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.*

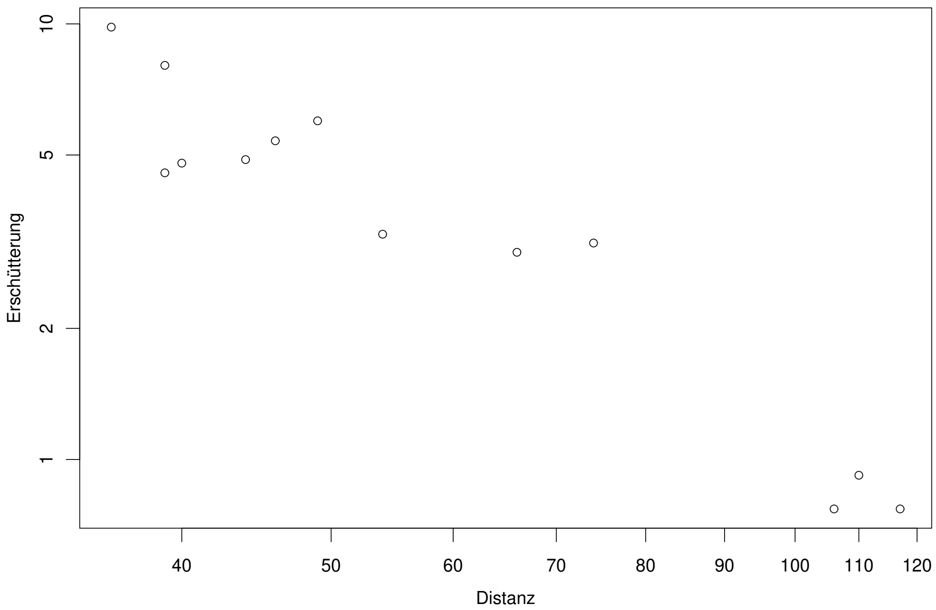

Abbildung 2.1.a: Distanz und Erschütterung bei Sprengungen mit Ladung 3.12. Die Achsen sind logarithmisch dargestellt

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 Ausgangs-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 Ausgangsgrösse 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.

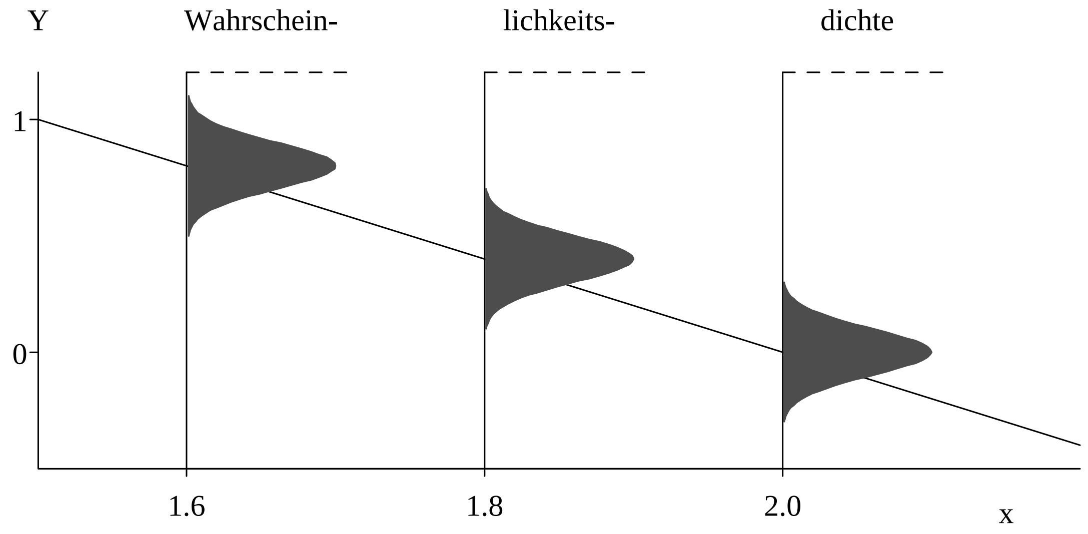

Abbildung 2.1.h: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ für drei Beobachtungen Y_1, Y_2 und Y_3 zu den x-Werten $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$

i Als zweite Veranschaulichung wollen wir **Zufallszahlen** gemäss unserm Modell ziehen und darstellen, also Beobachtungen, die dem Modell entsprechen, **simulieren**. Drei standardnormalverteilte 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

$$-0.419,\ -1.536,\ -0.671\ ; \qquad 0.253,\ -0.587,\ -0.065\ ;$$
$$1.287,\ 1.623,\ -1.442\ ; \qquad -0.417,\ 1.427,\ 0.897\ .$$

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.

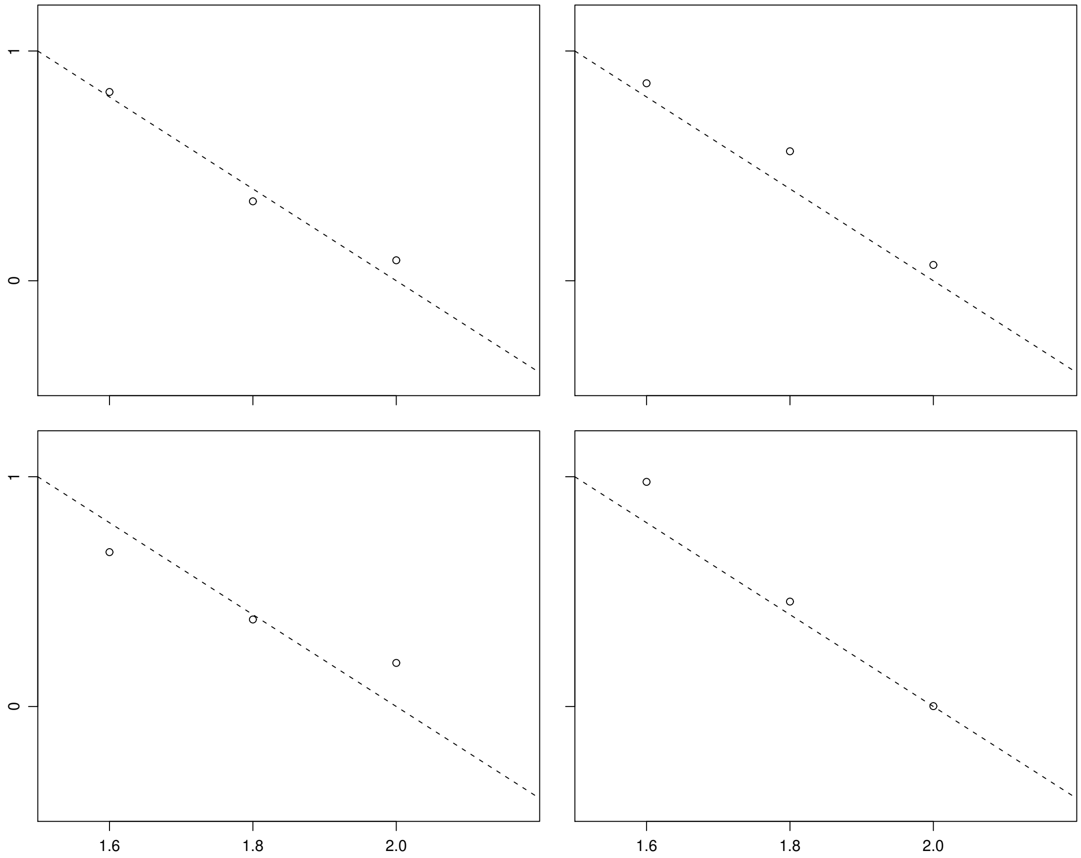

Abbildung 2.1.i: Vier simulierte Ergebnisse für drei Messungen gemäss dem Modell $Y_i = 4-2x_i+E_i$. Die gestrichelten Geraden stellen den hier bekannten „wahren" Zusammenhang $y = 4 - 2x$ dar.

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.*

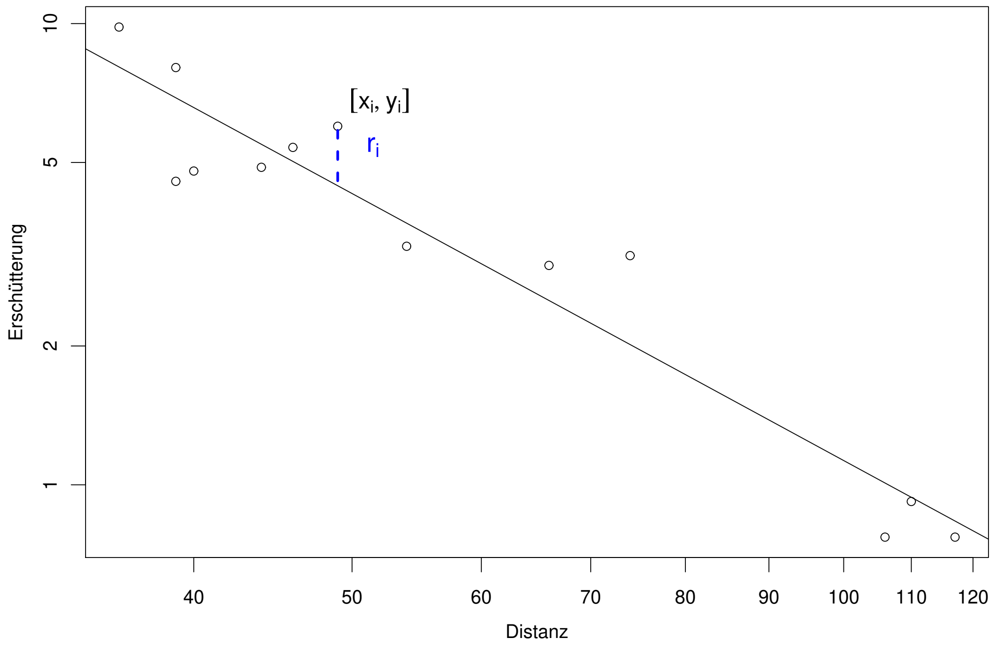

Abbildung 2.2.a: Geschätzte Gerade für das Beispiel der Sprengungen

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.

Die Schätzfunktionen lauten dann

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x} \,.$$

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. $\widehat{\alpha}$, $\widehat{\beta}$.

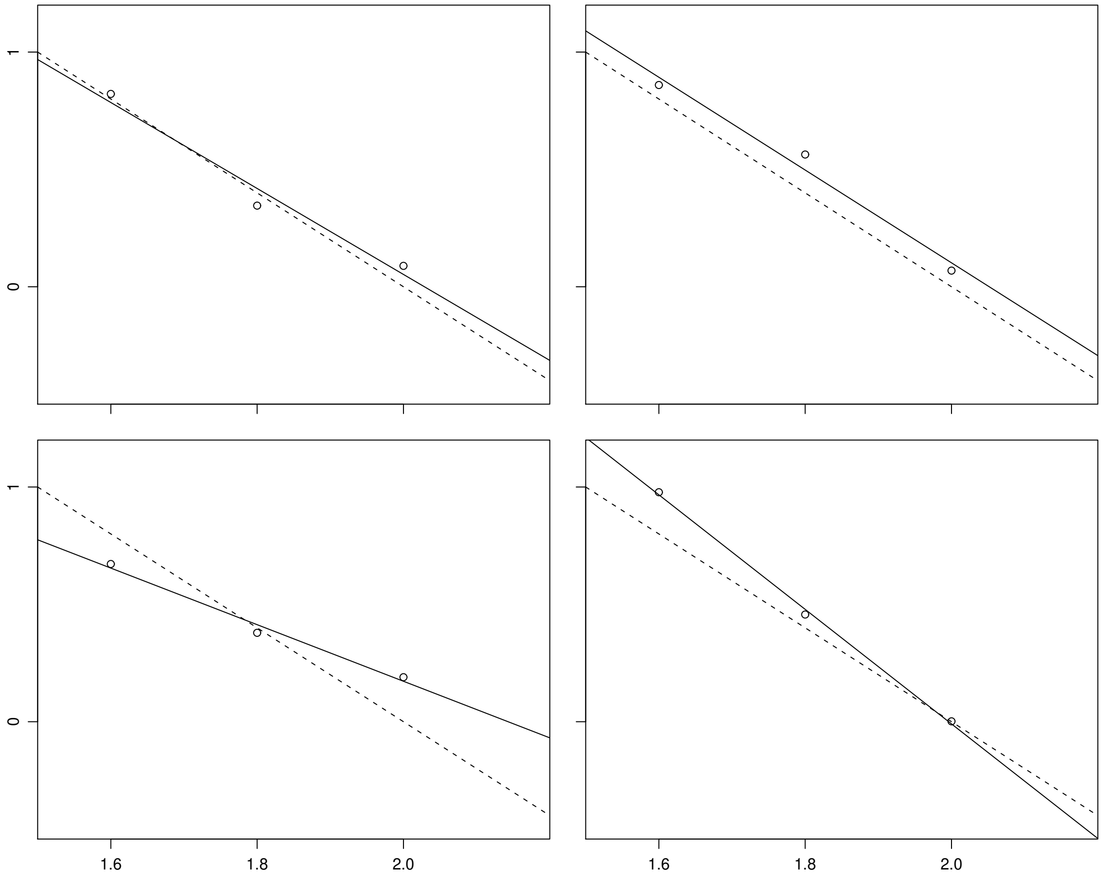

Abbildung 2.2.e: Vier simulierte Ergebnisse für drei Messungen mit den geschätzten (ausgezogenen) Geraden

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.

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.

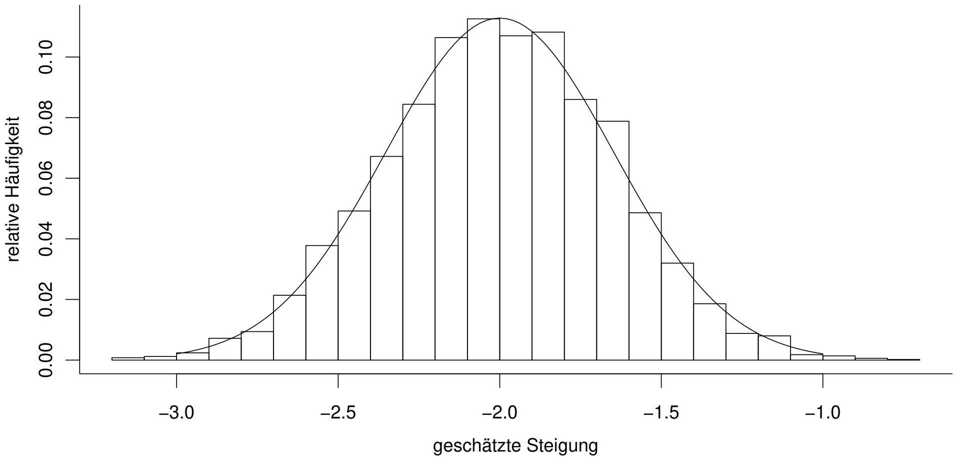

Abbildung 2.2.g: Simulierte und theoretische Verteilung der Schätzung $\widehat{\beta}$ der Steigung

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

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

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

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

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

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 $\mathrm{var}\left\langle \widehat{\beta} \right\rangle = \sigma^2/\mathrm{SSQ}^{(X)}$ angegeben. (Wenn eine Schätzung $\widehat{\theta}$ nicht erwartungstreu ist, ist der **Mittlere Quadratische Fehler**, englisch *mean squared error*, $\mathrm{MSE} = \mathcal{E}\left\langle (\widehat{\theta} - \theta)^2 \right\rangle$ 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 $\mathcal{N}\left\langle 0, \sigma^2 \right\rangle$ 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.)

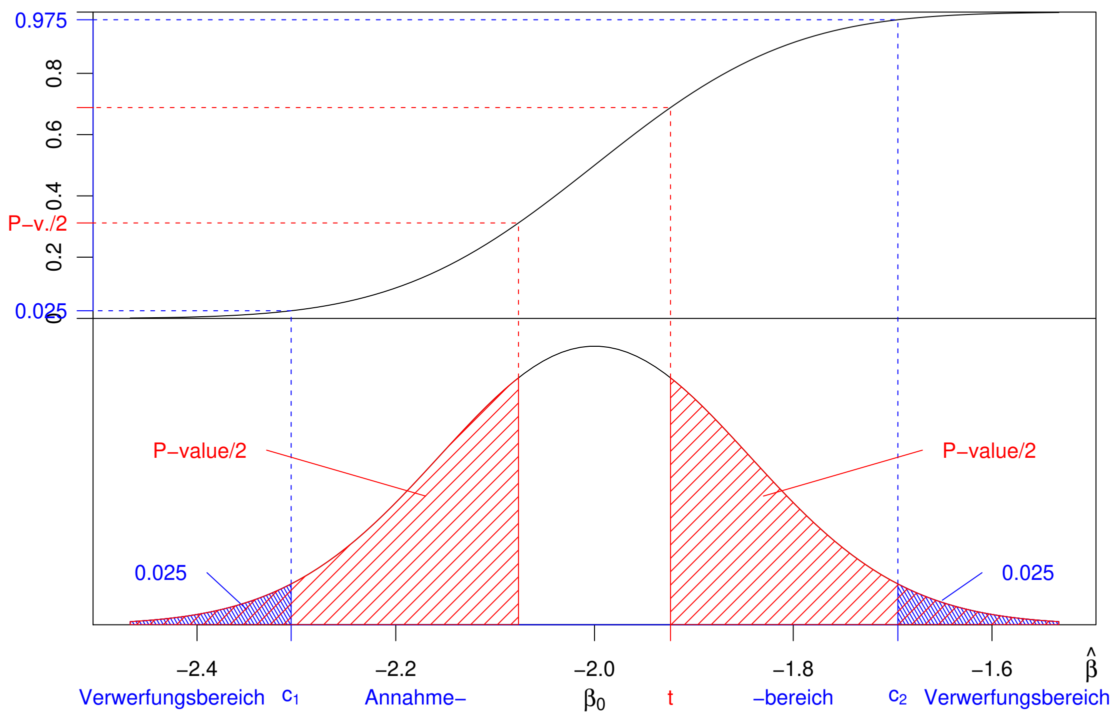

Abbildung 2.3.c: Veranschaulichung des P-Wertes und des Verwerfungsbereiches für einen zweiseitigen Test. Die obere Kurve stellt die kumulative Verteilungsfunktion, die untere die Dichte der Verteilung der Teststatistik dar.

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 Ausgangsgrösse 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% liegt. Die Grösse „P-Wert" ist gerade so konstruiert, dass für die Entscheidungsregel „signifikant falls P-Wert $\leq$ 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	$\text{se}^{(\alpha)} = 0.3156$	$T^{(\alpha)} = 12.36$	0
Slope	$\widehat{\beta} = -1.9235$	$\text{se}^{(\beta)} = 0.1783$	$T^{(\beta)} = -10.79$	0

R-squared = 0.9136 = r_{XY}^2
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

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

wobei $q = q_{0.975}^{t_{n-2}}$ 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 50m beträgt. Zunächst fragen wir nach dem Erwartungswert der Erschütterung bei 50m 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 $\mathrm{cov}\left\langle\overline{Y}, \widehat{\beta}\right\rangle = 0$ ist,

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

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.

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.

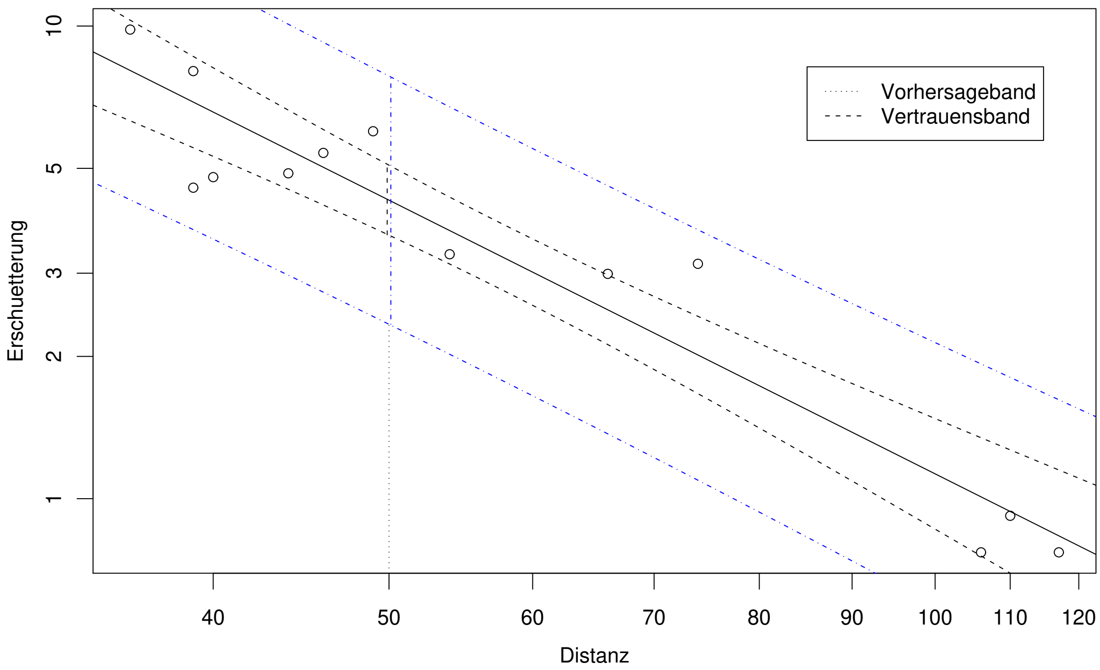

Abbildung 2.4.c: Vertrauensband für den Funktionswert $h\langle x\rangle$ und Vorhersage-Band für eine weitere Beobachtung im Beispiel der Sprengungen

e* Herleitung: Die Zufallsvariable Y_0 sei also der Wert der Zielgrösse bei einer Beobachtung mit Ausgangsgrösse 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 geschätzten Geraden führten. Beide haben Erwartungswert 0; die Varianzen addieren sich zu

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

Daraus ergibt sich das Vorhersage-Intervall

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

wobei wieder $q = q_{0.975}^{t_{n-2}}$ 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 (1998, §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

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

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

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

und setzen die Ableitung null; wir erhalten

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

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, $\mathrm{var}\langle X + Y\rangle = \mathrm{var}\langle X\rangle + \mathrm{var}\langle Y\rangle$,

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

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,

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

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 $\gamma = \alpha + \beta\overline{x}$; β ist als Parameter beibehalten worden. Ebenso hängen natürlich die Schätzungen zusammen,

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

die zweite Gleichheit erhält man aus 2.2.c.

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

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

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

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

Daraus ergibt sich

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

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 $\widehat{\alpha} = \widehat{\gamma} - \widehat{\beta}\,\overline{x}$ einfach zu bestimmen: Es ist die Normalverteilung mit $\mathcal{E}\langle\widehat{\alpha}\rangle = \mathcal{E}\langle\widehat{\gamma}\rangle - \overline{x}\,\mathcal{E}\langle\widehat{\beta}\rangle = \gamma - \overline{x}\beta = \alpha$ und

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

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

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

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 Ausgangsgrössen 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 Ausgangsgrössen 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")
```