Modell aufzunehmen. (In der Varianzanalyse werden wir auf das Problem zurückkommen und auch andere Abhilfen diskutieren.)

h ▷ Die numerischen Ergebnisse zeigt Tabelle 3.2.h. Die t- und P-Werte, die zu den „dummy“ Variablen `St2` bis `St4` angegeben werden, haben wenig Bedeutung. Bei unserer Wahl von $\gamma_1 = 0$ zeigen sie, ob der Unterschied zwischen der entsprechenden Stelle und Stelle 1 signifikant sei.

Coefficients:

	Value	Std. Error	t value	Pr(> |t|)	Signif
(Intercept)	2.51044	0.28215	8.90	0.000	***
log10(dist)	-1.33779	0.14073	-9.51	0.000	***
log10(ladung)	0.69179	0.29666	2.33	0.025	*
St2	0.16430	0.07494	2.19	0.034	*
St3	0.02170	0.06366	0.34	0.735	
St4	0.11080	0.07477	1.48	0.146	

Residual standard error: 0.1468 on 42 degrees of freedom
Multiple R-Squared: 0.8322
F-statistic: 41.66 on 5 and 42 degrees of freedom
the p-value is 3.22e-15

Tabelle 3.2.h: Computer-Ausgabe im Beispiel Sprengungen mit 3 Eingangs-Variablen

i ▷ Um die Idee grafisch veranschaulichen zu können, unterdrücken wir die Variable `ladung`, indem wir nur Beobachtungen mit `ladung`=2.6 berücksichtigen. Abbildung 3.2.i zeigt die Beobachtungen und das angepasste Modell: **Für jede Stelle** ergibt sich **eine Gerade**, und da für die verschiedenen Stellen im Modell die gleiche Steigung bezüglich der Variablen `log(dist)` vorausgesetzt wurde, sind die angepassten Geraden **parallel**. ◁

j Es gibt eine sehr nützliche vereinfachte **Notation**, in der solche Modelle aufgeschrieben werden, die „**Modell-Formeln**“. Das Modell im Beispiel wird geschrieben als

$$\texttt{log10(ersch)} \sim \texttt{log10(dist)} + \texttt{log10(ladung)} + \texttt{St}\ .$$

Die Indices, die Koeffizienten und der Fehlerterm werden weggelassen. Das Plus-Zeichen hat jetzt natürlich eine andere Bedeutung als üblich; es verbindet nicht mehr Zahlen, sondern Eingangs-Variable – in ursprünglicher oder transformierter Form.

Die Sprache der Modell-Formeln eignet sich zur Eingabe in Programm-Pakete. Für die Variable `St` muss dem Programm bekannt sein, dass es sich um eine nominale Variable oder einen so genannten Faktor (siehe Varianzanalyse) handelt. Es konstruiert sich dann die entsprechenden dummy Variablen selber. `St` ist also ein **Term** in der Modell-Formel, der eine ganze Gruppe von X-Variablen umfasst, die in ihrer Bedeutung zusammengehören.

In einigen Programmen können in der Modellangabe keine Transformationen festgelegt werden. Man muss dann zuerst transformierte Variable `lersch`=`log10(ersch)` und analog `ldist` und `lladung` erzeugen. Das Modell lautet dann `lersch` $\sim$ `ldist` + `lladung` + `St`.

k Die „X-Variablen“ erscheinen nun in verschiedenen Formen, die wir mit verschiedenen Ausdrücken bezeichnen wollen: Eine **Eingangsgrösse** oder **Eingangs-Variable** ist eine Grösse, von der angenommen wird, dass sie mit der Zielgrösse zusammenhängt, und für die deshalb eine geeignete Form gesucht wird, in der sie in das lineare Regressionsmodell einbezogen werden soll. Das kann in transformierter Form geschehen oder, wenn es eine nominale Variable ist, in Form mehrerer dummy-Variablen. Die X-Variablen, wie sie im linearen Modell erscheinen, nennt man auch **Regressoren**. Ein **Term** in der Modell-Formel kann ein einzelner Regressor sein oder eine Gruppe von zusammengehörigen Regressoren, die als Einheit betrachtet werden. Neben den Faktoren werden solche Gruppen vor allem Wechselwirkungen mit Faktoren sein, die bald eingeführt werden (3.2.t).

l Man wird die Frage stellen, ob die Messstelle (`St`) überhaupt einen Einfluss auf die Erschütterung habe. „Kein Einfluss“ bedeutet, dass die Koeffizienten aller entsprechenden Indikator-Variablen null sind, $\gamma_1 = 0,\ \gamma_2 = 0,\ \gamma_3 = 0,\ \gamma_4 = 0$. Den üblichen Test für diese Hypothese wollen wir allgemeiner aufschreiben.

m **F-Test zum Vergleich von Modellen**. Die Frage sei, ob die q Koeffzienten β_{j_1}, β_{j_2}, ..., β_{j_q} in einem linearen Regressionsmodell gleich null sein könnten.

- Nullhypothese: $\beta_{j_1} = 0$ und $\beta_{j_2} = 0$ und ... und $\beta_{j_q} = 0$
- Teststatistik:
$$T = \frac{(\mathrm{SSQ}^{(E)*} - \mathrm{SSQ}^{(E)})/q}{\mathrm{SSQ}^{(E)}/(n-p)} \; ;$$
$\mathrm{SSQ}^{(E)*}$ ist die Quadratsumme des Fehlers im „kleinen“ Modell, die man aus einer Regression mit den verbleibenden $m-q$ X-Variablen erhält, und p die Anzahl Koeffizienten im „grossen“ Modell ($= m+1$, falls das Modell einen Achsenabschnitt enthält, $= m$ sonst).
- Verteilung von T unter der Nullhypolthese: $T \sim \mathcal{F}_{q,n-p}$, F-Verteilung mit q und $n-p$ Freiheitsgraden.

Der Test heisst F-Test zum Vergleich von Modellen. Allerdings kann nur ein kleineres Modell mit einem grösseren verglichen werden, in dem alle X-Variablen des kleinen wieder vorkommen, also mit einem „umfassenderen“ Modell. Der früher besprochene F-Test für das gesamte Modell (3.1.e) ist ein Spezialfall: das „kleine“ Modell besteht dort nur aus dem Achsenabschnitt β_0.

n Zurück zur Prüfung des Einflusses einer nominalen erklärenden Variablen: Die besseren Programme liefern den entsprechenden Test gleich mit, indem sie in einer Tabelle den F-Test für die einzelnen Terme in der Modellformel zusammenstellen (Tabelle 3.2.n).

	Df	Sum of Sq	RSS	F Value	Pr(F)
log10(dist)	1	1.947	2.851	90.4	4.9e-12
log10(ladung)	1	0.117	1.022	5.44	0.025
Stelle	3	0.148	1.052	2.283	0.093

Tabelle 3.2.n: Tests für die Effekte der einzelnen Terme im Beispiel der Sprengungen

Für die ersten beiden erklärenden Variablen gibt diese Tabelle die gleiche Auskunft wie die vorhergehende (3.2.h). Der „F Value“ ist gleich dem quadrierten „t value“ von damals, und die entsprechenden Tests sind äquivalent. Die dritte Zeile vergleicht das umfassende Modell mit dem Modell ohne `St` als erklärende Variable. Sie zeigt, dass der Einfluss der Stelle nicht signifikant ist.

o* Achtung! Oft wird in einer genau gleich aussehenden Tabelle ein anderer Test durchgeführt, der im Allgemeinen wenig Bedeutung hat. Es wird nämlich in der eingegebenen Reihenfolge der Terme im Regressionsmodell schrittweise geprüft, ob der betreffende Term eine Verbesserung gegenüber dem vorhergehenden Modell, ohne diesen Term, bringt. Nur für den letzten Term in der Tabelle erhält man also den gewünschten Test.

p ▷ Wenn kontinuierliche Variable und Faktoren als Eingangsgrössen im Modell stehen, muss man üblicherweise die nützliche Information aus zwei verschiedenen Tabellen zusammensuchen: Aus Tabelle 3.1.d, liest man die Koeffizienten der kontinuierlichen Variablen ab und schaut sich auch ihren P-Wert für den Test gegen $\beta_j = 0$ an, und in der vorhergehenden Tabelle (3.2.n), die man extra verlangen muss, sucht man den P-Wert für die Faktoren. Das Resultat der Funktion `regr` zeigt beides in einer Tabelle (Tabelle 3.2.p). Die geschätzten Koeffizienten des Faktors erscheinen unterhalb der Haupttabelle. ◁

q In den üblichen Darstellungen der Resultate (3.2.h) werden Koeffizienten für Faktoren in der gleichen Tabelle wie für kontinuierliche Variable gezeigt. Je nach „Codierung“ sind diese aber nicht die Effekte γ_k der einzelnen Werte des Faktors (3.2.g), sondern kaum interpretierbare Grössen, die als Koeffizienten von erzeugten Variablen auftreten. Für die Koeffizienten werden dann, wie für die kontinuierlichen Variablen, t- und P-Werte angegeben, die nur bei geeigneter Codierung

```
Call:
regr(formula = log10(ersch) ~ log10(dist) + log10(ladung) + Stelle,
     data = d.spreng14)
Terms:
                 coef  stcoef  signif    R2.x df p.value
(Intercept)    2.5104  0.0000  4.4090      NA  1   0.000
log10(dist)   -1.3378 -0.7993 -4.7106 0.24825  1   0.000
log10(ladung)  0.6918  0.1510  1.1555 0.02409  1   0.025
Stelle             NA      NA  0.8986 0.08884  3   0.093

Coefficients for factors:
$Stelle
     1      2      3      4
0.0000 0.1643 0.0217 0.1108

St.dev.error:  0.147   on 42 degrees of freedom
Multiple R^2:  0.832  Adjusted R-squared:   NA
F-statistic:   41.7   on 5 and 42 d.f.,  p.value: 3.22e-15
```

Tabelle 3.2.p: Ergebnisse der Funktion `regr` für das Beispiel der Sprengungen

(„treatment“ oder „sum“ in S) mit der entsprechenden Vorsicht sinnvoll zu interpretieren sind.

r* Die Spalte „signif“ in der in 3.1.l eingeführten Darstellung der Resultate liefert für eine kontinuierliche Variable, wie beschrieben (3.1.l), das Verhältnis $\widetilde{T}_j$ zwischen dem geschätzten Koeffizienten und seiner Signifikanzgrenze. Die Grösse soll für Faktoren so definiert sein, dass sie eine ähnliche anschauliche Bedeutung erhält. Es sei (für irgendeinen Test) die „**z-ratio**“ das Quantil der Standard-Normalverteilung, das dem P-Wert entspricht, dividiert durch den entsprechenden kritischen Wert $q^{(\mathcal{N})}\langle 0.95\rangle = 1.96$,

$$\widetilde{T} = q^{(\mathcal{N})}\langle 1-p\rangle \Big/ q^{(\mathcal{N})}\langle 0.95\rangle \quad .$$

(Die t-ratio für kontinuierliche Variable ist zwar nicht genau gleich diesem Wert, aber für nicht allzu kleine Anzahlen von Freiheitsgraden sehr ähnlich.)

Fox and Monette (1992) verallgemeinern den Variance Inflation Factor für Faktoren. Hier wird dieser verallgemeinerte VIF verwendet und „in die R^2-Skala umgerechnet nach der Formel $R^2 = 1 - 1/\mathrm{VIF}$.

s* Allgemeinere Vergleiche von Modellen können nicht automatisch erfolgen, da es zu viele Möglichkeiten gibt und das Programm die interessanten kaum erraten kann. In umfassenden Programmen kann man die interessierenden Vergleiche angeben und erhält dann die gewünschten Testergebnisse. Sonst muss man sich die nötigen Quadratsummen aus zwei Computer-Ausgaben heraussuchen und mit der obenstehenden Formel den Wert der Testgrösse und den P-Wert bestimmen.

t Im Modell 3.2.f zeigt sich der Einfluss der Stelle nur durch eine additive Konstante. Der Wechsel von einer Messstelle zu einer anderen „darf“ also nur zur Folge haben, dass sich die logarithmierten Erschütterungen um eine Konstante vergrössern oder verkleinern; die Geraden in 3.2.d müssen **parallel** sein. Es ist natürlich denkbar, dass der Zusammenhang zwischen Erschütterung einerseits und Distanz und Ladung andererseits sich zwischen den Stellen auf kompliziertere Art unterscheidet.

Eine nahe liegende Variante wäre, dass sich die Steigungskoeffizienten β_1 und β_2 für verschiedene Messstellen unterscheiden. Man spricht dann von einer **Wechselwirkung** zwischen Distanz und Stelle oder zwischen Ladung und Stelle. Das ist eine allgemeinere Frage als die folgende einfache, die immer wieder auftaucht.

u **Sind zwei Geraden gleich?** Oder unterscheiden sie sich im Achsenabschnitt, in der Steigung oder in beidem? Um diese Frage zu untersuchen, formulieren wir als Modell

$$Y_i = \alpha + \beta\, x_i + \Delta\alpha\, g_i + \Delta\beta\, x_i g_i + E_i$$

wobei g_i die „Gruppenzugehörigkeit“ angibt: $g_i = 0$, falls die Beobachtung i zur einen Geraden, $g_i = 1$, falls sie zur anderen gehört. Für die Gruppe mit $g_i = 0$ entsteht die Gerade $\alpha + \beta x_i$, für $g_i = 1$ kommt $(\alpha + \Delta\alpha) + (\beta + \Delta\beta)x_i$ heraus. Die beiden Geraden stimmen in der Steigung überein, wenn $\Delta\beta = 0$ ist. Sie stimmen gesamthaft überein, wenn $\Delta\beta = 0$ und $\Delta\alpha = 0$ gelten. (Der Fall eines gleichen Achsenabschnitts bei ungleicher Steigung ist selten von Bedeutung.)

Das Modell sieht zunächst anders aus als das Grundmodell der multiplen Regression. Wir brauchen aber nur $x_i^{(1)} = x_i$, $x_i^{(2)} = g_i$ und $x_i^{(3)} = x_i g_i$ zu setzen und die Koeffizienten α, β, $\Delta\alpha$, $\Delta\beta$ als β_0, β_1, β_2, β_3 zu bezeichnen, damit wieder die vertraute Form dasteht.

Die Nullhypothese $\Delta\beta = 0$ lässt sich mit der üblichen Tabelle testen. Der Test für „$\Delta\alpha = 0$ und $\Delta\beta = 0$“ ist ein weiterer Fall für den F-Test zum Vergleich von Modellen.

v Das Beispiel zeigt, dass die x-Variablen im Modell in irgendeiner Weise aus ursprünglichen erklärenden Variablen ausgerechnet werden können. So darf beispielsweise auch $X^{(2)} = (X^{(1)})^2$ sein. Das führt zur **quadratischen Regression**,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i\ .$$

Abbildung 3.2.v zeigt die Anpassung dieses Modells im Beispiel der basischen Böden (Beobachtungen mit pH> 8.5 wurden weggelassen).
In gleicher Weise können auch höhere Potenzen eingeführt werden, was zur **polynomialen Regression** führt.

* Da jede glatte Funktion sich durch eine Polynom-Reihe annähern lässt, wird die polynomiale Regression oft eingesetzt, wenn man über die Art der Abhängigkeit zwischen einer erklärenden Variablen und einer Zielgrösse „keine“ Annahmen treffen will. Es gibt dafür aber unter dem Stichwort **Glättung** oder **smoothing** oder **nichtparametrische Regression** geeignetere Methoden.

w Nun geraten die Begriffe durcheinander: Eine quadratische Regression wird als (multiple) lineare Regression bezeichnet! – **Das Wort *linear* im Begriff der multiplen linearen Regression bezieht sich *nicht* auf eine lineare Beziehung zwischen Y und den $X^{(j)}$, sondern darauf, dass die *Koeffizienten* linear in der Formel vorkommen!**

x Dieser Abschnitt hat gezeigt, dass das Modell der multiplen linearen Regression viele Situationen beschreiben kann, wenn man die X-Variablen geeignet wählt:

- Transformationen der X- (und Y-) Variablen können aus ursprünglich nicht-linearen Zusammenhängen lineare machen.
- Ein Vergleich von zwei Gruppen lässt sich mit einer zweiwertigen X-Variablen, von mehreren Gruppen mit einem „Block“ von dummy Variablen als multiple Regression schreiben. Auf diese Art werden nominale erklärende Variable in ein Regressionsmodell aufgenommen.
- Die Vorstellung von zwei verschiedenen Geraden für zwei Gruppen von Daten kann als ein einziges Modell hingeschrieben werden – das gilt auch für mehrere Gruppen. Auf allgemeinere Wechselwirkungen zwischen erklärenden Variablen kommen wir zurück (4.6.g).
- Die polynomiale Regression ist ein Spezialfall der multiplen linearen (!) Regression.

3.3 Multiple Regression ist viel mehr als viele einfache Regressionen

a Die multiple Regression wurde eingeführt, um den Einfluss mehrerer erklärender Grössen auf eine Zielgrösse zu erfassen. Ein verlockender, einfacherer Ansatz zum gleichen Ziel besteht darin, für jede erklärende Variable eine einfache Regression durchzuführen. Man erhält so ebenfalls je einen geschätzten Koeffizienten mit Vertrauensintervall. In der Computer-Ausgabe der multiplen Regression stehen die Koeffizienten in einer einzigen Tabelle. Ist das der wesentliche Vorteil?

Die Überschrift über diesen Abschnitt behauptet, dass der Unterschied der beiden Ansätze – mehrere einfache gegen eine multiple Regressionsanalyse – viel grundlegender ist. Das soll im Folgenden begründet werden.

b ▷ **Modifiziertes Beispiel der Sprengungen.** Um Unterschiede der beiden möglichen Arten der Auswertungen zu demonstrieren, wurde der Datensatz der Sprengungen auf die Stellen 3 und 6 und Distanzen kleiner als 100 m eingeschränkt. Tabelle 3.3.b zeigt die numerischen Resultate der einfachen Regressionen der logarithmierten Erschütterung auf die logarithmierte Distanz und zum Vergleich das Resultat der multiplen Regression mit den erklärenden Variablen log(Distanz), log(Ladung) und Stelle.

```
------------------------------------------------------------------------
(i)
lm(formula = log10(ersch) ~ log10(dist), data = dd)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8976     0.5736   1.565    0.127
log10(dist)  -0.1316     0.3260  -0.404    0.689

Residual standard error: 0.2134 on 32 degrees of freedom
Multiple R-Squared: 0.00507,    Adjusted R-squared: -0.02602
F-statistic: 0.1631 on 1 and 32 degrees of freedom,     p-value: 0.689
------------------------------------------------------------------------
(ii)
lm(formula = log10(ersch) ~ log10(dist) + log10(ladung) + stelle,
   data = dd)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.19297    0.58161   2.051  0.04908 *
log10(dist)        -0.72687    0.35503  -2.047  0.04947 *
log10(ladung)       1.49261    0.44162   3.380  0.00203 **
stelle6             0.16956    0.08604   1.971  0.05803 .

Residual standard error: 0.1813 on 30 degrees of freedom
Multiple R-Squared: 0.3269,     Adjusted R-squared: 0.2596
F-statistic: 4.856 on 3 and 30 degrees of freedom,  p-value: 0.00717
------------------------------------------------------------------------
```

Tabelle 3.3.b: Ergebnisse für die (i) einfache Regressionen der logarithmierten Erschütterung auf die logarithmierte Distanz und für die (ii) multiple Regression mit Distanz, Ladung und Stelle.

Die einfache Regression liefert einen völlig unplausiblen Wert für den Koeffizienten der logarithmierten Distanz, mit einem Vertrauensintervall von $[-0.1316 \pm 2.037 \cdot 0.3260] = [-0.80, 0.53]$. Mit dem multiplen Modell ergibt sich für diesen Koeffizienten ein Intervall von $[-0.72687 \pm 2.042 \cdot 0.35503] = [-1.45, -0.002]$, das mit den Ergebnissen verträglich ist, die der gesamte Datensatz lieferte (3.2.h).

In Abbildung 3.3.b sind geschätzte Steigungen für die einfache Regression eingezeichnet – sowohl für beide Stellen zusammen als auch für die getrennte Auswertung. Die beiden weiteren, parallelen Geraden haben die Steigung, die sich aus der multiplen Regression ergibt, und geben die angepassten Werte für eine mittlere Ladung wieder. (Die Wechselwirkung zwischen log10(Distanz) und der Stelle, die einer unterschiedlichen Steigung der beiden Geraden entspricht, erwies sich als nicht signifikant.)

◁

c ▷ An künstlichen Beispielen lassen sich solche Effekte noch klarer veranschaulichen. In Abbildung 3.3.c sind für den Fall einer kontinuierlichen erklärenden Variablen $X^{(1)}$ und einer Gruppierungsvariablen $X^{(2)}$ vier mögliche Fälle aufgezeichnet. Die gestrichelten Geraden zeigen das Modell, nach dem die Beobachtungen erzeugt wurden: Zwei parallele Geraden mit Steigung β_1 und einem vertikalen Abstand von β_2. Die Beobachtungen der beiden Gruppen tragen verschiedene Symbole. Die ausgezogene Gerade stellt das Resultat einer einfachen Regression von Y auf $X^{(1)}$ dar; das schmale Rechteck am rechten Rand zeigt den Unterschied zwischen den Gruppenmittelwerten der Zielgrösse, was der einfachen Regression von Y gegen $X^{(2)}$ entspricht. Die Gerade und das Rechteck zeigen also das Resultat, das man erhält, wenn man die beiden Regressoren $X^{(1)}$ und $X^{(2)}$ je mit einfacher Regression „abhandelt“.

Die Ergebnisse der multiplen Regression sind nicht eingezeichnet; sie widerspiegeln das Modell ziemlich genau. Die vier Fälle zeigen die Schwierigkeiten der Interpretation von einfachen Regressionen drastisch:

(A) Beide Variablen haben einen positiven Effekt, $\beta_1 > 0$, $\beta_2 > 0$. Die geschätzte Steigung und der Unterschied der Gruppenmittelwerte werden zu gross.

(B) Kein Effekt der kontinuierlichen erklärenden Variablen $X^{(1)}$. Die geschätzte Gerade erhält ihre Steigung durch den Unterschied zwischen den Gruppen.

(C) Entgegengesetzte Effekte, $\beta_1 < 0$, $\beta_2 > 0$. Die geschätzte Steigung zeigt einen positiven Effekt der kontinuierlichen erklärenden Variablen $X^{(1)}$ auf die Zielgrösse, während er in Wirklichkeit negativ ist!

(D) Hier sind die Effekte so eingerichtet, dass sie sich gegenseitig aufheben. Man wird fälschlicherweise schliessen, dass keine der beiden Variablen einen Einfluss auf Y hat. ◁

d Wenn wir uns das Modell der multiplen Regression vergegenwärtigen, wird klar, wie der Unterschied zu den Ergebnissen der einfachen Regression entsteht: Der Koeffizient β_1 beispielsweise gibt an, um wie viel sich der erwartete Wert der Zielgrösse erhöht, wenn $X^{(1)}$ um 1 erhöht wird – und alle anderen erklärenden Variablen gleich bleiben. Im Beispiel bleibt die Ladung und die Stelle gleich; wir erhalten also die Steigung der Geraden innerhalb der Stelle bei konstanter Ladung – und gehen, wenn die Wechselwirkung im Modell fehlt, davon aus, dass diese für beide Stellen gleich ist.

Betrachten wir die einfache Regression der Zielgrösse auf $X^{(1)}$, dann wird sich die Bedeutung von β_1 ändern. Die zweite ausgewählte Stelle wurde bei grösseren Distanzen erfasst als die erste und führte trotzdem tendenziell zu gleich hohen Erschütterungen. Teilweise lag das daran, dass auch stärker geladen wurde. Wenn $X^{(1)}$ um 1 erhöht wird, kommen im Datensatz tendenziell Beobachtungen mit höherer Ladung und anderer Stellenzugehörigkeit zum Zuge, und daher sinkt der Erschütterungswert kaum. Die Effekte der erklärenden Variablen werden vermischt.

e Ist eine kontinuierliche erklärende Variable $X^{(2)}$ mit $X^{(1)}$ positiv korreliert, dann wird sich bei einer Erhöhung von $X^{(1)}$ um 1 erwartungsgemäss auch $X^{(2)}$ erhöhen, was einen zusätzlichen Effekt auf die Zielgrösse hat. (* Der Effekt, ausgedrückt durch den Koeffizienten β_2 im multiplen Modell und dem „Regressionskoeffizienten von $X^{(2)}$ auf $X^{(1)}$, $\beta_{21} = \mathrm{cov}\langle X^{(1)}, X^{(2)}\rangle / \mathrm{var}\langle X^{(1)}\rangle$, beträgt $\beta_2\beta_{21}$.) Analoges gilt, wenn $X^{(1)}$ sich für die verschiedenen Werte einer nominalen erklärenden Grösse $X^{(2)}$ im Mittel wesentlich unterscheidet.

Diese Betrachtung zeigt allgemeiner, dass die **Bedeutung der Regressionskoeffizienten** prinzipiell davon abhängt, welche erklärenden Grössen im Modell auftreten.

Beachten Sie, dass wir vom Modell gesprochen haben, dass also dieses Problem nicht mit der

Schätzung zusammenhängt.

f Grundlegend für alle Wissenschaften ist die Suche nach **Ursache-Wirkungs-Beziehungen**. Bekanntlich kann aus statistischen Korrelationen nicht auf solche Beziehungen geschlossen werden. Dennoch besteht eine wichtige Anwendung der Regression darin, Indizien für solche Beziehungen zu sammeln. Zwei Arten von Schlüssen sind üblich:

g Erste Schlussweise: Falls ein Koeffizient in einem Regressionsmodell **signifikant** von Null verschieden ist und eine ursächliche Wirkung der Zielgrösse auf die erklärende Grösse aus prinzipiellen Überlegungen heraus ausgeschlossen werden kann (die Erschütterung kann die Distanz zum Sprengort nicht beeinflussen!), dann wird dies als **Nachweis für eine vermutete ursächliche Wirkung** der erklärenden Grösse auf die Zielgrösse interpretiert.

h Oft kommt aber eine Korrelation zwischen einer erklärenden Variablen und der Zielgrösse dadurch zustande, dass **beide von einer dritten** Grösse Z verursacht werden.

Dies ist besonders häufig, wenn die Daten als **Zeitreihe** entstehen. Die Zahl der Neugeborenen hat im 20. Jahrhundert in den hochentwickelten Ländern abgenommen. Das lässt sich gut mit der Abnahme der Störche erklären... Die Zeit ist hier nicht die eigentliche Ursache der beiden Phänomene, sondern die Ursachen für den Niedergang der Anzahl Störche und der Anzahl Babies haben sich mit der Zeit ebenfalls verändert. Die Zeit kann dann die Ursachen in dieser Betrachtung (teilweise) vertreten.

Solche Situationen werden auch als **indirekte Zusammenhänge**, indirekte Korrelationen oder **Schein-Korrelationen** bezeichnet.

i Wenn die Grösse Z im Modell als erklärende Variable auftaucht, dann verfälschen die durch sie erfassten indirekten Wirkungen die Koeffizienten der anderen erklärenden Variablen nicht. Im Idealfall wird man also **alle denkbaren ursächlichen Variablen** für die betrachtete Zielgrösse als erklärende Variable **ins Modell aufnehmen**; dann stellt ein signifikanter Koeffizient von $X^{(1)}$ ein starkes Indiz für eine Ursache-Wirkungsbeziehung dar.

j Eine noch bessere Basis für eine solche Interpretation bilden, wenn sie möglich sind, **geplante Versuche**, in denen unter sonst gleichen Bedingungen nur die fragliche Variable $X^{(1)}$ variiert wird. Dann kann man die Wirkung direkt messen. Am überzeugendsten ist aber natürlich immer noch der konkrete **Nachweis eines Wirkungs-Mechanismus**.

k Zweite Schlussweise: Wenn ein Koeffizient **nicht signifikant** ist, wird dies oft als Nachweis betrachtet, dass die entsprechende erklärende Grösse **keinen Einfluss** auf die Zielgrösse habe. Dies ist in mehrfacher Hinsicht ein Fehlschluss:

- Wie bei allen statistischen Tests ist die Beibehaltung der Nullhypothese kein Beweis, dass sie gilt.
- Die vorher erwähnten Effekte von nicht ins Modell einbezogenen Einflussgrössen können auch dazu führen, dass eine ursächliche Wirkung durch indirekte Zusammenhänge gerade **kompensiert** wird (vergleiche das Beispiel!).
- Der Einfluss einer erklärenden Grösse kann nicht-linear sein. Dann kann man mit einer geeigneten Transformation (4.4, 4.6.c) oder mit Zusatztermen (4.6.d) zu einem genaueren Modell kommen.

l Die **am klarsten interpretierbare Antwort** auf die Frage nach einer Wirkung einer erklärenden Variablen auf die Zielgrösse erreicht man also, wenn man

- in einem geeignet **geplanten Versuch** die Variable gezielt verändert.

... oder, falls das nicht geht,

- möglichst alle denkbaren ursächlichen Grössen ins Modell aufnimmt,
- die Linearität der Zusammenhänge überprüft (siehe 4.4, 4.2.h),
- ein *Vertrauensintervall* für den Koeffizienten liefert – statt eines P-Wertes. Dieses gibt bei fehlender Signifikanz an, wie gross der Effekt dennoch sein könnte.

m Indirekte Effekte, wie sie hier als Gründe für falsche Interpretationen angeführt wurden, können nicht vorkommen, wenn die **erklärenden Grössen** selbst **nicht zusammenhängen** – wenigstens nicht linear – genauer: wenn sie „orthogonal“ sind. Wir könnten von *unkorreliert* reden, wenn die erklärenden Grössen Zufallsvariable wären. „Orthogonal“ heisst also: wenn wir trotz allem die empirische Korrelation zwischen den Variablen ausrechnen, so erhalten wir null. Wir kommen auf die Schwierigkeiten von „korrelierten“ erklärenden Variablen in 5.4 zurück.

Wenn das möglich ist – namentlich bei geplanten Versuchen – ist deshalb sehr zu empfehlen, die $x_i^{(j)}$-Werte so zu wählen, dass die Orthogonalität erfüllt wird. Näheres wird in der Versuchsplanung besprochen.

n Wenn alle erklärenden Variablen in diesem Sinne orthogonal zueinander sind, dann kann man zeigen, dass die *Schätzungen* der Koeffizienten der einfachen Regressionen genau die geschätzten Werte des multiplen Modells geben müssen. Trotzdem lohnt sich das multiple Modell, da die geschätzte Standardabweichung der Fehler kleiner wird und daduch die **Vertrauensintervalle kürzer** und die **Tests eher signifikant** werden.

o Zusammenfassend: Ein multiples Regressionsmodell sagt mehr aus als viele einfache Regressionen – im Falle von korrelierten erklärenden Variablen sogar **viel mehr**.

3.4 Modell und Schätzungen in Matrix-Schreibweise

a Es ist Zeit, wieder etwas Theorie zu behandeln. Es wird sich lohnen, auch für praktisch orientierte Leute. Sie wollen ja nicht nur Rezepte auswendig lernen. Für Rezepte gibt es Bücher. Theorie stellt Zusammenhänge her. Etliche Probleme, die in der praktischen Anwendung der Regression auftreten können, lassen sich mit Hilfe der Theorie besser verstehen.

Die Theorie, die hier folgt, zeigt die Nützlichkeit von Linearer Algebra, von Matrizen und Vektoren. Sie werden die hier eingeführten Begriffe und Methoden in der multivariaten Statistik und bei den Zeitreihen wieder antreffen.

Bevor wir zufällige Vektoren und Matrizen betrachten, empfiehlt es sich, die gewöhnliche Vektor- und Matrixalgebra in Erinnerung zu rufen. Was für die folgenden Abschnitte wichtig ist, fasst Anhang 3.A zusammen.

b Das Modell der multiplen Regression, $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i$, wollen wir mit Hilfe von Vektoren und Matrizen formulieren.

Dazu müssen wir zuerst den Begriff des „Vektors von Zufallsvariablen“ oder der „vektoriellen Zufallsvariablen“ oder des „**Zufallsvektors**“ einführen: Es handelt sich einfach um eine Zusammenfassung von mehreren Zufallsvariablen,

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \text{und} \qquad \underline{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} .$$

Man verwendet also Spaltenvektoren. (Drucktechnisch platzsparender wären Zeilenvektoren, und deshalb schreibt man oft den transponierten Vektor hin, $\underline{Y} = [Y_1, \ldots, Y_n]^T$; ^{T} steht für transponiert.)

c Die Koeffizienten β_j können wir auch als Vektor schreiben, und die erklärenden Variablen $x_i^{(j)}$ zu einer Matrix zusammenfassen:

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \qquad \text{und} \qquad \boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ \vdots & & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{bmatrix} .$$

Schliesslich brauchen wir noch den Vektor, der aus lauter Einsen besteht, $\underline{1} = [1, 1, \ldots, 1]^T$.

Jetzt wird das Regressionsmodell einfach zu

$$\underline{Y} = \beta_0 \underline{1} + \boldsymbol{X}\,\underline{\beta} + \underline{E} .$$

Was heisst das? Auf beiden Seiten des Gleichheitszeichens stehen Vektoren. Das i-te Element des Vektors rechts ist $\beta_0 \cdot 1 + \sum_j \beta_j x_i^{(j)} + E_i$, und das ist laut Modell gleich dem i-ten Element von $\underline{Y}$.

d Die Vektor-Gleichung ist noch nicht ganz einfach genug! Damit β_0 noch verschwindet, erweitern wir $\boldsymbol{X}$ um eine Kolonne von Einsen und $\underline{\beta}$ um das Element β_0:

$$\widetilde{\boldsymbol{X}} = [\underline{1}\ \boldsymbol{X}\ ] = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ & \vdots & & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{bmatrix} \qquad \widetilde{\underline{\beta}} = \begin{bmatrix} \beta_0 \\ \underline{\beta} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

Jetzt gilt

$$\underline{Y} = \widetilde{\boldsymbol{X}}\,\widetilde{\underline{\beta}} + \underline{E} .$$

Wenn das Modell keinen Achsenabschnitt enthält, setzen wir $\widetilde{\boldsymbol{X}} = \boldsymbol{X}$ und $\widetilde{\underline{\beta}} = \underline{\beta}$.

e Auf das Modell folgt die **Schätzung**. In der einfachen Regression haben wir das Prinzip der Kleinsten Quadrate angewandt. Die **Residuen**, die zu einem Parameter-Vektor $\widetilde{\underline{\beta}}^*$ gehören, sind

$$R_i = Y_i - (\beta_0^* + \sum\nolimits_j \beta_j^* x_i^{(j)}) .$$

Wir können auch sie zu einem Vektor zusammenfassen und erhalten

$$\underline{R} = \underline{Y} - \widetilde{\boldsymbol{X}}\,\widetilde{\underline{\beta}}^* .$$

(Wenn $\widetilde{\underline{\beta}}^* = \widetilde{\underline{\beta}}$ ist, sind die R_i gerade die Zufalls-Fehler E_i.)

Die Summe der Quadrate $\sum_i R_i^2$ kann man schreiben als

$$Q\langle\underline{\widetilde{\beta}}^*\rangle = \sum\nolimits_i R_i^2 = \underline{R}^T\underline{R}$$

(und das ist auch die quadrierte Norm des Vektors $\underline{R}$). Diesen Ausdruck wollen wir also minimieren. Dass dies aus dem Prinzip der Maximalen Likelihood folgt, wurde in 2.A.0.a gezeigt.

f Wir wollen dasjenige $\underline{\widetilde{\beta}}^*$ finden, für das $Q\langle\underline{\widetilde{\beta}}^*\rangle$ minimal wird, und es als Schätzung von $\underline{\widetilde{\beta}}$ verwenden. Eine klare Schreibweise für diese Aufgabe, die man vermehrt verwenden sollte, ist

$$\underline{\widehat{\beta}} = \arg\min_{\underline{\widetilde{\beta}}} \left\langle Q\langle\underline{\widetilde{\beta}}\rangle \right\rangle \ .$$

Minimieren läuft oft über Ableiten und null Setzen. Man kann Regeln für Ableitungen von und nach Vektoren herleiten und einsetzen. Wir kommen aber auch mit gewöhnlichen Ableitungen durch, wenn es auch etwas mühsam wird. Es ist

$$\partial Q\langle\underline{\widetilde{\beta}}\rangle/\partial\beta_j = \sum\nolimits_i \partial R_i^2/\partial\beta_j = 2\sum\nolimits_i R_i \partial R_i/\partial\beta_j$$

und

$$\partial R_i/\partial\beta_j = \partial\left(Y_i - (\beta_0 + \sum\nolimits_j \beta_j x_i^{(j)})\right)\Big/\,\partial\beta_j = -x_i^{(j)}$$

(wenn man $x_i^{(0)} = 1$ setzt, gilt dies auch für $j = 0$), also

$$\partial Q\langle\underline{\widetilde{\beta}}\rangle/\partial\beta_j = -2\sum\nolimits_i R_i x_i^{(j)} = -2\,(\widetilde{\boldsymbol{X}}^T\underline{R})_j \ .$$

Die Ableitungen (für $j = 0, 1, \ldots, m$) sollen gleich 0 sein.

g Das können wir gleich als Vektor hinschreiben, $\widetilde{\boldsymbol{X}}^T\underline{R} = \underline{0}$. Einsetzen führt zu

$$\widetilde{\boldsymbol{X}}^T(\underline{Y} - \widetilde{\boldsymbol{X}}\,\underline{\widehat{\beta}}) = \underline{0} \quad \Rightarrow \quad \widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\,\underline{\widehat{\beta}} = \widetilde{\boldsymbol{X}}^T\underline{Y} \ .$$

Die letzte Gleichung hat einen Namen: Sie heisst „die **Normal-Gleichung*en***“ – es sind ja p Gleichungen, in eine Vektoren-Gleichung verpackt.

Links steht eine quadratische, symmetrische Matrix,

$$\boldsymbol{C} = \widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} \ ,$$

multipliziert mit dem gesuchten Vektor $\underline{\widehat{\beta}}$, rechts ein Vektor, $\widetilde{\boldsymbol{X}}^T\underline{Y}$.

Bei der Auflösung dieser Gleichung macht sich die lineare Algebra erstmals richtig bezahlt: Wir multiplizieren die Gleichung von links mit der Inversen von $\boldsymbol{C}$, $\boldsymbol{C}^{-1}$, und erhalten

$$\underline{\widehat{\beta}} = \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T\underline{Y} \ .$$

h Dazu müssen wir voraussetzen, dass $\boldsymbol{C}$ invertierbar oder nicht-singulär (oder regulär oder von vollem Rang) ist. Sonst? Sonst ist die Lösung des Problems der Kleinsten Quadrate nicht eindeutig, und man muss mit komplizierteren Methoden dahintergehen (mit verallgemeinerten Inversen).

Das Prinzip der Kleinsten Quadrate führt also nicht immer zu einer eindeutigen Lösung.

Das ist nicht nur ein theoretisches Problem! Wenn $\boldsymbol{C}$ nicht invertierbar ist, heisst das, dass das Regressions-Modell selbst schlecht formuliert ist, dass nämlich die Parameter nicht eindeutig sind, also verschiedene Parameter-Kombinationen genau das gleiche Modell festlegen. Man spricht von **nicht identifizierbaren Parametern**. Das Modell wird dann besser so geändert, dass man wieder eindeutig weiss, was ein Parameter bedeuten soll. (Einen solchen Fall haben wir in 3.2.g angetroffen.)

Das Problem kann auch „fast“ auftreten. Wir kommen darauf unter dem Stichwort „Kollinearität“ zurück (5.3.m).

i Schreiben Sie die letzte Formel für die einfache lineare Regression (2.2.c) auf und zeigen Sie, dass sie mit 2.2.c übereinstimmt! Das ist nützlich, um die allgemeinere Formel besser zu verstehen und um etwas lineare Algebra zu üben.

3.5 Verteilung der geschätzten Regressionskoeffizienten

a Die geschätzten Regressionskoeffizienten lassen sich also in Matrixform sehr kurz schreiben,

$$\underline{\widehat{\beta}} = \widetilde{\boldsymbol{C}}\underline{Y} \,, \qquad \widetilde{\boldsymbol{C}} = \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T \,.$$

Wenn wir jetzt ein Element $\widehat{\beta}_j$ des Vektors $\underline{\widehat{\beta}}$ herausgreifen, so lässt sich dieses also auch als Summe ausdrücken,

$$\widehat{\beta}_j = \sum\nolimits_{i=1}^{n} \widetilde{\boldsymbol{C}}_{ji} Y_i \,.$$

Die $\widetilde{\boldsymbol{C}}_{ji}$ sind feste Zahlen, die Y_i Zufallsvariable. Wie in der Einführung über Wahrscheinlichkeitsrechnung gezeigt wird, ist eine solche „Linearkombination“ von normalverteilten Zufallsvariable wieder normalverteilt, und es bleibt noch, den Erwartungswert und die Varianz zu bestimmen.

b Der Erwartungswert ist gemäss der allgemeinen Formel $\mathcal{E}\langle\sum_i a_i Y_i\rangle = \sum_i a_i \, \mathcal{E}\langle Y_i\rangle$ gleich

$$\mathcal{E}\langle\widehat{\beta}_j\rangle = \sum\nolimits_{i=1}^{n} \widetilde{\boldsymbol{C}}_{ji} \, \mathcal{E}\langle Y_i\rangle = \sum\nolimits_{i=1}^{n} \widetilde{\boldsymbol{C}}_{ji} \sum\nolimits_k X_i^{(k)} \beta_k \,.$$

Das sieht sehr kompliziert aus. Wir nehmen wieder die Matrixrechnung zu Hilfe. Die Doppelsumme ist gleich dem $j$ten Element von

$$\widetilde{\boldsymbol{C}}\,\boldsymbol{X}\underline{\beta} = \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{X}\underline{\beta} = \boldsymbol{C}^{-1}\boldsymbol{C}\underline{\beta} = \underline{\beta} \,,$$

also gleich β_j.

c Für die Varianz einer Summe von unabhängigen Zufallsvariablen lautet die allgemeine Formel $\mathrm{var}\langle\sum_i a_i Y_i\rangle = \sum_i a_i^2 \, \mathrm{var}\langle Y_i\rangle$. Einsetzen ergibt

$$\mathrm{var}\langle\widehat{\beta}_j\rangle = \sum\nolimits_{k=1}^{n} \left(\widetilde{\boldsymbol{C}}_{jk}\right)^2 \mathrm{var}\langle Y_k\rangle = \sigma^2 \sum\nolimits_{k=1}^{n} \left(\widetilde{\boldsymbol{C}}_{jk}\right)^2 \,.$$

Die Summe der Quadrate ist gleich dem $j$ten Diagonalelement von

$$\begin{aligned} \widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{C}}^T &= \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T(\boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T)^T = \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}(\boldsymbol{C}^{-1})^T \\ &= \boldsymbol{C}^{-1}\boldsymbol{C}(\boldsymbol{C}^{-1})^T = (\boldsymbol{C}^{-1})^T \,. \end{aligned}$$

Da $\boldsymbol{C}$ symmetrisch ist (und wir sowieso nur die Diagonalelemente betrachten), kann man das Transponieren weglassen. Also ist

$$\mathrm{var}\langle\widehat{\beta}_j\rangle = \sigma^2 \left(\boldsymbol{C}^{-1}\right)_{jj} \,.$$

d Mit etwas mehr Theorie kann man auch Kovarianzen zwischen den geschätzten Koeffizienten $\widehat{\beta}_j$ erhalten. Diese Überlegungen gehören zum Thema der Multivariaten Statistik und werden im entsprechenden Block behandelt.

3.A Anhang: Grundbegriffe der Linearen Algebra

a **Matrizen.** Matrix, genauer $n \times m$-Matrix:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}$$

Zeilen $i = 1, ..., n$, Spalten $j = 1, ..., m$. Elemente a_{ij}.

Quadratische Matrix: Gleiche Anzahl Zeilen und Spalten, $n = m$.

Symmetrische Matrix: Es gilt $a_{ij} = a_{ji}$.

Diagonale einer quadratischen Matrix: Die Elemente $[a_{11}, a_{22}, ..., a_{nn}]$.

Diagonalmatrix: Eine, die „nur aus der Diagonalen besteht“, $d_{ij} = 0$ für $i \neq j$.

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix}$$

b **Transponierte Matrix**: Wenn man Zeilen und Spalten einer Matrix $\boldsymbol{A}$ vertauscht, erhält man die transponierte Matrix $\boldsymbol{A}^T$:

$$\boldsymbol{A}^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{bmatrix}$$

Bemerkungen:

1. Es gilt offensichtlich $(\boldsymbol{A}^T)^T = \boldsymbol{A}$ (vgl. die zweimal gewendete Matratze).
2. Für symmetrische Matrizen gilt $\boldsymbol{A}^T = \boldsymbol{A}$.

c **Vektoren.** Vektor, genauer Spaltenvektor: n Zahlen, unter einander geschrieben.

$$\underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ : \\ b_n \end{bmatrix}$$

Elemente b_i.

d **Transponierte Vektoren**: Spaltenvektoren werden zu Zeilenvektoren, wenn man sie transponiert:

$$\underline{b}^T = \begin{bmatrix} b_1 \\ b_2 \\ : \\ b_n \end{bmatrix}^T = [b_1, b_2, ..., b_n] \, .$$

Drucktechnisch platzsparender als Spaltenvektoren sind Zeilenvektoren, und deshalb schreibt man Spaltenvektoren oft als transponierte Zeilenvektoren hin: $\underline{b} = [b_1, b_2, ..., b_n]^T$.

e **Einfache Rechenoperationen.** Addition und Subtraktion: Geht nur bei gleichen Dimensionen. Man addiert oder subtrahiert die einander entsprechenden Elemente.

Multiplikation mit einer Zahl (einem „Skalar“): Jedes Element wird multipliziert. Division durch eine Zahl ebenso.

Recht oft trifft man in der Statistik und anderswo auf so genannte **Linearkombinationen** von Vektoren. Das ist ein schöner Name für Ausdrücke der Form

$$\lambda_1 \underline{b}_1 + \lambda_2 \underline{b}_2$$

+ eventuell weitere solche Terme – man addiert Vielfache der beteiligten Vektoren.

f **Matrix-Multiplikation.** Matrizen können nur multipliziert werden, wenn die Dimensionen passen: $\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B}$ ist definiert, wenn die Anzahl Spalten von $\boldsymbol{A}$ gleich der Anzahl Zeilen von $\boldsymbol{B}$ ist. Dann ist

$$c_{ik} = \sum\nolimits_{j=1}^{m} a_{ij} b_{jk}$$

Beispiel:

$$\begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 1 \\ 4 & -2 \end{bmatrix} = \begin{bmatrix} 2\cdot 3 + 1 \cdot 4 & 2 \cdot 1 + 1 \cdot (-2) \\ (-1)\cdot 3 + 0 \cdot 4 & (-1) \cdot 1 + 0 \cdot (-2) \\ 3 \cdot 3 + 1 \cdot 4 & 3 \cdot 1 + 1 \cdot (-2) \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ -3 & -1 \\ 13 & 1 \end{bmatrix}$$

Bemerkungen:

1. Im Beispiel ist $\boldsymbol{B} \cdot \boldsymbol{A}$ nicht definiert, da $\boldsymbol{B}$ 2 Spalten, $\boldsymbol{A}$ aber 3 Zeilen hat.
2. Wenn $\boldsymbol{A} \cdot \boldsymbol{B}$ und $\boldsymbol{B} \cdot \boldsymbol{A}$ beide definiert sind, sind die beiden im allgemeinen verschieden, $\boldsymbol{A} \cdot \boldsymbol{B} \neq \boldsymbol{B} \cdot \boldsymbol{A}$! Matrizen dürfen nicht vertauscht werden.
3. Es kann $\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{0}$ sein, obwohl weder $\boldsymbol{A} = \boldsymbol{0}$ noch $\boldsymbol{B} = \boldsymbol{0}$ ist.
4. Es gilt das Assoziativgesetz: $(\boldsymbol{A} \cdot \boldsymbol{B}) \cdot \boldsymbol{C} = \boldsymbol{A} \cdot (\boldsymbol{B} \cdot \boldsymbol{C})$
5. Es gilt das Distributivgesetz: $\boldsymbol{A} \cdot (\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{A} \cdot \boldsymbol{B} + \boldsymbol{A} \cdot \boldsymbol{C}$ und ebenso $(\boldsymbol{A} + \boldsymbol{B}) \cdot \boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{C} + \boldsymbol{B} \cdot \boldsymbol{C}$.
6. Transponieren eines Produktes: Es ist

   $$(\boldsymbol{A} \cdot \boldsymbol{B})^T = \boldsymbol{B}^T \cdot \boldsymbol{A}^T$$

   Man muss also beim Transponieren die Reihenfolge vertauschen!
7. Das Produkt $\boldsymbol{A} \cdot \boldsymbol{A}^T$ ist immer symmetrisch.

g All das gilt auch für Vektoren: Wenn $\underline{a}$ und $\underline{b}$ Spaltenvektoren sind, ist

$$\underline{a} \cdot \underline{b}^T = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \dots & a_1 b_m \\ a_2 b_1 & a_2 b_2 & \dots & a_2 b_m \\ \vdots & & & \vdots \\ a_n b_1 & a_n b_2 & \dots & a_n b_m \end{bmatrix} .$$

Wenn sie gleiche Länge haben, ist

$$\underline{a}^T \cdot \underline{b} = \sum\nolimits_i a_i \cdot b_i \, .$$

„Matrix mal Spaltenvektor“ ergibt (falls definiert) einen Spaltenvektor: $\boldsymbol{A} \cdot \underline{b} = \underline{c}$.

h Die **Länge eines Vektors** ist die Wurzel aus $\sum_i a_i^2$. Man bezeichnet sie oft mit $\|\underline{a}\|$. Man kann schreiben

$$\|\underline{a}\|^2 = \underline{a}^T \cdot \underline{a} \, .$$

i Die **Einheitsmatrix** (der Dimension m) ist definiert als Diagonalmatrix mit lauter Einsen:

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Sie lässt bei Multiplikation Matrizen unverändert: $\boldsymbol{I} \cdot \boldsymbol{A} = \boldsymbol{A}$, $\boldsymbol{A} \cdot \boldsymbol{I} = \boldsymbol{A}$.

j **Inverse Matrix.** Wenn $\boldsymbol{A}$ quadratisch ist und $\boldsymbol{B} \cdot \boldsymbol{A} = \boldsymbol{I}$ gilt, heisst $\boldsymbol{B}$ die zu $\boldsymbol{A}$ inverse Matrix; man schreibt $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

Bemerkungen:

1. Es gilt dann auch $\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{I}$. Wenn also $\boldsymbol{B} = \boldsymbol{A}^{-1}$ ist, ist auch $\boldsymbol{A} = \boldsymbol{B}^{-1}$.
2. Es gibt nicht zu jeder quadratischen Matrix $\boldsymbol{A}$ eine Inverse. Wenn es eine gibt, heisst $\boldsymbol{A}$ **regulär**, und es gibt nur eine Inverse. Wenn es keine Inverse gibt, heisst $\boldsymbol{A}$ **singulär**.
3. Es ist $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
4. Inverses eines Matrix-Produkts: Wenn $\boldsymbol{A}$ und $\boldsymbol{B}$ quadratisch sind, ist

   $$(\boldsymbol{A} \cdot \boldsymbol{B})^{-1} = \boldsymbol{B}^{-1} \cdot \boldsymbol{A}^{-1}$$

   Die Reihenfolge muss also vertauscht werden, wie beim Transponieren!
5. Es ist $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$. Man schreibt oft kurz $\boldsymbol{A}^{-T}$.

k **Lineares Gleichungssystem.** Kurz zusammengefasst: Das Gleichungssystem

$$\begin{aligned} a_{11}\beta_1 + a_{12}\beta_2 + ... + a_{1m}\beta_m &= y_1 \\ a_{21}\beta_1 + a_{22}\beta_2 + ... + a_{2m}\beta_m &= y_2 \\ \dots & .. \\ a_{m1}\beta_1 + a_{m2}\beta_2 + ... + a_{mm}\beta_m &= y_m \end{aligned}$$

(für die β_j) lässt sich schreiben als

$$\boldsymbol{A}\underline{\beta} = \underline{y}$$

(für $\underline{\beta}$). Es hat genau eine Lösung, wenn $\boldsymbol{A}$ regulär ist, also wenn die Inverse $\boldsymbol{A}^{-1}$ existiert. Dann ist

$$\underline{\beta} = \boldsymbol{A}^{-1}\underline{y}$$

diese Lösung.

l Wenn die **Matrix $\boldsymbol{A}$ singulär** ist, dann gibt es eine Zeile $[a_{i1}, a_{i2}, ..., a_{im}]$, die sich als Linearkombination der andern schreiben lässt. Die entsprechende Gleichung führt entweder zu einem Widerspruch (keine Lösung) oder ist überflüssig (unendlich viele Lösungen). Man spricht von **linearer Abhängigkeit** der Zeilen der Matrix oder der Gleichungen.

(Wenn die Matrix singulär ist, gibt es auch eine Spalte, die sich als Linearkombination der andern schreiben lässt. Es sind also auch die Spaltenvektoren linear abhängig.)

3.S S-Funktionen

a **Modell-Formeln** dienen dazu, Modelle von Regressionen und Varianzanalysen aller Art und auch Modelle der multivariaten Statistik festzulegen. Sie sind dadurch gekenntzeichnet, dass sie das Zeichen $\sim$ enthalten. Solche Ausdrücke bilden eine spezielle Klasse von S-Objekten, genannt `formula`-Objekte. Regressions- und Varianzanalyse-Funktionen verlangen jeweils als erstes Argument eine solche `formula`.

Bei Regressions- und Varianzanalyse-Modellen steht links von diesem Zeichen die Zielgrösse und rechts die Eingangsgrössen. In der einfachsten Form lautet ein multiples Regressionsmodell

```
y ~ x1 + x2
```

Das Zeichen + erhält hier eine neue Bedeutung. Es werden nicht `x1` und `x2` zusammengezählt, sondern die beiden Variablen werden als Eingangsvariable im Modell erkannt. In mathematischer Schreibweise entsteht also der Ausdruck β_1`x1` + β_2`x2`. Automatisch wird ein Fehlerterm $+E$ hinzugefügt. Ebenso ein **Achsenabschnitt** β_0, wenn man ihn nicht ausdrücklich unterdrückt, indem man -1 einfügt, also beispielsweise `y ~ -1 + x1 + x2` schreibt. So entspricht also der Ausdruck `y ~ x1 + x2` dem Regressionsmodell

$$\mathrm{y}_i = \beta_1 \mathrm{x1}_i + \beta_2 \mathrm{x2}_i + E_i \ .$$

Wie schon in 2.S.0.c erwähnt, können **Transformationen** direkt in die Formel geschrieben werden,

```
log10(ersch) ~ log10(dist) + log10(ladung)
```

b **Faktoren** oder nominale Eingangsgrössen können (wie in 3.2.j erwähnt) ebenfalls direkt in die S-Formel geschrieben werden. Die Regressionsfunktion verwandelt solche Variable zuerst in die entsprechende Anzahl von Dummy-Variablen (3.2.h). Normalerweise sind solche Variable im `data.frame` als `factor` gekenntzeichnet und werden deshalb automatisch richtig behandelt. Wenn eine numerische Variable, beispielsweise mit den Werten 1, 2, 3, 4, als Faktor interpretiert werden soll, braucht man die Funktion `factor`. Wäre die Stelle im Beispiel in `d.spreng` nicht als Faktor gespeichert, so könnte man durch

```
log10(ersch) ~ log10(dist) + log10(ladung) + factor(St)
```

das richtige Modell dennoch erhalten.

In 3.2.g von **Nebenbedingungen** gesprochen, die nötig sind, um bei Faktoren zu einem eindeutigen Modell zu kommen. Diese können verschieden gewählt werden. Die dort erwähnte Lösung, für die einfach die erste Dummy-Variable weggelassen wird, ist die Default-Methode. Eine andere, die für die Interpretation nützlich ist, erhält man über das Argument `contrasts="sum"`. Genauer wird dies in der Varianzanalyse diskutiert.

c **Wechselwirkungen** zwischen Variablen (3.2.t) können in der `formula` ebenfalls einfach angegeben werden, und zwar mit einem Ausdruck der Form `x1:x2`,

```
log10(ersch) ~ log10(dist) + St + log10(dist):St
```

Da in den Modellen Wechselwirkungen immer nur zwischen Variablen einbezogen werden sollen, die auch als Einzelterme („Haupteffekte“ im Gegensatz zu Wechselwirkungen) auftreten, gibt es eine Kurzschreibweise. `x1*x2` bedeutet das Gleiche wie `x1+x2+x1:x2`. Das vorhergende Modell kann deshalb kurz als

```
log10(ersch) ~ log10(dist) * St
```

angegeben werden.

d Wie man sieht, erhält nicht nur das Zeichen + eine neue Bedeutung, wenn es in einer `formula` erscheint, sondern auch * und : ; sie bezeichnen Wechselwirkungen. (In der Varianzanalyse werden auch `^` und `/` für Abkürzungen üblicher Modellstrukturen benützt werden.) Manchmal möchte man aber `*` auch als Multiplikationszeichen verstanden wissen. Wenn man beispielsweise eine in cm gemessene Variable in inches ausdrücken will, braucht man 2.51*`x` als Eingangsgrösse. Man kann diese einfache Transformation mit Hilfe der Funktion `I()` angeben durch `y`∼`I(2.51*x)`.

e **Funktion `lm, summary`.** Die Funktionen `lm` und `summary` produzieren die gleichen Resultate wie in der einfachen Regression (2.S.0.g), mit zusätzlichen Zeilen in der Koeffizienten-Tabelle, die dem erweiterten Modell entsprechen.

f **Funktion `drop1`.** Wenn eine Eingangsgrösse und damit ein Term in der Modell-Formel einen Faktor beinhaltet, sind die Tests für die einzelnen Koeffizienten nicht sinnvoll. Ihre Bedeutung hängt nämlich von den Nebenbedingungen, also von den `contrasts` ab. Der sinnvolle Test, der prüft, ob der ganze Term nötig sei (3.2.m), wird von der Funktion `drop1` durchgeführt.

```
> drop1(r.lm, test="F")
```

Die Funktion berechnet primär ein Kriterium mit Namen AIC, das wir später für die Modellwahl brauchen werden (5.2.e). Wenn das Argument `test` nicht angegeben wird, wird kein Test durchgeführt.

g Einige Eigenheiten dieser „Funktionen-Familie“ erscheinen dem Autor dieser Beschreibung wenig benutzerfreundlich. Beispielsweise ist nicht einzusehen, weshalb das Objekt, das `lm` produziert, wenig Nützliches zeigt, wenn man es direkt ausgibt, und dass deshalb zuerst die generische Funktion `summary` darauf angewendet werden muss. Will man die Resultate weiter verwenden, so sind einige interessante Ergebnisse, wie die geschätzte Standardabweichung $\widehat{\sigma}$ der Fehler, nicht im Ergebnis von `lm` enthalten, sondern erst im Ergebnis von `summary(r.lm)`, und es ist nicht trivial, das herauszufinden. Leider enthält auch das `summary` nicht das, was für die Interpretation gebraucht wird. Vertrauensintervalle, standardisierte Koeffizienten und die R_j^2-Werte müssen mit zusätzlichen Funktionen ermittelt werden. Für nominale Eingangsgrössen muss, wie erwähnt, `drop1` aufgerufen werden.

Ich habe daher eine neue grundlegende Funktion geschrieben, die eine Klasse von Objekten erzeugt, welche wiederum durch verbesserte Methoden der generischen Funktionen `print` und `plot` dargestellt werden. Die neuen Funktionen beruhen selbstverständlich auf den grundlegenden Funktionen von R. (Die neue Klasse „erbt“ auch die Methoden von `lm`, soweit keine speziellen Methoden zu generischen Funktionen nötig wurden.)

h **Funktion `regr`** (package regr0). Die Funktion `regr` hat die gleichen Argumente wie `lm` (und einige mehr, da sie auch andere Regressionsmodelle anpasst). Sie erzeugt ein Objekt der Klasse `regr`, das alle interessanten Resultate der Anpassung enthält.

```
> r.regr <- regr(log10(ersch)∼log10(dist)+log10(ladung)+stelle,
  data=d.spreng)
```

Die wichtigsten Resultate sieht man durch Eintippen von

```
> r.regr
```

Das Hauptresultat ist eine Tabelle, die für alle erklärenden Variablen den Test für die Nullhypothese „kein Einfluss“ prüft. Für Variable mit einem Freiheitsgrad wird neben dem geschätzten Koeffizienten die standardisierte Version angegeben. Statt dem Standarfehler wird eine nützliche Grösse angegeben, mit der das Vertrauensintervall einfach berechnet werden kann (3.1.l).

Für Terme mit mehreren Freiheitsgraden wird in der Haupttabelle nur der F-Test angegeben. Die geschätzten Koeffizienten folgen anschliessend an die Tabelle. Sie sind direkt interpretierbar, ohne dass bekannt sein muss, mit welchen Kontrasten Faktoren codiert werden.

Weitere Vorteile der Funktion `regr` werden sich bei der Residuen-Analyse und bei den Methoden

für andere Regressionsmodelle zeigen.

i **Resultate von `regr`**

- Aufruf, mit dem das Objekt erzeugt wurde;
- „Haupttabelle“ mit den Spalten
  - `coef`: die geschätzten Koeffizienten $\widehat{\beta}_j$ für Variable mit einem einzigen Freiheitsgrad,
  - `stcoef`: die standardisierten Koeffizienten $\widehat{\beta}_j^* = \widehat{\beta}_j \cdot \mathrm{sd}\langle X^{(j)} \rangle / \mathrm{sd}\langle Y \rangle$,
  - `Rx2`: Das Mass R_j^2 für Kollinearität,
  - `df`: Anzahl Freiheitsgrade,
  - `signif`: Für Variable mit einem einzigen Freiheitsgrad wird hier die t-ratio $= T/q_{0.975}^{(t_k)}$, der Quotient aus der klassischen t-Test-Statistik und ihrer Signifikanzgrenze, angegeben. Die Nullhypothese $\beta_j = 0$ wird abgelehnt, wenn die t-ratio betragsmässig grösser als 1 ist.

    Für Faktoren und andere Terme mit mehr als einem Freiheitsgrad liefert die Spalte eine monotone Transformation der Teststatistik des F-Tests, deren Wert ebenfalls mit 1 verglichen werden kann, siehe 3.2.r.
  - `p value`: Der P-Wert für den durchgeführten Test.
- Falls Faktoren oder andere Terme mit mehr als einem Freiheitsgrad vorkommen, folgen die geschätzten Koeffizienten.
- Es folgen die Angaben über die geschätzte Standardabweichung des Zufallsterms (mit einer sinnvollen Bezeichnung!), das Bestimmtheitsmass und der Gesamt-Test.
- Falls das Argument `correlation=TRUE` gesetzt wird, folgt die Korrelationsmatrix der geschätzten Koeffizienten (siehe `summary.lm`)

j **Funktionen `residuals, fitted.`** Die Residuen und die angepassten Werte sind als Komponenten in der Resultat-Liste von `lm` oder `regr` enthalten. Man kann sie also als `t.r$residuals` resp. `t.r$fitted.values` ansprechen. Eleganter, weil auch in anderen Modellen anwendbar und im Fall von fehlenden Werten angemessen, ist die Anwendung der Funktionen („Extraktor-Funktionen“) `residuals` und `fitted` (oder synonym `resid`, `fitted.values`. Man schreibt also beispielsweise `residuals(t.r)`, um die Residuen zu erhalten. Achtung: Bei `lm` ist, wenn die Daten fehlende Werte (NA) enthalten, der Residuen-Vektor kürzer als die Daten, ausser wenn `na.action=na.replace` gesetzt wurde. Dann enthält der Residuenvektor selbst NAs für jene Beobachtungen, die für die Regressionsrechnung nicht verwendet wurden.

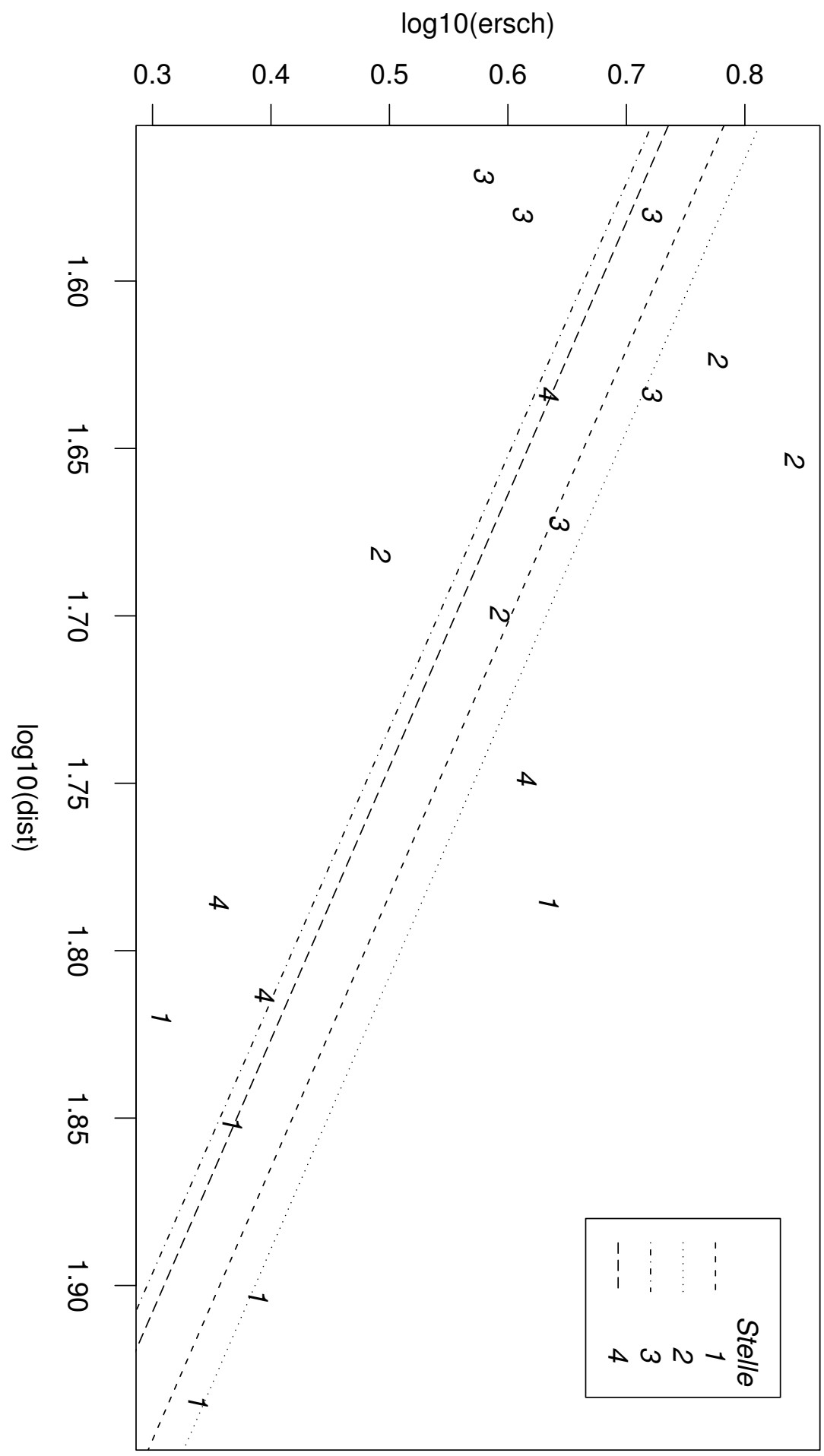

Abbildung 3.2.i: Beobachtungen und geschätzte Geraden im Beispiel der Sprengungen

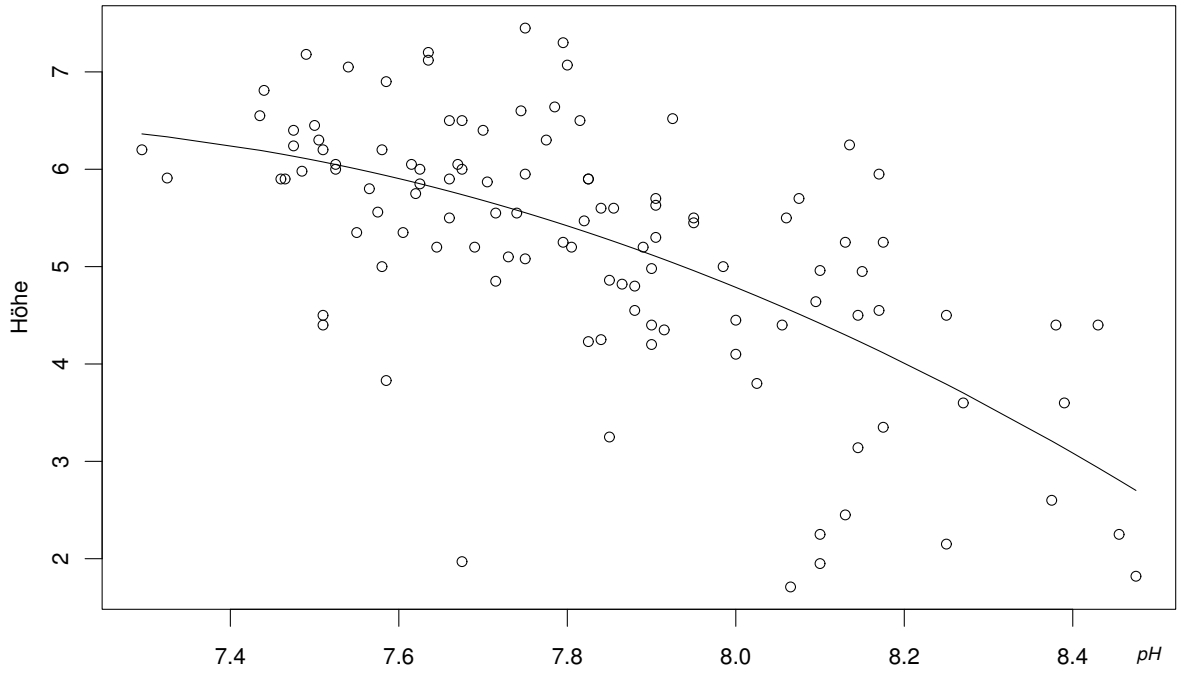

Abbildung 3.2.v: Quadratische Regression im Beispiel der basischen Böden

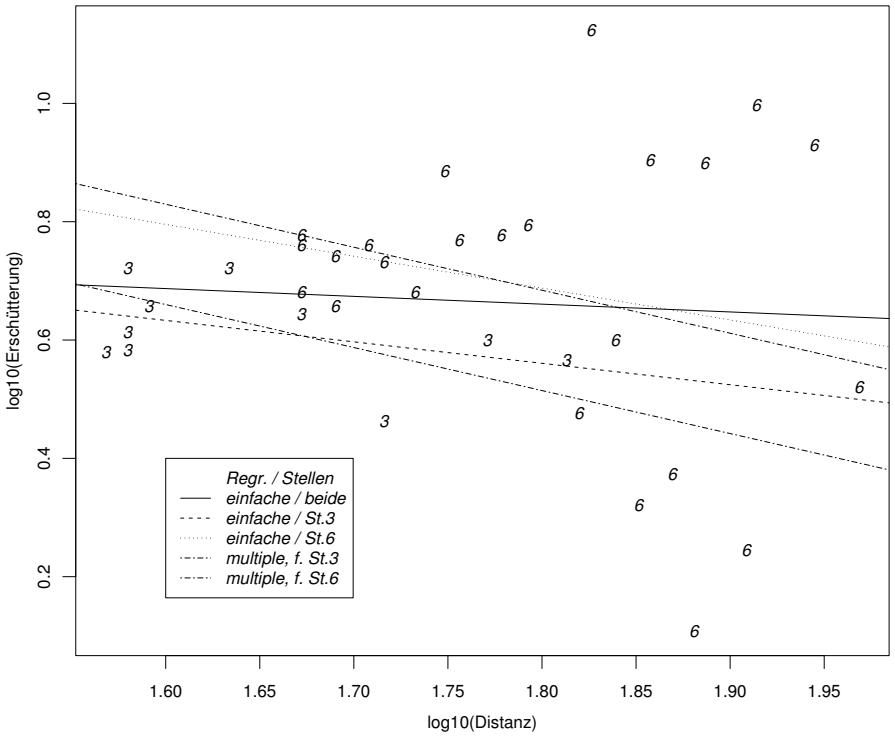

Abbildung 3.3.b: Daten des eingeschränkten Beispiels der Sprengungen (Stellen 3 und 6) mit geschätzten Regressionsgeraden: Die eingezeichneten Geraden stehen einerseits für die einfachen Regressionen, für beide Stellen zusammen wie auch separat gerechnet; andererseits erscheinen zwei parallele Geraden, die die angepassten Werte gemäss multipler Regression für eine mittlere Ladung für die beiden Stellen wiedergeben.

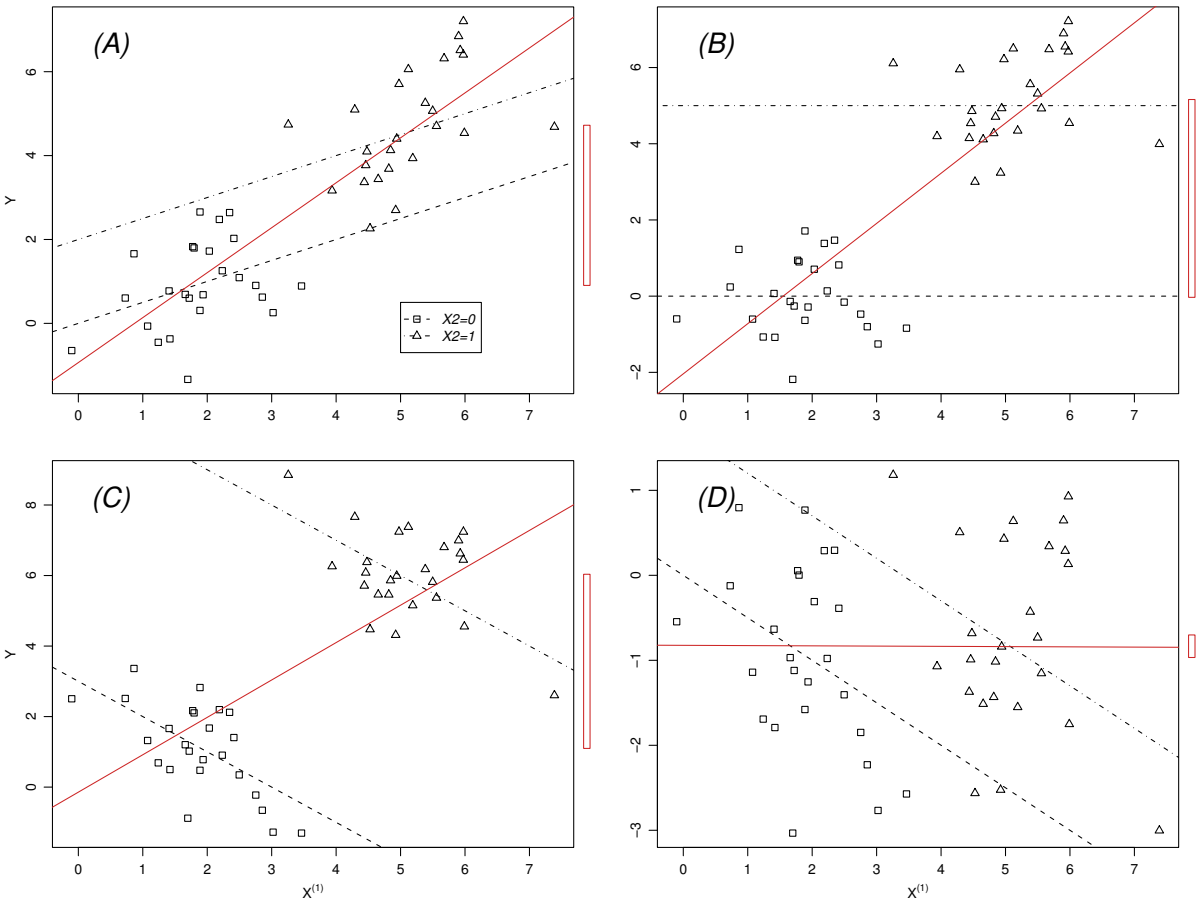

Abbildung 3.3.c: Einfache und multiple Regression für eine Gruppierungsvariable (binäre Variable) und eine kontinuierliche erklärende Variable

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:

(a) Der Erwartungswert der E_i ist $\mathcal{E}\langle E_i \rangle = 0$,

(b) sie haben alle die gleiche theoretische Varianz $\mathrm{var}\langle E_i \rangle = \sigma^2$,

(c) sie sind normalverteilt

(d) 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 Eingangs-Variable beurteilt werden. Für die multiple Regression entsteht eine ebenso anschauliche Darstellung, wenn auf der horizontalen Achse die **angepassten Werte** $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i^{(1)} + \widehat{\beta}_2 x_i^{(2)} + \ldots + \widehat{\beta}_m x_i^{(m)}$ 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 ($\widehat{y}_i$ 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 $\widehat{y}_i \approx 0.73$, der nach unten abweicht, und einen bei $\widehat{y}_i \approx 0.6$, 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** $R_i = Y_i - \widehat{y}_i$. 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 **wissen-**

schaftlichen Fragestellung ab. Wenn es um eine präzise Schätzung des Effekts einer Eingangs-Variablen auf die Zielgrösse in einem gut fundierten Modell geht, sind die Voraussetzungen kritischer, als wenn es darum geht, in einer Vielzahl von möglichen Eingangs-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.

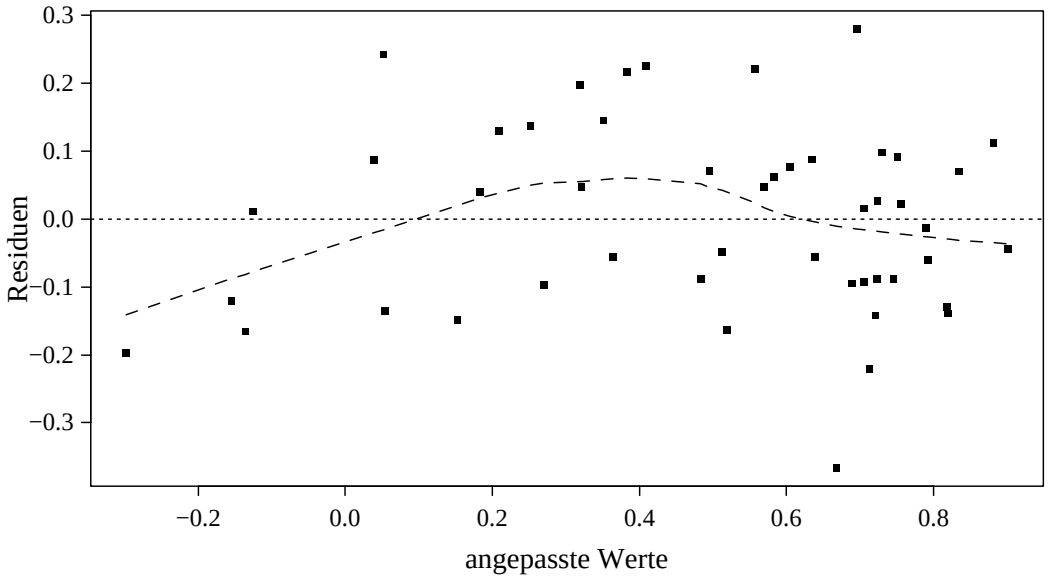

Abbildung 4.2.f: Tukey-Anscombe-Diagramm für das Beispiel der Sprengungen, mit einer Glättung und der Referenzgeraden $Y = \overline{Y}$

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**.

i Die Voraussetzung (a) lautet ja: $\mathcal{E}\langle E_i\rangle = 0$. Wenn wir nun einige Beobachtungen mit ähnlichem $\widehat{y}_i$ 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 veritkaler 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 kurze Beschreibung sollte nur die Grundidee des Glättens mit der wohl einfachsten Idee erklären. Das Verfahren kann ist leicht zu verbessern und sollte deshalb nicht verwendet werden. Genaueres zu Glättungmethoden bringt das Kapitel über „Nichtparametrische Regression".

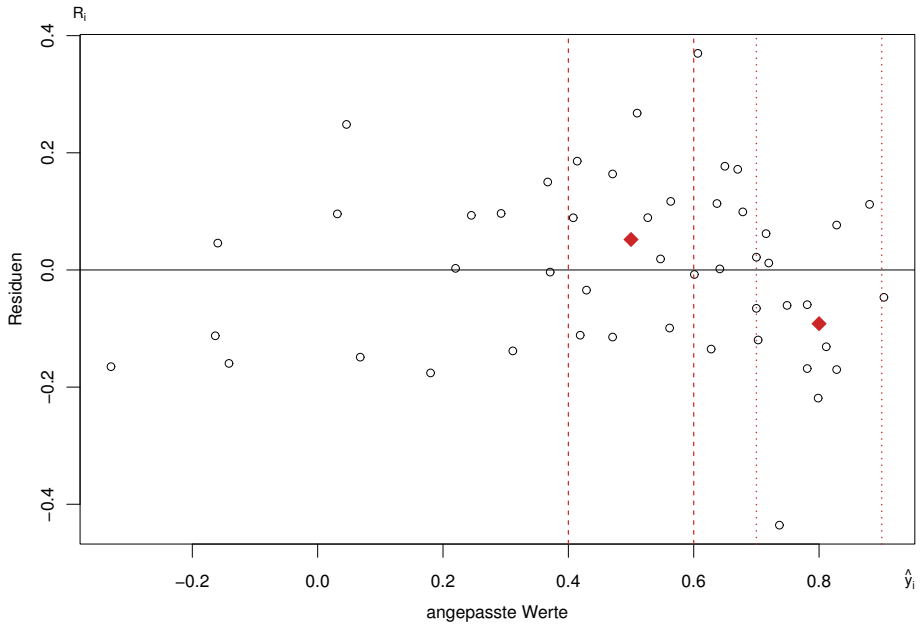

Abbildung 4.2.i: Bestimmung des gleitenden Mittels: Mittelwerte für zwei vertikale Streifen.

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 „loess".

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 benützen, 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 $Y_i^* = \widehat{y}_i + \widehat{\sigma}E_i^*$ bestimmt.

Schritt (2): Man führt die Regressionsrechnung mit den im Datensatz gegebenen Eingangs-

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 benützt 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.

l 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.

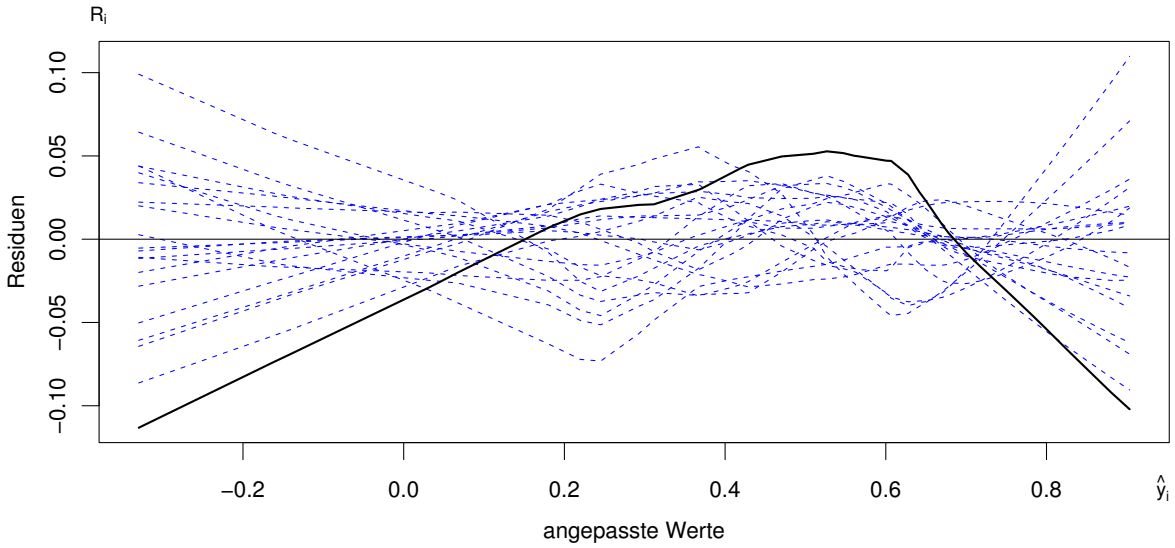

Abbildung 4.2.l: Die Glättung für die Residuen im Tukey-Anscombe-Diagramm (—) mit 19 simulierten Glättungskurven (- - -)

In Abbildung 4.2.l 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 $\widehat{y}$ 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 $\widehat{y}_i$ (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.

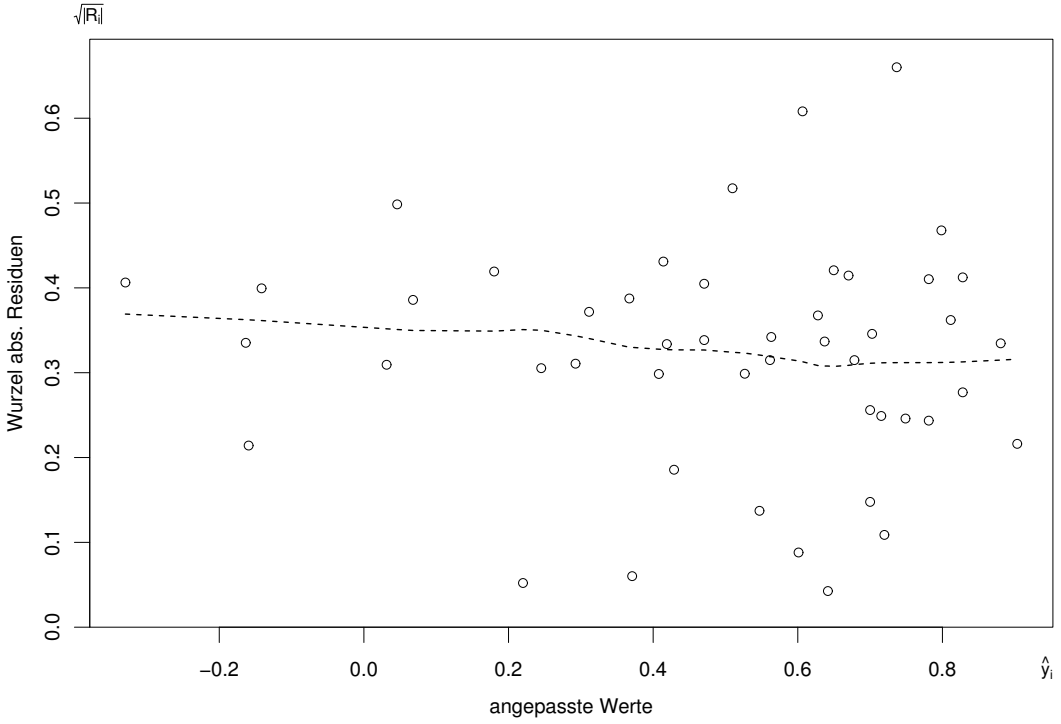

Abbildung 4.2.o: Wurzel-transformierte absolute Residuen $|R_i|$ gegen angepasste Werte im Beispiel der Sprengungen

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.

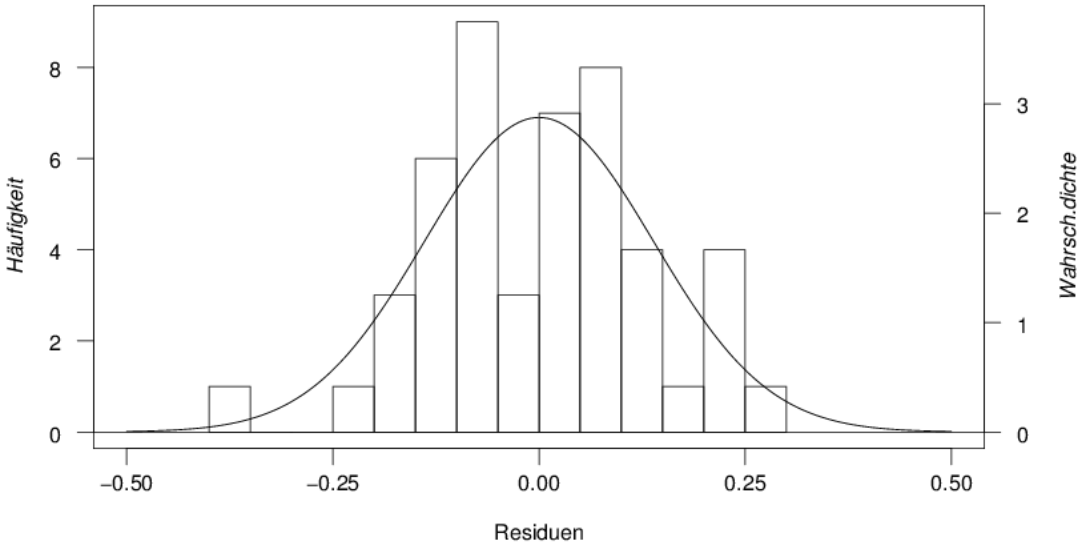

Abbildung 4.3.a: Histogramm der Residuen für das Beispiel der Sprengungen.

* Die empirische Varianz der Residuen ist nicht gleich der geschätzten Varianz $\widehat{\sigma}^2$ der Fehler, sondern gleich $(\sum R_i^2)/(n-1) = \widehat{\sigma}^2(n-p)/(n-1)$. 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 (2007), 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.

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

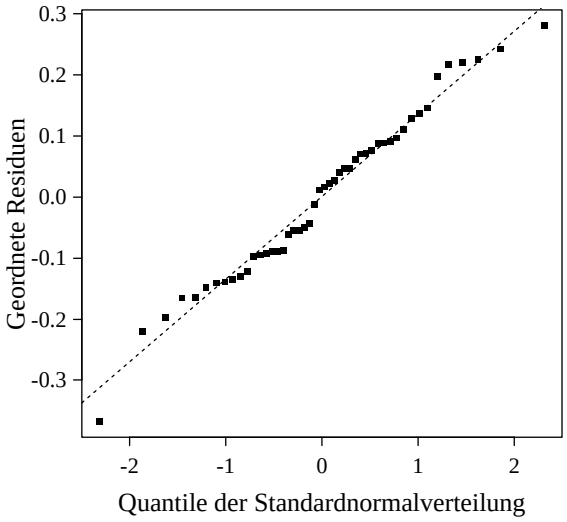

Abbildung 4.3.c: Normal plot der Residuen für das Beispiel der Sprengungen.

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.

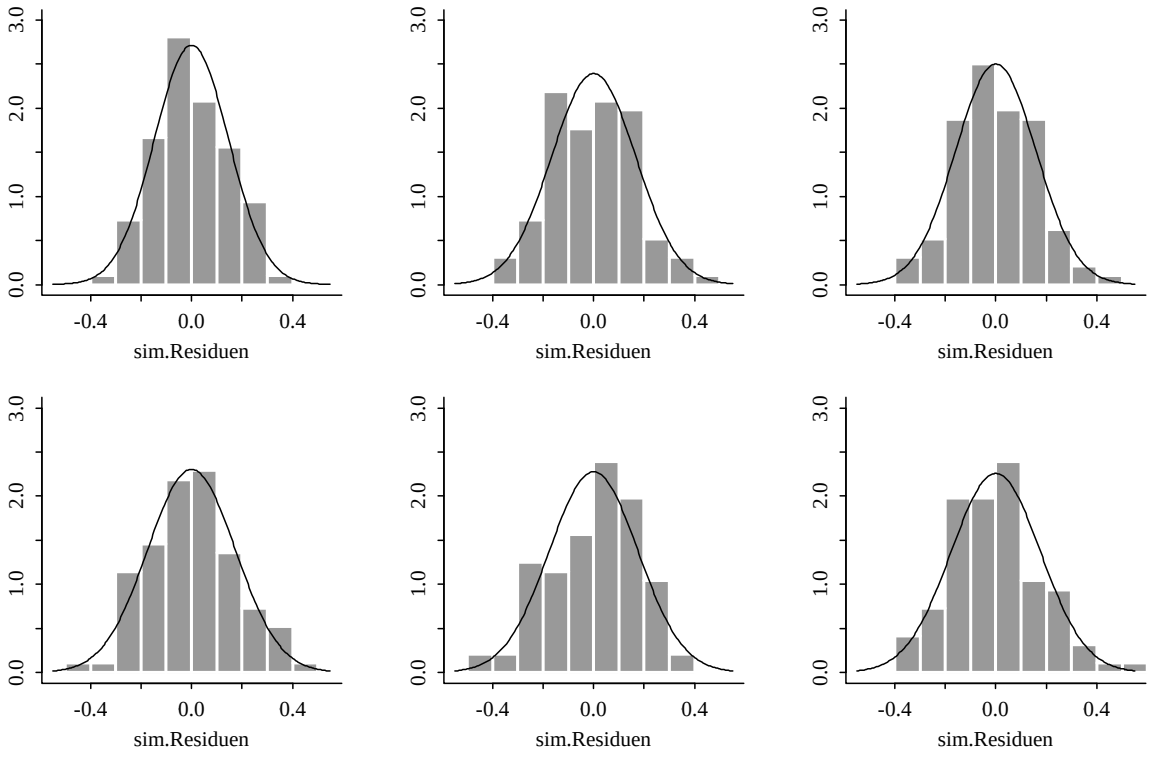

Abbildung 4.3.e (i): Histogramme von Residuen aus 6 simulierten Sätzen von Y-Werten im Beispiel der Sprengungen

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.

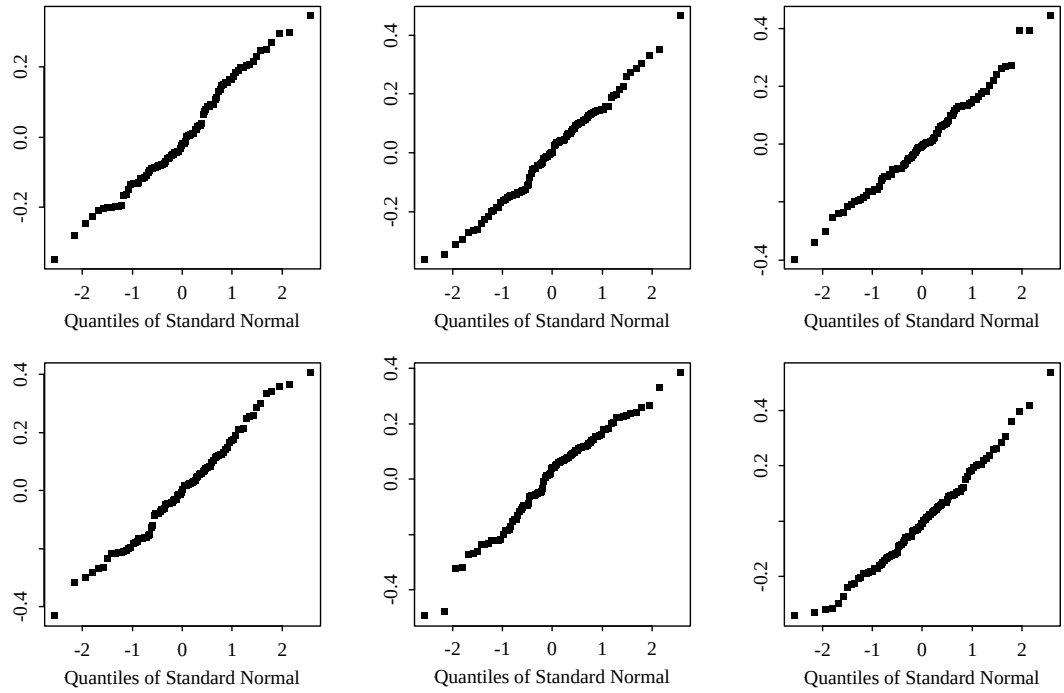

Abbildung 4.3.e (ii): Quantil-Quantil-Diagramme von Residuen aus 6 simulierten Sätzen von Y-Werten im Beispiel der Sprengungen

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 Kleinste-Quadrate-Schätzung ebenfalls. Aber sie haben nicht die gleiche **theoretische Varianz**, auch wenn die Fehler dies erfüllen; $\mathrm{var}\langle R_i\rangle$ 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

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

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 $\widehat{y}_i$. 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 $(1/n) + (x_i - \bar{x})^2/\mathrm{SSQ}^{(X)}$, 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**

$$\widetilde{R}_i = R_i \Big/ \left(\widehat{\sigma}\sqrt{1 - H_{ii}}\right)$$

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 $\mathrm{var}\langle R_i \rangle$ 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! ◁

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**

$$\widetilde{Y} = g\langle Y \rangle$$

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.

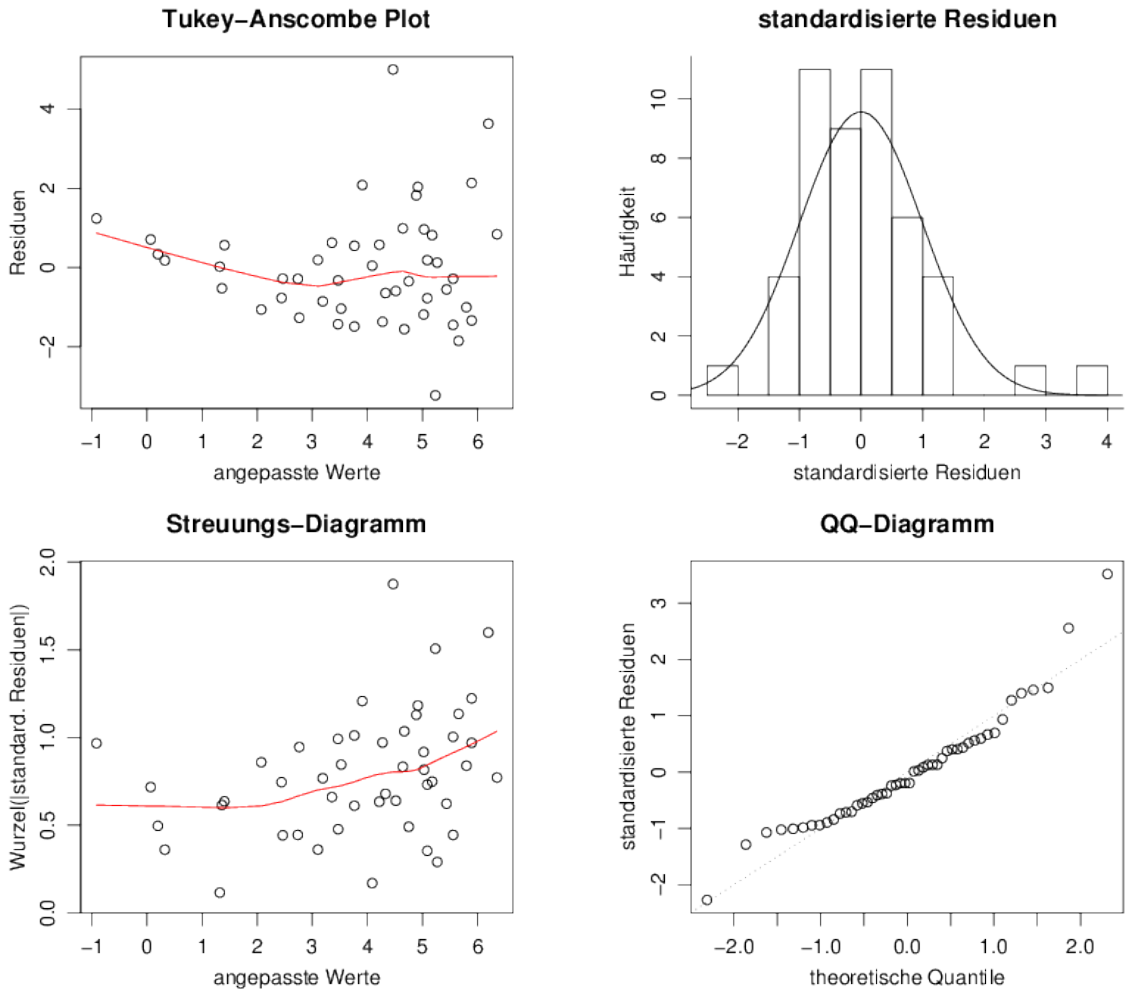

Abbildung 4.4.a: Tukey-Anscombe-Diagramm mit Streuungs-Diagramm und Histogramm und Normalverteilungs-Diagramm der standardisierten Residuen

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 Zusammenhanges 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 $\widehat{y} > 4$) 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 $\widetilde{Y} = Y^2$. Das Tukey-Anscombe-Diagramm zeigt danach keine Abweichungen von den Modellannahmen mehr. Die Residuen sind etwa symmetrisch verteilt.

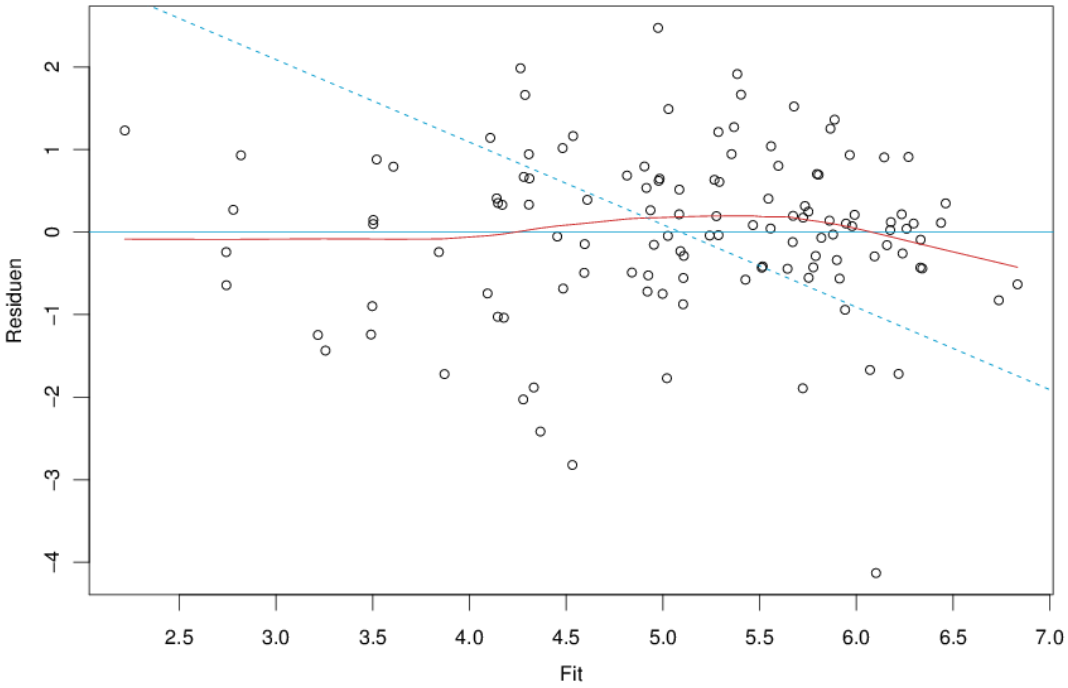

Abbildung 4.4.d: Tukey-Anscombe-Diagramm für das Beispiel der basischen Böden

* Die Transformation $\widetilde{Y} = Y^2$ 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.

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 **Zähldaten** und
- die so genannte Arcus-Sinus-Transformation $\widetilde{y} = \arcsin\sqrt{y}$ 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 Gegengründe gibt – und zwar auch für Eingangs-Variable.

h Wenn in einer einfachen Regression sowohl die Eingangs-Variable als auch die Zielgrösse Konzentrationen sind, führt die Regel zu $\widetilde{Y} = \log_{10}\langle Y\rangle$ und $\widetilde{X} = \log_{10}\langle X\rangle$. Aus $\widetilde{Y} = \alpha + \beta\widetilde{x}_i + E_i$ 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 Eingangsgrösse 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, wegfallen zu lassen, müsste zu einer systematischen Verfälschung der Resultate führen!

Die einfachste Formel zur Abänderung der Logarithmus-Funktion lautet $\widetilde{Y} = \log\langle Y + c\rangle$ 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 $c = \mathrm{med}\langle Y_k\rangle / s^{2.9}$ mit $s = \mathrm{med}\langle Y_k\rangle / q_{0.25}\langle Y_k\rangle$ 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

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

(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 $\lambda \to 0$ 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 gilt beispielsweise für die **Schadstoffe im Tunnel** (1.1.f): 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 Eingangs-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 in einer einfachen Regression die wahre Regressionsfunktion quadratisch ist (vergleiche 3.2.v), dann ergibt sich eine gekrümmte Glättung. Wenn die Funktion im Bereich der Daten ein Maximum oder ein Minimum zeigt, dann bleibt das auch erhalten, wenn man die Zielgrösse (monoton) transformiert.

Eine monotone Transformation der Zielgrösse kann einen Zusammenhang mit einer Eingangsgrösse nur dann linear machen, wenn dieser Zusammenhang selbst monoton ist. Nun sind im Tukey-Anscombe-Diagramm in vertikaler Richtung die *Residuen* 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 gleichen 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 $Y =$ 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 Eingangs-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 $\widehat{Y}$ eine **Eingangs-Variable** gewählt wird.

▷ Abbildung 4.6.a zeigt diese Streudiagramme für die zwei kontinuerlichen Eingangsgrössen im Beispiel der Sprengungen. Wieder wurden zur Beurteilung der Glättung 19 „zufällige Glättungen“ eingezeichnet. ◁

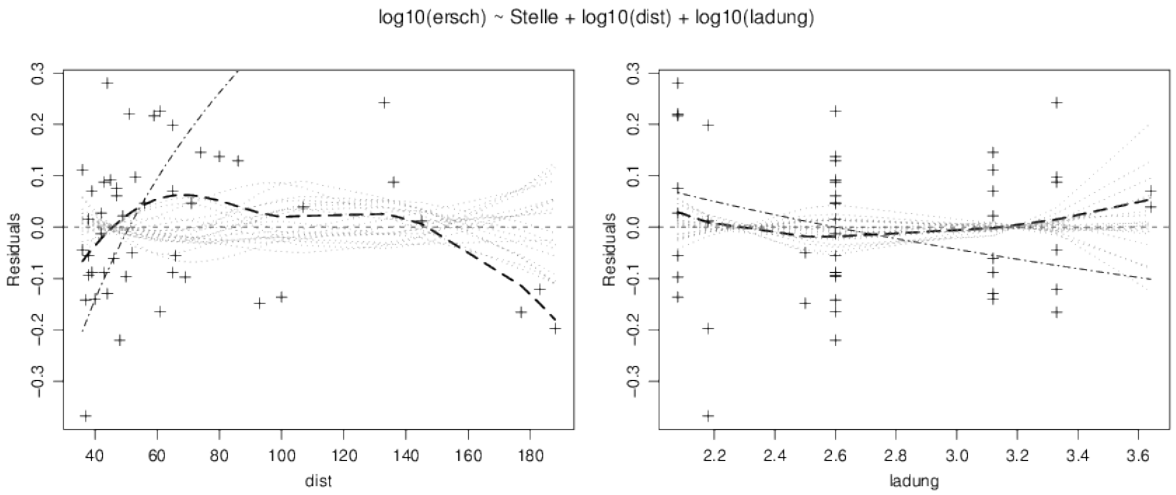

Abbildung 4.6.a: Streudiagramme der Residuen gegen zwei Eingangs-Variable, mit Glättung $(- - -)$ und Referenzlinie $Y =$konstant $(- \cdot - \cdot -)$

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 Eingangs-Variablen $X^{(j)}$ und den Residuen, 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) Eingangs-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 Eingangs-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 Eingangsgrösse $U^{(j)}$ 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. Wenn diese Differenz nicht linear, aber immerhin monoton zunimmt oder monoton abnimmt, kann eine monotone Transformation der Eingangs-Variablen helfen.

▷ 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 – die Differenz zwischen Glättung und Referenzlinie – immerhin monoton mit der Eingangs-Variablen abnimmt, hat man mit einer monotonen Transformation dieser Variablen eine Chance, die Krümmung weg zu bringen.

Die Transformation müsste grosse Werte der Eingangs-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 Eingangsgrösse 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.

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 untransfor-

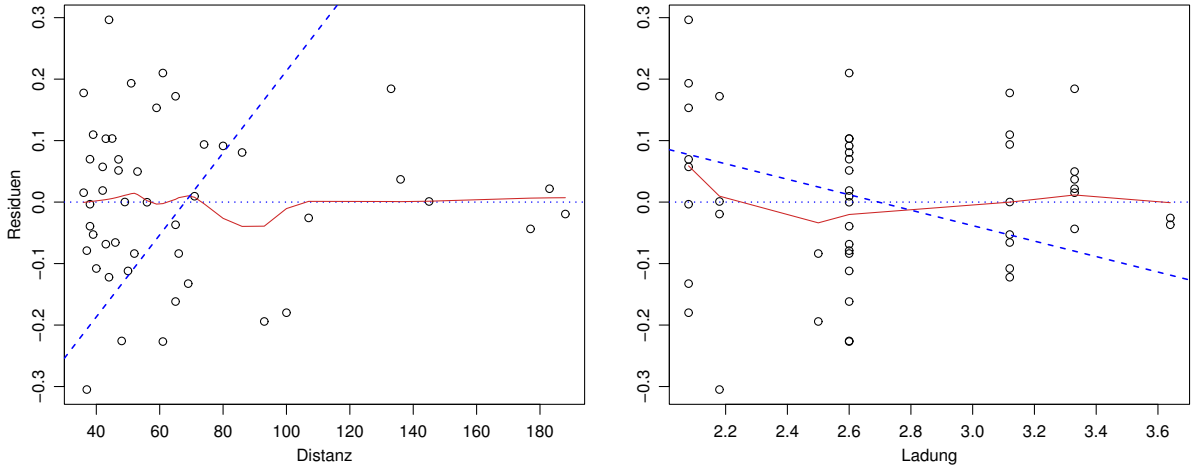

Abbildung 4.6.c: Streudiagramm der Residuen gegen die Eingangsgrössen Distanz und Ladung, die hier unlogarithmiert im Modell stehen

mierte Eingangsgrösse 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 Eingangsgrössen berechnet werden), muss neu geklärt werden, wie der Effekt einer Eingangsgrösse $U^{(j)}$ gemessen werden soll. Antwort: Man setzt alle anderen Eingangs-Variablen auf einen „typischen Wert“ u_k (Median für kontinuierliche und Modus für kategorielle Variable) und verwendet die Vorhersage $\widehat{y}\langle u_1, ..., u_{j-1}, U^{(j)}, u_{j+1}, ...\rangle$ als Funktion des varierenden $U^{(j)}$ als „component effect“ $\widehat{\gamma}^{(j)}$.

f Im Modell wird als nächstes vorausgesetzt, dass die **Effekte von zwei Eingangs-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.

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 Eingangs-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 Eingangs-Variable, mit verschiedenen Symbolen für die verschiedenen Faktor-

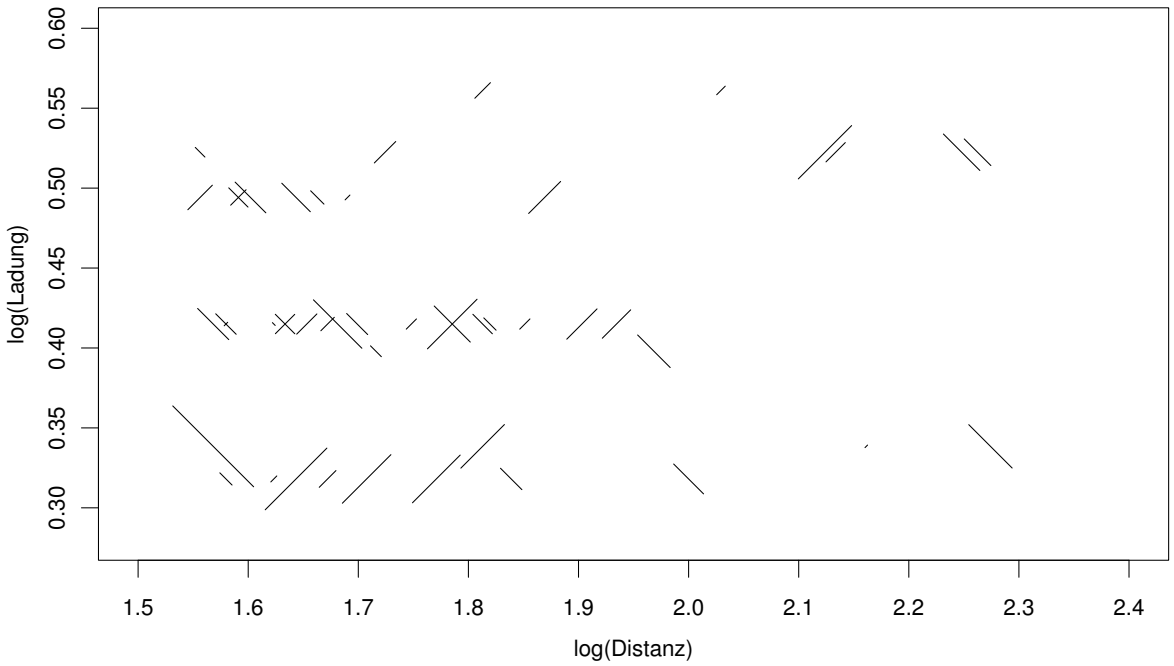

Abbildung 4.6.f (i): Residuen in Abhängigkeit von zwei Eingangs-Variablen im Beispiel der Sprengungen

werte (Abbildung 4.6.g (ii)).

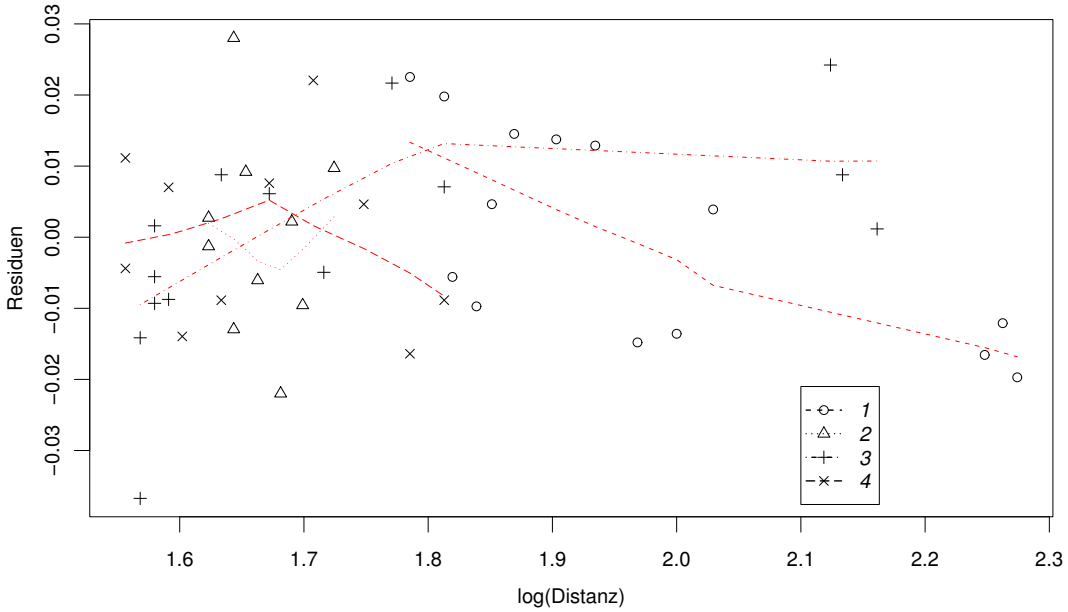

Abbildung 4.6.g (ii): Residuen gegen eine Eingangs-Variable, mit verschiedenen Symbolen und Glättungen für die verschiedenen Werte eines Faktors

h In den Streudiagrammen der Residuen gegen die Eingangs-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 $\sigma_i^2 = \mathrm{var}\langle E_i\rangle$ 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 $\mathrm{SSQ}^{(E)}$ 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 (2005, 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 Zufalls-Abweichungen 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 $\mathrm{var}\langle E_i\rangle = \sigma^2 v_i$ 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 Eingangsgrösse $U^{(j)}$ zeigen kann, dass die Streuung von $U^{(j)}$ abhängt. Dann kann man versuchen, eine Funktion v anzugeben, die diese Abhängigkeit beschreibt, für die also $\mathrm{var}\langle E_i\rangle \approx \sigma^2\, v\langle u_i^{(j)}\rangle$ 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 $\widehat{y}_i$ abhängt. Man geht dann oft so vor, dass man zuerst das Modell ohne Gewichte anpasst und die so berechneten

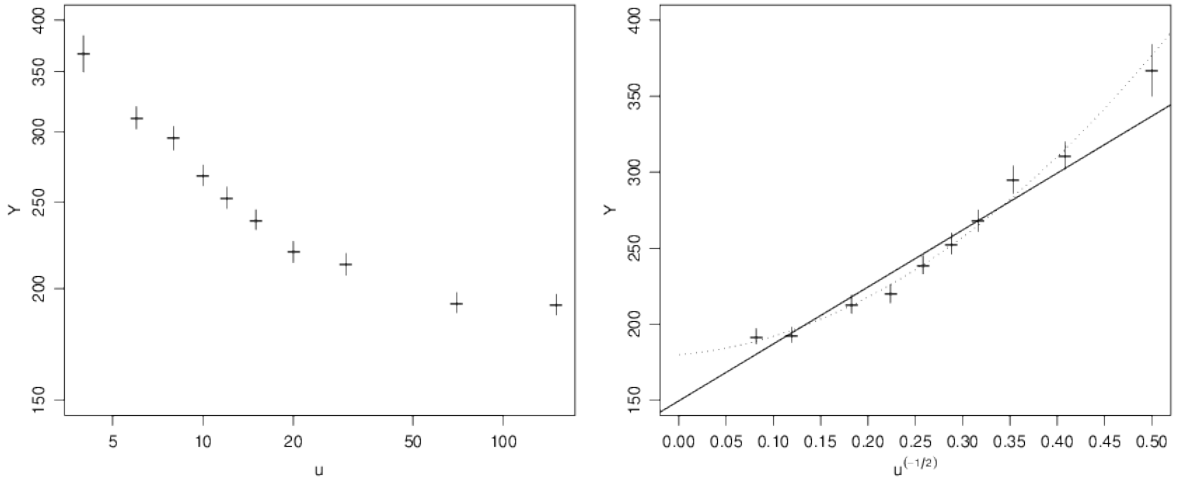

Abbildung 4.7.b: Daten des Beispiels der starken Wechselwirkung mit logarithmischen Achsen (links) und mit transformierter Energie (rechts). Im zweiten Fall sind die geschätzten Regressionsfunktionen mit linearem Modell (entsprechend der physikalischen Theorie) und quadratischem Modell eingezeichnet.

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.

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

$$\underline{\widehat{\beta}} = (\widetilde{\boldsymbol{X}}^T \boldsymbol{W} \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{W} \underline{Y} \; .$$

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

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

$$\begin{aligned} \widetilde{R}_i &= R_i \Big/ \left(\widehat{\sigma} \sqrt{1/w_i - (\boldsymbol{H}_W)_{ii}} \right) \qquad \text{mit} \\ \boldsymbol{H}_W &= \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \; . \end{aligned}$$

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 Streuungs-Diagramm 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.

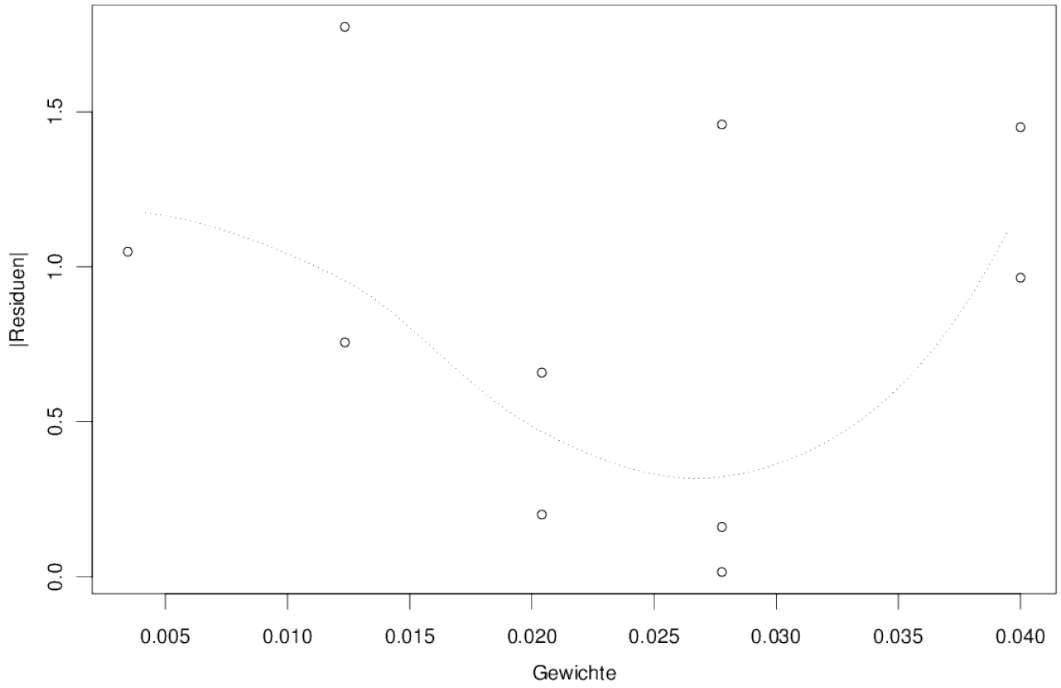

Abbildung 4.7.f: Absolute Residuen aus dem quadratischen Modell gegen Gewichte im Beispiel der starken Wechselwirkung

◁

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 genug 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 $\widehat{\sigma}^2$ der Fehler mit einer anderen Schätzung $\widetilde{\sigma}^2$ zu vergleichen, die unabhängig davon gewonnen wird. Falls das Modell stimmt, sollte $\widehat{\sigma}^2 \approx \widetilde{\sigma}^2$ sein. Andernfalls ist $\widehat{\sigma}^2$ grösser, weil die Residuen R_i zusätzlich zur zufälligen Streuung noch einen systematischen Fehler enthalten.

Die Testgrösse ist jeweils das Verhältnis $T = \widehat{\sigma}^2/\widetilde{\sigma}^2$. 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 $\sigma_i^2 = \mathrm{var}\langle E_i\rangle$ gegeben, dann lautet die Testgrösse $T = \sum_i R_i^2/\sigma_i^2$; sie ist chiquadrat-verteilt, $\sim \chi^2_{n-p}$, 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** $[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}]$ **mehrere Beobachtungen** $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ 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

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

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

Die Testgrösse

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

hat unter der Nullhypothese eine F-Verteilung mit $g-p$ und $n-g$ Freiheitsgraden. (Falls $g < p$ ist, sind die Paramter 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 $[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}]$ und $[x_h^{(1)}, x_h^{(2)}, \ldots, x_h^{(m)}]$ gesucht werden. Die quadrierten Differenzen $(R_i - R_h)^2$ der entsprechenden Residuen sollte im Mittel etwa $2\widehat{\sigma}^2$ betragen. Man kann dies grafisch überprüfen, indem man $(R_i - R_h)^2$ gegenüber einem geeigneten Distanzmass $d\big\langle x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; x_h^{(1)}, x_h^{(2)}, \ldots, x_h^{(m)}\big\rangle$ in einem Streudiagramm aufträgt. Der Vorschlag stammt von Daniel and Wood (1980, Abschnitt 7.10), die

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

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. ◁

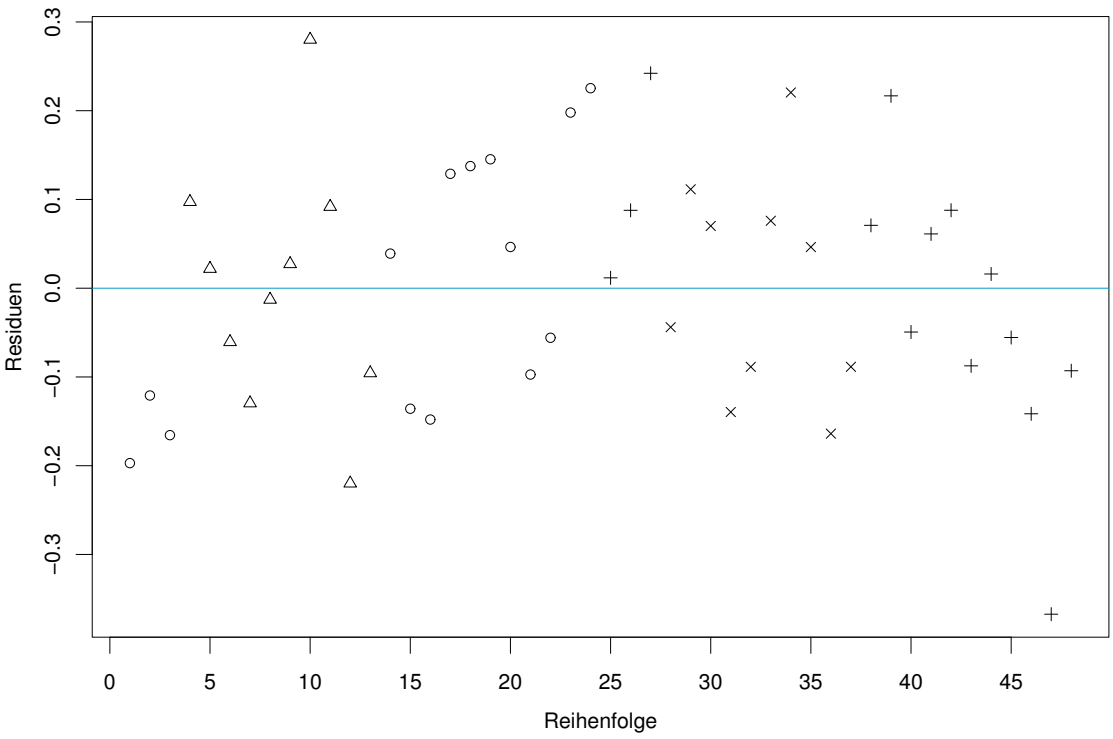

Abbildung 4.9.a: Residuen gegen Reihenfolge im Beispiel der Sprengungen. Die verschiedenen Stellen sind mit verschiedenen Symbolen dargestellt.

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

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

in diesem Fall klein ausfallen. Leider ist die Verteilung der Teststatistik unter der Nullhypothese der Unabhängigkeit der E_i von der Design-Matrix $\widetilde{\boldsymbol{X}}$ 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.

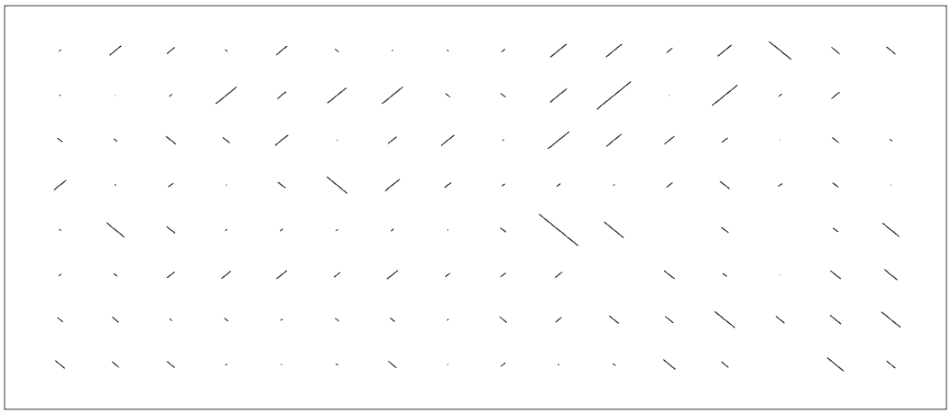

Abbildung 4.9.c: Residuen und räumliche Anordnung der Beobachtungen im Beispiel der basischen Böden

Benachtbarte 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

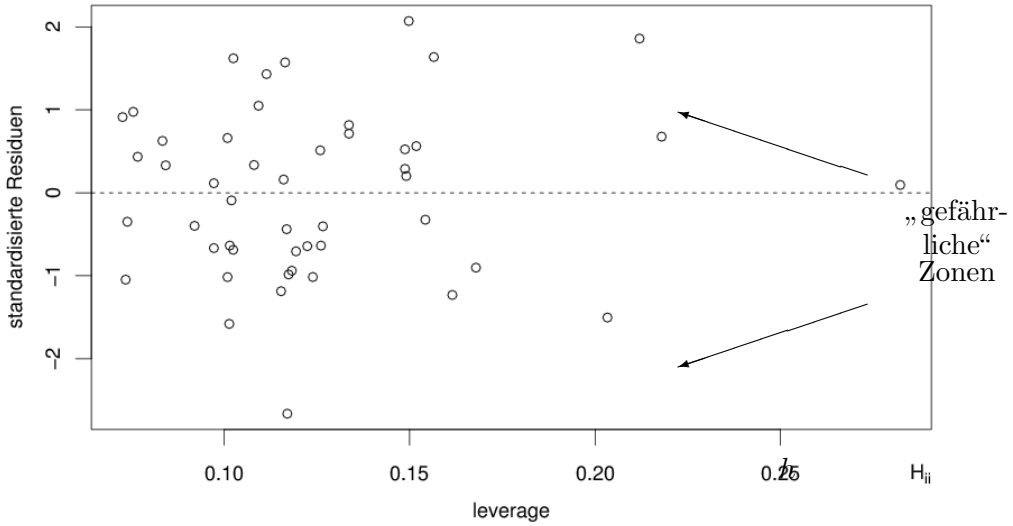

Abbildung 4.10.c: Hebelarm-Diagramm für das Beispiel der Sprengungen

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 $\widetilde{R}_i$, die ja selbst aus R_i, H_{ii} und $\widehat{\sigma}$ berechnet werden (4.3.i). In einem Streudiagramm der $\widetilde{R}_i$ 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 $\widehat{\sigma}$ ergeben.

e Die **Distanz von Cook** fasst die Veränderungen aller angepassten Werte $\widehat{y}_i$ beim Weglassen der $i$ten Beobachtung zu einer Zahl zusammen (nämlich zu ihrer Quadratsumme $(\underline{\widehat{y}}_{(-i)} - \underline{\widehat{y}})^T (\underline{\widehat{y}}_{(-i)} - \underline{\widehat{y}})$, dividiert durch $p\widehat{\sigma}^2$). Sie lässt sich schreiben als

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

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* Die „leverage“ ist ein Mass für die „Extremheit“ der Beobachtung i, in das auch Variable eingehen, die sich als unwichtig für das Modell erweisen. Als Ergänzung dazu kann die in eingeführte Distanz von Daniel and Wood (1980), angewandt zwischen $\underline{x}_i$ und dem Schwerpnukt $\underline{\overline{x}}$,

$$d\langle \underline{x}_i - \underline{\overline{x}} \rangle = \sum_j \left(\widehat{\beta}_j (x_i^{(j)} - \bar{x}^{(j)})\right)^2 \Big/ \widehat{\sigma}^2$$

dienen. Sie besteht aus der Quadratsumme der „component effects“ $\widehat{\beta}_j(x_i^{(j)} - \bar{x}^{(j)})$ und berücksichtigt die Wichtigkeit der Variablen.

g 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 Regressions-Analyse ohne die entsprechende Eingangs-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 Eingangs-Variablen – mit der Bildung solcher Residuen schaltet man die „indirekten Einflüsse“von $X^{(j)}$ auf Y aus.

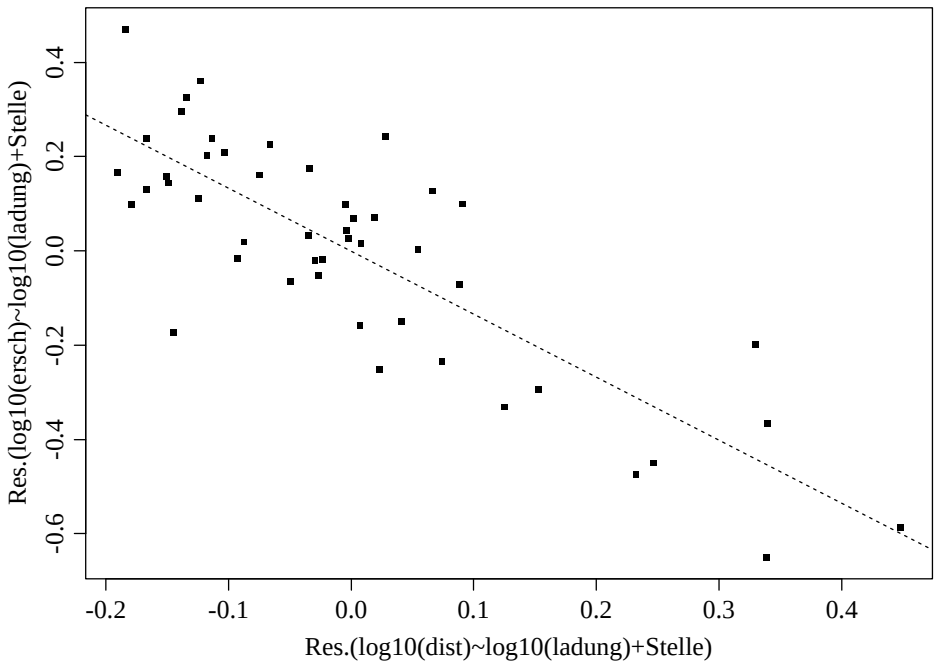

Abbildung 4.10.g: Added variable plot für die logarithmierte Distanz im Beispiel der Sprengungen

Wenn man in diesem Streudiagramm eine Gerade (mit Kleinsten Quadraten) anpasst, so hat sie genau die Steigung $\widehat{\beta}_j$, 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.g 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.

4.A Theoretische Verteilung der Residuen

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

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

Die Matrix $\boldsymbol{H}$ 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 $\boldsymbol{H}$ 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 $\widehat{y}_i$.

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,

$$\mathcal{E}\langle\underline{R}\rangle = \mathcal{E}\langle\underline{Y}\rangle - \widetilde{\boldsymbol{X}}\,\mathcal{E}\langle\underline{\widehat{\beta}}\rangle = \widetilde{\boldsymbol{X}}\underline{\widetilde{\beta}} - \widetilde{\boldsymbol{X}}\underline{\widetilde{\beta}} = \underline{0} \,.$$

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

$$\underline{R} = \underline{Y} - \underline{\widehat{y}} = \boldsymbol{I}\underline{Y} - \boldsymbol{H}\underline{Y} = (\boldsymbol{I} - \boldsymbol{H})\,\underline{Y}$$

und erhalten daraus

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

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

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

Also gilt

$$\mathrm{var}\langle\underline{R}\rangle = \sigma^2\,(\boldsymbol{I} - \boldsymbol{H})\,.$$

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

c Die Gleichung $\underline{R} = (\boldsymbol{I} - \boldsymbol{H})\,\underline{Y}$ 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 $\mathrm{var}\langle\underline{R}\rangle = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$ sind die Residuen korreliert,

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

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

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

zu minimieren. Es ergeben sich die Normalgleichungen

$$\widetilde{\boldsymbol{X}}^T \boldsymbol{W}\, \underline{R} = \underline{0} \qquad \text{oder} \quad \widetilde{\boldsymbol{X}}^T \boldsymbol{W}\, (\underline{Y} - \widetilde{\boldsymbol{X}}\, \underline{\widehat{\beta}}) = \underline{0} \quad \Rightarrow \quad \widetilde{\boldsymbol{X}}^T \boldsymbol{W}\, \widetilde{\boldsymbol{X}}\, \underline{\widehat{\beta}} = \widetilde{\boldsymbol{X}}^T \boldsymbol{W}\, \underline{Y}$$

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

$$\underline{\widehat{\beta}} = \boldsymbol{C}_W^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{W}\, \underline{Y} \; .$$

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

$$\begin{aligned} \mathrm{var}\langle\underline{\widehat{\beta}}\rangle &= \boldsymbol{C}_W^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{W} \cdot \sigma^2 \boldsymbol{W}^{-1} \cdot \boldsymbol{W} (\boldsymbol{C}_W^{-1} \widetilde{\boldsymbol{X}}^T)^T = \sigma^2\, \boldsymbol{C}_W^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{W} \widetilde{\boldsymbol{X}} (\boldsymbol{C}_W^{-1})^T \\ &= \sigma^2\, (\widetilde{\boldsymbol{X}}^T \boldsymbol{W} \widetilde{\boldsymbol{X}})^{-1} \; . \end{aligned}$$

f Die Residuen sind jetzt gleich

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

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

$$\begin{aligned} \mathrm{var}\langle\underline{R}\rangle &= (\boldsymbol{I} - \boldsymbol{H}_W \boldsymbol{W}) \cdot \sigma^2 \boldsymbol{W}^{-1} \cdot (\boldsymbol{I} - \boldsymbol{H}_W \boldsymbol{W})^T \\ &= \sigma^2\, (\boldsymbol{W}^{-1} - \boldsymbol{H}_W \boldsymbol{W} \boldsymbol{W}^{-1} - \boldsymbol{W}^{-1} \boldsymbol{W} \boldsymbol{H}_W + \boldsymbol{H}_W \boldsymbol{W} \boldsymbol{W}^{-1} \boldsymbol{W} \boldsymbol{H}_W) \\ &= \sigma^2\, (\boldsymbol{W}^{-1} - \boldsymbol{H}_W) \; . \end{aligned}$$

Die standardisierten Residuen sind also

$$\widetilde{R}_i = R_i \Big/ \Big(\widehat{\sigma} \sqrt{1/w_i - (\boldsymbol{H}_W)_{ii}}\Big) \; .$$

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. Als vierte Grafik folgt der leverage plot (Residuen gegen Hebelwerte H_{ii}). Einflussreiche Beobachtungen befinden sich rechts oben und unten.

b Wenn die Regression mit `regr` angepasst wurde, werden als Nächstes 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 Benützer nämlich darauf, anzusehen, was die Funktion `plot` automatisch liefert, und das ist bei Verwendung von `lm` zu wenig.

c **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** `termplot`. Residuen, genauer partial residuals, werden gegen die Eingangsgrössen aufgetragen.

e **Funktion** `plresx` (Zusatzfunktion zu `regr`). Diese Funktion leistet Ähnliches wie `termplot`: 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 für `regr`-Objekte rufen für jede grafische Darstellung die Funktion `stamp` auf, die zur Dokumentation des grafischen Outputs dient. Sie fügt in der rechten unteren Ecke das Datum und einen allfälligen Projekttitel (`userOptions(project=projecttitle, step=stepname)`) ein.

5 Modell-Entwicklung

5.1 Problemstellung

a Von der wissenschaftlichen **Fragestellung** und vom Vorwissen her gibt es verschiedene Arten, die Regressions-Analyse einzusetzen:

1. Im „Idealfall“ ist bereits klar, dass die Zielgrösse Y von den gegebenen Regressoren $X^{(1)}, \ldots, X^{(m)}$ linear abhängt. Man interessiert sich für eine klassische Fragestellung über die Koeffizienten der Regressoren, also für einen Test einer Nullhypothese (z. B. $\beta_j = 0$), eine Punkt- oder Intervallschätzung für einen oder mehrere Koeffizienten oder allenfalls für Vorhersage-Intervalle. Die entsprechenden Methoden haben wir behandelt.

2. Im anderen Extremfall dient die Studie dazu, Zusammenhänge zwischen der Zielgrösse Y und den Eingangs-Variablen überhaupt erst zu erforschen. Man weiss nicht, ob und in welcher Form die Eingangs-Variablen die Zielgrössen beeinflussen. Oft hat man dann für eine recht grosse Zahl potentieller Einflussgrössen „vorsorglich“ Daten erhoben.

3. Manchmal liegt die Fragestellung dazwischen:

   - Man ist eigentlich nur am Einfluss eines einzigen Regressors interessiert, aber unter Berücksichtigung der Effekte von anderen Eingangs-Variablen (um indirekte Einflüsse zu vermeiden). Beispiel: Wirkung eines Medikamentes.

   - Man weiss einiges aus früheren Studien und aus theoretischen Überlegungen und will zusätzliche Erkenntnisse gewinnen.

In 2. und 3. stellt sich – in unterschiedlichem Ausmass – die Frage der **Modellwahl**: Welche Eingangs-Variablen sollen in welcher Form in der Modell-Gleichung der linearen Regression erscheinen?

b ▷ **Beispiel Baukosten von Atomkraftwerken**. Die Baukosten von 32 Kernkraftwerken, die in den Jahren 1967-71 in den USA entstanden, wurden untersucht (Quelle: Cox and Snell (1981)). Eine Fragestellung war, ob eine partielle Kostengarantie des Generalunternehmers zu Einsparungen führe. Als weitere erklärende Angaben für die Baukosten wurden die in Tabelle 5.1.b aufgeführten Variablen notiert. – Das Beispiel ist zwar schon in die Jahre gekommen, und die Anzahl Beobachtungen ist prekär klein. Es zeigt aber die Chancen und Schwierigkeiten der Modellwahl recht schön. ◁

c Erinnern Sie sich, dass die $x^{(j)}$ in der Modellgleichung $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i$ nicht unbedingt die **ursprünglich** beobachteten oder gemessenen Grössen, die wir zur Unterscheidung mit $u^{(k)}$ bezeichnen wollen, sein müssen; es können **transformierte** Grössen (z. B. $x^{(j)} = \log_{10}\langle u^{(j)} \rangle$) sein oder Funktionen von mehreren ursprünglichen Grössen (z. B. $x^{(j)} = u^{(k)} \cdot u^{(\ell)}$). Auch die Zielgrösse Y kann durch geeignete Transformation oder Standardisierung aus einer oder mehreren ursprünglich gemessenen Variablen gewonnen werden.

d ▷ Im Beispiel führen allgemeine Überlegungen (siehe 4.4.g) zu den in Tabelle 5.1.b aufgeführten Transformationen als Eingangsgrössen. Die Wartezeit und die Bauzeit wurden, obwohl es sich um Beträge (positive Zahlen) handelt, nicht logarithmiert, da es gemäss Zinseszins-Rechnung sinnvoll ist, einen linearen Einfluss dieser Zeiten auf die logarithmierten Kosten anzunehmen. Es sind auch andere Transformationen denkbar, und solche sollen ja auf Grund der Residuenanalyse immer wieder in Betracht gezogen werden.

Das lineare Regressionsmodell mit allen transformierten Variablen, das „volle Modell“, lautet im

Bez.	Bedeutung	Typ	Transf.
K	Baukosten	Betrag	log
G	Grösse	Betrag	log
D	Datum der Baubewilligung	kontin.	–
WZ	Wartezeit zwischen Antrag und Baubewilligung	Betrag	–
BZ	Bauzeit: Zeit bis Inbetriebnahme	Betrag	–
Z	Zweitwerk: früheres Werk auf gleichem Gelände	binär	–
NE	Werk steht im Nordosten der USA	binär	–
KT	Werk arbeitet mit Kühlturm	binär	–
BW	Reaktor hergestellt durch Babcock-Wilcox	binär	–
N	Anzahl Werke, die das gleiche Ingenieur-Team bereits erbaut hat, +1	Anzahl	Wurzel
KG	Partielle Kostengarantie des Generalunternehmers	binär	–

Tabelle 5.1.b: Die Variablen des Beispiels Baukosten

Beispiel also in Modellschreibweise

$$\texttt{log10(K)} \sim \texttt{log10(G)} + \texttt{D} + \texttt{WZ} + \texttt{BZ} + \texttt{Z} + \texttt{NE} + \texttt{KT} + \texttt{BW} + \texttt{sqrt(N)} + \texttt{KG}$$

oder ausführlich

$$\begin{aligned} \log_{10}\langle \texttt{K}_i \rangle \;=\; & \beta_0 + \beta_1 \log_{10}\langle \texttt{G}_i \rangle + \beta_2 \texttt{D}_i + \beta_3 \texttt{WZ}_i + \beta_4 \texttt{BZ}_i \\ & + \beta_5 \texttt{Z}_i + \beta_6 \texttt{NE}_i + \beta_7 \texttt{KT}_i + \beta_8 \texttt{BW}_i + \beta_9 \sqrt{\texttt{N}_i} + \beta_{10} \texttt{KG}_i + E_i \; . \end{aligned}$$

e ▷ Tabelle 5.1.e zeigt die Computer-Ausgabe für das Beispiel. Es können mindestens 5 Variable als überflüssig angesehen werden. Auch die Kostengarantie ist „schwach nicht-signifikant“. Ist die Frage damit schon beantwortet? Wir werden das Beispiel noch weiter verfolgen. Schliesslich kann es um viel Geld gehen. ◁

Coefficients:

	Value	Std. Error	t value	Pr(> |t|)	Signif
(Intercept)	-6.02586	2.34729	-2.57	0.018	*
log10(G)	0.69254	0.13713	5.05	0.000	***
D	0.09525	0.03580	2.66	0.015	*
WZ	0.00263	0.00955	0.28	0.785	
BZ	0.00229	0.00198	1.16	0.261	
Z	-0.04573	0.03561	-1.28	0.213	
NE	0.11045	0.03391	3.26	0.004	**
KT	0.05340	0.02970	1.80	0.087	.
BW	0.01278	0.04537	0.28	0.781	
sqrt(N)	-0.02997	0.01780	-1.68	0.107	
KG	-0.09951	0.05562	-1.79	0.088	.

Tabelle 5.1.e: Computer-Ausgabe für das volle Modell im Beispiel Baukosten

5.2 Wichtigkeit eines einzelnen Terms

a Ist ein bestimmter Term $\beta_j x^{(j)}$ im Modell nötig? nützlich? überflüssig? – Die Beantwortung dieser Frage bildet einen **Grundbaustein für die Modellwahl**.

Als Hypothesen-Prüfung haben wir diese Frage schon gelöst: Wir wissen, wie man die Nullhypothese $\beta_j = 0$ prüft (mit dem t-Test). Diese Antwort tönt aber besser, als sie ist, denn es ergibt sich das Problem des **multiplen Testens**.

b Bei der Suche nach einem geeigneten Modell werden meistens einige bis viele Entscheidungen der erwähnten Art getroffen. Extremfall: Man habe 20 Regressoren („X-Variable“), und ein einziger Koeffizient sei „signifikant“ (auf dem 5%-Niveau) von 0 verschieden. Dann entspricht das auf Grund der Wahrscheinlichkeit eines „Fehlers erster Art“ der Erwartung für den Fall, dass überhaupt kein Regressor einen Einfluss auf Y hat!

c Dazu kommt ein weiteres, kleineres Problem: Man müsste die Voraussetzungen der Normalverteilung und der Unabhängigkeit der Fehler prüfen, wenn man die P-Werte der t-Tests zum Nennwert nehmen wollte.

d Man kann also nicht behaupten, dass ein Term mit signifikantem Test-Wert einen „statistisch gesicherten“ Einfluss auf die Zielgrösse habe.

Statt die Tests für strikte statistische Schlüsse zu verwenden, begnügen wir uns damit, die P-Werte der t-Tests für die Koeffizienten (oder direkt die t-Werte) zu benützen, um die *relative* Wichtigkeit der entsprechenden Regressoren anzugeben, insbesondere um die „wichtigste“ oder die „unwichtigste“ zu ermitteln.

e Eine nominale Variable (ein „Faktor“, also eine Variable mit mehreren möglichen Werten, die keine natürliche Ordnung zeigen) kann, wie in 3.2.e erklärt, in mehrere Indikator-Variable oder *dummy variables* verwandelt werden; wir reden von einem **Block von Indikator-Variablen**.

▷ (Das Beispiel enthält (leider) keine nominale Variable. Die fünf binären Variablen sind zwar Indikator-Variable, aber nicht im Sinne der „dummy variables“ eines Faktors verknüpft.) ◁

Wenn gefragt wird, ob man eine nominale Eingangs-Variable ins Modell einbeziehen soll oder nicht, muss man für den ganzen Block der entsprechenden Indikator-Variablen prüfen, ob alle weggelassen werden können. Das geschieht mit dem F-Test zum Vergleich von Modellen (3.2.m). Sein P-Wert kann mit den P-Werten der anderen Variablen „notfalls“ verglichen werden. (Besser eignet sich ein Vergleich mit den so genannten C_p-Werten, die in 5.3.g eingeführt werden.)

5.3 Automatisierte Verfahren zur Modellwahl

a Mit Hilfe eines Masses für die relative Nützlichkeit eines einzelnen Terms in der Regressionsgleichung können Strategien der Modellwahl formuliert werden:

• **Schrittweise rückwärts**. Man geht vom Modell aus, in dem alle in Frage kommenden Regressoren enthalten sind. (Das ist nur möglich, wenn die Zahl dieser Variablen kleiner ist als die Zahl der Beobachtungen – sie sollte bedeutend kleiner sein, sagen wir mindestens fünfmal kleiner.) Nun kann man schrittweise den „unwichtigsten“ wegnehmen, solange er unwichtig genug erscheint. Wo die entsprechende Grenze der „Wichtigkeit“, also des P-Wertes, liegen soll, ist kaum generell festzulegen. Die Schranke 0.05 für den P-Wert ist wegen des Problems des multiplen Testens nicht sinnvoller als andere (niedrigere) Werte.

b ▷ Im **Beispiel der Baukosten** ist gemäss Tabelle 5.1.e die Variable `WZ` die unwichtigste. Wenn sie weggelassen wird, ergeben sich neue t- und P-Werte und damit eine neue Reihenfolge. Die P-Werte sind jetzt

log10(G)	0.000	Z	0.213	BW	0.852
D	0.000	NE	0.003	sqrt(N)	0.092
BZ	0.262	KT	0.082	KG	0.084

Das Maximum zeigt die Variable `BW`, die also als nächste zu eliminieren ist. So werden der Reihe nach zunächst die Variablen `BW, BZ, Z,` $\sqrt{\texttt{N}}$ und `KT` weggelassen. Nun ist, wie Tabelle 5.3.b zeigt, der Einfluss der Kostengarantie hochsignifikant. Also doch! ◁

Coefficients:

	Value	Std. Error	t value	Pr($>\|t\|$)	Signif
(Intercept)	-3.4612	1.1458	-3.02	0.005	**
log10(G)	0.6629	0.1295	5.12	0.000	***
D	0.0610	0.0160	3.82	0.001	***
NE	0.0831	0.0330	2.52	0.018	*
KG	-0.1844	0.0424	-4.35	0.000	***

Tabelle 5.3.b: Computer-Ausgabe für das durch schrittweise Elimination reduzierte Modell im Beispiel Baukosten

c • **Schrittweise vorwärts**. Analog zum schrittweisen Rückwärts-Verfahren kann man vom „leeren“ Modell (kein Regressor) zu immer grösseren kommen, indem man schrittweise einen zusätzlichen Term (einen Regressor oder einen Faktor in Form des entsprechenden Blockes von dummy Variablen) hinzunimmt, und zwar in jedem Schritt denjenigen, der (von den verbleibenden) am „wichtigsten“ ist. Dieses Verfahren hatte in den Anfangszeiten der multiplen Regression eine grundlegende Bedeutung, da es einen minimalen Rechenaufwand erfordert.

d ▷ Im **Beispiel** zeigt die Kostengarantie `KG` die grösste einfache Korrelation mit den logarithmierten Baukosten und wird deshalb als erste Variable ins Modell aufgenommen! Es folgen $\log_{10}\langle\texttt{G}\rangle$, `D, NE` und `KT`. Der letzte Schritt führt zu einem formal nicht-signifikanten Koeffizienten. Wir lassen also `KT` wieder weg und haben das gleiche Modell wie vorher erreicht.

Nun sind wir von der Bedeutsamkeit der Kostengarantie überzeugt, nicht wahr? ◁

e • **„Alle Gleichungen“** ***(all subsets)***. Gehen wir wie beim Rückwärts-Verfahren von einem festen Satz von m möglichen Regressoren aus. Mit diesen Variablen lassen sich prinzipiell 2^m mögliche lineare Modell-Gleichungen bilden; man kann für jede Variable wählen, ob sie in der Gleichung erscheinen soll oder nicht. Der Computer kann alle möglichen Gleichungen an die Daten anpassen und nach einem geeigneten Kriterium die beste oder die paar besten suchen. (Intelligente Algorithmen vermeiden es, alle Gleichungen durchzurechnen.)

Im Folgenden bezeichnen wir die Anzahl Regressoren in einem in Frage stehenden Modell mit m'. Analog zu früher sei $p' = m' + 1$, falls das Modell einen Achsenabschnitt β_0 enthält und $= m'$ im gegenteiligen Fall.

f Als **Kriterien** können die folgenden Grössen verwendet werden:

1. „Bestimmtheitsmass“ R^2 oder multiple Korrelation R,
2. Wert der Teststatistik für das gesamte Modell (F-Test),
3. zur F-Teststatistik gehöriger P-Wert,
4. geschätzte Varianz $\widehat{\sigma}^2$ der Fehler (oder Standardabweichung $\widehat{\sigma}$).

Für eine feste Anzahl m' von Regressoren führen alle diese (und auch die unten aufgeführten) Kriterien zur gleichen Ordnung unter den $\binom{m}{m'}$ möglichen Modellen (da jedes sich aus jedem andern – für festes m' – über eine monotone Funktion ausrechnen lässt); es werden also von

allen die gleichen Modelle als die besten ausgewählt.

g Beim Vergleich zwischen **Modellen mit verschieden vielen Koeffizienten** gibt es Unterschiede:

Das Bestimmtheitsmass R^2 kann nicht abnehmen, wenn ein Term zur Modellgleichung hinzugefügt wird.

* Es misst ja im grösseren Modell das Quadrat der maximalen Korrelation zwischen Y und einer geschätzten Regressions-Funktion $\beta_0+\beta_{j_1}x^{(j_1)}+\ldots+\beta_{j_{m'+1}}x^{(j_{m'+1})}$. Die Variable $x^{(j_{m'+1})}$ weglassen heisst $\beta_{j_{m'+1}}=0$ setzen. Das Maximum unter dieser Nebenbedingung kann nicht grösser sein als ohne Bedingung.

Trotzdem ist ein grösseres Modell ja nicht unbedingt besser als ein kleineres. Sonst wäre ja das vollständige Modell immer das beste. Es sind deshalb Kriterien vorgeschlagen worden, die automatisch auch unter Gleichungen mit verschieden vielen Termen eine sinnvolle Wahl der besten vornehmen:

5. **Korrigiertes Bestimmtheitsmass** R^2 (adjusted R^2): $R^2_{\text{adj}} = 1-\frac{n-1}{n-p'}(1-R^2)$

6. C_p **von Mallows**. Dieses verbreitete Kriterium minimiert in gewisser Weise einen mittleren Vorhersagefehler. Es ist definiert als

$$C_{p'} := \text{SSQ}^{(E)}/\widehat{\sigma}^2_m + 2p' - n = (n-p')(\text{SSQ}^{(E)}/\widehat{\sigma}^2_m - 1) + p' \,,$$

wobei $\text{MSQ}^{(E)} = \text{SSQ}^{(E)}/(n-p')$ das „mittlere Quadrat des Fehlers“ ist und $\widehat{\sigma}_m$ die Schätzung von σ im grössten Modell.

7. Das Informations-Kriterium AIC von Akaike (und Varianten davon). Es ist $\text{AIC} = n\log\left\langle \text{MSQ}^{(E)}\right\rangle + kp'$ mit $k=2$, was $\approx C_{p'}$ plus eine Konstante ergibt (???)

Diese Kriterien zeichnen jeweils ein Modell als das beste aus. Oft sind sie sich nicht einig in bezug auf die Anzahl Terme. Innerhalb der Gleichungen mit gleicher Anzahl Terme führen sie, wie erwähnt, zur gleichen Ordnung wie die erste Liste, sind sich also auch untereinander einig.

Häufig, aber nicht immer, ist jedes dieser „besten“ auch unter den Modellen zu finden, die die schrittweisen Verfahren liefern.

h* !!! überprüfen!!! Die F-Statistik, die zum Testen der formalen Signifikanz eines einzelnen Koeffizienten gebraucht wird, ist

$$\begin{aligned}
F &= \frac{(n-p+1)\text{MSQ}^{(E)} - (n-p)\widehat{\sigma}^2_m}{\widehat{\sigma}^2_m} \\
&= (n-p+1)\left(\frac{\text{MSQ}^{(E)}}{\widehat{\sigma}^2_m} - 1 + 1 - \frac{n-p}{n-p+1}\right) \\
&= (n-p+1)\left(\frac{\text{MSQ}^{(E)}}{\widehat{\sigma}^2_m} - 1 - \frac{1}{n-p+1}\right) \\
&\approx (n-p+1)\left(\log\left\langle \text{MSQ}^{(E)}\right\rangle + \frac{p-1}{n-p+1} - \left(\log\langle\widehat{\sigma}^2_m\rangle + \frac{p}{n-p+1}\right)\right)
\end{aligned}$$

was einer Differenz von AIC-Werten mit $k\approx 1$ entspricht. Der Test ist dann signifikant, wenn die F-Statistik grösser als der kritische Wert $c = q^{(F_{1,n-p})}\langle 0.95\rangle$ ausfällt. Das trifft gemäss Näherung dann ein, wenn die Differenz

$$\log\left\langle \text{MSQ}^{(E)}\right\rangle + \frac{p-1}{n-p+1}(1+c) \;-\; \log\langle\widehat{\sigma}^2_m\rangle + \frac{p}{n-p+1}(1+c)$$

grösser als 0 ist, was einem AIC mit $k\approx 1+c$ entspricht.

i Einen grafischen Überblick über die Modelle und die zugehörigen Kriterienwerte vermittelt ein Streudiagramm der Kriterienwerte gegen die Anzahl Koeffizienten p' im Modell (Abbildung 5.3.i). Da dies für das Kriterium C_p eingeführt wurde (Daniel and Wood, 1980) wird die Grafik „C_p-Plot“ genannt.

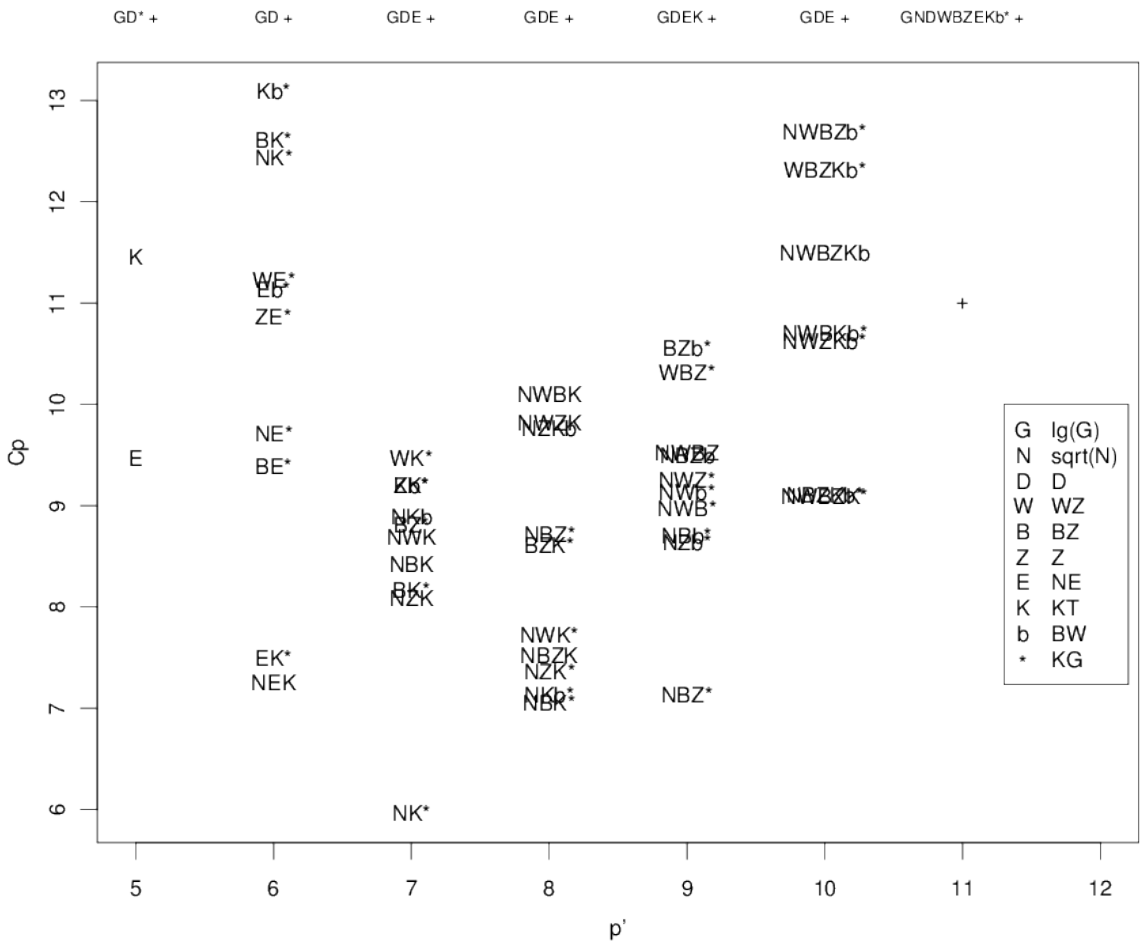

Abbildung 5.3.i: C_p-Plot für das Beispiel der Baukosten

j ▷ Im Beispiel würden laut dem C_p-Kriterium zusätzlich zu den in Tabelle 5.3.b erwähnten Variablen noch `KT` und $\sqrt{\texttt{N}}$ ins Modell einbezogen. In diesem Modell beträgt der P-Wert für die Kostengarantie 0.049 – ein nur noch ganz knapp signifikantes Resultat also! Die Frage, ob die Kostengarantie zu Einsparungen führe, wird also verschieden beantwortet, je nach den zusätzlichen erklärenden Variablen im Modell. Wir kommen auf diesen Punkt zurück (5.5.g). ◁

k Das „beste“ Modell ist aber noch lange nicht das „richtige“ oder das „wahre“ Modell! Wenn man Daten auf Grund eines bestimmten Modells simuliert, werden (je nach Streuung der Fehler, Anzahl Beobachtungen, Grösse der Modell-Koeffizienten und „Verteilung“ der Regressoren, genannt „design“) mehr oder weniger oft andere Modelle als „beste“ ausgelesen. **Das „beste Modell“ wird also vom Zufall mitbestimmt!**
Deshalb soll man immer **mehrere Modelle in Betracht ziehen**, die von den Kriterien als „gut“ – nicht viel schlechter als das „beste“ – bewertet werden.

Wie viel schlechter? Leider gibt die Statistik darauf keine Antwort. (Eine kleine Hilfe ist der Test für einzelne Koeffizienten, siehe oben.)

l* Eher peinlich berührt es, zu erwähnen, dass die meisten Programme zur Modellwahl mit den in 5.2.e erwähnten **Blöcken von Indikator- oder dummy-Variablen** (und anderen Variablen-Blöcken) nicht richtig umgehen. Es werden die einzelnen Indikator-Variablen als völlig unzusammenhängend behandelt. Die „beste" Gleichung enthält daher oft eine oder einige, aber nicht alle Indikator-Variablen eines Blocks – ein unsinniges Ergebnis.

m **Hohe Korrelationen zwischen Regressoren** oder allgemeinere Formen von **Kollinearität** führen zwar zu Problemen mit der Interpretation, sind aber von der Theorie her zugelassen. Im Vorwärts- und Rückwärts-Verfahren ist es in solchen Fällen häufig vom Zufall abhängig, welche der beteiligten Variablen als erste weggelassen respektive aufgenommen wird. Wenn alle Gleichungen untersucht werden, gibt es in diesem Fall jeweils Gruppen von ähnlich geeigneten.

Wir untersuchen diese Erscheinung im nächsten Abschnitt noch genauer. Eine ausführlichere Diskussion des Problems und von Lösungsmöglichkeiten findet man in Kap. 8 von Hocking (1996).

n Als Ergebnis der Modellwahl kann man die Teilmenge der ausgewählten Terme aus allen Termen des vollständigen Modells ansprechen – eine zufällige Menge also. Wenn man die Daten leicht verändert, wird diese Teilmenge in gewissen Fällen sprunghaft ändern, indem beispielsweise ein Regressor $X^{(j)}$ wegfällt. Man kann auch sagen, der entsprechende Koeffizient β_j springe auf 0. Das ist keine wünschenswerte Eigenschaft. Es gibt deshalb Verfahren, für die die Koeffizienten stetig von den Daten abhängen.

o Die Idee des Verfahrens namens **Lasso** (siehe Hastie, Tibshirani and Friedman, 2001) besteht darin, das Kriterium „Kleinste Quadrate", das ja bei der Bestimmung der Koeffizienten minimiert wird, durch einen „**Bestrafungsterm**" für die Grösse der Koeffizienten zu versehen. Man spricht im Englischen von „penalized regression". Damit die Grössen der Koeffizienten vergleichbar sind, benützt man standardisierte Koeffizienten β_j^* (siehe 3.1.m). Hier wird ausnahmsweise keine Quadratsumme als Mass der Grösse benützt, sondern die Summe der Absolutbeträge. Man minimiert also

$$Q\langle \underline{\beta}; \lambda \rangle = \sum\nolimits_i R_i^2 + \lambda \sum\nolimits_j |\beta_j^*| \; .$$

Die Grösse λ steuert, wie stark die Grösse der Koeffizienten gegenüber der Residuen-Quadratsumme ins Gewicht fallen soll.

Man kann das Problem der Minimierung von Q auch formulieren als Minimierung der Quadratsumme der Residuen unter Einhaltung einer Schranke für die Grösse der Koeffzienten. Man minimiert also $\sum_i R_i^2$ unter einer Nebenbedingung der Form $\sum_j |\beta_j^*| < c$. Jeder Lösung dieses zweiten Problems, mit bestimmtem c, entspricht eine Lösung des ersten Problems mit einem gewissen λ, das von c abhängt. Die Gesamtheit der Lösungen für alle verschiedenen c im zweiten Fall ist also gleich der Gesamtheit der Lösungen für alle verschiedenen λ im ersten Fall.

Wenn c so gross ist, dass die Kleinste-Quadrate-Schätzungwerte $\widehat{\beta}_j$ die Nebenbedingung erfüllen, also $\sum_j |\widehat{\beta}_j| \leq c$, dann ergibt sich keine Änderung. Wird c kleiner gewählt, dann werden die Koeffizienten demgegenüber verkleinert oder „gegen 0 geschrumpft". Um c in einem sinnvollen Bereich zu wählen, setzt man deshalb besser $b = c/\sum_j |\widehat{\beta}_j|$ fest auf einen Wert zwischen 0 und 1.

Die Art der Nebenbedingung führt dazu, dass bald der erste Koeffizient exakt gleich 0 wird und mit kleineren c-Werten immer mehr Koeffizienten verschwinden. Dadurch entsteht eine Modellselektions-Reihe wie in einem schrittweisen Rückwärts-Verfahren.

p ▷ Abbildung 5.3.p zeigt, wie die geschätzten standardisierten Koeffizienten von der relativen Schranke b abhängen. Wenn man von $b = 1$ rückwärts geht, wird zunächst der Koeffizient von `BW` gleich 0, dann derjenige von `WZ`, dann `Z`, $\sqrt{\texttt{N}}$ und `BZ`. Ein merkwürdiges Verhalten zeigt ausgerechnet der Koeffizient der Kostengarantie `KG`: Er ist im Bereich von mittleren Schranken am bedeutendsten. ◁

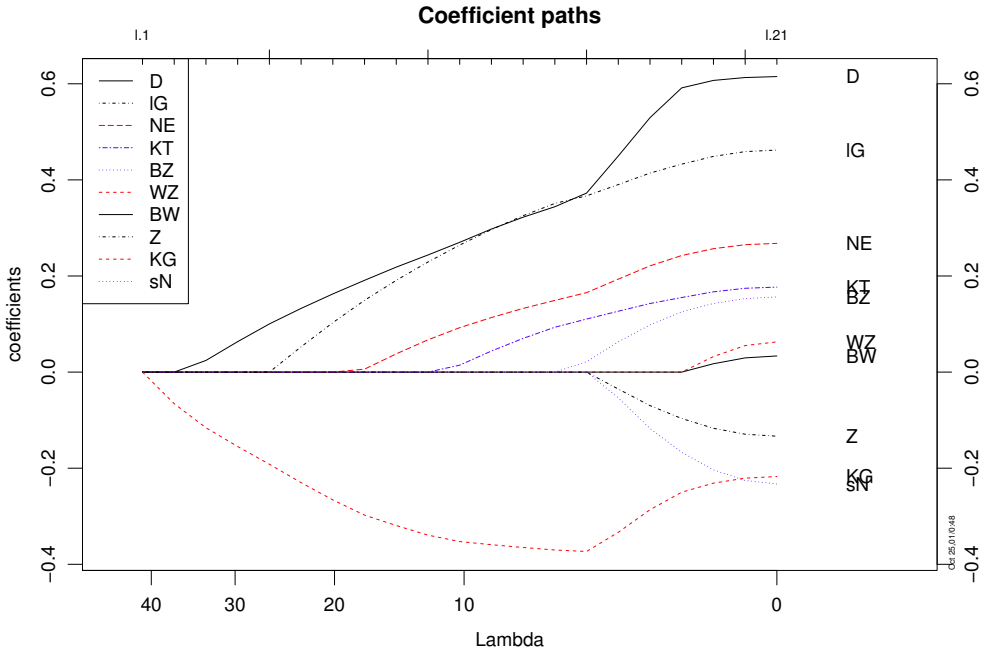

Abbildung 5.3.p: Lasso im Beispiel der Kernkraftwerke: standardisierte Koeffizienten in Abhängigkeit der relativen Schranke b

5.4 Kollinearität

a Der Begriff der Kollinearität stammt aus der linearen Algebra. Das Modell lautete in Matrix-Schreibweise $\underline{Y} = \widetilde{\boldsymbol{X}}\,\underline{\widetilde{\beta}} + \underline{E}$ (3.4.d), und die Schätzung war $\underline{\widehat{\beta}} = \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}^T\underline{Y}$ (3.4.g). Man braucht also die Inverse der Matrix $\boldsymbol{C} = \widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}$.

Die Matrix $\boldsymbol{C}$ ist **singulär**, wenn die Spalten der **Design-Matrix** $\widetilde{\boldsymbol{X}}$ **kollinear** sind,

$$\begin{aligned} \boldsymbol{C} \text{ singulär} \quad &\Longleftrightarrow \quad \text{es gibt Zahlen } \underline{\widetilde{c}} = [c_0,\, c_1, \ldots, c_p] \text{ mit } \widetilde{\boldsymbol{X}}\underline{\widetilde{c}} = \underline{0} \quad (\underline{\widetilde{c}} \neq \underline{0}) \\ &\Longleftrightarrow \quad \text{es gibt ein } j \text{ und Zahlen } [c_0,\, c_1, \ldots, c_p] \text{ mit } \quad \widetilde{x}_i^{(j)} = \sum_{k \neq j} \widetilde{c}_k \widetilde{x}_i^{(k)} \,. \end{aligned}$$

In diesem Fall sind die Parameter im Modell nicht eindeutig zu bestimmen. Wegen

$$\widetilde{\boldsymbol{X}}\,\underline{\widetilde{\beta}} = \widetilde{\boldsymbol{X}}\,(\underline{\widetilde{\beta}} + \gamma\underline{\widetilde{c}}) \quad \text{mit beliebigem } \gamma$$

gilt: Wenn $\underline{\widehat{\beta}}$ ein Schätzwert von $\underline{\widetilde{\beta}}$ ist, dann führt $\underline{\widehat{\beta}} + \gamma\underline{\widetilde{c}}$ zu den gleichen Abweichungen $\underline{R}$ und ist deshalb ein gleich guter Schätzwert. Die Kleinste-Quadrate-Schätzung ist also nicht eindeutig, und etliche Programme steigen aus.

b Das Problem kann gelöst werden, indem man eine x-Variable, $x^{(j)}$, also eine Spalte in der Design-Matrix, streicht – falls die verbleibende Matrix immer noch singulär ist, streicht man eine weitere, usw. (Man muss jeweils eine Spalte $\widetilde{x}^{(j)}$ wählen, für die die erwähnte Gleichung $\widetilde{x}_i^{(j)} = \sum_{k\neq j} \widetilde{c}_k \widetilde{x}_i^{(k)}$ erfüllt ist.) Die Verteilungen, die das Modell beschreibt, bleiben damit eigentlich die gleichen, nur die Parametrisierung ändert, und damit die Interpretation der Parameter.

c Wenn solche lineare Beziehungen zwischen den x-Variablen nicht exakt, aber näherungsweise gelten, sind die Parameter zwar formell identifizierbar, aber „schlecht bestimmt“. Man spricht dann in der Statistik immer noch von **Kollinearität**.

Ein anschauliches einfaches Beispiel bilden zwei stark korrelierte x-Variable, z. B. $x^{(1)}$ und $x^{(2)}$. Abbildung 5.4.c zeigt einen solchen Datensatz.

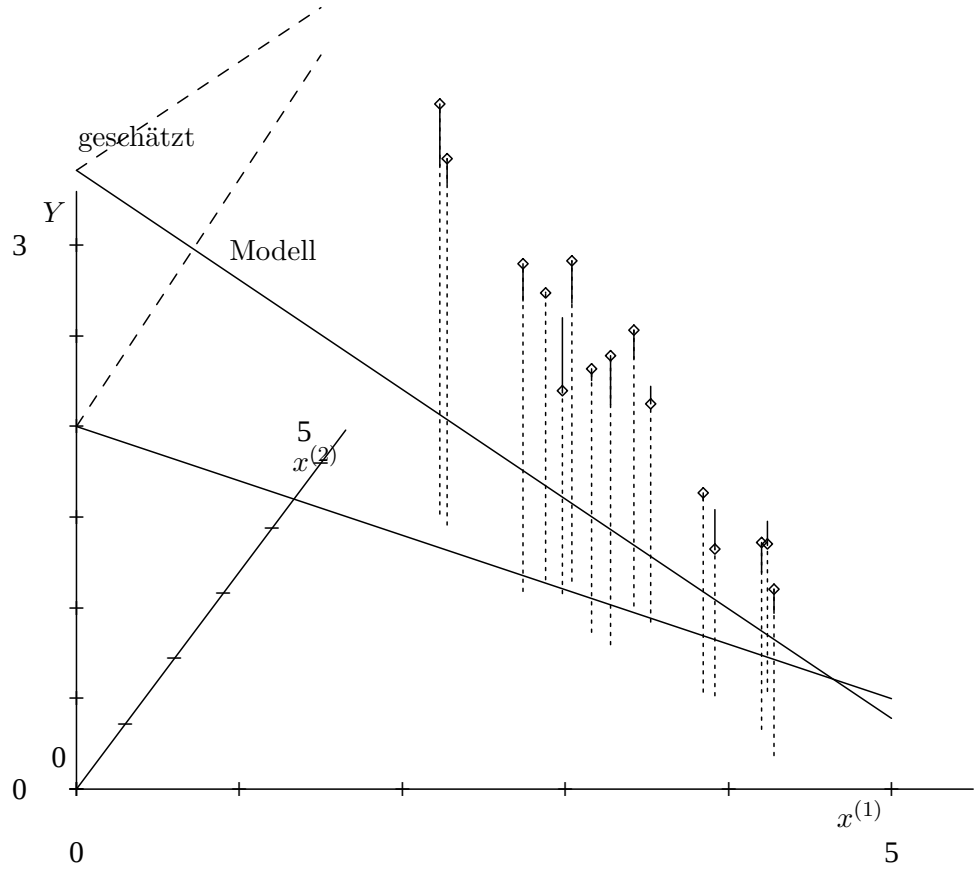

Abbildung 5.4.c: Kollinearität durch zwei stark korrelierte x-Variable. Die Y-Werte sind entsprechend dem „Modell“ simuliert. Eingezeichnet ist auch die „geschätzte“ Ebene.

d **Welches sind die Auswirkungen von Kollinearität?**
Im dargestellten Beispiel ist die Ebene, die dem linearen Regressionsmodell entspricht, in der einen Richtung, „entlang des Zauns“ gut, in der anderen (quer zum „Zaun“) schlecht bestimmt. Die Koeffizienten von $x^{(1)}$ und $x^{(2)}$, die Steigungen der Schnittgeraden der Ebene mit der „Aufriss-“ und „Seitenriss-Ebene“ ($x^{(1)}$-Y- und $x^{(2)}$-Y-Ebene), sind dann ebenfalls mit grosser Unsicherheit behaftet. Das führt zu grossen Standardfehlern für die geschätzten Koeffizienten. Deshalb kann man auf Grund des t-Tests (siehe 3.1.i) meistens die eine oder die andere Variable aus dem Modell streichen – aber oft nicht beide gleichzeitig!

e Die „Höhe“ der Ebene ist im Bereich der Daten mit der üblichen Genauigkeit durch diese bestimmbar, und in der Verlängerung des „Zauns“ recht gut extrapolierbar. An diesen Orten sind also Vorhersagen mit vernünftiger Genauigkeit anzugeben. Auf beiden Seiten des „Zauns“ nimmt aber die Genauigkeit rapide ab!

f **Wie entdeckt man Kollinearität?**
Die Probleme zeigen sich in den Standardfehlern, also auch in der Länge von Vertrauens-Intervallen und Prognose-Intervallen deutlich – sofern man darauf achtet!

Wir können aber auch direkter feststellen, ob eine Beziehung $\widetilde{x}_i^{(j)} \approx \widetilde{\sum}_{k\neq j} \widetilde{c}_k \widetilde{x}_i^{(k)}$ (annähernd) erfüllt ist. Das ist ein Regressionsproblem. Das Bestimmtheitsmass R_j^2 der Regression von $x^{(j)}$ auf alle übrigen erklärenden Variablen zeigt, wie stark eine solche Beziehung ist und ist also ein sinnvolles **Mass für Kollinearität**, das erst noch angibt, welche Variable „das Problem verursacht“.

Ein Mass, das man in Programmen findet, ist der so genannte **variance inflation factor** $\mathrm{VIF}_j = 1/(1 - R_j^2)$.

g **Was tun gegen Kollinearität?**
Wenn immer möglich, soll man die Beobachtungen so durchführen, dass das Problem vermieden wird. Bei Experimenten geben die x-Variablen die Versuchsbedingungen an. Kollinearität lässt sich durch geeignete Wahl der Versuchsbedingungen vermeiden.

h Können die Versuchsbedingungen nicht gewählt werden, dann kann man zu anderen X-Variablen übergehen, die besser bestimmte Koeffizienten ergeben.

Im Beispiel der beiden stark korrelierten Variablen ersetzt man diese durch ihre Summe und Differenz oder durch andere einfache Linearkombinationen, die nicht-kollineare neue Variable liefern.

Es gibt immer viele Möglichkeiten von linearen Transformationen, die zu „unkorrelierten“ x-Variablen führen. Für die Anwendung ist wesentlich, dass die neuen x-Variablen und damit ihre Koeffizienten leicht **interpretierbar** bleiben.

i Immer hilft das folgende Rezept:

- Die wichtigste Variable, sagen wir $x^{(1)}$, wird beibehalten;
- $x^{(2)}$ wird durch die Residuen einer Regression von $x^{(2)}$ auf $x^{(1)}$ ersetzt, also durch „den Teil von $x^{(2)}$, der von $x^{(1)}$ nicht erklärt wird“;
- Wenn die Kollinearität nicht von einem Paar von stark korrelierten Variablen stammt, sondern drei oder mehr Variable beteiligt sind, kann man allgemein die x-Variable mit dem höchsten R_j^2 wählen und durch Residuen bezüglich der Regression auf die anderen erklärenden Variablen ersetzen – und auch hier Modellwahl anwenden.

j Eine einfachere Lösung besteht darin, dass man die Variable mit dem höchsten R_j^2 aus dem Modell entfernt. (Das wird man oft auf Grund des t-Tests sowieso tun, siehe 5.3.m.)

k* In der Literatur wird auch ein Verfahren unter dem Namen „**ridge regression**“ vorgeschlagen. Ich finde es wenig hilfreich; die Ergebnisse sind schlecht interpretierbar.

5.5 Strategien der Modell-Entwicklung

a Die automatisierten Verfahren zur Modellwahl genügen für eine befriedigende explorative Analyse aus verschiedenen Gründen nicht:

- Wie erwähnt (5.3.k), ist die Auswahl der Variablen in der besten Gleichung entsprechend jedem Kriterium selbst **vom Zufall abhängig**, und man muss zumindest neben diesem „besten“ Modell die „**fast gleich guten**“ in Betracht ziehen. Um diese zu finden, ist die „all subsets“-Rechnung immerhin sehr hilfreich.
- Wir sind von einem festen Satz von Regressoren ausgegangen. Im Kapitel Residuen-Analyse haben wir gesehen, dass oft **Variable transformiert oder quadratische oder Wechselwirkungsterme eingeführt** werden sollten. Wollte man alle diese Möglichkeiten von Anfang an zum „festen Satz von Regressoren“ hinzufügen, dann würde dies schon bei wenigen ursprünglichen Eingangs-Variablen zu einer übergrossen Zahl von Regressoren führen. Solche Zusatzterme müssen daher mit anderen Mitteln auf ihre Eignung geprüft werden.
- Manchmal liefern die Verfahren Modelle, die mit dem gesicherten **Fachwissen** nicht übereinstimmen. Beispielsweise kann der geschätzte Koeffizient eines Regressors ein Vorzeichen haben, das „nicht stimmen kann“. Bevor man eine ganz neue Theorie entwickelt, wird man weitere Modelle prüfen wollen.

b Zur Modellwahl braucht es also eine **„Strategie“**, die allerdings noch selten formuliert und diskutiert wird. Sie wird eher als Kunst angesehen, die allenfalls durch Beispiele zu vermitteln sei.

c Die Modellwahl findet innerhalb eines gesamten Ablaufs der Datenanalyse statt, deren „nullter“ Schritt immer lautet:

0. Daten kennenlernen und bereinigen. Man macht sich mit der genauen **Bedeutung aller Variablen** bekannt und legt kurze, informative **Variablennamen** fest, die alle Beteiligten gut interpretieren können. Dann überprüft man **unmögliche oder unplausible Werte** und Ausreisser für alle Variablen im Datensatz, korrigiert wenn nötig und setzt verbleibende unmögliche Werte auf „fehlend“. In diesem Zusammenhang bewährt es sich (wenn die Zahl der Variablen nicht allzu gross ist), die **Streudiagramm-Matrix** aller Variablen (mindestens der Variablen mit stetigem oder geordnetem Wertebereich) zu studieren.

Schliesslich untersucht man die Häufigkeiten und Auffälligkeiten des Auftretens von **fehlenden Werten.** Wenn sie mit spürbarer Häufigkeit auftreten, muss eine eigene Strategie zu ihrer Behandlung festgelegt werden, die wir hier nicht besprechen wollen.

Wer hier zu wenig investiert, büsst später!

d Wir werden sehen, dass die geeignete Strategie vom Zweck der Studie abhängt (vergleiche 5.1.a). Gehen wir zunächst davon aus, dass es der Zweck der Studie sei, **die erklärenden Variablen zu identifizieren**, die die Zielgrösse beeinflussen.

Dieses Ziel ist nicht so klar, wie es zunächst tönt. Am befriedigsten wäre es, die **Ursachen** für die Werte der Zielvariablen zu finden. Das ist aber mit einer explorativen Analyse von Daten nicht zu erreichen, sondern nur mit geplanten Versuchen, soweit solche möglich sind (siehe Versuchsplanung).

Es geht also darum, ein Modell zu finden, das die vorliegenden Daten gut beschreibt und möglichst keine systematischen Abweichungen übriglässt – die zufälligen sind nicht zu vermeiden.

e **Eine Strategie** zur Analyse solcher Daten kann etwa so aussehen:

1. “First aid” Transformationen. Allgemeine statistische Gesichtspunkte (4.4.g) und spezifisches Fachwissen führen für jede Variable zu einer plausiblen „Skala“ – oft einer transformierten ursprünglichen Grösse (englisches Stichwort *re-expression*).

2. Ein grosses Modell. Man passt eine Gleichung an, die vermutlich zu viele erklärende Variable enthält, nämlich

- alle Variablen, falls deren Anzahl höchstens einen Fünftel der Anzahl Beobachtungen ausmacht (* allenfalls setzt man gar ein „general additive model“ an),
- alle Variablen, die entsprechend Plausibilitäts-Überlegungen und Fachwissen einen Einfluss auf die Zielgrösse haben könnten,
- die Variablen, die mit einem „Schrittweise-Vorwärts-Verfahren“ mit grosszügigem Abbruchkriterium (hohem P-Wert) ausgewählt werden.

Falls gemäss Fachwissen Wechselwirkungen zwischen erklärenden Variablen erwartet werden, sollen diese ebenfalls einbezogen werden.

Wenn möglich sollten robuste Schätzmethoden verwendet werden.

3. Überprüfung des zufälligen Teils:

- Ausreisser in den Residuen,
- Verteilung der Residuen,
- Gleichheit der Varianzen,

- Unabhängigkeit der Fehler.

Es kann auf Grund der Ergebnisse angezeigt sein,

- die Zielgrösse zu transformieren,
- Gewichte einzuführen,
- robuste(re) Methoden zu verwenden, soweit dies nicht schon sowieso geschieht,
- Blöcke in der zeitlichen Abfolge (oder geographischen Anordnung) zu bilden und eine entsprechende nominale erklärende Variable einzuführen, um serielle Korrelationen mit dem funktionalen Teil statt mit korrelierten Fehlern E_i zu beschreiben,
- Schätzmethoden zu verwenden, die den Korrelationen Rechnung tragen.

Allerdings müssen die Modell-Voraussetzungen für das angegebene Analyse-Ziel nur grob erfüllt sein.

4. Nicht-Linearitäten. Streudiagramme der Residuen gegen die erklärenden Variablen können zu Transformationen der erklärenden Variablen oder zu quadratischen Termen führen.

5. Automatisierte Variablen-Wahl mit „all subsets“, notfalls mit schrittweisem Rückwärts-Verfahren.

6. Variable hinzufügen. Streudiagramme der Residuen gegen die erklärenden Variablen, die nicht im Modell sind – auch gegen jene, die gerade eliminiert wurden – und wie in Schritt 4 verfahren.

7. Wechselwirkungen. Man prüft, ob Wechselwirkungsterme zwischen den Variablen, die bereits im Modell sind, zur Verbesserung der Anpassung führen. Wechselwirkungen mit Variablen, die mangels Einfluss auf die Zielgrösse nicht ins Modell aufgenommen werden, sind unerwünscht und selten nützlich (siehe Cox and Snell, 1981, S. 126). Wenn solche ins Modell aufgenommen werden, nimmt man auch die beteiligten (nicht-signifikanten) erklärenden Variablen wieder ins Modell auf

8. Einflussreiche Beobachtungen. Man sucht multivariate Ausreisser im Raum der x-Variablen, also hohe Leverage-Werte H_{ii}, und überprüft allgemein einflussreiche Beobachtungen (* mit robusten Methoden).

9. Kritik mit Fachwissen. Wenn das Modell Terme enthält, die unplausibel sind oder deren geschätzter Koeffizient das „falsche“ Vorzeichen hat, lässt man sie weg, sofern sich dadurch die Anpassung nicht allzu stark verschlechtert.

10. Anpassung prüfen. Man vergleicht die geschätzte Varianz der Fehler im Modell mit einer anderen Schätzung, beispielsweise einer minimalen, sicher vorhandenen Streuung (Messgenauigkeit) oder einer Schätzung aus wiederholten oder „benachbarten“ Messungen (4.8.a). Falls dieser Vergleich befriedigend ausfällt, kann man zu Schritt 12 gehen.

11. Revision. Falls sich das Modell seit Schritt 4 merklich verändert hat, geht man dorthin oder gar zu Schritt 3 zurück.

12. Entfernte Terme überprüfen. Wenn in Schritt 8 Terme unterdrückt wurden, muss man nochmals überprüfen, wie wichtig sie jetzt erscheinen.

f Die Strategie soll sich nach dem Zweck der Studie richten. Die Absicht sei nun, **eine Hypothese zu überprüfen**, genauer wollen wir beispielsweise überprüfen, ob der Koeffizient von $x^{(1)}$ null sein kann.

Dann wird man die Strategie anpassen:

1. Daten-Transformation (soweit von der Fragestellung her zugelassen), wie oben.

2-7. In gewissen Fällen ist auch hier eine Modellwahl möglich oder nötig. Man folgt dann den Schritten 2-7 der vorhergehenden Strategie, aber mit „Nebenbedingungen“:

- $X^{(1)}$ bleibt immer im Modell,
- man kümmert sich nur um Variable, die eine merkliche Vergrösserung von R^2 bewirken oder die mit $X^{(1)}$ korreliert sind,
- eventuell ist die Transformation der Zielgrösse und von $X^{(1)}$ von der Fragestellung her nicht erlaubt.

8. Kollinearitäten. Genaue Überprüfung der X-Variablen im Modell, die mit $X^{(1)}$ korreliert sind („kritische“ X-Variable). **Aufgepasst:** Die Fragestellung selbst ändert sich, wenn man Variable ins Modell einbezieht, die mit der zu testenden Variablen korreliert sind. Die Beurteilung des Modells vom Fachwissen her ist daher hier unumgänglich.

9. Annahmen über die Zufalls-Fehler überprüfen. Gegebenenfalls muss man die Testmethode anpassen (generalized least squares, robuster Test, ...). Die Einhaltung der Voraussetzungen ist hier wichtig.

10. Test-Resultate. Man berechnet die P-Werte für die Modelle mit und ohne kritische Variable.

g ▷ Im **Beispiel der Baukosten** liegt eine solche Fragestellung vor. Es soll ja herausgefunden werden, ob die Kostengarantie einen (vermindernden) Einfluss auf die Zielgrösse Kosten hat. Verschiedene Modelle haben zwiespältige Antworten geliefert. Die Variable `N`, die zählt, wie viele Werke das gleiche Ingenieur-Team bereits erbaut hat, ist eine „kritische“ Variable. Mit fachlicher Beurteilung kommt man zu einem überraschend klaren Ergebnis, das wir aber hier nicht ausführen wollen. ◁

h Ein dritter Zweck: **Vorhersage**. Hier ist noch keine Strategie formuliert. Es kommt bei dieser Fragestellung nur darauf an, gute angepasste Werte zu erhalten. Kollinearitäten sind unwichtig. Für Prognose-Intervalle ist die Form der Verteilung der Fehler wesentlich.

5.S S-Funktionen

a Die Wichtigkeit eines Terms in der Modellgleichung wird von `drop1` geprüft, siehe 3.S.0.f. Diese Funktion liefert nicht nur Test-Resultate (wenn man `test="F"` setzt), sondern (vor allem) einen AIC-Wert (5.3.g), der den Vergleich zwischen Modellen mit verschiedenen Anzahlen von Regressoren ermöglicht.

Analog zu `drop1` gibt es eine Funktion `add1`, die prüft, ob Terme zum bestehenden Modell hinzugefügt werden sollen.

b **Funktion `step`.** Die schrittweisen Verfahren sind in der Funktion `step` implementiert. Als erstes Argument verlangt `step` ein `lm`- (oder `regr`-) Resultat. Wenn nichts weiteres gegeben wird, dann läuft die Modellwahl schrittweise rückwärts. Man kann aber als Argument `scope=~.+X5+X6` zusätzliche Terme (`X5` und `X6`)angeben und auch festlegen, dass gewisse Terme in allen Modellen vorkommen müssen (`scope=list( lower=~X1, upper=~.+X5+X6)`). Will man ein Vorwärts-Verfahren vom „leeren Modell“ an durchführen, dann muss man zunächst „das leere Modell anpassen“, also `t.r <- lm(Y~1, data=...)` eingeben. Beispiel:

```
> t.r <- lm(K 1,data=d.nuk)

> t.rs <- step(t.r,
  scope=paste(" ",paste(names(d.nuk)[-1],collapse="+")))
```

Das schrittweise Verfahren stoppt, wenn die Grösse `AIC` nicht mehr abnimmt. Oft will man sehen,

welche Variablen in weiteren Schritten eliminiert würden. Dazu kann man das Argument `k=100` benützen. Dann ist zwar `AIC` nicht mehr, was es sein soll, aber das Rückwärts-Verfahren läuft weiter, meistens bis zum leeren Modell.

c **Funktion** `regsubsets` `library(leaps)`. Ermöglicht die Prüfung aller Gleichungen (all subsets).

```
> t.ras <- regsubsets(K .,data=d.nuk,nbest=3)
> summary(t.ras)
```

Mit `nvmax=` maximale Anzahl Regressoren und mit `force.in=` kann man den Aufwand reduzieren und deshalb (noch) grössere Modelle verarbeiten.

```
> t.ras <- regsubsets(x=d.nuk[,-1], y=d.nuk[,"K"],
  force.in=c("G","D"), nvmax=8,nbest=3)
```

d **Funktion** `update.` Die Idee der Funktion `update` ist es, einzelne Modell-Spezifikationen ändern zu können und auf einfache Art eine neue Modell-Anpassung zu erwirken. Beispielweise führt

```
> update(t.r,formula= .-BW)
```

zu einem Modell, das sich von dem in `t.r` abgespeicherten Modell-Ergebnis nur dadurch unterscheidet, dass der Term `BW` im Modell weggelassen wird. – Allerdings kann es gerade so effizient und transparent sein, mit „copy-paste“ den vorhergehenden Aufruf von `lm` zu duplizieren und abzuändern.

e Die **Lasso-Methode** ist im package `library(lasso2)` unter dem Namen `l1ce` implementiert. Die Standardisierung der Variablen muss man selber organisieren. Das Argument `bound` legt die relative Schranke b fest (ausser man setzt `absolute.t=TRUE`). Man kann diesem Argument mehrere Werte geben (einen Vektor), beispielsweise
`bound=seq(0.05,1,0.05)` und erhält dann eine ganze Liste von Regressionsresultaten. Mit `plot(...)` erhält man eine Darstellung der erhaltenen Koeffizienten in Abhängigkeit von der Schranke.

```
> t.r <- l1ce(K .,data=t.d, bound=seq(0.05,1,0.05))
> plot(t.r)
> summary(t.r[[5]])
```

6 Ergänzungen

6.1 Fehlerbehaftete erklärende Variable

a Die erklärenden Variablen erscheinen in den besprochenen Modellen nicht als Zufallsvariable, obwohl sie oft ebenso zufällig sind wie die Zielgrösse. Wir haben dies bisher vernachlässigt und immer so getan, als ob die x-Werte feste, vorgegebene Zahlen seien. Eine formale Begründung dafür besteht darin, dass die Verteilungen gemäss Modell als bedingte Verteilungen, gegeben die x_i-Werte, aufgefasst werden.

b Wir wollen nun untersuchen, was geschieht, wenn die erklärende Variable, deren Einfluss auf die Zielgrösse von Interesse ist, nur ungenau gemessen oder beobachtet werden kann. Wir stellen uns zwei **„latente" Variable** u und v vor, die deterministisch zusammenhängen – im einfachsten Fall linear,

$$v = \widetilde{\alpha} + \widetilde{\beta} u \ .$$

Sie können aber beide nicht exakt beobachtet werden, sondern nur mit zufälligen Fehlern, also

$$X_i = u_i + D_i \ , \qquad Y_i = v_i + E_i = \widetilde{\alpha} + \widetilde{\beta} u_i + E_i \ .$$

Die Fehler D_i sollen ebenso wie die Messfehler E_i normalverteilt sein,

$$D_i \sim \mathcal{N}\langle 0, \sigma_D^2 \rangle \ , \qquad E_i \sim \mathcal{N}\langle 0, \sigma_E^2 \rangle$$

– und unabhängig. Die u_i und damit auch die v_i seien feste Zahlen – wie es in der linearen Regression die x_i sind. Unser Interesse gilt dem Koeffizienten $\widetilde{\beta}$ und eventuell auch $\widetilde{\alpha}$.

Für $\sigma_D^2 = 0$ wird u_i gleich der beobachtbaren Variablen X_i, und man erhält das Modell der einfachen linearen Regression.

c Das beschriebene Modell ist der einfachste Fall einer Regression mit fehlerbehafteten erklärenden Variablen (**errors-in-variables regression**). Man spricht auch von einer **funktionalen Beziehung** (**functional relationship**). Wenn die wahren Werte u_i der erklärenden Variablen als zufällig statt als fest aufgefasst werden, dann heisst das Modell eine **structural relationship**.

d Den Unterschied zwischen dem Modell der funktionalen Beziehung und der einfachen linearen Regression wird in Abbildung 6.1.d an einem simulierten Beispiel gezeigt.

Vergleicht man die Beobachtungen mit den Punkten, die man erhalten hätte, wenn die erklärende Variable u ohne Messfehler verfügbar wäre, dann sieht man, dass sich die Streuung der Punkte in x-Richtung ausdehnt.

e Die Steigung der Regressionsgeraden, die mit Kleinsten Quadraten bestimmt wird, ist gleich

$$\widehat{\beta}_{LS} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\widehat{\mathrm{cov}}\langle X, Y \rangle}{\widehat{\mathrm{var}}\langle X \rangle} \ ,$$

also gleich dem Quotienten aus der (empirischen) Kovarianz zwischen X und Y und der (empirischen) Varianz von X. In Abbildung 6.1.d zeigt sich, dass die geschätzte Gerade viel flacher ist als die wahre. Ist das Zufall?

Um die gewünschte Steigung $\widetilde{\beta}$ zu bestimmen, müssten wir die X_i-Werte durch die u_i ersetzen können. Was würde sich ändern? Da die Zufallsfehler D_i unabhängig sind von den E_i und den u_i und damit auch von den $Y_i = \widetilde{\beta} u_i + E_i$, verändert sich die Kovarianz nicht (genauer: die empirische Kovarianz zwischen U und Y hat den gleichen Erwartungswert wie diejenige

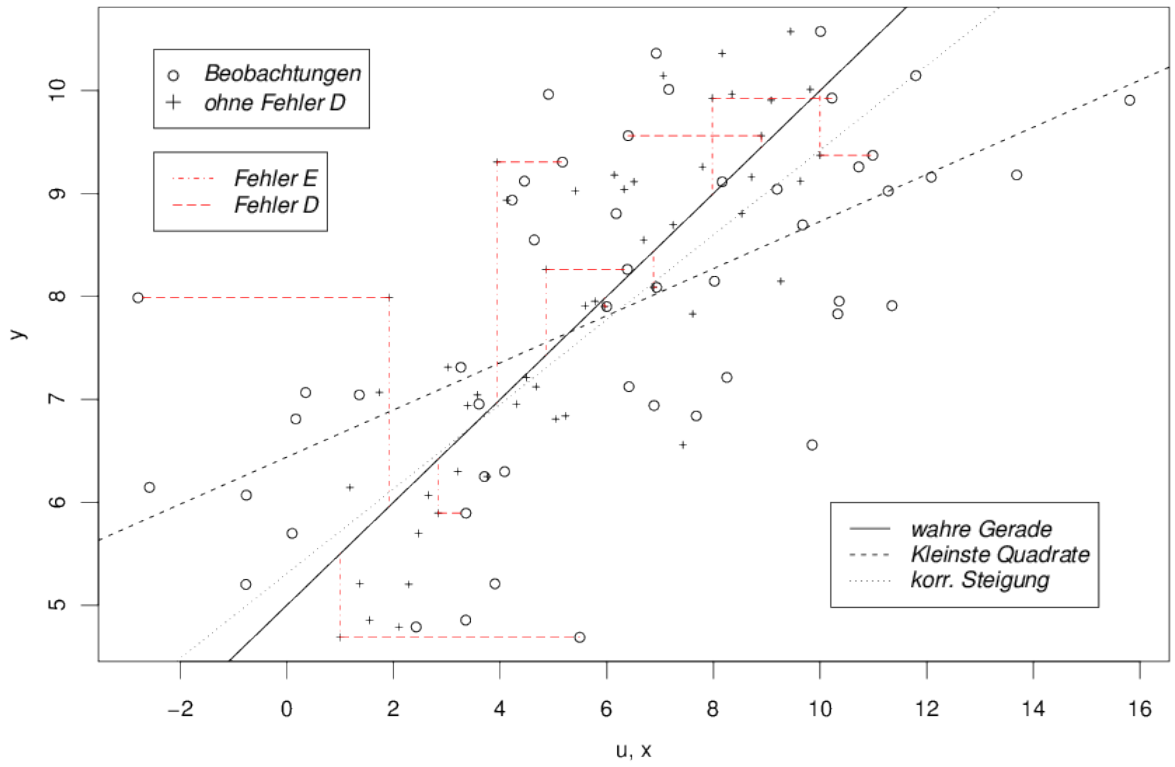

Abbildung 6.1.d: Veranschaulichung des Modells mit einer fehlerbehafteten erklärenden Variablen. 50 Beobachtungen wurden mit dem Modell $v = 5 + 0.5 \cdot u$, $\sigma_D = 3$ und $\sigma_E = 1$ simuliert. Die Beobachtungen ($\circ$) streuen in x-Richtung stärker als die „Beobachtungen ohne Fehler in x-Richtung"($+$), die aus der Simulation hier bekannt sind. Zusätzlich zur „wahren" Geraden sind die mit Kleinsten Quadraten geschätzte und die korrigierte Gerade eingezeichnet.

zwischen X und Y). Die empirische Varianz der u_i ist dagegen im Erwartungswert um σ_D^2 kleiner als die empirische Varianz der X_i. Deshalb wird der Nenner in der obigen Formel zu gross, während der Zähler den richtigen Erwartungswert hat. Das führt zu einer systematisch zu flachen Geraden.

Der systematische Fehler lässt sich aber leicht korrigieren, wenn σ_D bekannt ist: Wir setzen im Nenner $\widehat{\mathrm{var}}\langle X\rangle - \sigma_D^2$ statt $\widehat{\mathrm{var}}\langle X\rangle$ ein. Anders gesagt,

$$\begin{aligned} \widehat{\beta} &= \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2 - \sigma_D^2} = \widehat{\beta}_{LS}/\widehat{\kappa} \\ \widehat{\kappa} &= \frac{\widehat{\mathrm{var}}\langle X\rangle - \sigma_D^2}{\widehat{\mathrm{var}}\langle X\rangle} \end{aligned}$$

Die Grösse $\widehat{\kappa}$ schreiben wir mit Hut ($\widehat{\ }$), da sie (über die u_i) von der Stichprobe abhängt. Wenn die „wahren" Werte u_i der erklärenden Variablen selbst als Zufallsvariable modelliert werden, ist der Modellparameter, der durch $\widehat{\kappa}$ geschätzt wird gleich $\kappa = \mathrm{var}\langle U\rangle / \mathrm{var}\langle X\rangle$.

Die Grösse κ wird in der Literatur als „Abschwächungs-Koeffizient" *(attenuation coefficient)* bezeichnet. Er misst, wie viel flacher die mit der üblichen Methode geschätzte Steigung wird als die gesuchte Steigung $\widetilde{\beta}$. Er wird auch *reliability ratio* genannt, da er die „Verlässlichkeit" der Variablen X als Mass für die gewünschte Variable U misst.

f Den zweiten Parameter $\widetilde{\alpha}$, den Achsenabschnitt der gesuchten Geraden, schätzt man wie früher nach der Formel $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{X}$ (2.2.c) – hier natürlich mit der soeben eingeführten erwartungstreuen Schätzung $\widehat{\beta}$.

Bevor wir den Fall diskutieren, in dem σ_D nicht bekannt ist, soll ein Beispiel folgen.

g Im **Beispiel der Schadstoffe im Tunnel** (1.1.f) sollen die Emissionsfaktoren für die beiden Fahrzeugklassen „Personenwagen“ und „Lastwagen“ bestimmt werden. In der erwähnten Untersuchung im Gubrist-Tunnel konnte die Anzahl Fahrzeuge einer Fahrzeugklasse nicht genau bestimmt werden. Die systematische Abweichung (systematische Unterschätzung des Anteils der Lastwagen am Gesamtverkehr durch die Schlaufen-Klassierung) kann durch „Eichung“ (siehe 1.1.h und 6.2 unten) korrigiert werden, aber der Erfassungsfehler wird auch zufällig streuen. Die Daten, die zur Eichung dienen, liefern auch eine Schätzung der Varianz dieser zufälligen Fehler, also von σ_D^2, nämlich 0.0213^2.

Wenn die Schätzung diese zufälligen Fehler nicht berücksichtigt, wird die Gerade zu flach geschätzt, wie wir gesehen haben. Für Schadstoffe, die von den Lastwagen stärker emittiert werden, bewirkt das, dass ihre Emissionen unterschätzt und jene der Personenwagen überschätzt werden – und umgekehrt für Schadstoffe, die von Personenwagen in grösserer Menge ausgestossen werden. Abbildung 6.1.g zeigt die Daten der Studie, die für die Berechnung der Emissionsfaktoren brauchbar waren. In den Nachtstunden herrschte geringer Verkehr, was zu so kleinen Luftgeschwindigkeiten führt, dass die Emissionen nicht mehr richtig berechnet werden konnten. (Die Rechnung setzt laminare Luftströmung voraus.) Die flachere eingezeichnete Gerade resultiert aus einer robusten Schätzung ohne Berücksichtigung der Fehler der erklärenden Variablen; die steilere ist die korrigierte. Der Korrekturfaktor $1/\kappa$ für die Steigung beträgt 1.12. Der Achsenabschnitt, der den Emissionsfaktor für die Personenwagen misst, wird geringfügig von 1254 auf 1169 korrigiert, während der geschätzte Emissionsfaktor für die Lastwagen $(\widehat{\alpha} + \widehat{\beta})$ von 14580 um 10% auf 16065 klettert.

h Im Umweltbereich gibt es viele ähnliche Fragestellungen, vor allem auch auf dem Gebiet des Zusammenhangs von **Gesundheitsschäden** mit der **Exposition gegenüber Risikostoffen**: Die Schädigungen werden systematisch unterschätzt, wenn die Ungenauigkeit der Erfassung der Exposition nicht berücksichtigt wird.

i Statt der Ungenauigkeit der erklärenden Variablen X kann auch das **Verhältnis** $\gamma = \sigma_E/\sigma_D$ der Ungenauigkeiten von X und Y (näherungsweise) **bekannt** sein. Durch Umskalierung der einen Variablen ($X \to \gamma X$) lässt sich dann erreichen, dass beide gemäss Annahme die gleiche Genauigkeit aufweisen. Dann liefert die orthogonale Regression die richtige Schätzung.

j Die **orthogonale Regression** minimiert statt der Quadratsumme der vertikalen Abweichungen $r_i\langle a, b\rangle$ (Methode der Kleinsten Quadrate) diejenige der orthogonalen Abstände $d_i\langle a, b\rangle$ (Abbildung 6.1.j).

Das ergibt eine steilere Gerade als die Kleinsten Quadrate der r_i. (* Sie fällt mit der ersten **Hauptkomponente** einer Hauptkomponenten-Analyse zusammen – ein Thema der Multivariaten Statistik.)

Wenn die Masseinheit von X oder Y geändert wird, ändert sich die mit orthogonaler Regression bestimmte Gerade in einer Weise, die schwierig interpretierbar ist. (Probieren Sie Extremfälle aus!) Man soll diese Art der Regression daher nur auf geeignet standardisierte Daten anwenden.

Wenn X und Y auf empirische Standardabweichung 1 transformiert werden, ergibt sich immer eine Steigung von $+1$ oder -1 für die optimale Gerade, unabhängig von der „Stärke“ des Zusammenhangs. (Wenn die Korrelation 0 ist, ist die Gerade für standardisierte Variable unbestimmt.)

k Die bisher besprochenen Schätzmethoden setzen voraus, dass die Varianz σ_D^2 der Zufallsfehler D_i oder das Verhältnis σ_E/σ_D bekannt sei. Wenn über die **Varianzen** σ_D und σ_E **nichts**

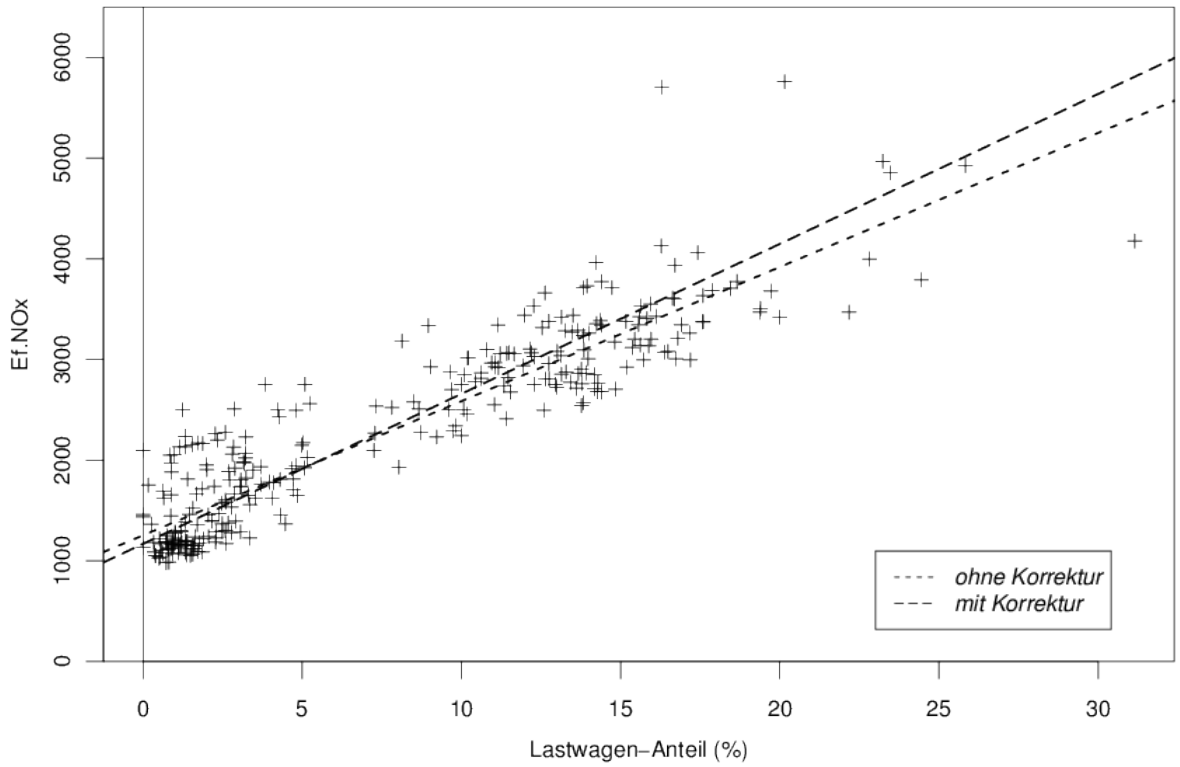

Abbildung 6.1.g: Emissionsfaktor für NO_X und Lastwagen-Anteil im Beispiel der Schadstoffe im Tunnel, für die Zeitabschnitte mit genügender Luftgeschwindigkeit. Die Geraden stellen die Schätzung mit und ohne Berücksichtigung der Messfehler des Lastwagen-Anteils dar.

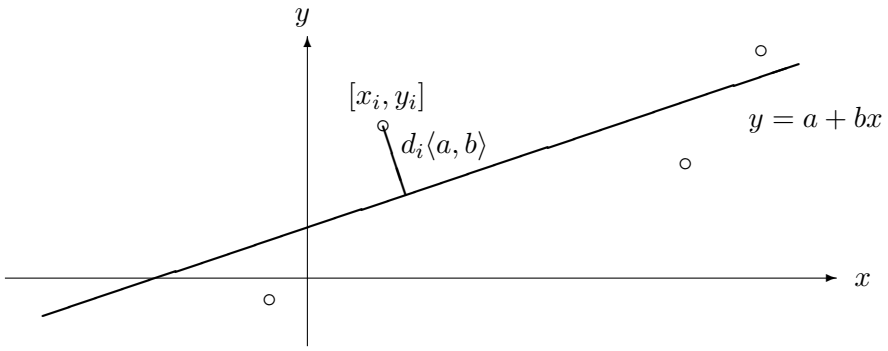

Abbildung 6.1.j: Zur Definition der orthogonalen Regression

bekannt ist, wird das Problem in einem grundlegenden Sinn schwierig. Wenn die wahren Werte u_i als normalverteilte Zufallsvariable $U_i \sim \mathcal{N}\langle\mu, \sigma_U^2\rangle$ modelliert werden, dann lässt sich zeigen, dass die Parameter auch mit unendlich vielen Beobachtungen nicht geschätzt werden können. Es führen dann nämlich verschiedene Parametersätze ($[\widetilde{\beta},\ \widetilde{\alpha},\ \sigma_D\ , \sigma_E\ , \sigma_U\ ]$) zur genau gleichen Verteilung der Beobachtungen $[X_i, Y_i]$. Das Modell ist „**nicht identifizierbar**“.

Bei anderen Annahmen über die u_i ist die Identifizierbarkeit zwar theoretisch gegeben, aber für vernünftige Stichprobenumfänge nicht wirklich erreichbar. Man braucht in der Praxis also eine zusätzliche Information.

Kennt man wenigstens eine obere Schranke („grösser als ... kann σ_D nicht sein“), dann kann man den schlimmsten Fall durchrechnen und aus dem Unterschied zu den Resultaten für $\sigma_D = 0$ abschätzen, ob das Problem bedeutsam sei oder nicht.

l Wieso wird diese Methodik so **selten behandelt** und noch weniger angewandt? Nicht nur wegen mangelndem Wissen!

Wenn man Y „**vorhersagen**“ oder interpolieren will, so macht dies meistens nur für gegebene X-Werte Sinn, nicht für gegebene u-Werte, da man diese ja nicht beobachten kann. Dann ist die gewöhnliche Regressionsrechnung angebracht. Allerdings muss gewährleistet sein, dass die X-Werte für die neuen Beobachtungen auf gleiche Weise zustande kommen wie die Daten, mit denen das Modell angepasst wurde.

Wenn die Frage interessiert, ob ein **Einfluss von** u **auf** Y (oder v) vorhanden sei, so muss man die Nullhypothese $\widetilde{\beta} = 0$ **testen**. Wenn die Hypothese gilt, ist auch die Steigung im Regressionsmodell von Y auf X null, und man kann den Test der gewöhnlichen Regressionsrechnung anwenden.

m *Literatur:* Wetherill (1986) gibt eine kurze, kritische Darstellung. Fuller (1987) ist ein umfassendes Werk über dieses Thema.

6.2 Eichung

a „Ausgleichs-Geraden“ werden oft verwendet, um eine Mess-Methode zu **eichen** oder um aus dem Resultat einer (billigen) Mess-Methode das Resultat einer anderen (teuren) zu „schätzen“.

Für die Bestimmung des Zusammenhangs geht man meist von bekannten „wahren“ Werten x_i (oder Werten der präzisen, teuren Mess-Methode) aus und bestimmt dazu die Werte Y_i der zu untersuchenden Methode. Es wird beispielsweise jeweils für eine chemische Lösung mit bekannter Konzentration die Absorption von Licht bei einer bestimmten Wellenlänge gemessen. (Meistens muss zunächst eine Reaktion durchgeführt werden, die die interessierende chemische Substanz in eine optisch erfassbare Substanz verwandelt.)

In der Anwendung der Eich-Geraden (oder -Kurve) ist umgekehrt der Wert Y der fraglichen Messmethode vorgegeben, und man will den zugehörigen wahren Wert x schätzen. Im Beispiel will man aus der Absorption die Konzentration der Lösung ausrechnen. Man verwendet die Regressions-Beziehung also in der „falschen“ Richtung. Daraus ergeben sich Probleme. Ihre Behandlung findet man auch unter dem Titel **inverse regression** oder **calibration**.

b Wir wollen hier eine einfache Behandlung vorstellen, die ein brauchbares Resultat ergibt, wenn der Zusammenhang eng (das Bestimmtheitsmass gross, beispielsweise über 0.95) ist.

Zunächst nehmen wir an, dass die x-Werte keine Messfehler aufweisen. Das erreicht man, indem man im Beispiel sehr sorgfältig erstellte Eich-Lösungen verwendet. Für mehrere solche Lösungen mit möglichst unterschiedlichen Konzentrationen führt man jeweils mehrere (möglichst) unabhängige Messungen (Aufbereitung und Ablesung des optischen Messgerätes) der Grösse Y durch. Daraus bestimmt man mit den besprochenen Methoden eine einfache lineare Regressionsgleichung – sofern Linearität vorhanden ist. Dies führt zu Schätzungen der Parameter α, β und σ und zu geschätzten Standardfehlern von $\widehat{\alpha}$ und $\widehat{\beta}$.

Wenn nun für eine zu messende Probe der Wert y abgelesen wird, ist klar, wie ein zugehöriger x-Wert bestimmt wird:

$$\widehat{x} = (y - \widehat{\alpha})/\widehat{\beta} \; .$$

c Die Frage stellt sich, wie genau dieser Wert ist.

Die Antwort lässt sich formulieren, indem wir x als Parameter ansehen, für den ein Vertrauensintervall gesucht ist. Ein solches Intervall ergibt sich (wie immer) aus einem Test. Nehmen wir als Nullhypothese $x = x_0$ an! Wie wir im Abschnitt über Vorhersage gesehen haben, liegt Y mit Wahrscheinlichkeit 0.95 in im Vorhersage-Intervall

$$\widehat{\alpha} \; + \; \widehat{\beta} x_0 \pm b \quad \text{mit } b = q_{0.975}^{t_{n-2}} \, \widehat{\sigma} \, \sqrt{1 + \tfrac{1}{n} + (x_0 - \bar{x})^2 / \mathrm{SSQ}^{(X)}} \; ,$$

das in Abbildung 6.2.c wie in Abbildung 2.4.c – gleich für alle möglichen x_0 – dargestellt ist.

Das Intervall bildet deshalb ein Annahmeintervall für die Grösse Y (die hier die Rolle einer Teststatistik spielt) unter der Nullhypothese $x = x_0$.

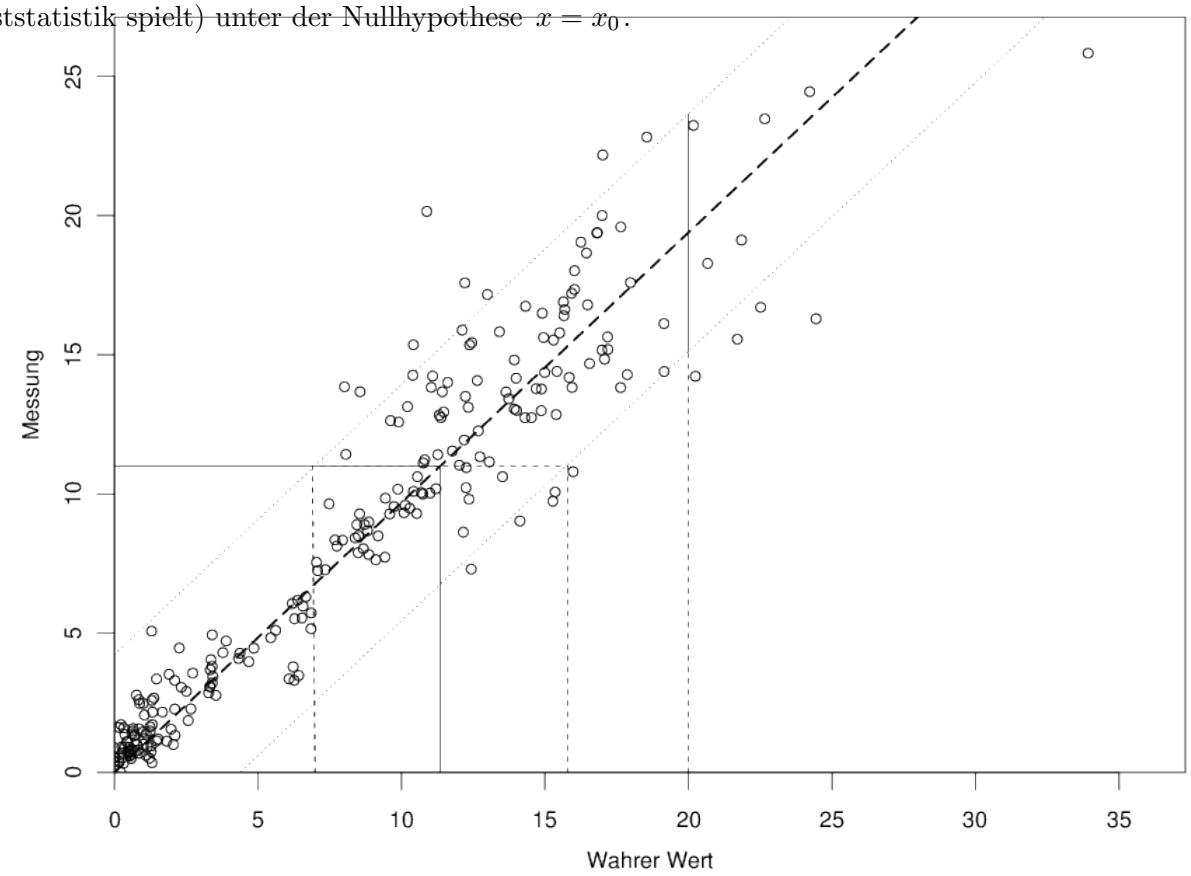

Abbildung 6.2.c: Veranschaulichung der Verwendung einer Eichgeraden für einen Messwert von 11. Zum Vergleich die Verwendung für eine Vorhersage des Messwertes bei einem wahren Wert von 20.

Die Abbildung veranschaulicht nun den weiteren Gedankengang: Messwerte y sind mit Parameterwerten x_0 vereinbar im Sinne des Tests, wenn der Punkt $[x_0, y]$ zwischen den eingezeichneten Kurven liegt. In der Figur kann man deshalb ohne Schwierigkeiten die Menge der x_0-Werte bestimmen, die mit der Beobachtung y verträglich sind. Sie bilden das eingezeichnete Intervall – das Vertrauensintervall für x_0. In sehr guter Näherung hat dies den Mittelpunkt $\widehat{x}$ und die Breite $2 \cdot b/\widehat{\beta}$, ist also gleich

$$(y - \widehat{\alpha})/\widehat{\beta} \pm b/\widehat{\beta} \; .$$

d* Einige weitere Stichworte:

- Fehlerbehaftete x-Werte: Man verwende eine Schätzung der „wahren Geraden“ $\widetilde{\alpha} + \widetilde{\beta}\, x$.
- Überprüfung der Linearität und anderer Modell-Annahmen ist wichtig!
- Periodische Eichung: sollte nicht mit Einzelmessungen erfolgen.

6.3 Allgemeinere Modelle für stetige Zielgrössen

a Das Modell der multiplen linearen Regression ist in mancher Hinsicht das einfachste, das eine Abhängigkeit einer Zielgrösse von mehreren erklärenden Variablen darstellt. In diesem und im nächsten Abschnitt sollen stichwortartig die bekannteren anderen Regressionsmodelle aufgeführt werden, um den Einstieg in die Spezialliteratur zu erleichtern.

b **Verteilung der Fehler.** Wenn man im linearen Regressionsmodell für die zufälligen Fehler nicht eine Normalverteilung, sondern irgendeine andere Verteilungsfamilie voraussetzt, führt die Maximierung der Likelihood nicht mehr zu Kleinsten Quadraten. Einige oft verwendete Familien werden durch die Verallgemeinerten Linearen Modelle abgedeckt; sie werden im nächsten Abschnitt behandelt (6.4.e). Andere Familien führen zu so genannten **M-Schätzungen** oder Huber-Typ-Schätzern, die bei geeigneter Wahl eine beschränkte **Robustheit** gegenüber Ausreissern aufweisen. Um gute Robustheit zu erreichen, braucht man allerdings andere Methoden, die unter dem Namen „Methoden mit **hohem Bruchpunkt**“ (high breakdown point regression) oder „mit beschränktem Einfluss“ (bounded influence regression) bekannt sind.

Literatur: Rousseeuw and Leroy (1987), Venables and Ripley (1997), Kap. 8.3-8.4.

c **Transformationen.** Im Modell der multiplen linearen Regression setzen wir voraus, dass die Regressionsfunktion h der Form nach bekannt sei, dass aber die Koeffizienten α, $\beta_1, \ldots,$ β_m unbekannt seien. Über sie will man aus den Daten Rückschlüsse ziehen. Wir haben diskutiert, dass man die erklärenden Variablen und eventuell die Zielgrösse auch **transformieren** darf. Die allgemeinste Form, die durch solche Veränderungen möglich wird, ist, wenn wir wieder die ursprünglichen erklärenden Variablen mit $U^{(k)}$ bezeichnen,

$$g\langle Y_i\rangle = \alpha + \sum_{j=1}^{m} \beta_j \; h_j\left\langle u_i^{(1)}, \ldots, u_i^{(m')}\right\rangle + E_i \; ,$$

wobei die Transformationen g und h_j als gegeben betrachtet werden. Viele Gesetzmässigkeiten, die zunächst nicht linear aussehen, lassen sich so auf eine multiple lineare Regression zurückführen.

d **Nicht-lineare Regression.** In der Enzym-Kinetik wird untersucht, welche Menge Enzym (Y) an „Bindungsstellen“ im Gewebe gebunden werden – in Abhängigkeit von der Konzentration x in der zugefügten Lösung. Eine alte Formel, die diese Abhängigkeit im einfachsten Fall gut beschreibt, lautet

$$Y_i = \frac{\theta_1}{(\theta_2/x_i)^{\theta_3} + 1} + E_i \; .$$

(Der Parameter θ_1 bedeutet die „Kapazität“, die Menge adsorbierten Enzyms bei grosser Konzentration; θ_2 und θ_3 bestimmen die Form der „Sättigungskurve“.) Diese Formel lässt sich mit keinen Transformationen auf die vorhergehende Form bringen.

Allgemein formuliert sei die Regressionsfunktion h bis auf einige Konstanten $\theta_1, \ldots, \theta_p$ bekannt,

$$Y_i = h\langle x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m)}; \theta_1, \theta_2, ..., \theta_p\rangle + E_i \; .$$

Die Parameter $\theta_1, \theta_2, ...\theta_p$ aus Daten zu bestimmen, ist die Aufgabe der **nicht-linearen Regression**. Meistens wird für den zufälligen Fehler wieder $E_i \sim \mathcal{N}\langle 0, \sigma^2\rangle$, unabhängig, angenommen, und für die Schätzung der Parameter $\theta_1, \ldots, \theta_p$ die Methode der Kleinsten Quadrate angewandt. Die Theorie, die korrekte Interpretation von Ergebnissen und selbst die Berechnung von Parametern werden wesentlich anspruchsvoller als für die multiple lineare Regression. Auch hier sind robuste Varianten möglich.

Literatur: Bates and Watts (1988), Chambers and Hastie (1992), Kap. 10, Venables and Ripley (1997), Kap. 9.

e* **Systemanalyse**. Die Funktion h kann im Prinzip von den erklärenden Variablen x und den Parametern $\theta_1, \dots \theta_p$ in beliebig komplizierter Weise abhängen. Man kann beispielsweise in der Atmosphärenphysik die Wolken- und Gewitterbildung oder Transportphänomene, in der Chemie Reaktionen und in der Ökologie die Entwicklung von Ökosystemen mit Hilfe von Differentialgleichungen beschreiben. Diese Gleichungen können Konstanten θ_k enthalten, die nicht oder ungenügend bekannt sind. Wenn man Anfangsbedingungen $x^{(j)}$ und eine Endgrösse Y messen kann, so kann man mit Hilfe der nicht-linearen Regression die Konstanten θ_k als Parameter des Modells schätzen. Zur Bestimmung eines Funktionswertes $h\langle x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m)}; \theta_1, \theta_2, ...\theta_p\rangle$ für bestimmte, mögliche Parameterwerte θ_j muss jeweils die Lösung des Differentialgleichungs-Systems für die Anfangsbedingungen $x_1^{(j)}, \dots, x_i^{(m)}$ bestimmt werden. In der Regel ist dies nur numerisch möglich. Mit genügendem Rechenaufwand können dennoch mittels nichtlinearer Regression diejenigen Konstanten gefunden werden, die mit gegebenen Endwerten Y_i bestmöglich übereinstimmen. Mit solchen Aufgaben befasst sich die Systemanalyse.

f **Glättung.** In all diesen Modellen ist die Regressionsfunktion h bis auf einige Konstanten bekannt. In vielen Fällen weiss man eigentlich nichts über die Funktion, ausser dass sie nicht „allzu wild“ sein kann, dass also h in irgendeinem festzulegenden Sinn **glatt** ist. Eine (allzu) einfache Methode, zu einer mehr oder weniger glatten geschätzten Funktion $\widehat{h}$ aufgrund der Daten zu gelangen, wurde in 4.2.i beschrieben. Es gibt viele **Glättungsverfahren** oder **smoother**. Die Methodik wird oft als **nicht-parametrische Regression** bezeichnet – ein eher missglückter Begriff, da zwar die Funktion h nicht durch wenige Parameter festgelegt ist, wohl aber für die Verteilung der Fehler oft die Normalverteilung vorausgesetzt wird. (Man kann sogar die ganze Funktion $h\langle x\rangle$ als „unendlich-dimensionalen Parameter“ auffassen. Dann müsste man von „superparametrischer Regression“ sprechen.)

Literatur: Hastie and Tibshirani (1990), Kap. 1; Chambers and Hastie (1992), Kap. 8.

g **Allgemeine additive Modelle.** Im Prinzip kann man auch für mehrere erklärende Variable nicht-parametrische Regressionsmethoden entwickeln. Allerdings machen heuristische Überlegungen und Erfahrung rasch klar, dass solche Methoden nur dann zu sinnvollen Resultaten führen können, wenn sehr viele Beobachtungen vorliegen oder die zufälligen Fehler klein sind. Je mehr Daten, desto weniger Annahmen sind nötig – und umgekehrt.

Eine sinnvolle Einschränkung liegt oft in der Annahme, dass die Effekte auf die Zielgrösse sich additiv verhalten. Sie führt auf ein allgemeines additives Modell *(general additive model, GAM)* mit $h\langle x^{(1)}, \dots, x^{(m)}\rangle = h_1\langle x^{(1)}\rangle + h_2\langle x^{(2)}\rangle + ... + h_m\langle x^{(m)}\rangle$. Wenn zusätzlich noch eine geeignete Transformation der Zielgrösse aus den Daten geschätzt wird, heissen die Methoden ACE und AVAS.

Literatur: Hastie and Tibshirani (1990); Chambers and Hastie (1992), Kap. 7; Venables and Ripley (1997), Kap. 11.1.+3.

h* **Projection Pursuit Regression.** Statt der einzelnen erklärenden Variablen kann je eine Linearkombination als Argument der glatten Funktionen eingesetzt werden,

$$h\langle x^{(1)}, \dots, x^{(m)}\rangle = h_1\Big\langle \sum\nolimits_{j=0}^{m} \alpha_j^{(1)} x^{(j)} \Big\rangle + h_2\Big\langle \sum\nolimits_{j=0}^{m} \alpha_j^{(2)} x^{(j)} \Big\rangle + ... \; .$$

Die Methodik der Projection Pursuit Regression schätzt sowohl die α_j als auch die Funktionen h_k aus den Daten, die dementsprechend zahlreich sein müssen.

Literatur: Venables and Ripley (1997), Kap. 11.2.

i* **Neuronale Netze.** Als Variante davon kann man für die h_k eine feste Funktion $\widetilde{h}$, beispielsweise die logistische, wählen, und erhält die Terme

$$z^{(k)} = \widetilde{h}\Big\langle \sum\nolimits_{j=0}^{m} \alpha_j^{(k)} x^{(j)} \Big\rangle$$

(wobei $x^{(0)} = 1$ ist). Aus diesen bildet man wieder eine Linearkombination und wendet, um konsistent zu bleiben, auch darauf die Funktion $\widetilde{h}$ an. So ergibt sich

$$h\langle x^{(1)}, \dots, x^{(m)}\rangle = \gamma_0 + \gamma_1 \widetilde{h}\Big\langle \sum\nolimits_{k} \beta_k z^{(k)} \Big\rangle \; .$$

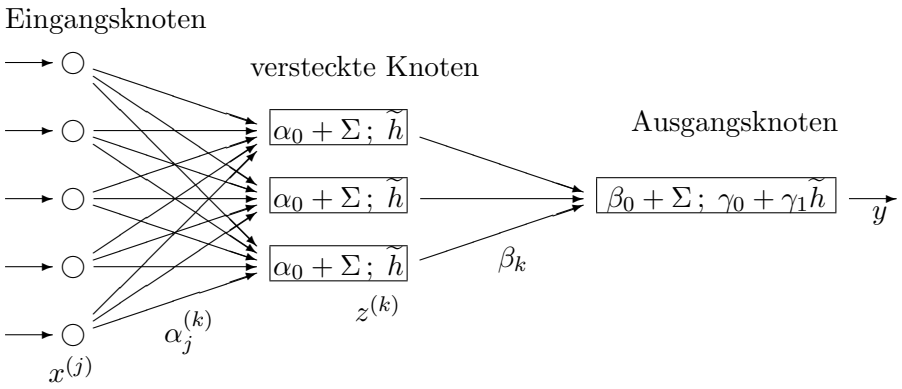

Abbildung 6.3.i: Schema eines Neuronalen Netzes mit einer „versteckten Schicht“ von Knoten

Dieses statistische Modell beschreibt das gebräuchlichste in der Klasse der „neuronalen Netze“, die sich an einem einfachen biologischen Modell der Funktion des Gehirns orientieren: Die Eingangszellen im Bild 6.3.i erhalten die Eingangssignale $X^{(j)}$, und das Augangssignal sollte Y sein. Das wird über dazwischengeschaltete „Nervenzellen“ oder Knoten in der „versteckten Schicht“ *(hidden layer)* bewerkstelligt; jede Zelle k empfängt die Signale $X^{(j)}$ der Eingangszellen mit verschiedenen Dämpfungen $\alpha_j^{(k)}$ und schickt die Summe, transformiert mit der nicht-linearen Funktion $\widetilde{h}$, weiter an die Ausgangszelle. Diese verarbeitet die Signale auf gleiche Weise zum geschätzten Ausgangssignal $\widehat{y}$. – Es können mehrere Ausgangszellen für mehrere Ausgangssignale angesetzt werden, und kompliziertere Netze können mehrere Schichten von versteckten Knoten enthalten.

Literatur: Ripley (1996).

j **Überlebens- oder Ausfallzeiten.** Wenn die Heilung von Patienten oder der Ausfall von Geräten untersucht wird, so kann man auch die Zeit bis zu diesem Ereignis messen. Beobachtungen dieser Art heissen **Überlebenszeiten** (englisch *survival* oder *failure time data*). Das bekannteste Modell zur Untersuchung der Abhängigkeit einer solchen Grösse von erklärenden Variablen heisst **Cox-Regression** und ist das einfachste Beispiel eines **proportional hazards** Modells.

Bei solchen Studien kann man meistens einige Überlebenszeiten nicht bis zu ihrem Ende abwarten; man muss **zensierte Daten** in Kauf nehmen. Die Regressionsmethoden für Überlebenszeiten können solche unvollständige Daten auswerten.

Literatur: Crowder, Kimber, Smith and Sweeting (1991), Collet (1994), Kalbfleisch and Prentice (2002). Überlebenszeiten werden unter dem allgemeinen Titel „Statistische Methoden im 2. Semester des Nachdiplomkurses besprochen.

k* Wenn für Patienten festgestellt wird, wie stark sie auf verschiedene Konzentrationen eines Medikamentes reagieren, so kann ein Modell sinnvoll sein, das die Reaktion eines Patienten (Y) als einfache lineare Regression auf die Dosis (x) beschreibt. Steigung und Achsenabschnitt für die Regressionsgerade des $i$ten Patienten kann man als Zufallsvariable modellieren. Das ergibt ein Modell

$$Y_{ih} = \mu + A_i + B_i\, x_{ih} + E_{ih}$$

mit **zufälligen Koeffizienten** A_i **und** B_i**.** Modelle mit zufälligen Koeffizienten (oder „Effekten“) werden auch in der Varianzanalyse (2. Teil) eingeführt. Man findet sie unter den Namen Varianz-Komponenten-Modelle und repeated measures oder split-plot designs.

Vergleicht man in dieser Situation die Wirkung zweier Medikamente, dann kommt noch ein fester Effekt für einen allfälligen systematischen Unterschied hinzu, zu schreiben als $+\gamma\widetilde{x}_i$, wobei $\widetilde{x}_i$ die Indikatorvariable für das eine der beiden Medikamente ist (vergleiche 3.2.e).

Modelle können also sowohl feste als auch zufällige Koeffizienten enthalten. Die erklärenden Variablen können Faktoren im Sinne der Varianzanalyse oder geordnete, oft kontinuierliche Variable im Sinne der Regression sein. So entstehen die so genannten **gemischten** oder **allgemeinen linearen Modelle** oder *mixed* respektive *general linear models.*

l **Multivariate Regression.** In den vorhergehenden Abschnitten wurde das Modell der *multiplen* linearen Regression behandelt. Irrtümlicherweise wird dafür ab und zu der Ausdruck **multivariate Regression** verwendet, der sich, richtig verwendet, auf Modelle bezieht, in denen gleichzeitig mehrere Zielgrössen $Y_i^{(1)}$, $Y_i^{(2)}, \dots$ in ihrer Abhängigkeit von (den gleichen) erklärenden Variablen $x_i^{(1)}$, $x_i^{(2)}, \dots$ beschrieben werden. Dies ist eine Problemstellung der multivariaten Statistik.

6.4 Ausblick auf die Verallgemeinerten linearen Modelle

a **Logistische Regression.** In toxikologischen Untersuchungen wird festgestellt, ob eine Maus bei einer bestimmten Giftkonzentration überlebt oder stirbt. In der Medizin denken wir lieber an den entgegengesetzen Fall: Wird ein Patient bei einer bestimmten Konzentration eines Medikaments in einer vorgegebenen Zeit gesund oder nicht?

Die **Zielgrösse** ist hier nicht mehr eine kontinuierliche, sondern eine **binäre** Variable (oder 0-1-Variable), die das Auftreten eines bestimmten Ergebnisses angibt. Es wird die Abhängigkeit dieses Ereignisses von einer oder mehreren erklärenden Variablen gesucht. Solche Situationen treten in vielen Gebieten auf: Ausfall von Geräten, Vorhandensein eines bestimmten Merkmals bei Lebewesen oder eines Fehlers an einem Produkt, Zugehörigkeit zu einer von zwei Gruppen (vergleiche Diskriminanz-Analyse in der Multivariaten Statistik) u.s.w.

b Ein Wahrscheinlichkeitsmodell für diese Situation trägt dem Umstand Rechnung, dass bei gegebener (mittlerer) Konzentration eines Giftes nicht jede Maus stirbt. Gesucht ist ein Modell für die Wahrscheinlichkeit, dass das Ereignis eintritt, also für $P\langle Y_i = 1\rangle$, in Abhängigkeit von der Konzentration oder, allgemein, von den Werten $x_i^{(1)}, \dots, x_i^{(m)}$ der erklärenden Variablen. Der einfachste Vorschlag, $P\langle Y_i = 1\rangle = \alpha + \beta_i x_i^{(1)} + \beta_2 x_i^{(2)} + \dots$ würde zu Wahrscheinlichkeitswerten ausserhalb des Intervalls von 0 bis 1 führen. Um dies zu vermeiden, transformiert man die Wahrscheinlichkeit, meistens mit der **„logit"-Transformation** $p \mapsto \ln\langle p/(1-p)\rangle$. So erhält man das Modell der logistischen Regression,

$$\ln\left\langle \frac{P\langle Y_i = 1\rangle}{1 - P\langle Y_i = 1\rangle} \right\rangle = h\langle x_i^{(1)}, \dots, x_i^{(m)}\rangle = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x^{(m)} \ .$$

Die Y_i sollen unabhängig sein.

Literatur: Cox (1989) behandelt die logistische Regression auf ansprechende Weise. Meist genügt aber ein Kapitel aus einem allgemeineren Lehrbuch.

c **Poisson-Regression**. Wovon hängen die Aufenthaltsorte von Seesternen ab? Um diese Frage zu untersuchen, wird man auf dem Meeresboden Flächen abgrenzen, für jede die Umweltvariablen $x^{(1)}, \dots x^{(m)}$ aufnehmen, und die Seesterne zählen. Die **Zielgrösse** Y_i ist eine **Anzahl**, für die man als einfachstes Modell annehmen kann, dass sie poissonverteilt ist (falls man direkte gegenseitige Beeinflussung vernachlässigen kann). Der Parameter λ_i soll wieder in transformierter Form linear von den erklärenden Variablen abhängen,

$$Y_i \sim \mathcal{P}\langle\lambda_i\rangle \ , \qquad \ln\langle\lambda_i\rangle = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)} \ .$$

So lautet das Modell der Poisson-Regression.

d **Log-lineare Modelle.** Auf ein ähnliches Modell führt die Analyse von **nominalen Zielgrössen**, beispielsweise die Untersuchung der Abhängigkeit der gewählten Partei von der gesellschaftlichen Klasse der Wählenden und eventuell von weiteren ihrer Merkmale. Solche Daten werden in (zweidimensionalen) **Kontingenztafeln** zusammengestellt. In Einführungsbüchern wird dafür der Chiquadrat-Test auf Unabhängigkeit behandelt. Die Frage, ob die gewählte Partei mit der gesellschaftlichen Klasse überhaupt zusammenhänge, kann damit beantwortet werden.

Zu einer genaueren Analyse führen die so genannten log-linearen Modelle. Sie erlauben es, die Abhängigkeit einer nominalen Zielgrösse von ebenfalls nominalen oder auch von stetigen erklärenden Variablen ebenso detailliert zu untersuchen, wie es bei stetigen Zielgrössen durch die multiple lineare Regression möglich ist. Beispielsweise kann man fragen, ob ein direkter Einfluss einer erklärenden Variablen, unter Ausschluss der indirekten Einflüsse anderer erklärender Variabler, auf die Zielgrösse vorhanden sei – anders gesagt: ob die bedingte gemeinsame Verteilung der Zielgrösse und der fraglichen erklärenden Variablen, gegeben alle anderen erklärenden Variablen, Unabhängigkeit zeige.

Solche genaueren Fragestellungen bilden eine wertvolle, oft unerlässliche Ergänzung der blossen Tests auf Unabhängigkeit in zweidimensionalen Kontingenztafeln, wie sie in der Auswertung von Umfragen üblich sind – genauso, wie die einfache Varianzanalyse und die einfache Regression nicht genügen, wenn mehrere erklärende Variable zur Verfügung stehen.

Literatur: Ein empfehlenswertes Buch zum Thema schrieb Agresti (2002).

e **Verallgemeinerte Lineare Modelle.** Die log-linearen Modelle, die logistische und die Poisson-Regression sind Beispiele einer grossen Klasse von Modellen, den **verallgemeinerten linearen Modellen** *(generalized linear models*, GLM, zu unterscheiden vom allgemeinen linearen Modell oder *general linear model*, siehe 6.3.k, das manchmal ebenfalls als GLM bezeichnet wird). Sie sagen, dass der Erwartungswert der Zielgrösse Y monoton von einer linearen Funktion der erklärenden Variablen $x^{(1)}, \ldots, x^{(m)}$ abhängt,

$$\mathcal{E}\langle Y_i\rangle = \widetilde{g}\Big\langle \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} \Big\rangle \ .$$

Die Varianz von Y muss ausserdem in der Form $\mathrm{var}\langle Y\rangle = \phi\ v\langle \mathcal{E}\langle Y\rangle\rangle$ vom Erwartungswert abhängen, wobei ϕ ein zusätzlicher Parameter und v eine gegebene Funktion ist. Drittens muss die Dichte von Y, gegeben die x-Werte, von einer bestimmten Form sein.

Obwohl diese Voraussetzungen recht kompliziert erscheinen, sind sie in wichtigen Fällen erfüllt. Neben den erwähnten Beispielen passt auch das Modell der multiplen linearen Regression mit normalverteilten Fehlern in diese allgemeine Form: Man setzt $\widetilde{g}\langle x\rangle = x$, $v\langle\mu\rangle = 1$ und $\phi = \sigma^2$. Die in technischen Anwendungen nützlichen Gamma- und Exponential-Verteilungen sind ebenfalls abgedeckt (nicht aber die Weibull-Verteilung).

Es zeigt sich, dass man mit der allgemeinen Form recht weitgehende theoretische Resultate, allgemeine Überlegungen zur Modellwahl und -Überprüfung und einheitliche Berechnungsmethoden erhalten kann. Deshalb werden sie in den Statistikprogrammen oft durch eine einzige Prozedur abgedeckt.

Literatur: Das klassische Werk über Theorie und Anwendung dieser Modelle stammt von McCullagh and Nelder (1989). Eine kurze anwendungsorientierte Beschreibung findet man in Kap. 6 von Chambers and Hastie (1992).

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn, Wiley, N.Y.

Agresti, A. (2007). *An Introduction to categorical data analysis*, Wiley Series in Probability & Math. Statistics, 2nd edn, Wiley, New York.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, N.Y.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove, Cal.

Chatterjee, S. and Price, B. (2000). *Regression Analysis By Example*, 3rd edn, Wiley, N.Y.

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.

Collet, D. (1994). *Modelling Survival Data in Medical Research*, Texts in Statistical Science, Chapman and Hall, London.

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.

Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991). *Statistical Analysis of Reliability Data*, Chapman and Hall.

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. (2004). *Probability and Statistics for Engineering and the Sciences*, 6th 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. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer-Verlag, New York.

Fox, J. (2002). *An R and S-Plus companion to applied regression*, Sage, Thousand Oaks, CA.

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.

Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**: 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, N.Y.

Harrell, F. E. J. (2002). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer Series in Statistics, Springer, NY. *Corrected second printing*

Hartung, J., Elpelt, B. und Klösener, K. (2002). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 13. Aufl., Oldenbourg, München.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, number 43 in *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

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.

Hocking, R. R. (1996). *Methods and Applications of Linear Models; Regression and the Analysis of Variance*, Wiley Series in Probability and Statistics, Wiley, N.Y.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn, Wiley, N.Y.

Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Mathemathical Statistics* **35**: 73–101.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, 2nd edn, Wiley, N.Y.

Kalbfleisch, J. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edn, Wiley, N.Y.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data*, number 15 in *Oxford Statistical Science Series*, Clarendon Press, Oxford.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics, Theory and Methods*, Wiley Series in Probility and Statistics, Wiley, Chichester, England.

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.

Pokropp, F. (1994). *Lineare Regression und Varianzanalyse*, Oldenbourg.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*, 3rd edn, Duxbury Press, Belmont, California.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, Cambridge, UK.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N.Y.

Ryan, T. P. (1997). *Modern Regression Methods*, Series in Probability and Statistics, Wiley, N.Y. *includes disk*

Sachs, L. (2004). *Angewandte Statistik*, 11. Aufl., Springer, Berlin.

Schlittgen, R. (2003). *Einführung in die Statistik. Analyse und Modellierung von Daten*, 10. Aufl., Oldenbourg, München. *schoen, inkl. Sensitivity und breakdown, einfache regr mit resanal*

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. (2007). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 5. Aufl., Vieweg, Wiesbaden.

Tableman, M. and Kim, J. S. (2003). *Survival Analysis Using S*, Texts in Statistical Science, Chapman & Hall/CRC. *with a contribution from Stephen Portnoy*

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data*, Springer, N.Y.

van der Waerden, B. L. (1971). *Mathematische Statistik*, 3. Aufl., Springer, Berlin.

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus*, Statistics and Computing, 2nd edn, Springer, Berlin.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edn, Wiley, N.Y.

Wetherill, G. (1986). *Regression Analysis with Applications*, number 27 in *Monographs on Statistics and Applied Probability*, Chapmann and Hall, London.