Regression

Skript zur Vorlesung im FS 2008

von Prof. Marloes Maathuis

Verfasst von Professor Hansruedi Künsch, unter Verwendung
der Mitschrift der Vorlesung von Professor Frank Hampel
durch Rolf Steiner (SS 94)

Seminar für Statistik
ETH Zürich

Februar 2008

Inhaltsverzeichnis

1 Lineare Regression **5**

1.1 Einführung: Fragestellung . . . 5

1.1.1 Beispiele, Historisches . . . 5

1.1.2 Lineares Modell, Beispiele . . . 7

1.2 Die Voraussetzungen des linearen Modells . . . 11

1.3 Kleinste Quadrate Schätzung . . . 15

1.3.1 Normalengleichungen . . . 15

1.3.2 Geometrische Interpretation . . . 16

1.3.3 Zusammenhang mit MLE bei Normalverteilung . . . 18

1.3.4 Warum nicht Regression auf jede Variable einzeln ? . . . 18

1.4 Eigenschaften der KQ-Schätzung . . . 19

1.4.1 Momente ohne Normalverteilungsannahmen . . . 20

1.4.2 Verteilungen unter Annahme der Normalverteilung . . . 21

1.4.3 Asymptotische Normalität . . . 22

1.5 Tests und Vertrauensintervalle . . . 23

1.5.1 Die grundlegenden Teststatistiken . . . 23

1.5.2 Vertrauensband für die ganze Hyperebene . . . 24

1.5.3 Vergleich zweier geschachtelter Modelle, Varianzanalyse . . . 25

1.5.4 Bestimmtheitsmass . . . 28

1.6 Einfache lineare Regression . . . 29

1.6.1 Resultate im Spezialfall der einfachen linearen Regression . . . 29

1.6.2 Regression und Korrelation . . . 30

1.6.3 Vertauschung von X und Y, Regression zum Mittel . . . 32

1.7 Residuenanalyse, Modellüberprüfung . . . 34

1.7.1 Normal plot . . . 34

1.7.2 Tukey-Anscombe Plot . . . 36

1.7.3 Zeitreihenplot, Durbin-Watson Test . . . 37

1.7.4 Interior Analysis . . . 39

1.7.5 Verallgemeinerte Kleinste Quadrate, Gewichtete Regression . . . 40

1.8 Modellwahl . . . 41

1.8.1 Modellwahl mit “stepwise regression” . . . 42

1.8.2 Modellwahlkriterien . . . 43

1.9 Das Gauss-Markov-Theorem . . . 47

2 Nichtlineare und nichtparametrische Methoden **51**

2.1 Robuste Methoden . . . 51

2.1.1 Einfluss einzelner Beobachtungen bei Kleinsten Quadraten . . . 51

2.1.2 Huber- und L_1-Regression . . . 53

2.1.3 Regressionsschätzer mit beschränktem Einfluss 55
2.1.4 Regressionsschätzer mit hohem Bruchpunkt 56
2.2 Nichtlineare Kleinste Quadrate 57
2.2.1 Asymptotische Vertrauensintervalle und Tests 59
2.2.2 Genauere Tests und Vertrauensintervalle 59
2.3 Verallgemeinerte Lineare Modelle 61
2.3.1 Logistische Regression 61
2.3.2 Allgemeiner Fall 62
2.4 Cox-Regression 63
2.5 Nichtparametrische Regression 64
2.5.1 Einige Verfahren im eindimensionalen Fall 65
2.5.2 Bias/Varianz-Dilemma 67
2.5.3 Fluch der Dimension 68

A Resultate aus der Wahrscheinlichkeitstheorie 1
A.1 Rechenregeln für Momente 1
A.2 Die Normalverteilung 2
A.2.1 Eindimensionale Normalverteilung 2
A.2.2 Mehrdimensionale Normalverteilung 4
A.2.3 Chiquadrat-, t- und F-Verteilung 6

B Literatur 9

Kapitel 1

Lineare Regression

1.1 Einführung: Fragestellung

1.1.1 Beispiele, Historisches

Bis zum Beginn der Neuzeit war die unsinnige Annahme weit verbreitet, dass beim Erhalt von verschiedenen Messergebnissen eines das völlig **richtige** sein müsse, und die anderen hingegen **falsch**. Wenn man beispielsweise die Koordinaten eines Sterns fünfmal gemessen hatte und fünf verschiedene Resultate erhielt, so war man der Ansicht, dass man viermal falsch und einmal richtig gemessen habe.

Erst vor etwa 400 Jahren änderte sich die Denkweise: Man führte das Konzept der **"zufälligen Fehler"** ein. Dieses Konzept entstand aus dem Gedanken, dass **alle Messresultate** einem kleinen **Fehler** unterworfen sind, so dass sie von der Wahrheit abweichen, jedoch alle **mindestens näherungsweise der Wirklichkeit entsprechen**. Es war die Geburtsstunde der **Stochastik**.

Damit wurde es insbesondere auch möglich, approximative/stochastische Beziehungen zwischen Variablen zu untersuchen, das Thema dieser Vorlesung.

Die Methode der Kleinsten Quadrate (abgekürzt KQ), auf englisch least squares, wurde 1805 von Legendre publiziert. Gauss diskutierte diese Methode ebenfalls in einem Buch von 1809 und erwähnte dabei, dass er diese seit 1795 gebraucht habe. Einen Beweis dieser Behauptung gibt es aber nicht, so dass nicht klar ist, wem die Ehre der Entdeckung gebührt.

Ursprünglich wurde diese Methode auf Probleme der Himmelsmechanik angewandt, wo man die Daten an theoretisch berechnete Bahnen anpasste.

Beispiel aus der Astronomie (Ceres): Aufgrund der Bode-Titius'schen Reihe, welche die Gesetzmässigkeiten der Abstände der Planeten von der Sonne empirisch beschreibt (entdeckt durch Titius 1766), erahnte man, dass zwischen den Planeten Mars und Jupiter noch ein weiterer Planet sein musste. Dieser wurde am 1.1.1801 schliesslich gefunden durch Giuseppe Piazzi und mit Ceres benannt. Ceres ist der grösste der Planetoiden.

Ceres bewegt sich relativ schnell, und er wurde rasch aus den Augen verloren. Es war Gauss, welcher dank der Methode der Kleinsten Quadrate aus den wenigen vorliegenden Beobachtungen eine so genaue Bahn berechnete, dass man dadurch Ceres wiederfinden konnte.

Später wurde die Methode aber in sehr viel allgemeinerem Rahmen angewendet, auch in den Sozialwissenschaften. So hat Yule im Jahr 1899 z.B. untersucht, ob es besser sei, armengenössige Leute in Armenhäusern einzusperren oder sie in ihrer gewohnten Umgebung zu unterstützen. Dazu verwendete er die Regressionsgleichung

$$\Delta Paup = a + b \cdot \Delta Out + c \cdot \Delta Old + d \cdot \Delta Pop + error.$$

Dabei bezeichnet Δ die Veränderung zwischen zwei Volkszählungen, "Paup" die Anzahl Leute, die Fürsorgeunterstützung erhalten, "Out" das Verhältnis der Anzahl Personen mit Fürsorgeunterstützung ausserhalb von Armenhäusern zur Anzahl Personen in Armenhäusern, "Old" der Anteil von über 65-jährigen in der Bevölkerung und "Pop" die Bevölkerungszahl. Diese Gleichung mit Hilfe der Daten von jeweils zwei Volkszählungen für verschiedene Verwaltungsbezirke ("unions") angepasst. Diese Bezirke waren in ihrer Sozialpolitik weitgehend autonom. Yule hat 4 Kategorien von unions gebildet (ländlich, gemischt, städtisch und grossstädtisch) und für jede Kategorie die Koeffizienten a, b, c und d separat geschätzt. So ergab sich etwa aufgrund der Volkszählungen 1871 und 1881 für grossstädtische Bezirke ein geschätzter Koeffizient $b = 0.755$. Das heisst also, eine Zunahme der Variable Out geht einher mit einer Zunahme der Anzahl armengenössiger Leute – selbst wenn man die andern Einflussgrössen wie "Old" berücksichtigt. Daraus schloss Yule, dass Fürsorge für Leute in ihrer gewohnten Umgebung zu mehr Armut führt.

Dies ist jedoch nicht stichhaltig. Es wurde nachgewiesen, dass die Bezirke mit effizienteren Verwaltungen in der Zeit auch mehr Armenhäuser bauten. Eine effizientere Verwaltung führt aber gleichzeitig zu einer Reduktion von Armut, d.h. die Effekte der effizienteren Verwaltung und der Einrichtung von Armenhäusern auf die Armut können nicht getrennt werden. Man sagt, dass die beiden Variablen "confounded" seien. Weiter sind auch ökonomische Faktoren mögliche confounders. Allgemein muss man beim Schluss von einem statistischen Zusammenhang (Assoziation) auf eine Kausalbeziehung äusserst vorsichtig sein. Insbesondere kann man im allgemeinen von einer Regressionsgleichung nicht auf die Wirkung einer Intervention (Änderung einer Variablen) schliessen. Für eine weitere Diskussion dieses Punktes verweise ich auf den Artikel von D. Freedman, "From association to causation: Some remarks on the history of statistics", Statistical Science 14 (1999), 243-258, aus dem obiges Beispiel stammt.

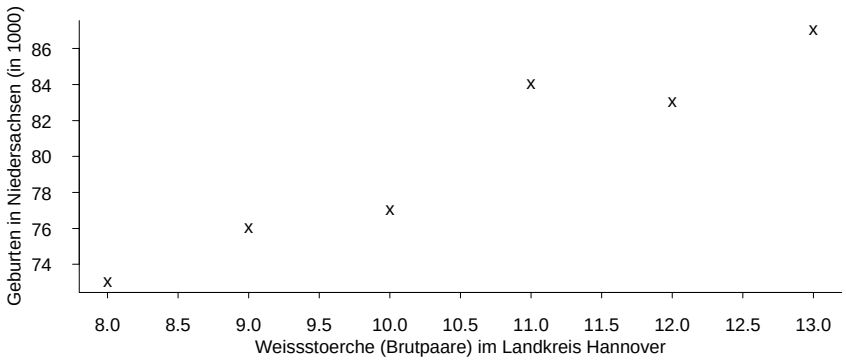

Abbildung 1.1: Zusammenhang zwischen Anzahl Geburten und Anzahl Störchen

Ein weiteres (fiktives) Beispiel: Gibt es einen Zusammenhang zwischen der Anzahl Störche

und der Anzahl Geburten beim Menschen? Die Daten dazu sind in der Abbildung 1.1 dargestellt.

Die Statistik liefert einen hochsignifikanten Zusammenhang zwischen der Anzahl Störche und der Geburtenzahl. Daraus könnte man fälschlicherweise schliessen, dass der Klapperstorch die Babies bringt. ("Ursache und Wirkung, Kausalzusammenhang").

Hier ist die confounding Variable die Zeit, was ziemlich häufig auftritt (**Unsinnkorrelation zwischen Zeitreihen**). In diesem Beispiel ist sofort offensichtlich, dass die Assoziation keinen Kausalzusammenhang impliziert. Betrachten wir aber zum Beispiel die Anzahl der Brutalitäten in TV-Sendungen und die Anzahl der Gewaltverbrechen, so finden wir ziemlich sicher einen statistischen Zusammenhang (beide nehmen im Verlauf der Zeit zu), und a priori ist auch ein Kausalzusammenhang möglich. Letzterer ist aber gar nicht so leicht zu beweisen!

1.1.2 Lineares Modell, Beispiele

Die multiple Regression:
Gegeben ist eine abhängige Variable (Zielvariable), die bis auf Messfehler (oder zufällige Schwankungen) linear von mehreren "unabhängigen" oder "erklärenden" Variablen (oder Versuchsbedingungen) abhängt.
Gesucht sind die Parameter, die diese Abhängigkeit beschreiben, sowie die Fehlervarianz.

Modell in Formeln:

$$\boxed{Y_i = \theta_1 x_{i1} + \ldots + \theta_p x_{ip} + \varepsilon_i \qquad (i = 1, \ldots, n)}$$

Bezeichnungen:

- Die $(Y_i; i = 1, \ldots, n)$ bilden den Vektor $\mathbf{y}$ der Realisierungen der **abhängigen Variable** (auch Zielvariable oder "response").
- Die $(x_{ij}; i = 1, \ldots, n)$ bilden den Vektor $\mathbf{x}^{(j)}$ der Realisierungen der j-ten **unabhängigen (erklärenden) Variable** (Versuchsbedingung) $(j = 1, \ldots, p)$.
- Die $(x_{ij}; j = 1, \ldots, p)$ bilden den Vektor $\mathbf{x}_i$ der erklärenden Variablen (Versuchsbedingungen) für die i-te Beobachtung $(i = 1, \ldots, n)$.
- Die $(\theta_j; j = 1, \ldots, p)$ bilden den Vektor $\boldsymbol{\theta}$ der **unbekannten Parameter**.
- Die $(\varepsilon_i; i = 1, \ldots, n)$ bilden den Vektor $\boldsymbol{\varepsilon}$ der (unbekannten) **Fehler**, die wir als zufällig annehmen.
- n ist die Anzahl Beobachtungen, p die Anzahl erklärender Variablen.

Die Grössen θ_j und ε_i sind unbekannt, während $x_{ij},\ y_i$ bekannt sind.

Modell in Vektorschreibweise:

$$\boxed{Y_i = \mathbf{x}_i^T \boldsymbol{\theta} + \varepsilon_i \qquad (i = 1, \ldots, n)}$$

bzw. in **Matrixschreibweise:**

$$\boxed{\begin{array}{ccccccc} \mathbf{Y} & = & X & \times & \boldsymbol{\theta} & + & \boldsymbol{\varepsilon} \\ n \times 1 & & n \times p & & p \times 1 & & n \times 1 \end{array}}$$

wobei dann X eine $(n \times p)$–Matrix ist mit Zeilen $\mathbf{x}_i^T$ und Spalten $\mathbf{x}^{(j)}$.

Meist ist die erste erklärende Variable einfach die Konstante, d.h. $x_{i1} = 1$ für alle i. Damit hat man einen Achsenabschnitt im Modell. Um in dem Fall den Parameter θ_1 interpretierbar zu machen, setzt man voraus, dass die Fehler ε_i den Erwartungswert null haben. Diese Annahme wird auch sonst meist gemacht.

Ferner setzt man üblicherweise voraus, dass man mehr Beobachtungen als Variablen hat $(p < n)$ und dass die Matrix X maximalen Rang p hat, d.h. die p Spalten von X sind linear unabhängig. Sonst sind die Parameter nicht identifizierbar (verschiedene Parameter ergeben das gleiche Modell). Manchmal verwendet man aber trotzdem Modelle mit linear abhängigen Spalten und erzwingt die Identifizierbarkeit mit Nebenbedingungen.

Ein Wort zur Notation: Wir bemühen uns, Vektoren konsequent fett zu schreiben. Hingegen sind wir weniger konsequent bei der Unterscheidung von Zufallsvariablen (gross) und deren Realisierungen (klein), weil man sehr oft zwischen den beiden Interpretationen wechselt und weil auch feste Matrizen gross geschrieben werden.

Beispiele:

(1) **Das Lokationsmodell:**

$$p = 1, \qquad X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \theta_1 = \mu.$$

(2) **Das 2–Stichproben–Modell:**

$$p = 2, \qquad X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \qquad \theta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

Die häufigsten Fragestellungen sind hier "Ist $\mu_1 = \mu_2$ plausibel ?", bzw. "Wie gross ist der Unterschied ?". Aus der Einführungsvorlesung sind dafür der 2–Stichproben–t–Test oder der 2–Stichproben–Wilcoxon–Test bekannt.

Statistisch gesehen ist das 2-Stichproben-Modell das einfachste, das man sich vorstellen kann. (In der Praxis meist einfacher als das 1–Stichproben–(Lokations–)Modell, da sich systematische und semisystematische Fehler oft herausheben.)

(3) **1–Weg (Einfache) Varianzanalyse mit k Niveaus (Gruppen)**
Dies ist eine Verallgemeinerung des vorherigen Beispiels von 2 auf k Gruppen. Dann ist $p = k$ und die Parameter sind die Erwartungswerte μ_j von Gruppe j $(1 \leq j \leq k)$ und die Matrix X sieht analog aus wie zuvor. Oft verwendet man aber eine andere Parametrisierung, nämlich

$$\mu_j = \mu + \alpha_j,$$

wobei μ den Gesamtmittelwert bezeichnet und α_j den Effekt der Gruppe j. Dann sind natürlich die $k+1$ Parameter nicht identifizierbar und die Spalten von X sind linear abhängig. Man erzwingt Identifizierbarkeit durch eine Nebenbedingung, z. B. $\sum \alpha_j = 0$.

Modelle dieser Art werden in der Vorlesung Varianzanalyse ausführlicher behandelt.

(4) **Die Regression durch den Nullpunkt:** $Y_i = \beta x_i + \varepsilon_i$ $(i = 1, \ldots n)$.

$$p = 1, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \theta_1 = \beta.$$

(5) **Einfache lineare Regression:** $Y_i = \alpha + \beta x_i + \varepsilon_i$ $(i = 1, \ldots n)$.

$$p = 2, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

(6) **Die quadratische Regression:** $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$ $(i = 1, \ldots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \theta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}.$$

Herkömmliche Interpretation der quadratischen Regression: Wir passen eine Parabel an die zweidimensionale Punktwolke $(x_1, y_1), \ldots, (x_n, y_n)$ an, vgl. Abb. 1.2, links.

Andere Interpretation: Die Parabel ist durch die Wertepaare $(x_1, x_1^2), \ldots, (x_n, x_n^2)$ fest vorgegeben. Wir suchen nun in der dritten Dimension (also y) eine geeignete Ebene, vgl. Abb. 1.2, rechts.

Fazit: Die Funktion, die wir anpassen, ist quadratisch in den bekannten Versuchsbedingungen, aber **linear** in den unbekannten Koeffizienten (und deshalb ein Spezialfall des allgemeinen linearen Modells).

(7) **Potenz-, bzw. exponentieller Zusammenhang:**
Eine Beziehung der Form $Y_i = \alpha x_i^\beta + \varepsilon_i$, bzw. $Y_i = \alpha \exp(\beta x_i) + \varepsilon_i$, mit unbekannten Parametern α und β passt nicht in unser Modell. Der deterministische Teil ändert sich jedoch nicht, wenn wir logarithmieren. Dies führt zum verwandten Modell

$$\log(Y_i) = \log(\alpha) + \beta \log(x_i) + \varepsilon_i.$$

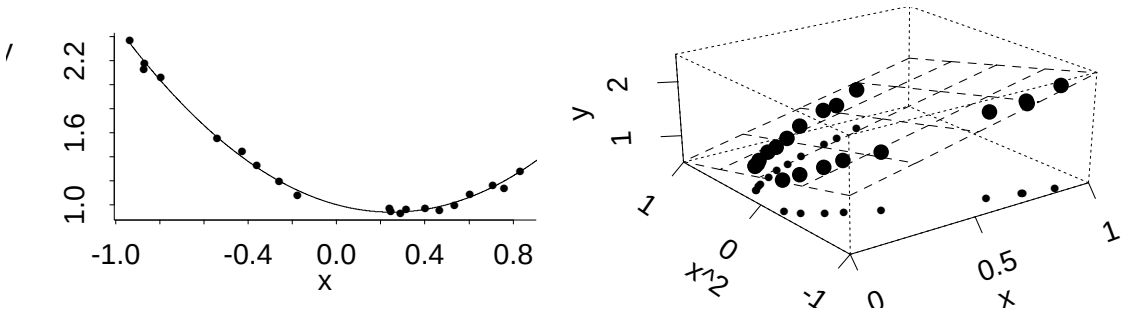

Abbildung 1.2: Quadratische Regression (links), interpretiert als multiple lineare Regression (rechts).

Dies ist ein Beispiel des allgemeinen linearen Modells mit Zielgrössen $\log(Y_i)$ und

$$p = 2 \qquad X = \begin{pmatrix} 1 & \log(x_1) \\ 1 & \log(x_2) \\ \dots & \\ 1 & \log(x_n) \end{pmatrix} \qquad \theta = \begin{pmatrix} \log(\alpha) \\ \beta \end{pmatrix}.$$

Wenn wir wieder zurücktransformieren, erhalten wir

$$Y_i = \alpha x_i^{\beta} \cdot \eta_i$$

mit $\eta_i = \exp(\varepsilon_i)$. Das heisst, auf der ursprünglichen Skala haben wir **multiplikative** und nicht **additive** Fehler. Bei Potenz- oder exponentiellen Zusammenhängen sind multiplikative Fehler meist plausibler, weil dann die Grösse der Fehler proportional zum mittleren Wert der Zielvariablen ist.

Als Beispiel mit konkreten Daten betrachten wir einen Datensatz, bei dem die Zielvariable die Erschütterung bei Sprengungen ist und man die Sprengladung und die Distanz zwischen dem Spreng- und dem Messort als erklärende Variablen verwendet. Die Daten sind in Figur 1.3 dargestellt. Es ist offensichtlich, dass die Abhängigkeit von der Distanz nicht linear ist. Auch scheint die Streuung mit der Distanz abzunehmen.

Aus physikalischen Gründen vermutet man, dass die Erschütterung umgekehrt proportional zur quadrierten Distanz ist. Wir hätten also ein Potenzmodell mit bekanntem β. In Figur 1.4 sind die Logarithmen der Erschütterungen gegen die Logarithmen der Distanzen bei einer festen Ladung dargestellt. Die Beziehung ist näherungsweise linear und die Streuung etwa konstant. Eine naheliegende Frage ist, ob die Steigung der Geraden wirklich gleich zwei ist, wie es von der physikalischen Theorie postuliert wird. Wir werden diese und weitere Fragen im Verlauf der Vorlesung untersuchen.

Wir haben in diesen Beispielen zwei wichtige Prinzipien gesehen:

- Das Modell heisst linear, weil es linear in den Parametern ist. Die ursprünglichen erklärenden Variablen können wir beliebig transformieren.
- Wir können oft ein lineares Modell erhalten, wenn wir beide Seiten der deterministischen Beziehung transformieren. Man muss sich dann aber Gedanken machen, ob additive Fehler in der transformierten Skala plausibel sind.

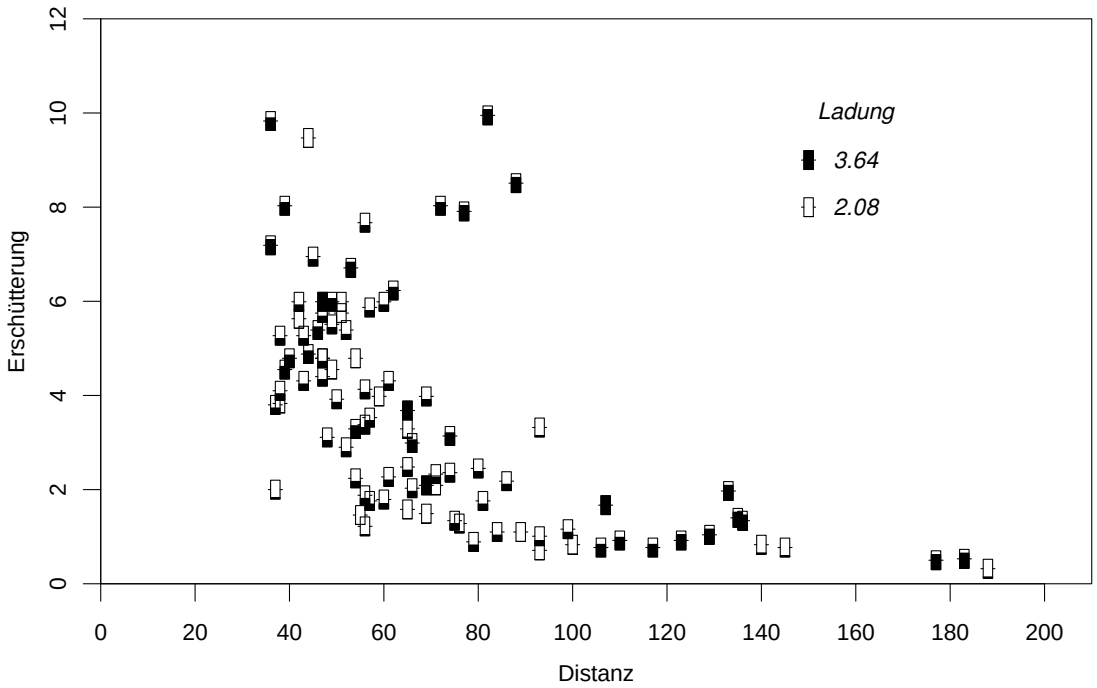

Abbildung 1.3: Erschütterung in Abhängigkeit der Distanz für verschiedene Ladungen

Unsere Ziele einer Regressionsanalyse:

- **Ein möglichst guter "fit".** Anpassung einer (Hyper–)Ebene über den Versuchsbedingungen durch den "response", so dass die Abweichungen klein sind. Das Standardverfahren ist die Methode der Kleinsten Quadrate, aber man kann die Grösse der Abweichungen auch anders quantifizieren.
- **Möglichst gute Parameterschätzungen.** Damit kann man die Frage beantworten: Wie ändert sich der response, wenn man eine erklärende Variable ändert ?
- **Eine möglichst gute Prognose.** Damit kann man die Frage beantworten: Welcher response ist unter neuer Versuchsbedingung zu erwarten ?
- **Eine Angabe der Unsicherheit für die drei vorangehenden Probleme** mittels Tests und Vertrauensintervallen.
- **Die Entwicklung eines einfachen und gut passenden Modells.** Dies geschieht meist in einem iterativen Prozess.

1.2 Die Voraussetzungen des linearen Modells

Wir benötigen gewisse Voraussetzungen, damit die Anpassung eines linearen Modells mit der Methode der kleinsten Quadrate sinnvoll ist, und damit die statistischen Tests und Vertrauensintervalle, die wir im folgenden herleiten, gültig sind. Bevor wir diese Bedingungen in absteigender Reihenfolge der Wichtigkeit auflisten, halten wir fest, dass das Modell keine Voraussetzungen über die erklärenden Variablen macht. Diese können stetig oder

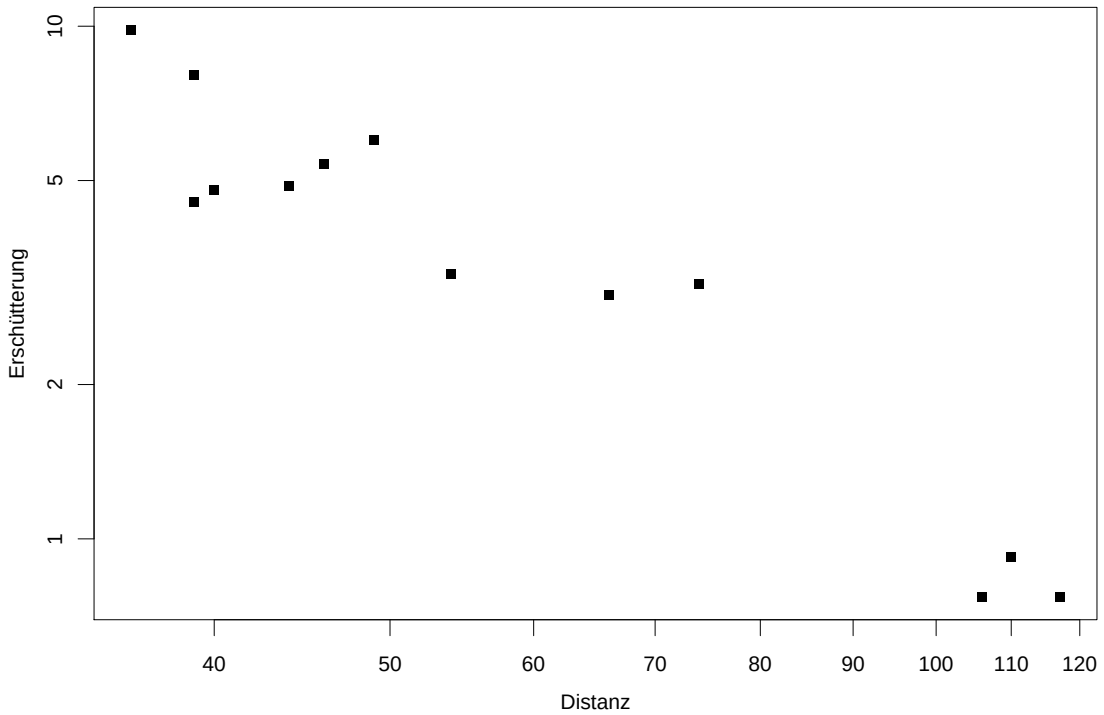

Abbildung 1.4: Distanz und Erschütterung bei Sprengungen mit Ladung 3.12. Die Achsen sind logarithmisch dargestellt

diskret sein, und sie können beliebig transformiert und kombiniert werden. Ausserdem spielt es im Prinzip keine Rolle, ob die Werte der erklärenden Variablen von der Person, die das Experiment durchführt, deterministisch festgelegt werden, oder ob sie selber auch Realisierungen von Zufallsvariablen sind. Die Theorie, die wir in den folgenden Abschnitten herleiten, betrachtet die erklärenden Variablen immer als deterministisch. Das heisst, dass unsere Aussagen im Fall, wo diese zufällig sind, als bedingte Aussagen gegeben die Werte der erklärenden Variablen zu verstehen sind.

Diese Voraussetzungen kann man teilweise mit statistischen Methoden nachprüfen, wie wir weiter unten sehen werden.

1. **Die Daten sind brauchbar für die gewünschte Information ("repräsentativ"; "sinnvoll"). Sie sind eine Zufallsauswahl aus der zu untersuchenden Population.**.

   - Ist diese Voraussetzung nicht erfüllt, so ist die ganze Analyse zum vornherein wertlos. Man kann die Daten genauso gut wegwerfen.
   - Um zu beurteilen, ob die Daten brauchbar sind, müssen wir Einsicht in das Problem haben. Es ist die reale Situation, die darüber entscheidet, und dies kann nicht mit statistischen Methoden überprüft werden.
   - Es kann sein, dass unsere eigentlichen Zielgrössen nicht genau messbar sind. Zum Beispiel: Wie misst man Intelligenz? Wir versuchen dies durch Prüfungen und Tests, die wir auswerten, und wir definieren dann die Intelligenz durch die Resultate aus Intelligenztests. Der Zusammenhang zwischen diesen beiden

Grössen ist aber eine offene Frage. Ein anderes Beispiel: Der Staat misst das Vermögen der Bürger durch das in der Steuererklärung angegebene Vermögen. **Die Übertragung der Ergebnisse von den messbaren Grössen auf die interessierenden Grössen ist ein eigenes Problem.**

2. **Die Regressionsgleichung ist korrekt.** Das heisst:

$$\boxed{\mathbf{E}\left[\varepsilon_i\right] = 0 \quad \forall i}$$

Insbesondere sollen keine wesentlichen erklärenden Variablen fehlen und die Beziehung zwischen der Zielgrösse und den erklärenden Variablen soll linear sein (nach geeigneten Transformationen).

3. **Die Fehler sind unkorreliert.** Das heisst (unter der Voraussetzung 2.):

$$\boxed{\mathbf{E}\left[\varepsilon_i \varepsilon_j\right] = 0 \quad \forall i, j \quad (i \neq j)}$$

Sind die Fehler korreliert, so bleibt die Anpassung mit Kleinsten Quadraten zwar noch brauchbar, aber die **Genauigkeit**, die wir zu haben glauben, entspricht nicht der wahren Genauigkeit: Wir erhalten falsche Niveaus für Tests und Konfidenzintervalle. Dies wird später ausführlicher diskutiert.

4. **Alle $\mathbf{x}_i$ sind exakt.**

Mit andern Worten sind alle erklärenden Variablen $\mathbf{x}_i$ ohne Fehler bekannt. Falls diese auch mit Beobachtungsfehlern versehen sind, so führt die Kleinste Quadrate Methode zu systematischen Fehlern. Es gibt dafür Korrekturen, wenn man zumindest das Verhältnis der Fehlervarianzen in den einzelnen Variablen kennt. Dies wird in der Literatur unter dem Stichwort "errors in variables models" diskutiert.

Das scheint zunächst ein Widerspruch zu sein zur vorherigen Aussage, dass die $\mathbf{x}_i$ durchaus auch zufällig sein können. Die Voraussetzung hier besagt aber, dass wir das $\mathbf{x}_i$ exakt kennen, welches zum entsprechenden Wert y_i geführt hat. Das ist etwas anderes als die Frage, wie der Wert von $\mathbf{x}_i$ zustande gekommen ist.

5. **Konstante Fehlervarianz ("homoscedasticity")** Das heisst:

$$\boxed{\mathbf{E}\left[\varepsilon_i^2\right] = \sigma^2 \quad \forall i}$$

Es sollen also alle Messungen die gleiche Genauigkeit haben. (Speziell sollen auch keine "groben Fehler" mit viel grösserer Varianz auftreten.) Oft kann man mittels einer einfachen Transformation der Zielgrösse eine konstante Varianz erhalten. Ist die Voraussetzung der Homoskedeastizität nicht erfüllt, so wird die Methode der kleinsten Quadrate bald einmal ungenau (verglichen mit andern Methoden). Dies wird ebenfalls später ausführlicher diskutiert.

6. **Die Fehler $(\varepsilon_i; i = 1, \ldots n)$ sind gemeinsam normalverteilt**

(Das Gleiche gilt dann auch für die Y_i's.) Dass die Annahme der Normalverteilung für Fehler oft naheliegend ist, geht aus den allgemeinen Eigenschaften der Normalverteilung hervor (Stichwort Hypothese der Elementarfehler, siehe Anhang). Sie muss jedoch trotzdem kritisch hinterfragt werden.

Die Voraussetzungen 2), 3), 5) und 6) kann man teilweise mit statistischen Methoden nachprüfen. Geeignete Verfahren werden wir weiter hinten im Skript besprechen.

Im Allgemeinen sind die obigen Voraussetzungen nur näherungsweise erfüllt. Die Kunst der Statistik besteht darin, ein Gefühl zu entwickeln, welche Abweichungen von den Voraussetzungen wesentlich sind und welche Aussagen und Verfahren auch noch sinnvoll sind, wenn das Modell nicht stimmt.

Wenn zum Beispiel $(X_1, X_2, \dots X_p, Y)$ ein $(p+1)$-dimensionaler Zufallsvektor ist mit beliebiger Verteilung und wir ein lineares Modell aufgrund von n unabhängigen Realisierungen dieses Zufallsvektors mit Kleinsten Quadraten anpassen, dann schätzen wir effektiv die Koeffizienten der besten linearen Prognose. Diese ist definiert als

$$\arg\min_{\theta_0,\dots\theta_p} \mathbf{E}\left[(Y - \theta_0 - \sum_{j=1}^{p} \theta_j X_j)^2\right].$$

In dem Sinne kann man Kleinste Quadrate fast immer verwenden, wenn man nur an Vorhersagen interessiert ist. Mit der Interpretation der Parameter und den Angaben zu deren Genauigkeit muss man jedoch aufpassen.

Zum Schluss noch ein Beispiel zur Verletzung der Voraussetzung 3) (sowie auch 1) und 2)): Die abhängige Variable ist die Anzahl Lebendgeburten in der Schweiz seit 1930, die erklärende Variable ist die Zeit (sowie gewisse Transformationen davon, wenn man quadratische Trends betrachten will).

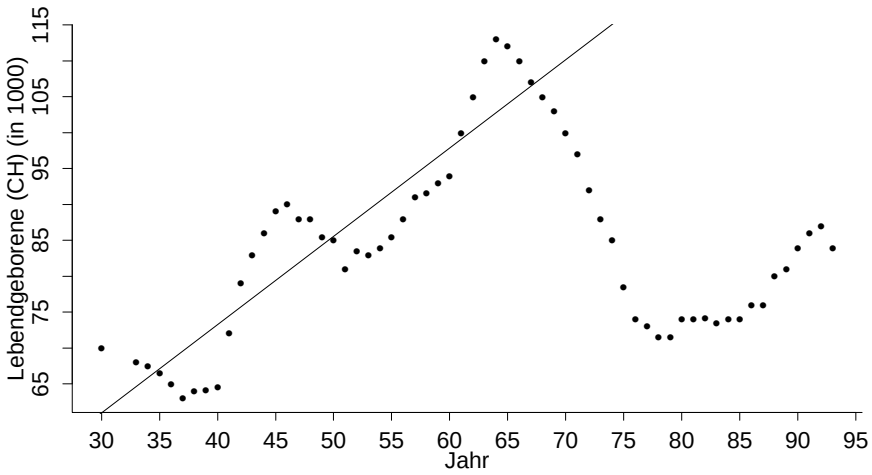

Abbildung 1.5: Der Pillenknick

Wie Abbildung 1.5 zeigt, lassen sich die Daten der Lebendgeburten in der Schweiz nach dem 2. Weltkrieg bis zum "Pillenknick" im Jahre 1964 in erster Näherung durch eine Gerade anpassen. Allerdings sieht man beim genaueren Betrachten, dass die Datenpunkte nicht symmetrisch um die Regressionsgerade verteilt sind; es gibt "Gruppen" links und rechts der Geraden. Maxima und Minima folgen sich im Abstand von etwa zwanzig Jahren (einer Generation).

Ausserdem sind die Jahre bis 1964 nicht "repräsentativ" für die folgenden Jahre, bzw. das Modell stimmt dann nicht mehr. Allgemein ist es gefährlich, ein angepasstes lineares Modell auf einen Bereich zu extrapolieren, wo man keine erklärenden Variablen beobachtet hat.

1.3 Kleinste Quadrate Schätzung

Es sei: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$

Wir suchen eine "möglichst gute" Schätzung von $\boldsymbol{\theta}$. Die Kleinste Quadrate Lösung $\widehat{\boldsymbol{\theta}}$ ist definiert als derjenige Wert, welcher die L_2–Norm des Fehlers minimiert:

$$\|\mathbf{y} - X\widehat{\boldsymbol{\theta}}\| = \min_{\boldsymbol{\theta}} \|\mathbf{y} - X\boldsymbol{\theta}\|.$$

Wir minimieren also den Euklid'schen Abstand der Abweichungen $\mathbf{y} - X\boldsymbol{\theta}$ vom Nullvektor.

1.3.1 Normalengleichungen

Wir berechnen die partiellen Ableitungen von $\|\mathbf{y} - X\boldsymbol{\theta}\|^2$ nach $\boldsymbol{\theta}$ (was einen Vektor ergibt) und setzen sie gleich Null, wodurch wir erhalten:

$$(-2)\ X^T(\mathbf{y} - X\widehat{\boldsymbol{\theta}}) = \mathbf{0} \qquad ((p \times 1) - \text{Nullvektor}),$$

das heisst: $\boxed{X^T X\widehat{\boldsymbol{\theta}} = X^T\mathbf{y}}$.

Dies sind die **Normalengleichungen**: Man hat p lineare Gleichungen für p Unbekannte (beachte, dass $X^T X$ eine $p \times p$-Matrix ist). Die Elemente von $X^T X$ sind die Skalarprodukte der Spalten von X. Die Lösung der Normalengleichungen werden also besonders einfach, wenn die Spalten $\mathbf{x}^{(j)}$ von X orthogonal sind. Eine andere Interpretation ergibt sich aus der Darstellung $X^T X = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$: $X^T X$ ist also n mal die Matrix der empirischen zweiten Momente der Versuchsbedingungen $(\mathbf{x}_i)$.

Wenn wir nun voraussetzen, dass die Matrix X den vollen Rang p hat, dann ist $X^T X$ invertierbar. In dem Fall ist also die Kleinste Quadrate Lösung eindeutig und hat die Darstellung

$$\boxed{\widehat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \mathbf{y}.}$$

Diese Formel ist nützlich für theoretische Überlegungen, aber für numerische Berechnung ist sie nicht zu empfehlen, da sie sehr anfällig auf Rundungsfehler ist. Ein numerisch stabiler Algorithmus benützt die QR-Zerlegung mit Hilfe von Givens-Rotationen.

Spezialfall: Einfache lineare Regression $y_i = \alpha + \beta x_i + \varepsilon_i$. Dann ist

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

und Einsetzen in die Normalengleichungen ergibt

$$\boxed{\begin{array}{lcl} n\alpha + (\sum_{i=1}^n x_i) \cdot \beta & = & \sum_{i=1}^n y_i \\ (\sum_{i=1}^n x_i)\alpha + (\sum_{i=1}^n x_i^2) \cdot \beta & = & \sum_{i=1}^n x_i y_i \end{array}}.$$

Zur einfachen Lösung dieses Gleichungssystems wenden wir die Technik der "Orthogonalisierung" an, d.h. wir führen eine neue Variable ein

$$x \longrightarrow \widetilde{x} := x - \overline{x} \text{ (wobei: } \overline{x} := \frac{1}{n}\sum_{i=1}^n x_i = \text{"arithmetisches Mittel")}$$

Dann gilt $y_i = \widetilde{\alpha} + \beta\widetilde{x}_i$ mit $\widetilde{\alpha} = \alpha + \beta\overline{x}$. Wegen

$$\sum_{i=1}^{n} \widetilde{x}_i = \sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0$$

ergibt sich sofort:

$$\widehat{\widetilde{\alpha}} = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}\ ; \qquad \widehat{\beta} = \frac{\sum_{i=1}^{n} \widetilde{x}_i y_i}{\sum_{i=1}^{n} \widetilde{x}_i^2}.$$

Durch Rücktransformation ergeben sich schliesslich die gesuchten Grössen:

$$\boxed{\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\ ; \qquad \widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}}$$

Im Fall der multiplen linearen Regression kann man, sofern die erste Spalte von X aus lauter Einsen besteht (d.h. man hat einen Achsenabschnitt im Modell), analog die Orthogonalisierung:

$$y_i = \alpha + \sum_{j=2}^{p} \theta_j \widetilde{x}_{ij}$$

durchführen. Dabei ist $\widetilde{x}_{ij} = x_{ij} - \overline{x}_j$ mit $\overline{x}_j = \sum_{i=1}^{n} x_{ij}/n$ (für $j > 1$) und $\alpha = \theta_1 + \sum_{j=2}^{p} \theta_j \overline{x}_j$. Daraus folgt dann leicht, dass $\overline{y} = \widehat{\theta}_1 + \sum_{j=2}^{p} \widehat{\theta}_j \overline{x}_j$, das heisst der Punkt $(\overline{y}), \overline{x}_2, \ldots \overline{x}_p, \overline{y})$ liegt auf der angepassten Hyperebene.

1.3.2 Geometrische Interpretation

Es gibt eine zeilen- und eine spaltenweise Betrachtung. Wenn das Modell einen Achsenabschnitt enthält (d.h. $x_{i1} \equiv 1$) dann haben wir bei der zeilenweisen Betrachtung n Punkte $(y_i, x_{i2}, \ldots x_{ip})$ im p-dimensionalen Raum, welche zufällig um eine $(p-1)$-dimensionale Hyperebene streuen. (Bei Modellen ohne Konstante sind es n Punkte im $(p+1)$-dimensionalen Raum, die um eine Hyperebene durch den Ursprung streuen). Die zufällige Streuung erfolgt aber nur in der Richtung der y-Achse. Daher bestimmt die Kleinste Quadrate Schätzung die Hyperebene so, dass die Summe der quadrierten Abstände der Punkte von der Ebene *in y-Richtung* minimal wird.

Die spaltenweise Betrachtung ist im Allgemeinen ergiebiger. Dabei betrachten wir den Beobachtungsvektor $\mathbf{y}$ als einen Punkt im n-dimensionalen Raum $\mathbb{R}^n$. Wenn wir den Parameter $\boldsymbol{\theta}$ variieren, beschreibt $X\boldsymbol{\theta}$ einen p-dimensionalen Unterraum, d.h. eine p-dimensionale Hyperebene durch den Ursprung im $\mathbb{R}^n$. Dann ist es naheliegend, $\boldsymbol{\theta}$ so zu schätzen, dass $X\boldsymbol{\theta}$ der Punkt auf der Hyperebene ist, der am nächsten bei $\mathbf{y}$ liegt. Die Wahl der L_2-Norm als Abstand im $\mathbb{R}^n$ entspricht der Euklid'schen Distanz, und bedeutet geometrisch, dass wir $\mathbf{y}$ orthogonal auf diese Hyperebene projizieren. Insbesondere ist die Kleinste Quadrate Lösung charakterisiert durch die Eigenschaft, dass der Residuenvektor $\mathbf{r} = \mathbf{y} - X\widehat{\boldsymbol{\theta}}$ orthogonal auf allen Spalten von X steht:

$$(\mathbf{y} - X\widehat{\boldsymbol{\theta}})^T X = 0.$$

Dies ist eine geometrische Deutung der Normalengleichungen (vgl. Abbildung 1.6).

Die orthogonale Projektion $X\widehat{\boldsymbol{\theta}}$ ist dann die Schätzung von $\mathbf{E}[\mathbf{y}]$ in unserem Modell. Diese Grössen heissen die angepassten (gefitteten) Werte und werden üblicherweise mit $\widehat{\mathbf{y}}$ bezeichnet.

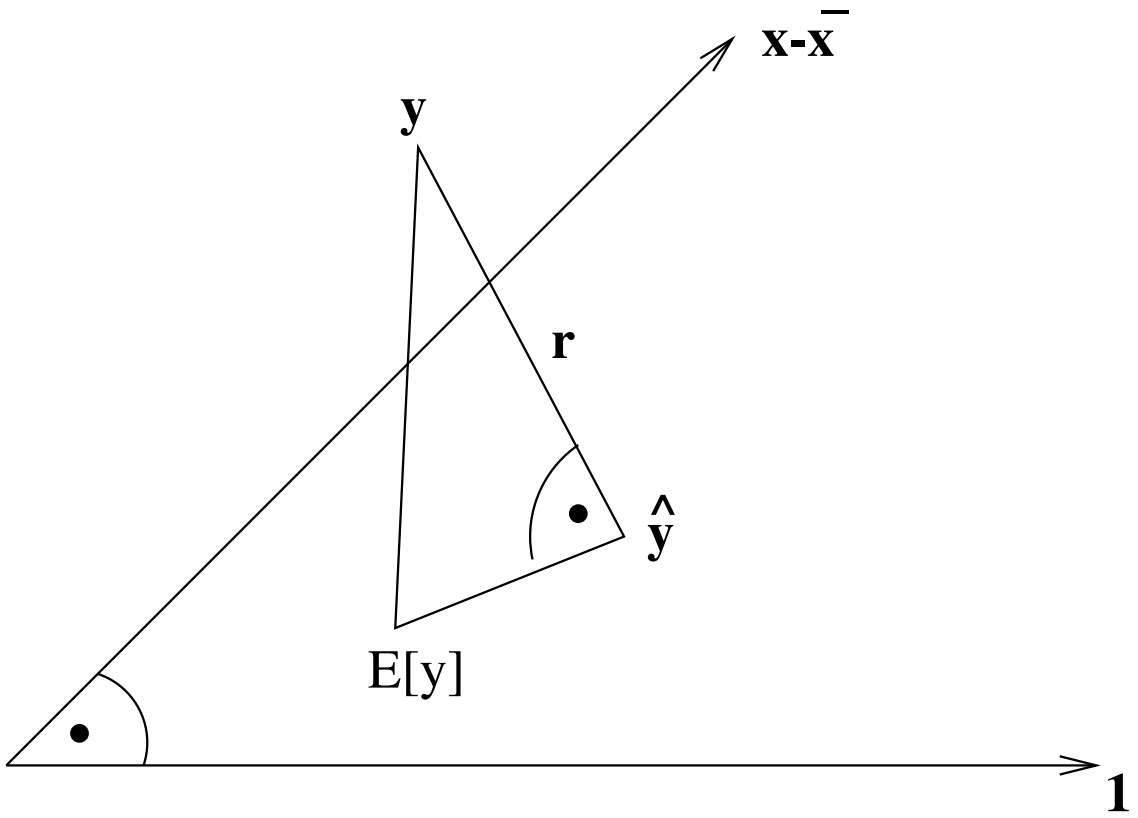

Abbildung 1.6: Der Residuenvektor $\mathbf{r}$ steht senkrecht auf der durch die Vektoren $\mathbf{1}$ und $\mathbf{x}$ aufgespannte Hyperebene.

In Formeln berechnet sich $\widehat{\mathbf{y}}$ als

$$\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\theta}} = \underbrace{X(X^TX)^{-1}X^T}_{=:P}\mathbf{y} = P\mathbf{y} \qquad \text{also:} \qquad \boxed{\widehat{\mathbf{y}} = P\mathbf{y}}$$

Man kann leicht nachprüfen, dass die Abbildungsmatrix P die Eigenschaften

$$P^T = P, \quad P^2 = P$$

hat, und dass

$$\sum_i P_{ii} = \mathrm{tr}(P) = \mathrm{tr}(X(X^TX)^{-1}X^T) = \mathrm{tr}((X^TX)^{-1}X^TX) = \mathrm{tr}(I_{p\times p}) = p.$$

Diese Eigenschaften sind notwendig und hinreichend dafür, dass P eine orthogonale Projektion vom $\mathbb{R}^n$ in den $\mathbb{R}^p$ beschreibt.

Die Matrix P hängt offensichtlich nur von den erklärenden Variablen (den Versuchsbedingungen) ab, und nicht von den Zielgrössen. Sie heisst auch Hut-Matrix (sie setzt $\mathbf{y}$ den Hut auf). Das Diagonalelement P_{ii} gibt an, wie stark der angepasste Wert $\widehat{y}_i$ an der Stelle $\mathbf{x}_i$ durch die Beobachtung y_i an der gleichen Stelle beeinflusst wird.

Die Residuen $\mathbf{r}$, die wir auch als $\widehat{\boldsymbol{\varepsilon}}$ bezeichnen, lassen sich in ähnlicher Art darstellen wie die angepassten Werte, es gilt nämlich:

$$\mathbf{r} = \widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = \underbrace{(I-P)}_{=:Q}\mathbf{y} = Q\mathbf{y} \qquad \text{also:} \qquad \boxed{\mathbf{r} = Q\mathbf{y}}$$

Q ist dann wiederum eine orthogonale Projektion, welche senkrecht zu P steht und in den $(n-p)$–dimensionalen Raum der Residuen führt:

$$Q^T = Q^2 = Q, \quad PQ = QP = 0 \ \ (0 \text{ als } (n \times n)\text{–Matrix}), \quad tr(Q) = n - p.$$

1.3.3 Zusammenhang mit MLE bei Normalverteilung

Die (bedingte) Dichte von $y_1, ..., y_n$ (gegeben die erklärenden Variablen) ist gemäss den Voraussetzungen unseres Modells (Unabhängigkeit und Normalverteilung der Fehler) gleich

$$L_{\mathbf{y},X}(\boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma} \varphi((y_i - \sum_{j=1}^{p} \theta_j x_{ij})/\sigma).$$

Bei der Maximum Likelihood Schätzung fassen wir die Dichte als Funktion der Parameter σ und $\boldsymbol{\theta}$ auf (bei festen y_i und x_{ij}). Diese Funktion bezeichnen wir als Likelihoodfunktion, und wir bestimmen die Parameter $\boldsymbol{\theta}$ und σ so, dass die Likelihoodfunktion (oder äquivalent dazu ihr Logarithmus) maximal wird. Das Resultat heisst der Maximum Likelihood Schätzer (MLE).

Man sieht sofort, dass die Maximierung nach $\boldsymbol{\theta}$ nicht vom Wert von σ abhängt und äquivalent ist zur Minimierung von $\|\mathbf{y} - X\boldsymbol{\theta}\|^2$. Wenn die Fehler ε_i i.i.d. und $\mathcal{N}(0, \sigma^2)$-verteilt sind, dann sind der MLE und der Kleinste Quadrate Schätzer für $\boldsymbol{\theta}$ identisch.

Als MLE für σ^2 findet man

$$\widehat{\sigma}^2_{ML} = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}.$$

Üblicherweise verwendet man aber nicht diesen Schätzer, sondern dasjenige Vielfache, das zu einem erwartungstreuen Schätzer führt. Wir werden sehen, dass der richtige Faktor $n/(n-p)$ ist, d.h. wir verwenden als Schätzer der Fehlervarianz

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n-p}.$$

1.3.4 Warum nicht Regression auf jede Variable einzeln ?

Das folgende (konstruierte) Beispiel soll zeigen, warum die multiple Regression nicht einfach durch mehrere einfache Regressionen ersetzt werden kann.

Wir haben 2 Versuchsbedingungen x_1, x_2 und zugehörigen Beobachtungen mit den folgenden Werten

x_1	0	1	2	3	0	1	2	3
x_2	−1	0	1	2	1	2	3	4
y	1	2	3	4	−1	0	1	2

Linkes Diagramm in Abb. 1.7: Wir tragen die y-Werte über den Versuchsbedingungen x_1 und x_2 auf. Man findet eine Ebene, welche auf die 8 gegebenen Punkte (im dreidimensionalen Raum) **exakt** passt:

$$\boxed{y = 2x_1 - x_2 \qquad (\widehat{\sigma}^2 = 0)}$$

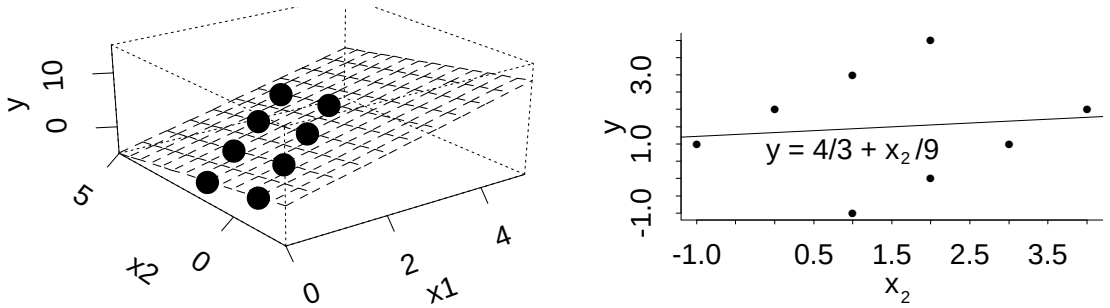

Abbildung 1.7: Multiple Regression gegenüber einfacher Regression

Die Koeffizienten 2, bzw. -1 geben an, wie sich y bei Veränderung von x_1 bzw. x_2 verhält, wenn dabei jeweils die andere Variable festgehalten wird.

Wir stellen fest: y nimmt bei zunehmendem $\mathbf{x_2}$ ab (x_2 grösser $\Rightarrow$ y kleiner).

Rechtes Diagramm in Abb. 1.7: Wir machen eine einfache lineare Regression von y auf x_2, wobei wir uns nicht um die x_1–Werte kümmern (sie werden also nicht konstant gehalten): Die sich ergebende Regressionsgerade lautet dann:

$$\boxed{y = \frac{1}{9}x_2 + \frac{4}{3} \qquad (\widehat{\sigma}^2 = 1.72)}$$

Diese Gerade gibt an, wie sich y bei Veränderung von x_2 verhält, falls x_1 variabel ist.

Wir stellen fest: y nimmt bei zunehmendem $\mathbf{x_2}$ zu (x_2 grösser $\Rightarrow$ y grösser).

Der Grund dafür, dass sich in den beiden Diagrammen die Variable y bezüglich der Variable x_2 ganz unterschiedlich verhält, liegt daran, dass x_1 und x_2 **sehr stark korreliert sind.** Konkret: Wenn x_2 wächst, so wächst auch x_1 mit.

Zusammenfassend halten wir fest:

Mehrere einfache Regressionen liefern (mit der Methode der kleinsten Quadrate) im allgemeinen nur dann dasselbe Ergebnis wie die multiple Regression, falls die erklärenden Variablen orthogonal sind.

1.4 Eigenschaften der KQ-Schätzung

Zuerst machen wir einige intuitive Überlegungen zur Genauigkeit der Regressionsebene: Nehmen wir einen bekannten linearen Zusammenhang an und simulieren wir sodann eine zufällige Punktwolke, die diesem Zusammenhang entspricht. Wir berechnen anhand dieser Punktwolke unsere Regressionsebene mit der Methode der kleinsten Quadrate. Wenn wir eine zweite Punktwolke (oder dritte, ...) nehmen, so wird sich für dieselbe theoretische Ebene immer wieder eine andere geschätzte Regressionsebene ergeben (vgl. Abb. 1.8). Mit andern Worten: Die geschätzten Parameter und die geschätzte Regressionsebene sind zufällig ! Man benötigt daher eine Angabe der Genauigkeit.

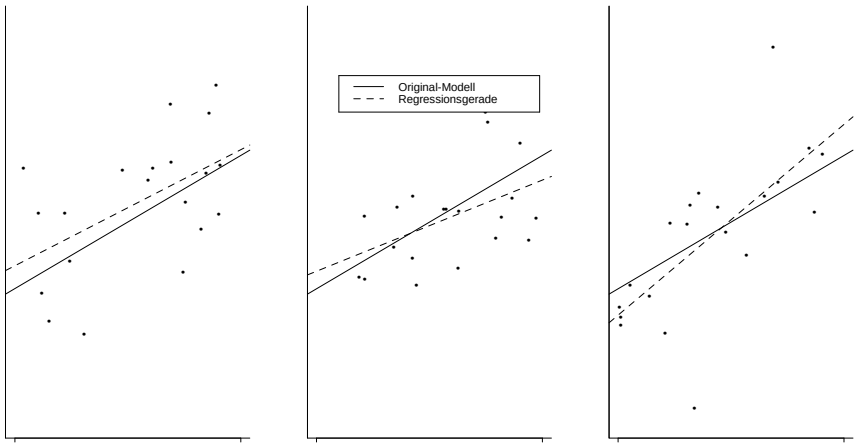

Abbildung 1.8: Drei verschiedene Regressionsgeraden zum selben Original-Modell.

1.4.1 Momente ohne Normalverteilungsannahmen

Für die folgenden Resultate braucht man keine Normalverteilung der Fehler ε_i. Wir vewenden die folgenden **Voraussetzungen für diesen Abschnitt:**

Übliches Modell: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$ **mit:** $\boxed{\begin{array}{l} \mathbf{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0} \\ \mathrm{Cov}[\boldsymbol{\varepsilon}] = \mathbf{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \sigma^2 I_{n\times n} \end{array}}$

Resultate:

(i) $\mathbf{E}\left[\widehat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}$, denn
$\mathbf{E}\left[\widehat{\boldsymbol{\theta}}\right] = \mathbf{E}\left[(X^TX)^{-1}X^T\mathbf{y}\right] = \mathbf{E}\left[(X^TX)^{-1}X^T(X\boldsymbol{\theta}+\boldsymbol{\varepsilon})\right] = \boldsymbol{\theta} + \mathbf{0} = \boldsymbol{\theta}$.

(ii) $\mathbf{E}\left[\widehat{\boldsymbol{\varepsilon}}\right] = \mathbf{0}$, $\mathbf{E}\left[\widehat{\mathbf{y}}\right] = \mathbf{E}\left[\mathbf{y}\right] = X\boldsymbol{\theta}$.

(iii) $\mathrm{Cov}[\widehat{\boldsymbol{\theta}}] = \sigma^2(X^TX)^{-1}$, denn
$\mathbf{E}\left[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\right] = \mathbf{E}\left[(X^TX)^{-1}X^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^TX(X^TX)^{-1}\right] = \sigma^2(X^TX)^{-1}$.

(iv) $\mathrm{Cov}[\widehat{\mathbf{y}}] = \mathrm{Cov}[P\mathbf{y}] = \sigma^2PP^T = \sigma^2P$ (weil P eine Projektionsmatrix ist).

(v) $\mathrm{Cov}[\widehat{\boldsymbol{\varepsilon}}] = \sigma^2Q$ (analog).

(vi) $\mathrm{Cov}[\widehat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{y}}] = 0$, (weil $QP = 0$).

Die beiden Kovarianzmatrizen in (iv) und (v) sind nur positiv semidefinit. Wie (v) zeigt, sind die Residuen $r_i = \widehat{\varepsilon}_i$ im Unterschied zu den wahren Fehlern korreliert, und ihre Varianz ist nicht konstant, sondern

$$\mathrm{Var}[\widehat{\varepsilon}_i] = \sigma^2(1 - P_{ii}).$$

Ausserdem folgt daraus, dass:

$$\begin{aligned} \mathbf{E}\left[\sum_{i=1}^n r_i^2\right] &= \sigma^2\sum_{i=1}^n(1-P_{ii}) \\ &= \sigma^2(n - \mathrm{tr}(P)) = \sigma^2(n-p). \end{aligned}$$

Daher ist:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-p} = \frac{\|\mathbf{y} - X\widehat{\boldsymbol{\theta}}\|^2}{n-p}$$

eine **erwartungstreue Schätzung** für σ^2 wie früher behauptet.

Man beachte, dass wir keine Aussagen machen können über die Varianz von $\widehat{\sigma}^2$. Dazu bräuchten wir das vierte Moment der Fehler ε_i.

1.4.2 Verteilungen unter Annahme der Normalverteilung

Voraussetzungen für diesen Abschnitt:

Übliches Modell: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$ **wobei jetzt:** $\boxed{\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0},\ \sigma^2 I_{n\times n})}$

Resultate:

(i) $\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta},\ \sigma^2(X^TX)^{-1})$ (denn $\widehat{\boldsymbol{\theta}}$ ist als Linearkombination normalverteilter Grössen selbst wiederum normalverteilt).

(ii) $\widehat{\mathbf{y}} \sim \mathcal{N}_n(X\boldsymbol{\theta},\ \sigma^2 P)$, $\widehat{\boldsymbol{\varepsilon}} \sim \mathcal{N}_n(\mathbf{0},\ \sigma^2 Q)$ (gleiche Begründung wie oben).

iii) $\widehat{\mathbf{y}}$ und $\widehat{\boldsymbol{\varepsilon}}$ sind **unabhängig** (da unkorreliert und zusätzlich normalverteilt).

(iv)

$$\frac{\sum_{i=1}^n r_i^2}{\sigma^2} \sim \chi^2_{n-p}.$$

(siehe unten).

(v) $\widehat{\sigma}^2$ ist unabhängig von $\widehat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\mathbf{y}$ (Dies folgt aus iii), denn es gilt auch $\widehat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\widehat{\mathbf{y}}$).

Beweis von iv): Betrachte ein Koordinatensystem mit orthogonalen Basisvektoren, so dass die ersten p Basisvektoren gerade den Spaltenraum von X aufspannen. Die zugehörige Transformationsmatrix bezeichnen wir mit A, d.h. die Spalten von A enthalten die Koordinaten der neuen Basisvektoren im alten System. Dann ist A orthogonal, und wenn wir einen Stern für das neue Koordinatensystem verwenden, dann gilt $\mathbf{y}^* = A^T\mathbf{y}$, $\boldsymbol{\varepsilon}^* = A^T\boldsymbol{\varepsilon}$, etc.. Nach Konstruktion ist klar, dass

$$\begin{aligned}\widehat{\mathbf{y}}^* &= (y_1^*, y_2^*, \dots y_p^*, 0, \dots 0)^T,\\ \widehat{\boldsymbol{\varepsilon}}^* &= (0, \dots, 0, \varepsilon_{p+1}^*, \dots \varepsilon_n^*)^T\end{aligned}$$

gilt (man kann es auch nachrechnen – dann muss man beachten, dass die letzten $n-p$ Zeilen von A^TX alle gleich null sind). Also gilt wegen der Orthogonalität von A

$$\sum_{i=1}^n \widehat{\varepsilon}_i^2 = \sum_{i=1}^n \widehat{\varepsilon}_i^{*2} = \sum_{i=p+1}^n \varepsilon_i^{*2}.$$

Daraus folgt die Behauptung, denn wegen der Orthogonalität von A ist $\boldsymbol{\varepsilon}^*$ ebenfalls $\mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n\times n})$-verteilt.

1.4.3 Asymptotische Normalität

Die obigen Resultate über die Verteilung der Schätzer bilden den Schlüssel für Unsicherheitsangaben, d.h. Vertrauensintervalle oder Tests. Man fragt sich daher, wie entscheidend die Voraussetzung der Normalverteilung für die Fehler ist. Es zeigt sich, dass die Resultate genähert richtig bleiben, wenn die Fehler nicht exakt normalverteilt sind. Mathematisch kann man das formulieren, indem man die Verteilung untersucht im Grenzfall, wo die Anzahl Beobachtungen gegen unendlich geht.

Wir betrachten die folgende Situation: Wir haben n Datenpunkte $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, welche dem linearen Modell genügen. Dabei ist wie zuvor $\mathbf{x}_i$ jeweils ein p-dimensionaler Spaltenvektor (d.h. $\mathbf{x}_i^T$ ist die i-te Zeile von X). Wir nehmen an, dass die Fehler ε_i i.i.d., aber nicht unbedingt normalverteilt sind und betrachten den Grenzfall $n \longrightarrow \infty$.

Für die Gültigkeit der asymptotischen Näherung brauchen wir schwache Bedingungen an die erklärenden Variablen $\mathbf{x}_i$:

- Der kleinste Eigenwert $\lambda_{\min,n}$ von $X^T X = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ konvergiert gegen ∞.
- $\max_j P_{jj} = \max_j \mathbf{x}_j^T (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \mathbf{x}_j$ konvergiert gegen null.

Die erste Bedingung besagt, dass man mit wachsendem n immer mehr Information bekommt. Die zweite Bedingung besagt, dass kein $\mathbf{x}_j$ die andern dominiert.

Satz 1.4.1. *Wenn die Fehler ε_i i.i.d. sind mit Erwartungswert 0 und Varianz σ^2 und $(\mathbf{x}_i)$ die obigen Bedingungen erfüllt, dann sind die KQ-Schätzer $\widehat{\boldsymbol{\theta}}$ konsistent (für $\boldsymbol{\theta}$) und die Verteilung von*

$$(X^T X)^{1/2} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

konvergiert schwach gegen $\mathcal{N}_p(\mathbf{0}, \sigma^2 I)$.

Beweis: Die i-te Komponente $\widehat{\theta}_i$ ist erwartungstreu und hat die Varianz $\sigma^2((X^T X)^{-1})_{ii}$, welche auf Grund der ersten Annahme gegen null konvergiert. Also folgt die Konsistenz aus der Chebyshev-Ungleichung.

Um schwache Konvergenz der Verteilungen eines Zufallsvektors im $\mathbb{R}^p$ zu beweisen, genügt es die schwache Konvergenz der Verteilungen von Linearkombinationen nachzuweisen (Satz von Cramér-Wold, siehe Literatur). Also betrachten wir

$$\mathbf{c}^T((X^T X)^{1/2})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{c}^T((X^T X)^{-1/2}) X^T \boldsymbol{\varepsilon} = \mathbf{a_n}^T \boldsymbol{\varepsilon} = \sum_{i=1}^n a_{ni} \varepsilon_i$$

mit

$$\mathbf{a_n} = X((X^T X)^{-1/2})^T \mathbf{c}.$$

Wir haben also eine Summe von n unabhängigen, aber nicht identisch verteilten Summanden $a_{ni}\varepsilon_i$. Ausserdem ändert sich die Verteilung der Summanden mit n. Das ist genau die Situation, für die der Satz von Lindeberg geschaffen wurde (siehe Einführungsvorlesung). Es gilt

$$\operatorname{Var} \sum_{i=1}^n a_{ni} \varepsilon_i = \sigma^2 \sum_{i=1}^n a_{ni}^2 = \sigma^2 \mathbf{a_n}^T \mathbf{a_n} = \sigma^2 \mathbf{c}^T \mathbf{c}.$$

Wir können ohne Einschränkung annehmen, dass diese Varianz gleich 1 ist. Also müssen wir nur noch die Bedingung

$$\sum_{i=1}^{n} a_{ni}^2 \mathbf{E}\left[\varepsilon_i^2 1_{[|\varepsilon_i|>\eta/a_{ni}]}\right] \to 0$$

für alle $\eta > 0$ nachprüfen. Weil alle ε_i die gleiche Verteilung und endliches zweites Moment haben, folgt

$$\mathbf{E}\left[\varepsilon_i^2 1_{[|\varepsilon_i|>d}\right] = \mathbf{E}\left[\varepsilon_1^2 1_{[|\varepsilon_1|>d]}\right] \stackrel{d\to\infty}{\longrightarrow} 0.$$

Weil ausserdem noch $\sum_i a_{ni}^2 = 1$ ist, genügt es zu zeigen, dass $\max_i |a_{ni}|$ gegen null geht. Mit der Schwarz'schen Ungleichung ist

$$a_{ni}^2 \leq \|\mathbf{c}\|^2 \, \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i.$$

Daher folgt die Behauptung direkt aus der zweiten Bedingung. □

Man kann auch die Konsistenz von $\widehat{\sigma}^2$ zeigen. Für die asymptotische Normalität von $\widehat{\sigma}^2$ würde man aber die Existenz des vierten Moments der ε_i brauchen, und die asymptotische Varianz hängt wesentlich vom Wert dieses vierten Moments ab.

Folgerungen:

Die Tests und Konfidenzintervalle für $\boldsymbol{\theta}$ und für die Erwartungswerte $\mathbf{E}[\mathbf{y}]$, die wir im nächsten Abschnitt unter der Annahme von normalverteilten Fehlern herleiten, haben auch ohne Normalverteilung approximativ das korrekte Niveau. Jedoch wissen wir nichts über die Effizienz der Verfahren, und die Vertrauensintervalle für σ, die man in der Literatur findet, haben ohne Normalverteilung meist ein grob falsches Niveau.

1.5 Tests und Vertrauensintervalle

1.5.1 Die grundlegenden Teststatistiken

Wir nehmen an, dass das lineare Modell mit $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_{n\times n})$ gilt. Wie wir im letzten Abschnitt gesehen haben ist unter diesen Annahmen $\widehat{\boldsymbol{\theta}}$ **exakt** $\mathcal{N}_p(\boldsymbol{\theta}, \sigma^2 (X^T X)^{-1})$**–verteilt** und $\widehat{\sigma}^2$ ist **unabhängig** von $\widehat{\boldsymbol{\theta}}$. Daraus erhalten wir

(a) Für jeden einzelnen Parameter

$$\frac{\widehat{\theta}_i - \theta_i}{\widehat{\sigma}\sqrt{((X^T X)^{-1})_{ii}}} \sim t_{n-p}.$$

(b) Für den ganzen Parametervektor $\boldsymbol{\theta}$

$$\frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T (X^T X)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{p\widehat{\sigma}^2} \sim F_{p,n-p}$$

(c) Für eine Linearkombination $\boldsymbol{\vartheta} = B\boldsymbol{\theta}$, wobei B eine $(q \times p)$–Matrix ist:

$$\frac{(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})^T V^{-1} (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})}{q\widehat{\sigma}^2} \sim F_{q,n-p},$$

mit $V = B(X^T X)^{-1} B^T$.

(d) Für den Erwartungswert der i-ten Beobachtung (die wahre Höhe der Hyperebene über der i–ten Versuchsbedingung)

$$\frac{\widehat{y}_i - \mathbf{E}\left[y_i\right]}{\widehat{\sigma}\sqrt{p_{ii}}} \sim t_{n-p} \qquad \text{wobei } p_{ii} := (P)_{ii}$$

(e) Für den Erwartungswert einer neuen Beobachtung bei einer beliebigen Versuchsbedingung $\mathbf{x}_0$ (die wahre Höhe der Hyperebene über der Versuchsbedingung $\mathbf{x}_0$)

$$\frac{\widehat{y}_0 - \mathbf{E}\left[y_0\right]}{\widehat{\sigma}\sqrt{\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}} \sim t_{n-p}$$

(f) Für eine zufällige neue Beobachtung $y_0 = y_0(\mathbf{x}_0)$ unter den Versuchsbedingungen $\mathbf{x}_0$:

$$\frac{y_0 - \widehat{y}_0}{\widehat{\sigma}\sqrt{1 + \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}} \sim t_{n-p}.$$

Damit kann man in der üblichen Art Tests durchführen (die Grössen auf der linken Seite in den obigen Aussagen verwendet man als Teststatistik) und Vertrauensbereiche bilden für einzelne Parameter, Linearkombinationen von Parametern, die unbekannte wahre Höhe der Hyperebene an einer Stelle $\mathbf{x}_0$ etc. Ebenso erlaubt die Aussage (f), Prognoseintervalle für zukünftige Beobachtungen zu bilden.

Zur Illustration betrachten wir das Beispiel der Sprengungen. Tabelle 1.1 zeigt den Computeroutput mit der logarithmierten Erschütterung als Ziel- und der logarithmierten Distanz und der logarithmierten Ladung als erklärende Variablen .

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

Residual standard 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 1.1: Computer-Output für das Beispiel der Sprengungen

Es werden also die geschätzten Koeffizienten, die Standardfehler $\widehat{\sigma}\sqrt{((X^T X)^{-1})_{ii}}$ und die Testresultate für die Nullhypothesen $\theta_i = 0$ angegeben. "Residual standard error" bedeutet die Schätzung $\widehat{\sigma}$. Die andern Angaben werden in den folgenden Teilabschnitten erklärt. Die t-Verteilung mit 45 Freiheitsgraden ist sehr nahe bei der Normalverteilung. Daher ist es klar, dass der wahre Koeffizient der logarithmierten Distanz kleiner als zwei sein muss, während der wahre Koeffizient der Ladung ohne weiteres eins sein kann.

1.5.2 Vertrauensband für die ganze Hyperebene

Es ist auch möglich, beispielsweise einen 95%-Bereich anzugeben, in welchem die theoretische Hyperebene sich befindet. Wir erläutern vorerst ein naheliegendes Vorgehen, das

dann jedoch scheitert. Für einen beliebigen Punkt $\mathbf{x}_0$ (mit dem Wert $\widehat{y}(\mathbf{x}_0)$ auf der konstruierten Regressionsebene) können wir wie oben zwei Grenzen um den Wert $\widehat{y}(\mathbf{x}_0)$ konstruieren, innerhalb derer wir den Wert der theoretischen Hyperebene zu 95% erwarten. Wenn man dieses Vorgehen bei jedem Wert $\mathbf{x}_0$ durchführt, erhält man dann einen 95%–Konfidenzbereich für die theoretische Hyperebene ?

Die Antwort ist natürlich "nein". An einer Stelle $\mathbf{x}_0$ ist zwar die Wahrscheinlichkeit, dass die wahre Ebene durch das Intervall führt, genau 95%, falls wir aber zwei Intervalle haben mit je dieser 95%-Wahrscheinlichkeit, so haben wir für die Ebene eine Chance von mindestens 90% bis höchstens 95%, dass sie durch beide Bereiche führt.

Extremfälle: (entsprechend)	10 Punkte	$\Rightarrow$	50 - 95 %
	20 Punkte	$\Rightarrow$	0 - 95 %

Dies ist somit kein guter Weg, um auf einen Konfidenzbereich für die wahre Hyperebene zu kommen!

Ein besserer Weg ist der folgende. Die Schwarz'sche Ungleichung für das Skalarprodukt $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T (X^T X)^{-1} \mathbf{b}$ impliziert

$$\begin{aligned} |\widehat{y}_0 - \mathbf{E}\,[y_0]\,| &= |\mathbf{x}_0^T (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})| = |\mathbf{x}_0^T (X^T X)^{-1} (X^T X)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})| \\ &\leq (\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0)^{1/2} ((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T (X^T X)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))^{1/2}. \end{aligned}$$

Mit (b) von oben folgt daher, dass mit Wahrscheinlichkeit $1 - \alpha$ *simultan für alle* $\mathbf{x}_0$

$$(\widehat{y}_0 - \mathbf{E}\,[y_0])^2 \leq \widehat{\sigma}^2\, (\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0)\, p\, F_{p,n-p}(1 - \alpha).$$

gilt. Dies ergibt den gesuchten simultanen Bereich, ein Hyperboloid. Er ist die Envelope aller Hyperebenen, deren Parameter gemäss b) mit den Daten kompatibel sind.

1.5.3 Vergleich zweier geschachtelter Modelle, Varianzanalyse

Voraussetzungen:

"Rahmenhypothese"	$H : \mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$
	$(X : n \times p,\ Rg(X) = p,\ \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0,\ \sigma^2 I))$.
"spezielle Nullhypothese"	H_0: obiges und zusätzlich $B\boldsymbol{\theta} = \mathbf{b}$
	(mit $B : (p-q) \times p,\ Rg(B) = p - q < p$).

Zum Beispiel:

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{b} = \mathbf{0}.$$

In Worten lautet dann die Nullhypothese "Die ersten $p - q$ Koeffizienten θ_i sind gleich 0." Wir testen, ob die ersten $p - q$ Variablen im Modell überflüssig sind.

Gemäss Punkt (c) aus Abschnitt 1.5.1 ist

$$\frac{(B\widehat{\boldsymbol{\theta}} - \mathbf{b})^T (B(X^T X)^{-1} B^T)^{-1} (B\widehat{\boldsymbol{\theta}} - \mathbf{b})}{(p - q)\widehat{\sigma}^2}$$

eine geeignete Teststatistik der obigen Nullhypothese. Sie hat unter dieser Nullhypothese eine $F_{p-q,n-p}$-Verteilung. Die folgende geometrische Überlegung führt zu einer anderen Form und Interpretation dieser Teststatistik.

Wir nehmen an, dass $\mathbf{b} = \mathbf{0}$ (Dies ist keine wesentliche Einschränkung, weil wir statt der ursprünglichen Beobachtungen $\mathbf{y} - X\boldsymbol{\theta}$ betrachten können mit einem $\boldsymbol{\theta}$, welches $B\boldsymbol{\theta} = \mathbf{b}$ erfüllt). Dann kann man $\mathbf{y}$ zunächst in den p–dimensionalen Raum, der von den Spalten von X aufgespannt wird, projizieren und nachher noch weiter auf den q–dimensionalen Unterraum, der durch die zusätzliche Nebenbedingung $B\boldsymbol{\theta} = \mathbf{0}$ definiert wird.

Die dazugehörigen Quadratsummen der Residuen (unter H und H_0) seien SSE und SSE_0.

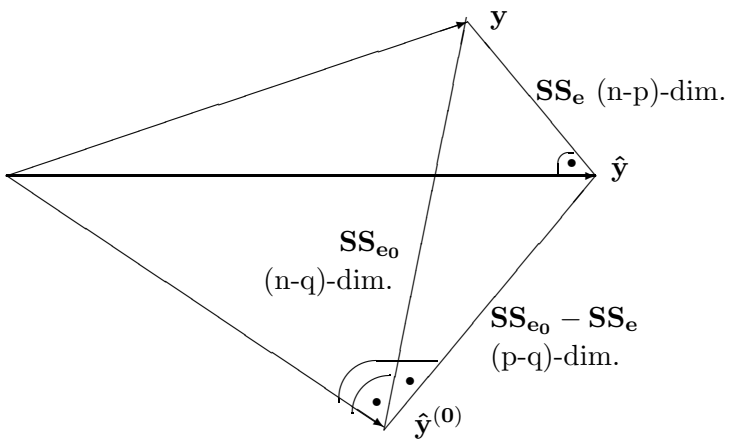

Abbildung 1.9: Modellvergleich

Resultate:

- $SSE/(n-p)$ liefert stets (unter der Rahmenhypothese H **und** unter der Nullhypothese H_0) eine **erwartungstreue** Schätzung für σ^2.
- SSE und $SSE_0 - SSE$ sind Quadratsummen in orthogonalen Unterräumen, und **unter der Nullhypothese** H_0 ist $(SSE_0 - SS_e)/(p-q)$ eine **erwartungstreue** Schätzung für σ^2. Wenn nur die Rahmenhypothese H gilt, ist der Erwartungswert dieser Differenz grösser als σ^2.
- Wegen der Orthogonalität der Unterräume gilt

$$\left\|\mathbf{y} - \widehat{\mathbf{y}}^{(0)}\right\|^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \left\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(0)}\right\|^2$$

.

Also ist unter $\mathbf{H_0}$:

$$\frac{(SSE_0 - SSE)/(p-q)}{SSE/(n-p)} = \frac{\left\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(0)}\right\|^2/(p-q)}{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2/(n-p)} \sim F_{p-q,n-p}$$

und wir können den Ausdruck auf der linken Seite als Teststatistik für H_0 verwenden.

Auf den ersten Blick sind die beiden oben hergeleiteten Teststatistiken verschieden (gleiche Verteilung heisst ja a priori nicht, dass die Zufallsvariablen identisch sind). Das folgende Lemma zeigt, dass die beiden Ausdrücke tatsächlich gleich sind.

Lemma 1.5.1. *Der Kleinste Quadrate Schätzer* $\widehat{\boldsymbol{\theta}}_{(0)}$ *unter der Nebenbedingung* $B\boldsymbol{\theta} = \mathbf{b}$ *ist gleich*

$$\widehat{\boldsymbol{\theta}}_{(0)} = \widehat{\boldsymbol{\theta}} - (X^T X)^{-1} B^T (B(X^T X)^{-1} B^T)^{-1} (B\widehat{\boldsymbol{\theta}} - \mathbf{b}).$$

Ferner gilt

$$SSE_0 = SSE + (B\widehat{\boldsymbol{\theta}} - \mathbf{b})^T (B(X^T X)^{-1} B^T)^{-1} (B\widehat{\boldsymbol{\theta}} - \mathbf{b}),$$

Beweis: Wir führen ein Vektor $\boldsymbol{\lambda}$ ein, der die $p-q$ Lagrange Multiplikatoren für die $p-q$ Nebenbedingungen enthält. Dann müssen wir

$$(y - X\boldsymbol{\theta})^T (y - X\boldsymbol{\theta}) + (B\boldsymbol{\theta} - \mathbf{b})^T \boldsymbol{\lambda}$$

bezüglich $\boldsymbol{\theta}$ und $\boldsymbol{\lambda}$ minimieren. Dies ergibt die Bedingungen

$$X^T (y - X\widehat{\boldsymbol{\theta}}_{(0)}) + B^T \boldsymbol{\lambda} = 0, \quad B\widehat{\boldsymbol{\theta}}_{(0)} = \mathbf{b}.$$

Man kann leicht nachprüfen, dass das angegebene $\widehat{\boldsymbol{\theta}}_{(0)}$ diese Bedingungen erfüllt. Mit dem Satz von Pythagoras folgt ferner

$$(y - X\widehat{\boldsymbol{\theta}}_{(0)})^T (y - X\widehat{\boldsymbol{\theta}}_{(0)}) = SSE + (X(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(0)}))^T (X(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(0)})).$$

Mit Einsetzen folgt daraus die zweite Behauptung. □

Genauso wie man multiple Regression nicht durch einfache Regressionen auf einzelne Variablen ersetzen kann, ergibt der Test der Nullhypothese $\beta_1 = \beta_2 = 0$ unter Umständen ganz andere Resultate als die beiden Tests für die Nullhypothesen $\beta_1 = 0$. bzw. $\beta_2 = 0$. Es kann zum Beispiel geschehen, dass die beiden letzteren Nullhypothesen ohne weiteres akzeptiert werden, aber die Nullhypothese $\beta_1 = \beta_2 = 0$ wuchtig verworfen wird. Das heisst man kann die erste oder die zweite erklärende Variable weglassen, aber nicht beide. Die Erklärung dieses scheinbaren Widerspruchs ist, dass beide Variablen stark korreliert sind. Daher kann die eine ohne weiteres an die Stelle der andern treten.

Oft hat man eine erklärende Variable, die verschiedene Kategorien darstellt (Herkunft, Typ, Farbe, Geschlecht, ...). Im Beispiel der Sprengungen wurde an sechs verschiedenen Stellen gesprengt, was über eine variierende Beschaffung des Untergrunds einen Einfluss haben kann. Eine solche Variable heisst auch ein **Faktor**. Das einfachste Modell postuliert für jede Kategorie einen andern Achsenabschnitt, während die andern Koeffizienten für alle Kategorien gleich sind. Zur Formulierung führt man Indikatorvariablen für jede Kategorie als zusätzliche erklärende Variablen ein. Damit die Matrix X vollen Rang hat, muss man dann entweder die erste Spalte $x_{ij} \equiv 1$ oder die Indikatorvariable für die erste Kategorie weglassen. Eine sinnvolle Nullhypothese bei einer solchen kategoriellen erklärenden Variablen lautet, dass die Koeffizienten für *alle* Indikatoren null sind, was man mit einem F-Test testet. Im Beispiel der Sprengungen ist das Ergebnis in Tabelle 1.2 zu sehen. Die dritte Zeile vergleicht das umfassende Modell mit dem Modell ohne den Faktor "St" als erklärende Variable. Sie zeigt, dass der Einfluss der Stelle extrem signifikant ist.

	Df	Sum of Sq	RSS	F Value	Pr(F)
log10(dist)	1	2.79	5.07	108	0
log10(ladung)	1	0.59	2.86	23	7.62e-06
St	5	2.10	4.38	16	0

Tabelle 1.2: Tests für die Effekte der einzelnen Terme im Beispiel der Sprengungen

Die Aufteilung

$$\left\|\mathbf{y}-\widehat{\mathbf{y}}^{(0)}\right\|^2 = \|\mathbf{y}-\widehat{\mathbf{y}}\|^2 + \left\|\widehat{\mathbf{y}}-\widehat{\mathbf{y}}^{(0)}\right\|^2$$

wird auch als Varianzanalysen-Zerlegung bezeichnet,und die Tabelle 1.2 heisst daher Varianzanalysetabelle, oder auf englisch ANOVA (Analyis of Variance)-Tabelle.

1.5.4 Bestimmtheitsmass

Ein bedeutender Spezialfall der Resultate im vorhergehenden Abschnitt ist der folgende: Man prüfe, ob überhaupt Abhängigkeit von den unabhängigen **x**-Variablen vorliegt.

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 1 & \dots & \dots & x_{np} \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \quad (p-1)\times p$$

$$\text{Unter Nullhypothese } H_0: \qquad \widehat{\boldsymbol{\theta}}_{(0)} = \begin{pmatrix} \overline{y} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \widehat{\mathbf{y}}^{(0)} = \begin{pmatrix} \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} =: \overline{\mathbf{y}} \quad (n\times 1)$$

$$\begin{aligned} SSE_0 &= \|\mathbf{y}-\widehat{\mathbf{y}}^{(0)}\|^2 = \|\mathbf{y}-\overline{\mathbf{y}}\|^2 \\ &= \|\mathbf{y}-\widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}}-\overline{\mathbf{y}}\|^2 \end{aligned}$$

Varianzanalyse (ANOVA)–Tabelle:

	Quadratsumme	Freiheitsgrade	Quadratmittel	$\mathbf{E}$ [Quadratmittel]
Regression	$\|\widehat{\mathbf{y}}-\overline{\mathbf{y}}\|^2$	$p-1$	$\|\widehat{\mathbf{y}}-\overline{\mathbf{y}}\|^2/(p-1)$	$\sigma^2 + \frac{\|\mathbf{E}[\mathbf{y}]-\mathbf{E}[\overline{\mathbf{y}}]\|^2}{p-1}$
Fehler	$\|\mathbf{y}-\widehat{\mathbf{y}}\|^2$	$n-p$	$\|\mathbf{y}-\widehat{\mathbf{y}}\|^2/(n-p)$	σ^2
Total um Gesamtmittel	$\|\mathbf{y}-\overline{\mathbf{y}}\|^2$	$n-1$	–	–

Man prüft die Signifikanz der Abhängigkeit von den erklärenden Variablen mit der Teststatistik

$$F = \frac{\|\widehat{\mathbf{y}}-\overline{\mathbf{y}}\|^2/(p-1)}{\|\mathbf{y}-\widehat{\mathbf{y}}\|^2/(n-p)}.$$

Diese ist unter H_0 $F_{p-1,n-p}$-verteilt. In der Vorlesung Varianzanalyse betrachtet man analoge, kompliziertere Zerlegungen der Quadratsumme $\|\mathbf{y}-\overline{\mathbf{y}}\|^2$ in speziellen linearen Modellen, zu denen verschiedene F-Tests gehören.

Man kann auch die Verteilung der Teststatistik F unter der Alternative $H\cap(\neg H_0)$ und damit die Macht des F-Tests studieren. Man erhält die sogenannte nichtzentrale F-Verteilung $F_{p-1,n-p,\delta^2}$ mit **Nichtzentralitätsparameter** $\delta^2 = \|\mathbf{E}[\mathbf{y}] - \mathbf{E}[\overline{\mathbf{y}}]\|^2$ erhält (die Definition des Nichtzentralitätsparameters ist aber nicht einheitlich in der Literatur). Genaueres findet man in der Literatur.

Eine wichtige Grösse ist der Quotient

$$R^2 := \frac{\|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2}{\|\mathbf{y} - \overline{\mathbf{y}}\|^2}.$$

Er heisst das **Bestimmtheitsmass (coefficient of determination)**, oder der Anteil der durch das Modell erklärten Varianz. Er misst die **Güte** der Anpassung des Modells mit den erklärenden Variablen $\mathbf{x}^{(j)}$. Es ist nicht schwierig zu sehen, dass R^2 auch der maximale quadrierte Korrelationskoeffizient von $\mathbf{y}$ mit einer beliebigen Linearkombination der Spalten $\mathbf{x}^{(j)}$ ist. Das Bestimmtheitsmass ist also auch das Quadrat des **multiplen Korrelationskoeffizienten** zwischen y und den $\mathbf{x}^{(j)}$. Die Linearkombination, welche maximale Korrelation mit y hat, ist gerade die Kleinste Quadrate Vorhersage $\widehat{\mathbf{y}}$.

Bemerkung: R^2 und F sind **zunächst** die **wichtigsten** Zahlen eines Computer-Outputs.

1.6 Einfache lineare Regression

1.6.1 Resultate im Spezialfall der einfachen linearen Regression

Die Kleinste Quadrate Schätzer haben wir schon früher hergeleitet. Wir geben hier noch die expliziten Formeln für die wichtigsten Teststatistiken bzw. Vertrauensintervalle an:

Test für die Nullhypothese $\beta = \beta_0$ zum Niveau γ: Man verwirft, falls

$$\frac{|\widehat{\beta} - \beta_0|}{\widehat{\sigma}/\sqrt{SS_X}} > t_{n-2;1-\gamma/2},$$

wobei

$$SS_X = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Das Vertrauensintervall für β lautet entsprechend

$$\widehat{\beta} \pm t_{n-2;1-\gamma/2} \cdot \frac{\widehat{\sigma}}{\sqrt{SS_X}}.$$

Das Vertrauensintervall für den Erwartungswert einer neuen Beobachtung an einer Stelle x_0 (d.h. den Wert der Regressionsgerade an der Stelle x_0) ist:

$$\widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2;1-\gamma/2} \cdot \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_X}}.$$

Der Vertrauensbereich für gesamte Regressionsgerade (simultan für alle x) ist

$$\widehat{\alpha} + \widehat{\beta} x \pm \sqrt{2F_{2,n-2;1-\gamma}} \cdot \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{SS_X}}.$$

Der simultane Vertrauensbereich ist natürlich breiter als der individuelle. Schliesslich lautet der Prognosebereich für eine neue Beobachtung an der Stelle x_0:

$$\widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2;1-\gamma/2} \cdot \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_X}}.$$

Der Prognosebereich ist wieder breiter als der Vertrauensbereich. Die Begrenzungen aller drei Bereiche sind jeweils Hyperbeln.

1.6.2 Regression und Korrelation

Früher sprach man mehr von **Korrelation** als von Regression.

Es seien Y **und** X Zufallsvariablen, das heisst die Daten $x_1, \ldots, x_n$ werden nun ebenfalls nicht mehr als fest vorgegeben angenommen.

Definition 1.6.1. ***Die Korrelation ("Pearson'sche Produktmomentenkorrelation") ist definiert als:***

$$\rho = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}} \quad (\textit{falls } \mathrm{Var}(X) \neq 0,\ \mathrm{Var}(Y) \neq 0)$$

Eigenschaften der Korrelation:

(i) $-1 \leq \rho \leq +1$ (Schwarz'sche Ungleichung)

(ii) $|\rho| = 1 \Leftrightarrow$ Die gemeinsame Verteilung der X und Y ist konzentriert auf einer Geraden (und das Vorzeichen von ρ ist gleich dem Vorzeichen der Steigung der Geraden).

(iii) Wenn $\rho = 0$, dann heissen X und Y unkorreliert.

(iv) ρ wird geschätzt durch:

$$r = \widehat{\rho} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \cdot \sum_{i=1}^n (y_i - \overline{y})^2}}$$

und für diese Schätzung $\widehat{\rho}$ gilt:

- $-1 \leq \widehat{\rho} \leq 1$
- $|\widehat{\rho}| = 1 \Leftrightarrow$ alle Punkte liegen auf einer Geraden
- sign $(\widehat{\rho})$ = sign $(\widehat{\beta})$

Typische Streudiagramme mit verschiedenen Korrelationskoeffizienten sind in Abbildung 1.10 gezeigt.

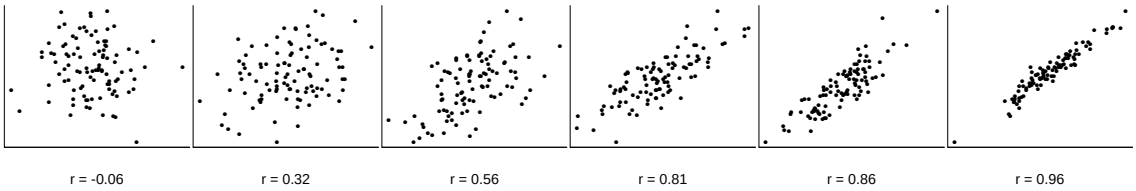

Abbildung 1.10: Streudiagramme mit verschiedenen Korrelationskoeffizienten.

Die z-Transformation ("varianzstabilisierende Transformation für den Korrelationskoeffizienten") (Fisher). Es seien (X,Y) gemeinsam normalverteilt. Definiere:

$$z := \tanh^{-1}(\widehat{\rho}) = \frac{1}{2} \log\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right)$$

dann gilt für beliebiges ρ in sehr guter Näherung (für ca. $n > 10$)

$$z \sim \mathcal{N}\left(\tanh^{-1}(\rho), \frac{1}{n-3}\right).$$

Graphische Interpretation der z-Transformation:

- Liegt das wahre ρ in der Nähe von 0, dann ist die Varianz von $\widehat{\rho}$ **gross**,
- Liegt das wahre ρ in der Nähe von ± 1, dann ist die Varianz von $\widehat{\rho}$ **klein**.

Die z-Transformation ändert die Skala nun so, dass die Varianz konstant wird (das heisst man "staucht in der Mitte" und "streckt an den Rändern").

Um $\rho = 0$ gegen $\rho \neq 0$ zu testen kann man aus 3 Tests auswählen:

1. Tabelle oder Diagramm (siehe Abbildung 1.11)
2. $t-$ bzw. F-Test für $\beta = 0$
3. $\tanh^{-1}$-Transformation.

Mit dem ersten und dritten Test kann man auch auf einen beliebigen festen Wert von ρ testen (und so Vertrauensintervalle konstruieren).

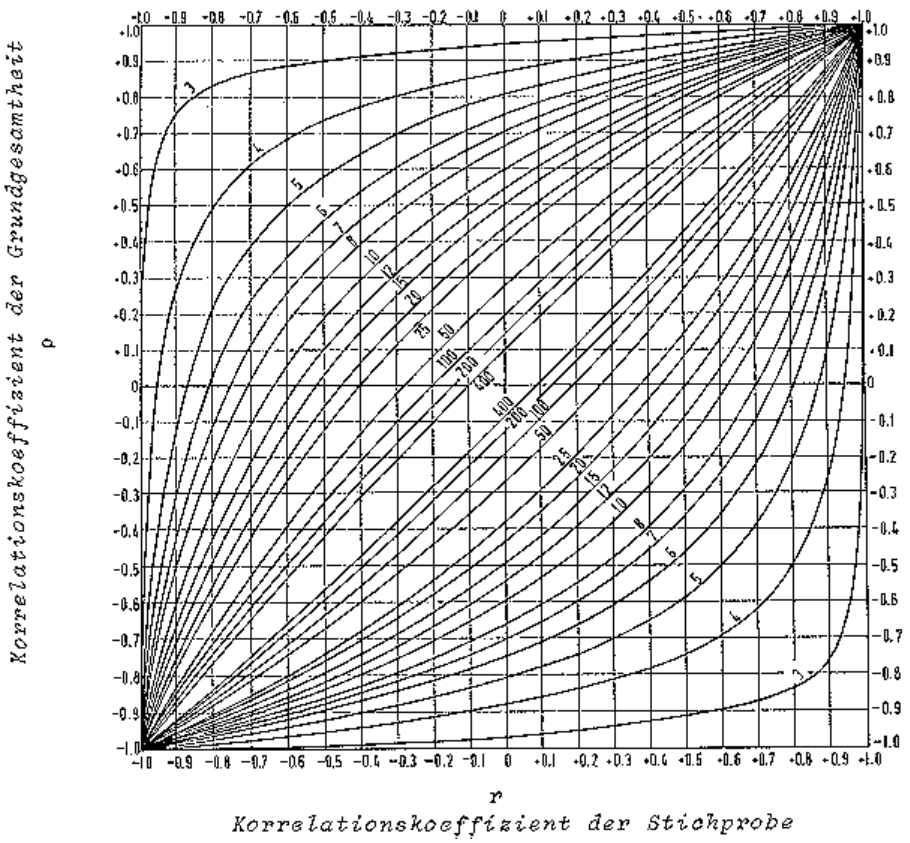

Abbildung 1.11: Vertrauensgrenzen des Korrelationskoeffizienten: 95%-Vertrauensbereich für ρ: Die Zahlen an den Kurven bezeichnen den Stichprobenumfang (aus F.N. DAVID: Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples, The Biometrika Office, London 1938)

Rangkorrelation: Da die Pearson'sche Korrelation nicht robust gegenüber Ausreissern ist (siehe Abbildung 1.12), benutzt man oft eine **Rangkorrelation**. Es gibt zwei Varianten, diejenige von Spearman und diejenige von Kendall:

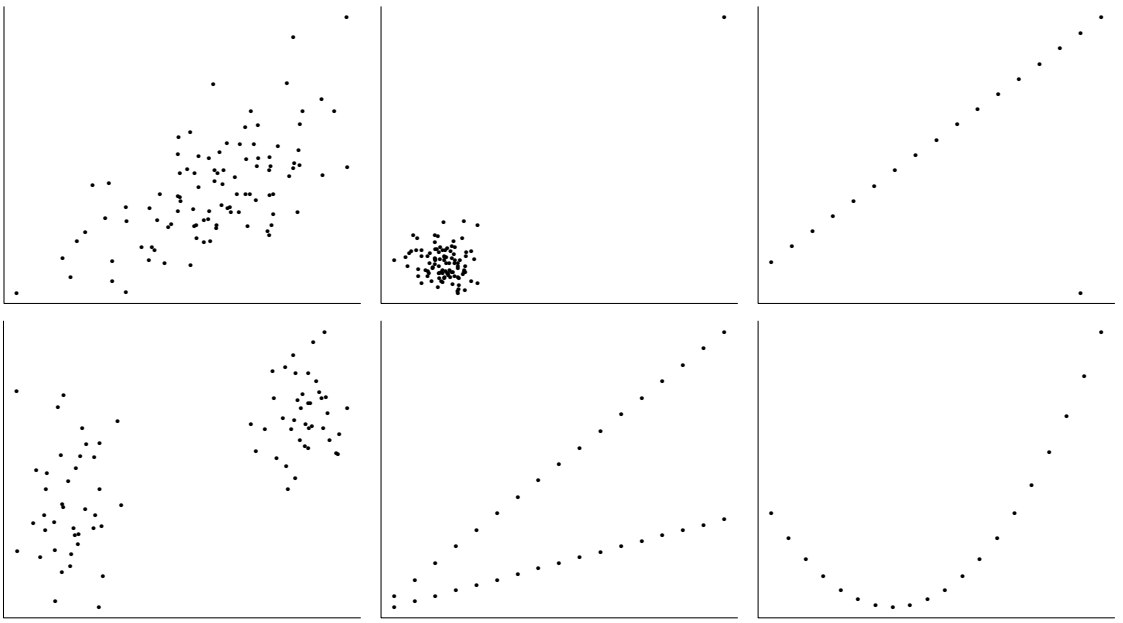

Abbildung 1.12: Verschiedenartige Punktwolken mit Korrelation r = 0.7

Die "Spearman'sche Rangkorrelation" ist einfach die Pearson'sche Korrelation der **Ränge** der X_i mit den **Rängen** der Y_i. Da die Summe der Ränge (oder deren Quadrate) einen festen Wert hat, lassen sich die Formeln vereinfachen. Man erhält:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad D_i := Rg(X_i) - Rg(Y_i)$$

Die Kendall'sche Rangkorrelation ist definiert als

$$r_K = 2 \cdot \frac{T_k - T_d}{n(n-1)}$$

wobei:
$$\begin{array}{llll} T_k & = \#\ \textbf{Konkordanzen} & = \#\ \text{Paare mit} & (x_i - x_j)(y_i - y_j) > 0 \\ T_d & = \#\ \textbf{Diskordanzen} & = \#\ \text{Paare mit} & (x_i - x_j)(y_i - y_j) < 0 \end{array}$$

Ergänzung: Partielle Korrelationen

Gegeben seien 3 Zufallsvariablen X, Y und Z. Dann ist die partielle Korrelation zwischen X und Y unter Festhalten von Z definiert als:

$$\rho_{XY.Z} := \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}} \quad \text{oder geschätzt:} \quad r_{XY.Z} := \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

Dies misst Stärke und Richtung des linearen Zusammenhangs von X und Y nach Ausschalten des linearen Zusammenhangs von X bzw. Y mit Z.

1.6.3 Vertauschung von X und Y, Regression zum Mittel

Wenn sowohl X als auch Y als zufällig angesehen werden, können wir die Kleinste Quadrate Gerade wie folgt schreiben:

$$y - \overline{y} = \widehat{\rho}\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X}(x - \overline{x}).$$

Da $\widehat{\rho}$ stets absolut kleiner ist als 1, ist die Prognose von Y stets näher beim Mittel als der zugehörige X-Wert, wenn wir bei beiden Grössen die Abweichungen in Standardeinheiten messen. Wenn z.B. $\widehat{\rho}$ positiv ist und wir einen X-Wert von z.B. einer Standardabweichung über dem Mittel betrachten, dann ist die Prognose des zugehörigen Y-Wertes **weniger** als eine Standardabweichung über dem Mittel. Das heisst, man prognostiziert stets eine Rückkehr zum Mittel, was zum Namen "Regression" geführt hat.

Dieses Phänomen wird immer wieder von neuem entdeckt und oft ausführlich in einem kulturpessimistischen Sinne interpretiert. Wie unsere Formeln zeigen ist dies jedoch ein ganz allgemeines Phänomen, das stets auftritt und keiner besonderen Interpretation bedarf. Es ist viel mehr eine Eigenschaft von Prognosen: Weil mehr Beobachtungen in der Mitte liegen, tendiert man bei Prognosen in die Richtung der Mitte.

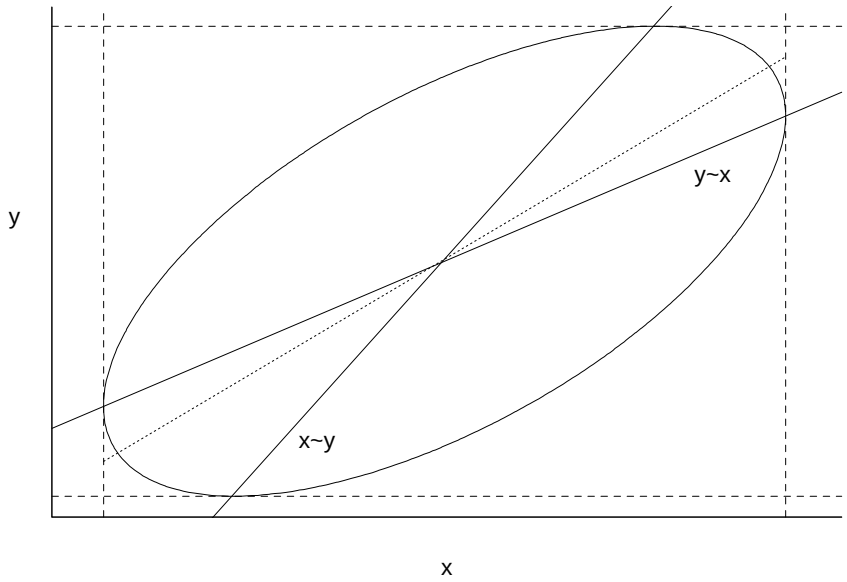

Abbildung 1.13: Regressionsgeraden "y gegen x" und "x gegen y"

Dass dieser Rückschritt zum Mittel nichts aussergewöhnliches ist, wird noch klarer, wenn wir die Rolle von X und Y vertauschen. Die Regressionsgerade von X auf Y hat aus Symmetriegünden die Form

$$x - \overline{x} = \widehat{\rho}\frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y}(y - \overline{y}).$$

Das heisst, wenn wir rückwärts schauen, und fragen, wie der X-Wert war, wenn der zugehörige Y-Wert eine Standardabweichung über dem Mittel ist, dann ergibt sich die Antwort "weniger als eine Standardabweichung". Man könnte dann versucht sein, dies als ein Zeichen von Fortschritt zu interpretieren !

Wenn wir die beiden Regressionsgerade im gleichen Plot einzeichnen, haben wir die beiden Steigungen

$$\widehat{\rho}\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X}, \text{bzw. } \frac{1}{\widehat{\rho}}\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X}.$$

Die beiden Regressionsgeraden sind also offensichtlich nicht gleich, es ergibt sich die sogenannte "Regressionsschere" (siehe Abbildung 1.13).

Wenn (X, Y) zweidimensional normalverteilt ist, dann sind die Höhenkurven der gemeinsamen Dichte Ellipsen. Die beiden Regressionsgeraden schneiden eine Höhenkurve dort,

wo die Tangenten vertikal bzw. horizontal sind, denn dort werden die bedingten Dichten von X gegeben $Y = y$, bzw. Y gegeben $X = x$ maximal.

1.7 Residuenanalyse, Modellüberprüfung, Vorgehen falls Voraussetzungen verletzt

Residuenanalyse ist die graphische (und zum Teil auch numerische) Untersuchung der Residuen, d.h. der geschätzten Fehler

$$r_i := \widehat{\varepsilon}_i = y_i - \widehat{y}_i,$$

zur nachträglichen Überprüfung der Modellannahmen und zur Entwicklung eines besseren Modells.

1.7.1 Normal plot

Verteilungsannahmen kann man allgemein mit dem Quantil-Quantil-Diagramm (Q-Q plot = quantile-quantile-plot) überprüfen. Wenn man speziell die Normalverteilung prüfen will, spricht man auch vom normal plot.

Wir definieren den normal plot zunächst für i.i.d. Zufallsvariablen $X_1, \ldots, X_n$. Die "empirische kumulative Verteilungsfunktion" ist definiert als:

$$u = F_n(x) = \frac{1}{n}\text{Anzahl}\{X_i \leq x\}.$$

Dies ist eine Treppenfunktion, die sich für grosses n der wahren Verteilungsfunktion annähert (Lemma von Glivenko-Cantelli). Insbesondere gilt

$$F_n(x) \longrightarrow \Phi(\frac{x-\mu}{\sigma}),$$

falls die X_i normalverteilt sind. Wenn wir daher

$$z := \Phi^{-1}(F_n(x))$$

setzen, dann ist für "genügend grosses" n

$$z \approx \frac{x-\mu}{\sigma}.$$

Beim normal plot tragen wir (an ausgewählten Punkten) x gegen z auf. Wenn die X_i tatsächlich normalverteilt sind, dann ergibt der normal plot ungefähr eine Gerade, und die Parameter μ und σ sind gerade gleich dem Achsenabschnitt und der Steigung. Allerdings ergeben sich durch die zufälligen Schwankungen der Daten Abweichungen. Eine Idee, wie gross diese etwa sein können, erhält man mit Hilfe von Simulationen, vgl.Abbildung 1.14.

Wenn die Annahme der Normalverteilung nicht zutrifft, zeigen sich im normal plot systematische Abweichungen von einer Geraden. Typische Fälle sieht man in der Abbildung 1.15. Die Interpretation ist jedoch nicht immer eindeutig, da die Übergänge fliessend sind. Die Situation in der Abbildung 1.16 kann man als eine Mischung aus zwei Gruppen oder als eine kurzschwänzige Verteilung interpretieren.

Es gibt auch einen formalen Test auf die Normalverteilung, der auf dem Normalplot beruht, und zwar den "Shapiro-Wilks-Test". Er misst im Wesentlichen die Korrelation der Punktwolke im normal plot.

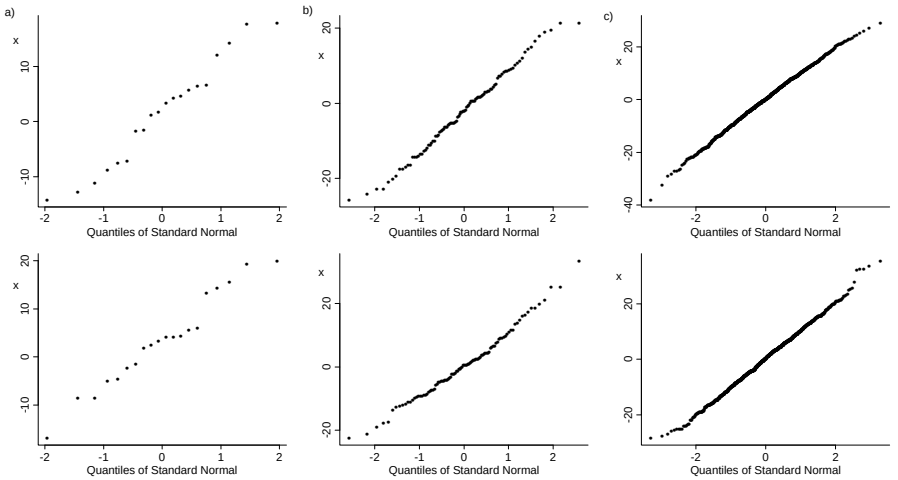

Abbildung 1.14: QQ-plots für normalverteilte Zufallsvariable X mit a) 20, b) 100 und c) 1000 Realisierungen.

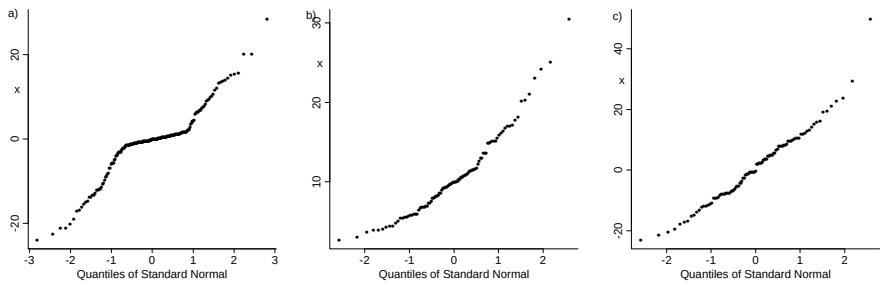

Abbildung 1.15: a) Langschwänzige Verteilung b) Schiefe Verteilung c) Ausreisser

Details zur Darstellung

- Wenn n klein ist (ca. $n \leq 100$):
  Trage alle Beobachtungen einzeln auf der Ordinate auf. Durch die Anordnung ergeben sich gerade die "Ordnungsgrössen" (order statistics, geordnete Beobachtungen) $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. Nach Definition ist $F_n(X_{(i)}) = i/n$, aber um die Effekte der Sprünge von F_n zu mildern, trägt man auf der Abszisse $\Phi^{-1}(\frac{i-1/2}{n})$ auf (oder $\Phi^{-1}(\frac{i}{n+1})$ oder $\Phi^{-1}(\frac{i-(3/8)}{n+(1/4)})$ oder $\Phi^{-1}(\frac{i-(1/3)}{n+(1/3)})$.

- Wenn n gross ist (ca. $n \geq 100$):
  Wähle irgendwelche, z.B. äquidistante x-Werte aus dem Wertebereich der Daten auf der Ordinate.

- Häufig beschriftet man die Abszisse auch mit $u = \Phi(z)$. Wegen der nichtlinearen Skala bleibt im Diagramm (u, x) eine Gerade. Dann kann man die Werte $F_n^{-1}(u)$ auf der Ordinate ohne Umrechnung leicht eintragen. Dies war vor allem in der Vor-Computerzeit wesentlich, wo man noch spezielles "Wahrscheinlichkeitspapier" verwendete.

- Häufig werden die Achsen auch vertauscht, d.h. man zeichnet z gegen x.

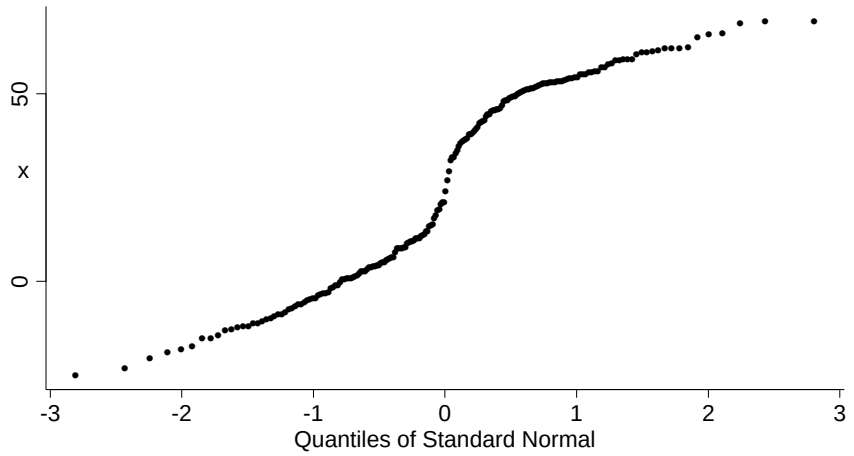

Abbildung 1.16: Zwei Gruppen oder kurzschwänzige Verteilung?

In der Regression haben wir die Fehler ε_i nicht beobachtet. Wir können aber den normal plot mit den Residuen $\widehat{\varepsilon}_i$, bzw. den standardisierten Residuen $\widehat{\varepsilon}_i/\sqrt{1-P_{ii}}$ verwenden. Zur Erinnerung: die Residuen $\widehat{\varepsilon}_i$ haben weder konstante Varianz noch sind sie unkorreliert. Die Standardisierung bewirkt, dass die Varianz wieder konstant wird. Meist ist aber dieser Effekt vernachlässigbar.

Zusammenfassend: Der **normal plot** prüft die Normalverteilung der Residuen gegenüber Schiefe, Langschwänzigkeit (oder Kurzschwänzigkeit) der Verteilung, Ausreissern und anderen Besonderheiten.

1.7.2 Tukey-Anscombe Plot

Der Tukey-Anscombe Plot ist eine Darstellung der Residuen r_i gegen die geschätzten Werte $\widehat{y}_i$. Es gilt stets $\sum r_i\widehat{y}_i = 0$, d.h. die Stichprobenkorrelation im Tukey-Anscombe Plot ist immer gleich null. Wenn in diesem Plot eine nichtlineare Struktur besteht, weist das auf eine Verletzung der Modellannahmen hin. Würde man die Residuen gegen die Werte y_i plotten, würde die Korrelation die Interpretation erschweren.

Bei einfacher Regression ist dies (im Wesentlichen) äquivalent zur Darstellung der r_i gegen die x_i (ganz anders in der multiplen Regression, denn dort ist der Plot r_i gegen $\widehat{y}_i$ informativer als die komponentenweisen Plots r_i gegen x_{ij}). Ebenfalls ist bei einfacher Regression der Plot ähnlich zur ursprünglichen Darstellung y_i gegen x_i. Man kann aber Abweichungen von der Horizontalen besser beurteilen als Abweichungen von einer schiefen Gerade.

Wie der Tukey-Anscombe Plot im Idealfall aussieht, ist in der Abbildung 1.17 dargestellt. Eine häufige Abweichung von der Annahme konstanter Fehlervarianz ist, dass die Varianz mit der Zielvariablen zunimmt. Wie sich das äussert im Tukey-Anscombe Plot, sieht man in Abb. 1.18 a)–c). Wenn sich im Tukey-Anscombe Plot eine Struktur in Form eines "Trends" zeigt, ist das ein Hinweis darauf, dass die Regressionsfunktion nicht die angenommene Form hat (d.h. der Erwartungswert der Fehler ist nicht null.) Abb. 1.18d ist ein typisches Beispiel dafür, dass wohl in der Modellannahme ein quadratischer Term vergessen wurde.

Wenn man im Tukey-Anscombe Plot einen systematischen Zusammenhang der Fehlervarianz mit $\widehat{y}_i$ oder den $\mathbf{x}$–Variablen entdeckt, dann soll man die Zielvariablen transformieren

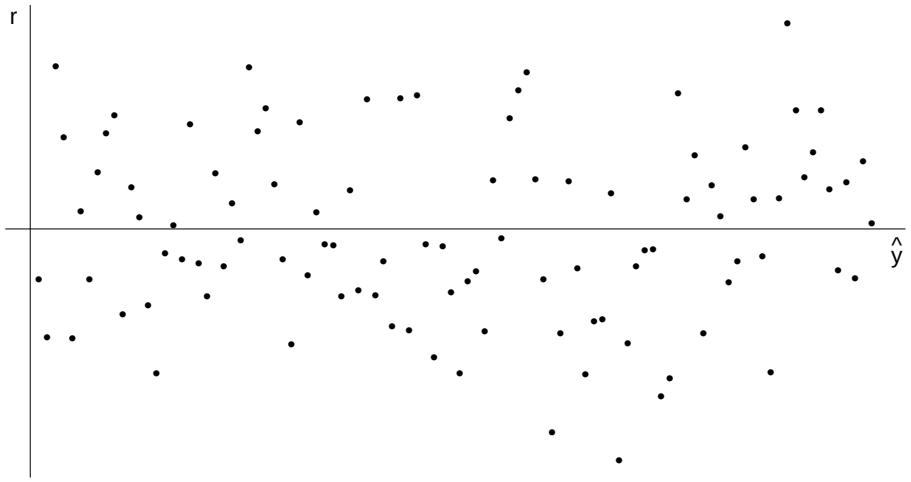

Abbildung 1.17: Tukey-Anscombe Plot in einem Fall, wo die Voraussetzungen des Modells erfüllt sind.

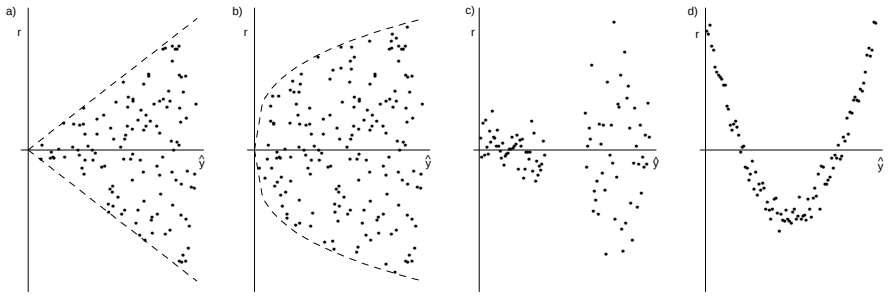

Abbildung 1.18: a) Lineare Zunahme der Standardabweichung, b) Nicht-lineare Zunahme der Standardabweichung, c) 2 Gruppen mit unterschiedlicher Varianz, d) Fehlender quadratischer Term im Modell.

oder eine "gewichtete Regression" (siehe Abschnitt 1.7.5) durchführen. Wenn die Streuung der Fehler linear mit den angepassten Werten zunimmt, stabilisiert die Logarithmustransformation die Varianz. Wenn die Streuung der Fehler ungefähr mit der Wurzel der angepassten Werte zunimmt, stabilisiert die Wurzeltransformation die Varianz. (Dies folgt aus einer Taylor-Entwicklung)

1.7.3 Zeitreihenplot, Durbin-Watson Test

Abhängigkeiten der Fehler führen dazu, dass die Niveaus der Tests und Vertrauensintervalle nicht mehr korrekt sind. Dies ist einfach einzusehen: Wenn $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \Sigma)$, dann folgt mit einfacher Rechnung, dass

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta}, (X^T X)^{-1}(X^T \Sigma X)(X^T X)^{-1}).$$

Wie gross die Effekte von Korrelationen der Fehler sind, hängt sowohl von den Versuchsbedingungen X als auch von der Gestalt der Kovarianzmatrix Σ ab. In vielen Fällen ist der Effekt jedoch wesentlich.

Ohne Annahmen über die Struktur der Abhängigkeiten kann man nicht viel machen: Selbst wenn man die Fehler ε_i kennen würde, kann man ohne zusätzliche Annahmen aus n Daten

nicht eine beliebige Kovarianzmatrix mit $n(n+1)/2$ Elementen schätzen.

Wenn die Beobachtungen in zeitlicher Reihenfolge gemacht werden, ist es oft so, dass die Kovarianz der Fehler eine (üblicherweise monoton fallende) Funktion der Zeit zwischen den Beobachtungen ist. Diesen Typ von Abhängigkeit kann man entdecken, indem man die Residuen r_i gegen die Beobachtungszeiten t_i der Beobachtung aufträgt. Falls diese unbekannt sind, kann man auch die serielle Nummer $k(i)$ der Beobachtung nehmen.

Wenn im Zeitreihenplot die Punkte zufällig um die Abszisse variieren, so ist das Bild in Ordnung. Falls dagegen benachbarte r_i einander besonders ähnlich sind, so deutet dies auf eine serielle Korrelation der Fehler hin. Manchmal beobachtet man auch einen Sprung im Niveau der Residuen. Dann hat sich das Modell offenbar zu einem bestimmten Zeipunkt schlagartig geändert.

Man kann auch die Unabhängigkeit gegen die Alternative einer seriellen Korrelation testen. Zwei mögliche Tests sind

(i) **Der Iterationstest** (run test) Der run-Test zählt die Anzahl der Teilfolgen ("runs") der Residuen mit jeweils gleichem Vorzeichen. Bei Unabhängigkeit darf man einerseits nicht zu wenige, andererseits aber auch nicht zu viele runs haben.

(ii) **Der Durbin-Watson-Test** Dieser Test verwendet die Teststatistik

$$T = \frac{\sum_{i=1}^{n-1}(r_{i+1} - r_i)^2}{\sum_{i=1}^{n} r_i^2}$$

Durch Ausmultiplizieren folgt

$$T \approx 2\left(1 - \frac{\sum_{i=1}^{n-1} r_i r_{i+1}}{\sum_{i=1}^{n} r_i^2}\right),$$

(Die Unterschiede sind nur Randeffekte). Der Quotient ist eine Schätzung der Korrelation von ε_i und ε_{i+1} (unter der Annahme, dass alle ε_i die gleiche Varianz haben). Wenn die ε_i unabhängig sind, ist T also ungefähr 2, und kleine Werte von T weisen auf positive Abhängigkeit hin.

Die Bestimmung der kritischen Werte für diesen Test wird dadurch kompliziert, dass die Verteilung der r_i und damit auch die Verteilung von T vom Versuchsplan (design), d.h. von den gewählten Stellen $\mathbf{x}_i$ abhängig ist. Der Test betrachtet nur die Extrema über alle designs, was dazu führt, dass man zwei Tabellenwerte (siehe z.B. Sen und Srivastava (1990), p. 326). hat. Ist T kleiner als der untere Tabellenwert, so wird die Nullhypothese "Unabhängigkeit" verworfen; ist hingegen T grösser als der obere Tabellenwert, dann ist man im Annahmebereich. Dazwischen hängt es von den $\mathbf{x}_i$ ab, man hat also gewissermassen "Stimmenthaltung".

Der Nachteil des Durbin-Watson-Tests ist, dass er nur die Korrelation zwischen unmittelbar aufeinanderfolgenden Beobachtungen anschaut.

Wie oben gesagt, haben die Versuchsbedingungen einen Einfluss darauf, wie stark sich Korrelationen der Fehler auswirken. Wenn man selber auswählen kann, in welcher Reihenfolge man die Beobachtungen macht, dann ist es im Fall einer erklärenden Variablen häufig verlockend, diese gerade nach aufsteigenden x-Werten zu machen, weil dies oft am wenigsten Aufwand erfordert. Diese Versuchsanordnung ist aber sehr ungünstig, weil sich

dann positive Korrelationen der Fehler besonders stark auf die Schätzung der Steigung auswirken. Etwas anders gesagt, eventuelle zeitliche Trends der Fehler werden dabei mit dem Effekt der unabhängigen Variablen vermischt.

Viel besser wählt man deshalb die x-Werte möglichst "orthogonal zur Zeit", d.h. , die x_i und die t_i (oder $k(i)$) sollen unkorreliert sein. Dann heben sich Trends bei der Schätzung der Steigung heraus (bei linearen Trends exakt, sonst genähert), und entsprechend haben positive Korrelationen einen kleinen Effekt auf die Varianz der geschätzten Steigung. Dies gilt jedoch nicht beim Achsenabschnitt: Die Varianz des arithmetischen Mittels wird gross, wenn die Korrelation der Beobachtungen gross ist. Genäherte Orthogonalität zur Zeit erhält man am einfachsten, indem man die Reihenfolge der x-Werte zufällig wählt (sogenanntes Randomisieren).

1.7.4 Interior Analysis

"Interior analysis" heisst lokale Schätzung der Fehlervarianz aus Replikaten oder "Fast-Replikaten". Unter "Replikaten" versteht man wiederholte Messungen an der gleichen Stelle $\mathbf{x}_i$, und unter "Fast-Replikaten" Messungen, die beinahe an der gleiche Stelle $\mathbf{x}_i$ liegen, vgl. die Abbildung 1.19. Die "Interior analysis" prüft, ob das Modell einen "lack of fit", also einen systematischen Fehler in der angepassten Modellklasse, aufweist. Mit andern Worten, man überprüft damit die Annahme $\mathbf{E}[\varepsilon_i] = 0$.

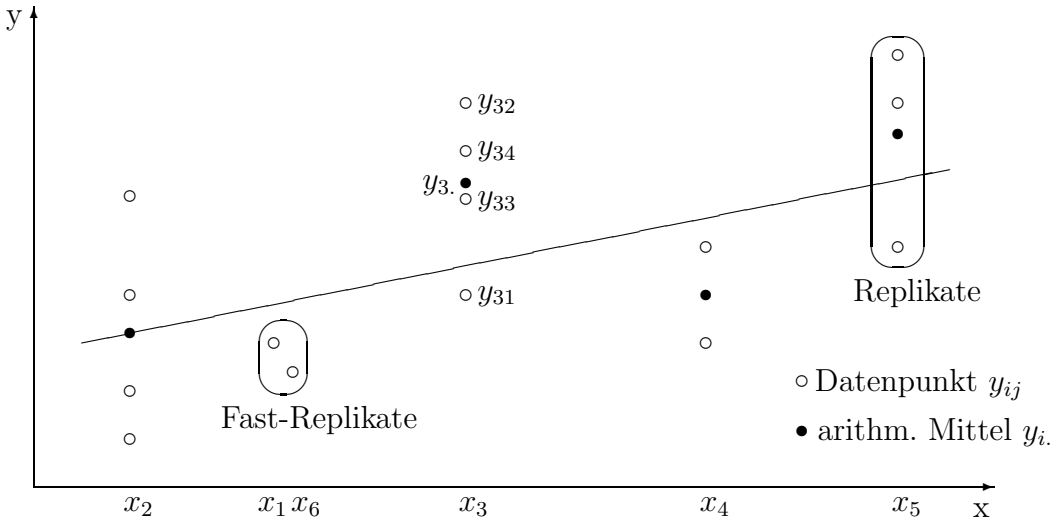

Abbildung 1.19: Replikate

Das Modell mit n_i Replikaten an der Stelle $\mathbf{x}_i$ lautet:

$$Y_{ij} = \mathbf{x}_i^T \theta + \varepsilon_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, n_i)$$

wobei ε_{ij} *i.i.d.* $\sim \mathcal{N}(0, \sigma^2)$. Die Anzahl Beobachtungen ist dann $n = \sum_{i=1}^{k} n_i$. Dieses Modell können wir mit dem grösseren Modell vergleichen, bei dem die Beziehung zwischen Y und X durch eine beliebige Funktion f gegeben ist. Wir führen die Werte $\mathbf{E}[Y_{ij}] = f(\mathbf{x}_i) = \mu_i$ als unbekannte Parameter ein:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, n_i),$$

Die Kleinste Quadrate Schätzung von μ_i ist dann einfach das arithmetische Mittel der Beobachtungen an der Stelle $\mathbf{x}_i$:

$$\widehat{\mu}_i = y_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i.$$

Wir haben also zwei ineinander geschachtelte Modelle und können den üblichen F-Test durchführen. Die orthogonale Zerlegung für die ANOVA-Tabelle lautet:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y}_i)^2 = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})^2}_{(1)} + \underbrace{\sum_{i=1}^{k} n_i(y_{i.} - \widehat{y}_i)^2}_{(2)}$$

- (1) misst den zufälligen Fehler. Die Anzahl Freiheitsgrade ist $\sum_{i=1}^{k}(n_i - 1) = n - k$.
- (2) misst den zufälligen Fehler und den "lack of fit". Die Anzahl Freiheitsgrade ist $k - p$

Bei Fast-Replikaten kann man eine Korrektur vornehmen: Zwei Fast-Replikate können parallel zur Regressionsgerade an die mittlere $\mathbf{x}$-Stelle der beiden verschoben werden, vgl. Abb. 1.20. Dies gibt uns dann zwei Replikate. Wenn p gross ist, ist es jedoch im Allgemeinen schwierig, Fast-Replikate zu finden.

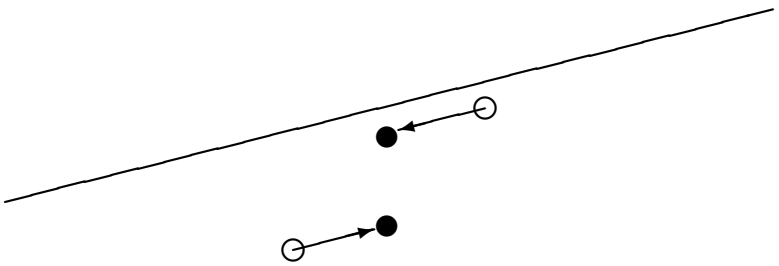

Abbildung 1.20: Korrektur für Fast-Replikate

1.7.5 Verallgemeinerte Kleinste Quadrate, Gewichtete Regression

Modell: $\boxed{\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}}$ wie bisher, **aber jetzt:** $\boxed{\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\ \sigma^2\Sigma)}$.

Wir nehmen an, dass Σ bekannt ist, aber σ^2 unbekannt, d.h. die Kovarianzmatrix der Fehler ist bis auf eine multiplikative Konstante bekannt. Zusätzlich nehmen wir an, Σ sei **positiv definit**. Dann existiert eine reguläre Matrix A so dass $AA^T = \Sigma$, d.h. A ist eine Quadratwurzel von Σ, siehe Anhang.

Rückführung auf unser bekanntes Modell: Bilde

$$\tilde{\mathbf{y}} := A^{-1}\mathbf{y} = A^{-1}(X\theta + \boldsymbol{\varepsilon}) = \underbrace{A^{-1}X}_{\tilde{X}}\boldsymbol{\theta} + \underbrace{A^{-1}\boldsymbol{\varepsilon}}_{\tilde{\boldsymbol{\varepsilon}}} = \tilde{X}\boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}$$

Dann gilt

$$\begin{aligned}
\mathbf{E}\left[\tilde{\varepsilon}\right] &= \mathbf{E}\left[A^{-1}\boldsymbol{\varepsilon}\right] = A^{-1}\mathbf{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0} \\
\mathrm{Cov}[\tilde{\varepsilon}] &= \mathrm{Cov}[A^{-1}\boldsymbol{\varepsilon}] = A^{-1}\,\mathrm{Cov}[\boldsymbol{\varepsilon}](A^{-1})^T \\
&= A^{-1}\sigma^2(AA^T)(A^{-1})^T = \sigma^2 I.
\end{aligned}$$

Das heisst, das "Tilde-Modell", das wir durch die lineare Transformation mit A^{-1} entsteht , erfüllt gerade die Voraussetzungen des bisherigen Modells der multiplen Regression. Dabei ist die Invertierbarkeit der Matrix A der zentrale Punkt (was durch die positive Definitheit von Σ gegeben ist).

Anwendung der bekannten Theorie auf das "Tilde–Modell":

Wir schätzen $\boldsymbol{\theta}$ mit Kleinsten Quadraten im "Tilde-Modell", d.h. wir minimieren

$$\|\tilde{\mathbf{y}} - \tilde{X}\boldsymbol{\theta}\|^2 = (\mathbf{y} - X\boldsymbol{\theta})^T A^{-T} A^{-1} (\mathbf{y} - X\boldsymbol{\theta}) = (\mathbf{y} - X\boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y} - X\boldsymbol{\theta}).$$

Dies entspricht einer Kleinste Quadrate Schätzung für die ursprünglichen Daten $(\mathbf{y}, X)$ mit einem andern Skalarprodukt. Wir erhalten

$$\widehat{\boldsymbol{\theta}} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\mathbf{y}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \mathbf{y},$$

die sogenannte **verallgemeinerte Kleinste Quadrate Schätzung** für $\boldsymbol{\theta}$. Diese hat die Eigenschaft, dass

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta},\ \sigma^2 (X^T \Sigma^{-1} X)^{-1}).$$

Mit den gleichen Überlegung wie früher erhält man Tests, Vertrauensintervalle etc. Aus den Resultaten des Abschnitts 1.9 wird folgen, dass verallgemeinerte Kleinste Quadrate eine kleinere Varianz haben als gewöhnliche Kleinste Quadrate falls $\Sigma \neq I$.

Ein wichtiger Spezialfall ist der, wo Σ **diagonal** ist, d.h. die Fehler sind unkorreliert, aber sie haben unterschiedliche Genauigkeit:

$$\Sigma = \begin{pmatrix} v_1 & & & 0 \\ & v_2 & & \\ & & \ddots & \\ 0 & & & v_n \end{pmatrix} \qquad (v_i > 0 \quad \forall i)$$

Dann führt man Gewichte w_i proportional zu $\frac{1}{v_i}$ ein, d.h. man minimiert $\sum_i w_i r_i^2$. Je genauer eine Beobachtung ist, ein desto grösseres Gewicht erhält sie in der verallgemeinerten Kleinste Quadrate Schätzung.

Die Fälle, wo man die Kovarianz der Fehler bis auf einen konstanten Faktor kennt, sind eher selten. Wenn das nicht der Fall ist, dann verwendet man oft zuerst gewöhnliche Kleinste Quadrate und schätzt eine Kovarianzmatrix $\widehat{\Sigma}$ auf Grund der Residuen. In einem zweiten Schritt verwendet man dann verallgemeinerte Kleinste Quadrate mit dieser geschätzten Kovarianzmatrix $\widehat{\Sigma}$. Vor allem bei zeitlich korrelierten Fehler ist ein solch zweistufiges Vorgehen sehr verbreitet und wird meist als Cochrane-Orcutt Verfahren bezeichnet.

1.8 Modellwahl

Wir nehmen an, dass unsere Beobachtungen erzeugt wurden gemäss dem Modell

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \qquad (i = 1, \ldots, n)$$

mit ε_i i.i.d., $\mathbf{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Als Modell für den Erwartungswert verwenden wir

$$f(\mathbf{x}_i) \approx \sum_{j=1}^{p} \theta_j x_{ij}.$$

Wir nehmen an, dass wir einen Achsenabschnitt im Modell haben, d.h. $x_{i1} = 1$ für alle i.

Fragestellung: Welche Variablen sollen in die Modellgleichung aufgenommen werden ?

Naive Antwort: "Je mehr, desto besser!".

Diese Antwort trifft nicht generell zu. Es kann vorkommen, dass von den p Variablen einige "überflüssig" sind oder nur einen kleinen Beitrag zu einer besseren Erklärung liefern. Je mehr Koeffizienten wir aber schätzen, desto grösser wird auch der zufällige Fehler der geschätzten Parameter und der Modellvorhersage sein. Die Antwort ist also falsch.

Die Variablen können durch theoretische Überlegungen des Fachgebietes (zum Beispiel in der Physik) gegeben sein. Dann treffen wir keine Modellwahl im Sinne der Statistik. Im Unterschied zu früher betrachten wir auch keine Transformationen für y und/oder $\mathbf{x}$, es geht nur noch darum, ob eine (allenfalls schon transformierte) Variable aufgenommen oder weggelassen wird.

Die Suche nach dem "besten Modell" hängt dabei ab von der Problemstellung:

(i) Regressionsmodell als erklärendes Modell

(ii) Regressionsmodell für Vorhersagen

Zunächst besprechen wir schrittweise Verfahren, danach Methoden, bei denen man alle möglichen 2^{p-1} Modelle anschaut und dann auf Grund eines geeigneten Kriteriums das "Beste" auswählt. Die schrittweisen Verfahren sind natürlich weniger aufwändig.

1.8.1 Modellwahl mit "stepwise regression"

(1) **Stepwise regression forward:**

Starte mit dem Modell, das nur die Konstante x_{i1} enthält:

$$y_i = \theta_1 + \varepsilon_i.$$

(Die Schätzung von θ_1 ist dann natürlich das arithmetische Mittel der Beobachtungen). Die Variablen werden nun einzeln (und "schrittweise") ins Modell hineingenommen, und zwar in jedem Schritt diejenige Variable, die den signifikantesten F–Wert im Vergleich zum vorherigen Modell liefert.

Stoppregel: Iteriere, bis die F–Statistik nicht mehr signifikant ist (d.h. man muss ein Signifikanzniveau vorgeben). Da man wiederholte Tests durchführt, ist bei der Interpretation dieses Niveaus jedoch Vorsicht geboten.

(2) **Stepwise regression backward:**

Starte mit dem vollen Modell:

$$y_i = \theta_1 + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \varepsilon_i$$

Die Variablen werden nun einzeln (und "schrittweise") aus dem Modell herausgenommen, und zwar in jedem Schritt diejenige Variable, welche den kleinsten F–Wert beim vergleichenden Test liefert, solange, bis dieses F signifikant ist.

Diskussion der stepwise regression:

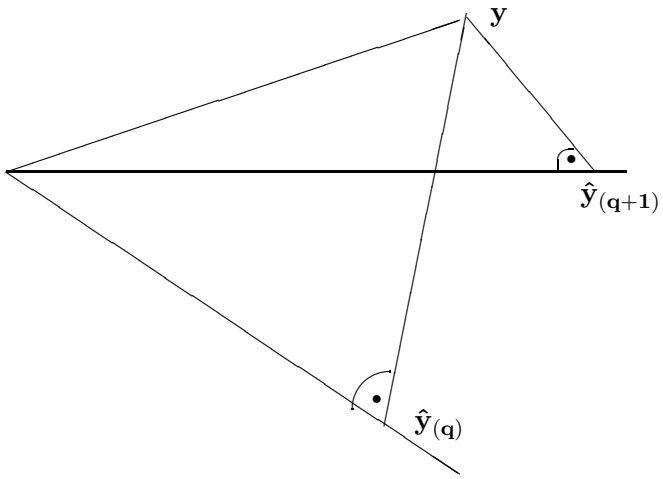

Abbildung 1.21: Bisheriges Modell (q-dimensional) und grösseres Modell (($q+1$)-dimensional, eine Variable mehr).

- Das "backward" ist aufwändiger und unter Umständen numerisch heikler (wenn $p \geq n$ ist, ist es gar nicht durchführbar). Es liefert aber mit grösserer Sicherheit ein gutes Modell.
- Oft werden "forward" und "backward" kombiniert (mit zwei verschiedenen Signifikanzniveaus, da man sonst mit der Wegnahme und Hinzunahme der $\mathbf{x}^{(j)}$ hin– und herpendeln würde).
- Die Stoppregel liefert nicht notwendig ein "bestes" Modell im Sinne der Kriterien, die wir im nächsten Abschnitt besprechen.
- Die Sequenz der hinzu– resp. herausgenommenen Variablen sollte nicht als Ordnung für die Wichtigkeit der Variablen interpretiert werden.
- Das "forward" und das "backward" Verfahren können ganz verschiedene Lösungen ergeben.

Beispiel zum letzten Punkt: Wir wählen drei erklärende Variablen, so dass

- X_1 und X_2 sind schlecht mit Y korreliert, aber Y ist (fast) genau eine Linearkombination von X_1 und X_2.
- X_3 ist stark korreliert mit Y.

Die forward regression wählt zuerst X_3 und stoppt dann (bzw. wählt noch $\{X_1, X_3\}$ oder $\{X_2, X_3\}$), während backward $\{X_1, X_2\}$ wählt und dann stoppt.

1.8.2 Modellwahlkriterien

Die C_p -Statistik von Mallows

C_p ist eine Schätzung für den mittleren quadratischen Vorhersagefehler eines angepassten Modells, gemittelt über die realisierten Versuchsbedingungen $\mathbf{x}_i$ ($i = 1, \ldots, n$). Es berücksichtigt auch den Bias eines nichtpassenden Modells; es benötigt aber eine gute (biasfreie und genügend genaue) Schätzung für σ^2 (z.B. von einem "vollen" Modell mit allen benötigten, aber eventuell zu vielen Variablen; aus Replikaten oder "Fast-Replikaten"; oder

aus Erfahrung). Es liefert eine automatische "Bestrafung" für überflüssige Variablen und kann als (geschätztes) Gesamtmass für die Güte einer Modellgleichung benutzt werden.

Gemäss den Annahmen zu Beginn dieses Abschnittes sind die y_i unabhängig mit $\mathbf{E}[y_i] = f(\mathbf{x}_i) = \mu_i$ und $\mathrm{Var}(y_i) = \sigma^2$. Ein Modell wird beschrieben durch die Teilmenge $M \subset \{1, 2, \ldots, p\}$ der aufgenommenen Variablen. Dabei soll $x_{i,1} \equiv 1$ gelten und jedes M soll 1 enthalten. Die zugehörige X-Matrix bezeichnen wir mit X^M, d.h.

$$X^M = (x_{ij}; 1 \leq i \leq n, j \in M).$$

Wir schätzen dann mit Kleinsten Quadraten den Parameter des Modells M

$$\widehat{\boldsymbol{\theta}}^M = ((X^M)^T X^M)^{-1} (X^M)^T \mathbf{y}$$

und den Vektor der Erwartungswerte $\mu_i = \mathbf{E}[y_i]$ entsprechend durch

$$\widehat{\mathbf{y}}^M = X^M \widehat{\boldsymbol{\theta}}^M.$$

Wir passen also ein möglicherweise falsches lineares Modell M mit $|M|$ Variablen durch Kleinste Quadrate an. Das Modell kann zum Beispiel eine Gerade sein, obwohl die $\mathbf{E}[y_i]$ auf einer Parabel liegen, siehe Abb. 1.22.

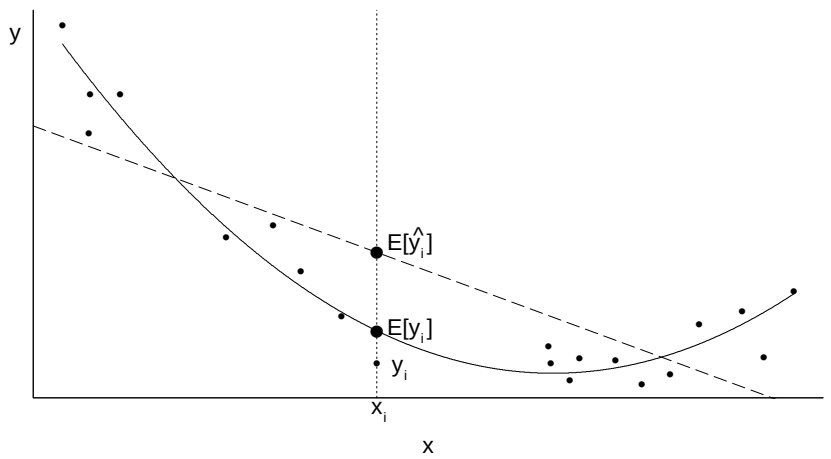

Abbildung 1.22: Theoretisch bestpassende Gerade (gestrichelt) zu in Wirklichkeit quadratischem wahrem Modell (ausgezogen).

Es gilt analog wie im Fall, wo das Modell als richtig angenommen war, dass

$$\mathbf{E}\left[\widehat{\mathbf{y}}^M\right] = X^M((X^M)^T X^M)^{-1}(X^M)^T \boldsymbol{\mu},$$

was nichts anderes ist als die beste Approximation von $\boldsymbol{\mu}$ durch die im Modell aufgenommenen Variablen. Diese Approximation wird natürlich immer besser, je mehr Variablen wir ins Modell aufnehmen (meist sogar echt besser, denn in der Praxis ist der Einfluss einer Variablen zwar oft klein, aber kaum je exakt null). Ferner gilt

$$\mathrm{Cov}(\widehat{\mathbf{y}}^M) = \sigma^2 X^M((X^M)^T X^M)^{-1}(X^M)^T,$$

d.h. die zufälligen Schwankungen sind gleich wie im Fall, wo das Modell richtig ist. Insbesondere ist

$$\sum_{i=1}^{n} \mathrm{Var}(\widehat{y}_i^M) = \sigma^2 tr(X^M((X^M)^T X^M)^{-1}(X^M)^T) = |M|\sigma^2.$$

Je mehr Variablen wir ins Modell aufnehmen, desto grösser wird also die Summe der Varianzen von $\widehat{y}_i^M$.

Als Mass für die Güte eines Modells verwenden wir die Summe der mittleren quadratischen Abweichungen (sum of mean square errors) der $\widehat{y}_i^M$ von den wahren Werten μ_i:

$$SMSE = SMSE(M) = \mathbf{E}\left[||\widehat{\mathbf{y}}^M - \boldsymbol{\mu}||^2\right] = \sum_{i=1}^n \mathbf{E}\left[(\widehat{y}_i^M - \mu_i)^2\right].$$

Weil für jede Zufallsvariable Z und jede Konstante c

$$\mathbf{E}\left[(Z-c)^2\right] = \mathbf{E}\left[((Z - \mathbf{E}[Z]) + (\mathbf{E}[Z] - c))^2\right] = \mathrm{Var}(Z) + (\mathbf{E}[Z] - c)^2 + 2 \cdot 0$$

gilt, haben wir

$$SMSE = \sum_{i=1}^n \mathrm{Var}(\widehat{y}_i^M) + \sum_{i=1}^n (\mathbf{E}\left[\widehat{y}_i^M\right] - \mu_i)^2 = |M|\sigma^2 + \sum_{i=1}^n (\mathbf{E}\left[\widehat{y}_i^M\right] - \mu_i)^2.$$

Der erste Term wird klein, wenn das Modell möglichst wenige Variablen enthält, der zweite hingegen dann, wenn das Modell möglichst viele Variablen enthält. Oft skaliert man $SMSE$ noch mit σ^2 und betrachtet

$$\Gamma_p(M) = \frac{SMSE(M)}{\sigma^2}.$$

Es gilt immer $\Gamma_p(M) \geq |M|$ mit Gleichheit genau dann, wenn das Modell keinen Bias hat (aber u.U. überflüssige Terme).

Wir können $\widehat{y}_i^M$ auch als Prognose für eine neue Beobachtung $Y_{n+i} = \mu_i + \varepsilon_{n+i}$ betrachten. Dann ist die Summe der mittleren quadratischen Prognosefehler (sum of prediction square errors) gleich

$$SPSE = \sum_{i=1}^n \mathbf{E}\left[(Y_{n+i} - \widehat{y}_i^M)^2\right] = \sum_{i=1}^n \mathbf{E}\left[(Y_{n+i} - \mu_i)^2\right] + \sum_{i=1}^n \mathbf{E}\left[(\widehat{y}_i^M - \mu_i)^2\right] = n\sigma^2 + SMSE.$$

(Genau genommen müsste man "sum of mean square prediction errors" sagen).

Minimierung von $SMSE$, bzw. Γ_p bzw. $SPSE$ führt also stets auf das gleiche Modell. Wir können aber keine dieser Grössen berechnen, wenn wir σ und $\boldsymbol{\mu}$ nicht kennen. Eine naive Schätzung von $SPSE$ wäre die Summe der quadrierten Residuen (sum of squared errors)

$$SSE(M) = ||\mathbf{y} - \widehat{\mathbf{y}}^M||^2 = \sum_{i=1}^n (y_i - \widehat{y}_i^M)^2,$$

d.h. einfach das Kriterium der Kleinsten Quadrate Schätzung. Diese Grösse wird jedoch immer kleiner, je mehr Variablen man dazunimmt, und sie unterschätzt $SPSE$. Es gilt nämlich

$$\begin{aligned}
\mathbf{E}\left[||\mathbf{y} - \widehat{\mathbf{y}}^M||^2\right] &= \sum_{i=1}^n \mathrm{Var}(y_i - \widehat{y}_i^M) + \sum_{i=1}^n (\mathbf{E}[y_i] - \mathbf{E}\left[\widehat{y}_i^M\right])^2 \\
&= (n - |M|)\sigma^2 + \sum_{i=1}^n (\mathbf{E}[y_i] - \mu_i)^2 = SPSE(M) - 2|M|\sigma^2.
\end{aligned}$$

Also ist es besser $SPSE$ durch $SSE(M) + 2|M|\widehat{\sigma}^2$ zu schätzen, wobei $\widehat{\sigma}^2$ eine Schätzung von σ^2 ist (z.B. aus dem vollen Modell $M = \{1, 2, \ldots p\}$), und dann dasjenige Modell zu wählen, welches diese Schätzung von $SPSE$ minimiert. Analog können wir die Schätzung

$$C_p(M) := \frac{SSE(M)}{\widehat{\sigma}^2} - n + 2|M|,$$

von Γ_p verwenden. Wegen der zufälligen Streuung kann C_p im Unterschied zu Γ_p auch kleiner als $|M|$ oder sogar negativ werden.

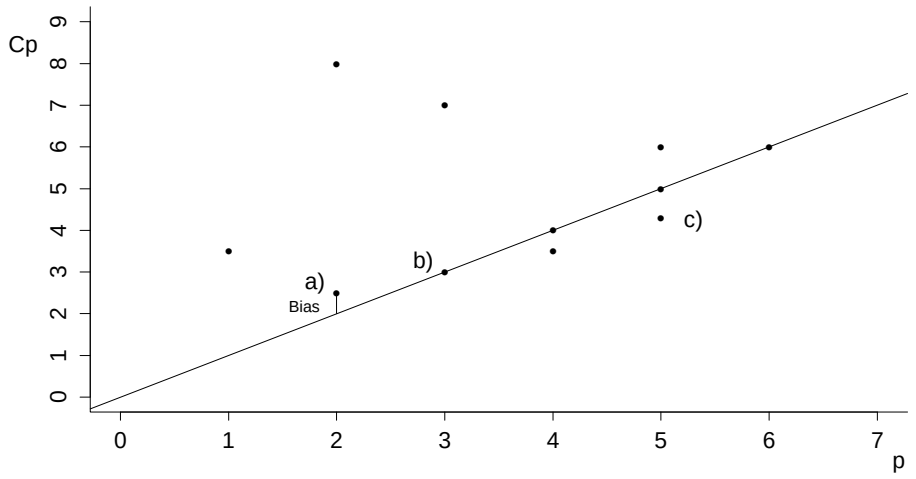

Abbildung 1.23: C_p-Plot: a) beste geschätzte Vorhersage, b) beste geschätzte biasfreie Vorhersage, c) Punkt unterhalb der Winkelhalbierenden aufgrund der zufälligen Streuung von C_p.

Das Informations-Kriterium von Akaike AIC

Man definiert für ein beliebiges Modell (nicht unbedingt ein Regressionsmodell) mit k Parametern die Grösse

$$AIC(\alpha) = -2\widehat{\ell}_k + \alpha\, k$$

wobei $\widehat{\ell}_k$ der maximalen Wert der log–Likelihood-Funktion dieses Modells ist (d.h. die log-Likelihood-Funktion beim MLE). Der erste Term misst, wie gut das Modell zu den Daten passt, und der zweite Term bestraft die Komplexität des Modells. Üblicherweise ist die multiplikative Konstante α im Bestrafungsterm gleich zwei. Unter verschiedenen Modellen, die zur Wahl stehen, nimmt man dann dasjenige mit **minimalem** AIC.

Wenn wir nun speziell ein lineares Modell mit normalverteilten Fehlern und einer Auswahl $M \subset \{1, 2, \ldots p\}$ von erklärenden Variablen haben, dann ist

$$\log f_M(\mathbf{y}, \boldsymbol{\theta}) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}(\mathbf{y} - X^M\boldsymbol{\theta}^M)^T(\mathbf{y} - X^M\boldsymbol{\theta}^M)$$

Wenn man σ^2 als gegeben anschaut (z.B. aus Erfahrung oder andern Untersuchungen), dann ist bis auf Konstanten

$$-2\widehat{\ell}_M = \frac{1}{\sigma^2}\underbrace{(\mathbf{y} - X^M\widehat{\boldsymbol{\theta}}^M)^T(\mathbf{y} - X^M\widehat{\boldsymbol{\theta}}^M)}_{SSE(M)},$$

also ist die Grösse $AIC(2)$ bis auf eine (für die Modellwahl irrelevante) Konstante gleich C_p, wenn beide das gleiche σ benutzen.

Wenn man auch σ schätzt mit dem Maximum Likelihood Schätzer

$$\widehat{\sigma}^2(M) = \frac{SSE(M)}{n},$$

dann ist das Akaikekriterium bis auf Konstanten gleich

$$AIC(\alpha) = n\log(\widehat{\sigma}^2(M)) + \alpha\,|M|.$$

Mit einer Taylor-Approximation des Logarithmus an einer Stelle σ^2 , folgt

$$AIC(2) \approx n\log(\sigma^2) + \frac{SSE(M)}{\sigma^2} - n + 2|M|.$$

Also sieht man, dass auch in diesem allgemeineren Fall AIC sehr ähnlich ist zu C_p, zumindest für die Modelle, bei denen $SSE(M)/n$ in der Nähe der Schätzung von σ^2 ist, die bei C_p verwendet wird.

1.9 Das Gauss-Markov-Theorem

Dieses Theorem sagt, dass die Kleinste Quadrate Schätzung in einem gewissen Sinn "optimal" ist. Es gibt eine Version ohne Annahme der Normalverteilung und eine mit dieser Annahme. Es unterscheiden sich aber nicht nur die Voraussetzungen, sondern auch die Behauptungen, und die Unterschiede sind wesentlich !

Wir beginnen mit dem Resultat ohne Annahme der Normalverteilung.

Satz 1.9.1 (Gauss-Markov)**.** *Es sei*

$$\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon} \qquad \mathbf{E}\,[\boldsymbol{\varepsilon}] = \mathbf{0} \qquad \mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 I \qquad \mathrm{Rg}[X] = p.$$

Ferner sei $\mathbf{c}$ *ein beliebiger* $(p \times 1)$*-Vektor und* $\widehat{\boldsymbol{\theta}}$ *die Kleinste Quadrate Schätzung. Dann hat* $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ *die kleinste Varianz unter* <u>*allen linearen*</u> *erwartungstreuen Schätzern von* $\mathbf{c}^T\boldsymbol{\theta}$*.*

Man sagt auch, die Kleinste Quadrate Schätzungen seien "BLUE" ("<u>b</u>est <u>l</u>inear <u>u</u>nbiased <u>e</u>stimators").

Die Version, welche Normalverteilung voraussetzt, lautet

Satz 1.9.2. *Es sei zusätzlich* $\boldsymbol{\varepsilon}$ *normalverteilt. Dann hat* $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ *sogar minimale Varianz unter* <u>*allen*</u> *erwartungstreuen Schätzern von* $\mathbf{c}^T\boldsymbol{\theta}$*.*

Man sagt, dass die Kleinste Quadrate Schätzungen dann sogar "UMVU" seien ("<u>u</u>niformly <u>m</u>inimum <u>v</u>ariance <u>u</u>nbiased"). (Uniform, weil das für beliebige Werte von $\boldsymbol{\theta}$ und σ^2 gilt.)

Dieser Satz (und nicht das Gauss–Markov-Theorem) liefert (neben der Einfachheit der Methode) die beste Rechtfertigung für Kleinste Quadrate. Allerdings ist auch Erwartungstreue nicht immer eine zwingende Forderung und wird zum Beispiel in Bayes'scher Regression und bei "ridge regression" verletzt. Diese Verfahren besprechen wir hier jedoch nicht.

Der Verzicht auf die Annahme der Normalverteilung im Gauss-Markov-Theorem wird erkauft durch die Einschränkung auf lineare Schätzer. Der springende Punkt ist dabei, dass **alle linearen** Schätzer eine **kleine Effizienz** (d.h. eine grosse Varianz) haben, selbst wenn die Abweichungen von der Normalverteilung noch relativ klein sind.

Begründende Beispiele: Wenn wir versuchen, die wahren Fehlerverteilungen von Daten hoher Qualität durch t–Verteilungen mit ν Freiheitsgraden zu approximieren, so finden wir $\nu = 5 - 9$ Freiheitsgrade für "sehr normal" aussehende Daten, ziemlich oft aber auch $\nu = 3$ Freiheitsgrade, und selbst $\nu = 1$ (Cauchy) kann gelegentlich am besten passen. Nun ist die asymptotische Effizienz (also z.B. das inverse Verhältnis der benötigten Stichprobenumfänge, um gleiche Genauigkeit zu erreichen) von Kleinsten Quadraten im Vergleich zu einer asymptotisch bestmöglichen Schätzung (z.B. maximum likelihood) unter einer t-Verteilung mit ν Freiheitsgraden $= 1 - 6/(\nu(\nu+1))$ (für $\nu \geq 2$). Damit ist die tatsächliche Effizienz von Kleinsten Quadraten für "sehr gut" normalverteilte Daten ($t_5 - t_9$) 80-93%, und für "ziemlich gut" normalverteilte Daten (t_3) ca. 50%. Für $\widehat{\sigma}^2$ sieht die Sache sogar noch wesentlich schlimmer aus !

Beweis des Gauss–Markov-Theorems Sei $\mathbf{a}$ ein $(n \times 1)$–Vektor und a_0 eine Konstante, so dass $\mathbf{a}^T\mathbf{y} + a_0$ eine erwartungstreue Schätzung von $\mathbf{c}^T\boldsymbol{\theta}$ ist. Dann muss

$$\mathbf{E}\left[\mathbf{a}^T\mathbf{y} + a_0\right] = \mathbf{a}^T X\boldsymbol{\theta} + a_0 = \mathbf{c}^T\boldsymbol{\theta}$$

gelten für alle $\boldsymbol{\theta}$. Daraus folgt $a_0 = 0$ und $\mathbf{a}^T X = \mathbf{c}^T \Leftrightarrow X^T\mathbf{a} = \mathbf{c}$.

Der Vektor, der zur Kleinste-Quadrate-Schätzung gehört, $\mathbf{a_{KQ}} = X(X^TX)^{-1}\mathbf{c}$, ist eine spezielle Lösung von $X^T\mathbf{a} = \mathbf{c}$. Ferner steht $\mathbf{a_{KQ}}$ senkrecht auf allen Lösungen $\mathbf{a_h}$ des homogenen Systems $X^T\mathbf{a} = \mathbf{0}$, denn $\mathbf{a_{KQ}}^T\mathbf{a_h} = \mathbf{c}^T(X^TX)^{-1}X^T\mathbf{a_h} = 0$. Weil $\mathrm{Cov}[\mathbf{Y}] = \sigma^2 I$ gilt daher

$$\mathrm{Var}((\mathbf{a_{KQ}} + \mathbf{a_h})^T\mathbf{Y}) = \mathrm{Var}(\mathbf{a_{KQ}}^T\mathbf{Y}) + \mathrm{Var}(\mathbf{a_h}^T\mathbf{Y}) \geq \mathrm{Var}(\mathbf{a_{KQ}}^T\mathbf{Y}).$$

□

Beweis der Variante des Gauss–Markov-Theorems (unter Benutzung der mehrdimensionalen Cramér-Rao-Ungleichung, welche in der Mathematischen Statistik bewiesen wird):

Wir betrachten die folgende allgemeine Situation: Sei $(f_{\boldsymbol{\eta}}(\mathbf{y}))$ eine parametrische Familie von strikt positiven Dichten im $\mathbb{R}^n$. Der Parameter $\boldsymbol{\eta}$ variiere in einer offenen Menge im $\mathbb{R}^k$ und $f_{\boldsymbol{\eta}}(\mathbf{y})$ sei differenzierbar bezüglich $\boldsymbol{\eta}$. Der Parameter von Interesse sei $g(\boldsymbol{\eta})$, wobei g eine beliebige, reellwertige Funktion von $\boldsymbol{\eta}$ ist. Dann gilt

Satz 1.9.3 (Cramér-Rao)**.** *Wenn $T(\mathbf{y})$ ein beliebiger erwartungstreuer Schätzer für $g(\boldsymbol{\eta})$ ist, d.h.*

$$\mathbf{E}_{\boldsymbol{\eta}}[T(\mathbf{y})] = g(\boldsymbol{\eta}) \quad \forall \boldsymbol{\eta},$$

dann ist g differenzierbar und es gilt

$$\mathrm{Var}_{\boldsymbol{\eta}}(T(\mathbf{y})) \geq \frac{\partial g}{\partial \boldsymbol{\eta}}^T I(\boldsymbol{\eta})^{-1} \frac{\partial g}{\partial \boldsymbol{\eta}}$$

wobei $I(\boldsymbol{\eta})$ die sogenannte Fisher-Informationsmatrix bezeichnet:

$$I(\boldsymbol{\eta}) = \mathbf{E}_{\boldsymbol{\eta}}\left[\frac{\partial \log f_{\boldsymbol{\eta}}(\mathbf{Y})}{\partial \boldsymbol{\eta}} \frac{\partial \log f_{\boldsymbol{\eta}}(\mathbf{Y})}{\partial \boldsymbol{\eta}}^T\right].$$

Dies wenden wir nun an auf

$$f_{\boldsymbol{\eta}}(\mathbf{y}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\theta})^T(\mathbf{y} - X\boldsymbol{\theta})),$$

$$\boldsymbol{\eta} = (\sigma^2, \boldsymbol{\theta}^T)^T$$

und

$$g(\boldsymbol{\eta}) = \mathbf{c}^T\boldsymbol{\theta}.$$

Mit einer Rechnung erhält man die Fisher-Information

$$I(\boldsymbol{\eta}) = \begin{pmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{1}{\sigma^2}X^TX \end{pmatrix},$$

d.h. der Kleinste Quadrate Schätzer $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ erreicht die untere Schranke der Cramér-Rao-Ungleichung. Damit hat er offensichtlich minimale Varianz. $\square$

Kapitel 2

Nichtlineare und nichtparametrische Methoden

2.1 Robuste Methoden

Allgemein heisst ein statistisches Verfahren für ein parametrisches Modell robust, wenn sich seine Eigenschaften nur wenig ändern bei kleinen Abweichungen vom Modell. Im linearen Modell sind Kleinste Quadrate nicht robust: Langschwänzige Fehlerverteilungen haben grosse Auswirkungen ! Die Verteilung der Schätzungen ist zwar einigermassen stabil (vgl. 1.4.3), aber schon bei relativ kleinen Abweichungen von der Normalverteilung gibt es bessere Schätzungen als Kleinste Quadrate. Als Konsequenz davon ist das Niveau der Tests und Vertrauensintervalle basierend auf Kleinsten Quadraten robust, nicht aber die Macht.

Ein damit verwandtes Phänomen ist, dass Kleinste Quadrate Schätzungen und die darauf basierenden Tests und Vertrauensintervalle sehr empfindlich sind auf vereinzelte Ausreisser. Langschwänzige Fehlerverteilungen produzieren ja gerade Beobachtungen, die wie Ausreisser aussehen.

Im Folgenden untersuchen wir zuerst genauer die Auswirkungen eines Ausreissers auf Kleinste Quadrate und diskutieren dann alternative Schätzer, welche robuster sind.

2.1.1 Einfluss einzelner Beobachtungen bei Kleinsten Quadraten

Wir untersuchen zunächst, wie sich das Weglassen bzw. Hinzufügen einzelner Beobachtungen auf die Kleinste Quadrate Schätzung auswirkt. Dazu ist folgendes Lemma nützlich, das auf Gauss zurückgeht:

Lemma 2.1.1. *Sei A eine invertierbare $p \times p$ Matrix und $\mathbf{a}$ und $\mathbf{b}$ zwei $p \times 1$ Vektoren mit $\mathbf{b}^T A^{-1}\mathbf{a} \neq -1$. Dann ist auch $A + \mathbf{a}\mathbf{b}^T$ invertierbar, und es gilt*

$$(A - \mathbf{a}\mathbf{b}^T)^{-1} = A^{-1} + \frac{1}{1 - \mathbf{b}^T A^{-1}\mathbf{a}} A^{-1}\mathbf{a}\mathbf{b}^T A^{-1}.$$

Beweis als Übung

Wir bezeichnen die Kleinste Quadrate Schätzung ohne die i-te Beobachtung mit $\widehat{\boldsymbol{\theta}}^{(-i)}$ und benutzen ferner die Abkürzungen

$$A = X^T X = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T, \quad \mathbf{c} = X^T y = \sum_{j=1}^{n} y_j \mathbf{x}_j^T.$$

Dann gilt mit obigem Lemma

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(-i)} &= (A - \mathbf{x}_i \mathbf{x}_i^T)^{-1}(\mathbf{c} - y_i \mathbf{x}_i) \\
&= A^{-1}\mathbf{c} - y_i A^{-1}\mathbf{x}_i + \frac{1}{1 - \mathbf{x}_i^T A^{-1}\mathbf{x}_i} A^{-1}\mathbf{x}_i \mathbf{x}_i^T A^{-1}(\mathbf{c} - y_i \mathbf{x}_i) \\
&= \widehat{\boldsymbol{\theta}} - y_i A^{-1}\mathbf{x}_i (1 + \frac{\mathbf{x}_i^T A^{-1}\mathbf{x}_i}{1 - \mathbf{x}_i^T A^{-1}\mathbf{x}_i}) + \mathbf{x}_i^T \boldsymbol{\theta} A^{-1}\mathbf{x}_i \frac{1}{1 - \mathbf{x}_i^T A^{-1}\mathbf{x}_i},
\end{aligned}$$

also

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}} &= -\frac{1}{1 - \mathbf{x}_i^T (X^T X)^{-1}\mathbf{x}_i} (X^T X)^{-1}\mathbf{x}_i (y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}) \\
&= -\frac{r_i}{1 - P_{ii}} (X^T X)^{-1}\mathbf{x}_i.
\end{aligned}$$

Wir sehen also, dass der Einfluss der i-ten Beobachtung sowohl vom i-ten Residuum als auch vom Diagonalelement P_{ii} der "Hutmatrix" (für alle Beobachtungen, inklusive der i-ten) abhängt. Um einflussreiche Beobachtungen zu entdecken, plottet man daher oft die Residuen r_i gegen die P_{ii}.

Die Differenz der geschätzten Parameter ist etwas schwierig zu interpretieren, weil es ein ganzer Vektor ist, der ausserdem von der Skalierung der erklärenden Variablen abhängt. Eine skalare und invariante Grösse erhält man, indem man die Länge von $\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}}$ in der Metrik berechnet, welche durch die geschätzte Kovarianzmatrix von $\widehat{\boldsymbol{\theta}}$ definiert ist:

$$D_i = \frac{(\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})^T (X^T X)(\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})}{p\widehat{\sigma}^2} = \frac{1}{p} \frac{r_i^2}{\widehat{\sigma}^2(1 - P_{ii})} \frac{P_{ii}}{1 - P_{ii}}.$$

Die Grösse D_i heisst Cook-Distanz. Sie ist eine einfache Funktion von P_{ii} sowie dem Quadrat des studentisierten Residuums $r_i/(\widehat{\sigma}\sqrt{1 - P_{ii}})$. Sie wird häufig als diagnostisches Hilfsmittel verwendet. Diejenigen Beobachtungen, bei denen D_i deutlich grösser ist als beim Rest, sollten besonders betrachtet, bzw. in der Analyse weggelassen werden.

Analog kann man die Änderung beim Hinzufügen einer Beobachtung an einer beliebigen Stelle $(y, \mathbf{x})$ betrachten:

$$\Delta\widehat{\boldsymbol{\theta}} = \frac{1}{1 + \mathbf{x}^T (X^T X)^{-1}\mathbf{x}} (X^T X)^{-1}\mathbf{x}(y - \mathbf{x}^T \widehat{\boldsymbol{\theta}}).$$

Man sieht daraus, dass man durch Hinzufügen einer einzigen Beobachtung die Kleinste Quadrate Schätzung beliebig verändern kann, d.h. die Kleinste Quadrate Schätzung ist nicht robust. Zudem hängt dieser Effekt sehr stark davon ab, wo man die Beobachtung hinzufügt. Die Formel wird noch etwas durchsichtiger, wenn man annimmt, dass die $\mathbf{x}_i$ zufällig und i.i.d sind. Dann erhält man für $n \to \infty$ in erster Näherung

$$\Delta\widehat{\boldsymbol{\theta}} \sim \frac{1}{n} (\mathbf{E}\left[\mathbf{x}_i \mathbf{x}_i^T\right])^{-1}\mathbf{x}(y - \mathbf{x}^T \boldsymbol{\theta}).$$

Die Entdeckung und spezielle Behandlung einflussreicher Beobachtungen mit Hilfe der Cook-Distanz hat jedoch zwei Nachteile: Erstens ist der Effekt des Weglassens von zwei oder mehr Beobachtungen nicht einfach die Summe der einzelnen Effekte (eine einflussreiche Beobachtung kann eine andere maskieren), und zweitens ist durch das Weglassen einflussreicher Beobachtungen die Gültigkeit der Tests und Vertrauensintervalle, basierend auf den verbleibenden Daten, in Frage gestellt.

2.1.2 Huber- und L_1-Regression

Einzelne Beobachtungen haben bei Kleinsten Quadraten einen grossen Einfluss, weil grosse Residuen unter einem quadratischen Kriterium sehr stark ins Gewicht fallen. Um dies zu vermeiden, kann man stattdessen z.B. den L_1-Schätzer betrachten:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \boldsymbol{\theta}|.$$

Historisch ist diese Methode sogar älter als Kleinste Quadrate: Sie wurde 1760 von Boscovich und 1789 von Laplace vorgeschlagen und diskutiert !

Im Lokationsfall, d.h. für $p = 1$ und $x_i \equiv 1$, ergibt das als Lösung den Median der Daten, der bei normalverteilten Daten wesentlich ungenauer ist als das arithmetische Mittel (d.h. der Kleinste Quadrate Schätzer): Der Median braucht für die gleiche Genauigkeit etwa 50% mehr Beobachtungen.

Einen Kompromiss zwischen der Minimierung des L_2- und des L_1-Abstandes stellt die Huber-Regression dar:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho_c(y_i - \mathbf{x}_i^T \boldsymbol{\theta}),$$

wobei

$$\rho_c(u) = \frac{1}{2}u^2 \ (|u| \leq c), \quad \rho_c(u) = c(|u| - \frac{c}{2}) \ (|u| \geq c),$$

vgl. Abbildung 2.1. Die Wahl $c = 0$ entspricht der L_1-Regression. Durch Ableiten und Nullsetzen erhält man die Gleichungen

$$\sum_{i=1}^{n} \psi_c(y_i - \mathbf{x}_i^T \boldsymbol{\theta}) \mathbf{x}_i = 0$$

wobei $\psi_c(u) = \rho_c(u)' = \operatorname{sign}(u) \min(|u|, c)$.

Die Huber-Regression macht jedoch nur Sinn, wenn der "Knickpunkt" c in Beziehung zu der Streuung der Residuen gewählt wird. Man betrachtet daher üblicherweise die Schätzer, welche durch die folgenden Gleichungen definiert sind:

$$\begin{aligned} \sum_{i=1}^{n} \psi_c \left(\frac{y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}}{\widehat{\sigma}} \right) x_i &= 0, \\ \sum_{i=1}^{n} \chi \left(\frac{y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}}{\widehat{\sigma}} \right) &= 0. \end{aligned}$$

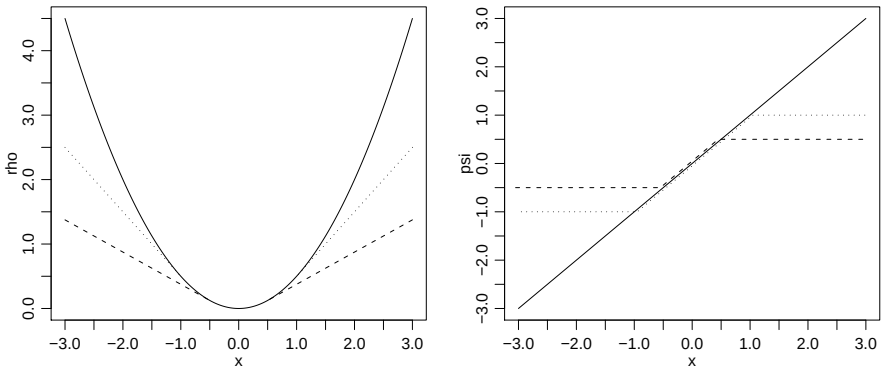

Abbildung 2.1: Die Huber -Funktion und ihre Ableitung für verschiedene c's

Die Funktion $\chi(u)$ wählt man entweder als $\psi_c(u)^2 - \beta$ oder als $\chi(u) = \text{sign}(|u| - \beta)$ (sign bezeichnet die Vorzeichenfunktion). Die Konstante β schliesslich ist festgelegt durch die Bedingung

$$\int \chi(u) \exp(-u^2/2) du = 0,$$

welche dafür sorgt, dass bei normalverteilten Fehlern $\widehat{\sigma}$ eine konsistente Schätzung für die Standardabweichung ist. Die erste Wahl von χ ist der sogenannte Proposal 2 von Huber, während die zeite Wahl $1/\beta$ mal den Median der Absolutbeträge der Residuen ergibt.

Die Berechnung des L_1- oder des Huber-Schätzers ist nicht mehr in geschlossener Form möglich, es gibt jedoch heute effiziente Algorithmen. Die Berechnung der L_1-Regression kann auf ein Problem der linearen Optimierung zurückgeführt werden, das mit "interior point" Methoden sogar schneller gelöst werden kann als die Berechnung der Kleinsten Quadrate Schätzer. Für die Huber-Regression verwendet man iterativ gewichtete Kleinste Quadrate mit den Gewichten

$$w_i \propto \frac{\psi_c((y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}})/\widehat{\sigma})}{y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}} \propto \min(1, \frac{c\widehat{\sigma}}{|y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}|})$$

solange, bis sich die Gewichte nicht mehr ändern.

Die exakte Verteilung der Schätzer bei der L_1- oder der Huber-Regression kann man nicht geschlossen angeben. Man stützt sich daher auf die Asymptotik ab, welche besagt, dass für zufällige, unabhängige und identisch verteilte $\mathbf{x}_i$ der standardisierte Vektor $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ genähert normalverteilt ist mit Erwartungswert null und Kovarianzmatrix

$$\frac{\mathbf{E}\left[\psi_c(\varepsilon_i/\sigma)^2\right]}{P[|\varepsilon_i| \leq c\sigma]^2} \sigma^2 \mathbf{E}\left[\mathbf{x}_i \mathbf{x}_i^T\right]^{-1}.$$

Dies ist bis auf einen Faktor die gleiche Kovarianzmatrix wie bei den Kleinsten Quadraten. Für c im Bereich $[1, 1.5]$ ist dieser Faktor bei langschwänzigen Verteilungen wesentlich kleiner als 1, während er bei Normalverteilung nicht viel grösser als 1 ist. Man muss bei numerischen Vergleichen jedoch beachten, dass für nicht normalverteilte ε_i der Parameter σ nicht mehr die Standardabweichung ist, sondern die Lösung von

$$\mathbf{E}\left[\chi(\varepsilon_i/\sigma)\right] = 0.$$

Diese asymptotische Näherung bildet auch die Grundlage für Tests und Vertrauensintervalle. Wir verzichten hier auf die Details.

Leider löst die Huber-Regression aber nicht alle Probleme mit einflussreichen Beobachtungen. Exakt kann man den Effekt des Hinzufügens oder Weglassens von Beobachtungen nicht mehr angeben. Näherungsweise ist aber die Differenz der Schätzungen nach Hinzufügen einer Beobachtung an einer Stelle $(\mathbf{x}, y)$ gegeben durch

$$\Delta\widehat{\boldsymbol{\theta}} \sim \frac{1}{nP[|\varepsilon_i| \leq c\sigma]}(\mathbf{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right])^{-1}\mathbf{x}\ \psi_c(\frac{y - \mathbf{x}^T\boldsymbol{\theta}}{\sigma})\sigma.$$

Der Einfluss grosser y-Werte ist also beschränkt für eine feste Stelle $\mathbf{x}$, aber durch Variation von $\mathbf{x}$ kann dieser Einfluss dennoch beliebig gross werden. Raffiniertere Methoden, die diesen Nachteil nicht haben, werden in den beiden nächsten Abschnitten besprochen.

Anstelle der Huber-Regression werden oft auch Schätzer betrachtet, welche durch obige implizite Gleichungen mit beliebigem ungeradem ψ und geradem χ definiert sind. Solche Schätzer heissen M-Schätzer. Besonders beliebt sind ψ-Funktionen, die für grosses $|r|$ gegen null gehen, weil dann extreme Ausreisser ganz verworfen werden. In diesem Fall haben die Gleichungen jedoch meist mehrere Lösungen. Welche der Lösungen gefunden wird, hängt dann meist vom Algorithmus und insbesondere vom gewählten Startwert ab.

2.1.3 Regressionsschätzer mit beschränktem Einfluss

Um den Einfluss sowohl von y als auch von $\mathbf{x}$ zu beschränken, betrachtet man Schätzer welche durch ein Gleichungssystem der Form

$$\sum_{i=1}^{n} \eta\left(\mathbf{x}_i, \frac{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\theta}}}{\widehat{\sigma}}\right)\mathbf{x}_i = 0$$

definiert sind. Zur Bestimmung von $\widehat{\sigma}$ kommt noch eine Gleichung ähnlich wie bei der Huber-Regression dazu. Verschiedene Formen von $\eta(\mathbf{x}, r)$ werden benutzt, unter anderem

$$\eta(\mathbf{x}, r) = \min(1, \frac{a}{||A\mathbf{x}||})\psi_c(r) \quad \text{(Mallows)}$$

und

$$\eta(\mathbf{x}, r) = \frac{1}{||A\mathbf{x}||}\psi_c((||A\mathbf{x}||)r) \quad \text{(Schweppe)}.$$

Die Matrix A wird so gewählt, dass $||A\mathbf{x}||$ ausdrückt, wie stark $\mathbf{x}$ von der Punktwolke der $(\mathbf{x}_i)_{1\leq i\leq n}$ abweicht. Dies wird zum Beispiel erreicht, wenn

$$||A\mathbf{x}||^2 = \text{const} \cdot \mathbf{x}^T(X^TX)^{-1}\mathbf{x},$$

aber diese Wahl kann selber wieder stark durch eine Beobachtung mit einem ungewöhnlichen $\mathbf{x}_i$ beeinflusst werden, weil X^TX ja geschrieben werden kann als $\sum_i \mathbf{x}_i\mathbf{x}_i^T$. Man ersetzt daher X^TX durch eine analoge, aber robuste Grösse. Dies ergibt zusätzliche Gleichungen für A, die wir hier nicht näher besprechen.

Im Vorschlag von Mallows werden Beobachtungen mit stark abweichenden erklärenden Variablen auf jeden Fall heruntergewichtet. Um Schweppe's Vorschlag besser zu verstehen, ist die Beziehung $\psi_c(dr)/d = \psi_{c/d}(r)$ nützlich. Es wird bei Schweppe's Vorschlag also nur der Knickpunkt der Huber-Funktion heruntergesetzt, so dass eine Beobachtung auch bei stark abweichenden erklärenden Variablen noch volles Gewicht bekommen kann, wenn das Residuum nahe bei null ist.

Näherungsweise ist bei diesen Verfahren die Differenz der Schätzungen nach Hinzufügen einer Beobachtung an einer Stelle $(\mathbf{x}, y)$ gegeben durch

$$\Delta\widehat{\boldsymbol{\theta}} \sim \frac{1}{n}\left(\mathbf{E}\left[\frac{\partial}{\partial r}\eta(\mathbf{x}_i, \frac{\varepsilon_i}{\sigma})\mathbf{x}_i\mathbf{x}_i^T\right]\right)^{-1} \mathbf{x}\ \eta\left(\mathbf{x}, \frac{y - \mathbf{x}^T\boldsymbol{\theta}}{\sigma}\right)\sigma.$$

Der Einfluss ist also bei beiden Vorschlägen für η sowohl in $\mathbf{x}$ als auch in y beschränkt.

Diese Schätzer sind jedoch auch nicht wirklich befriedigend, denn sie haben den Nachteil, dass ihr *"Bruchpunkt"* höchstens gleich $1/p$ ist. Der Bruchpunkt ist dabei definiert als der maximale Anteil von Ausreissern, den ein Schätzer verkraften kann, ohne zu divergieren.

2.1.4 Regressionsschätzer mit hohem Bruchpunkt

Schätzer, deren Bruchpunkt nicht von der Dimension abhängt, erhält man, indem man bei

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2 = \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2$$

das arithmetische Mittel ersetzt durch den Median:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \operatorname{median}((y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2)$$

(Least median of squares; Hampel 1975, Rousseeuw 1984). Das heisst man sucht unter allen Paaren von parallelen Hyperebenen, die 50% der Beobachtungen $(y_i, \mathbf{x}_i)$ enthalten, dasjenige Paar mit minimalem Abstand in y-Richtung. Dies ist in der Abbildung 2.2 links illustriert.

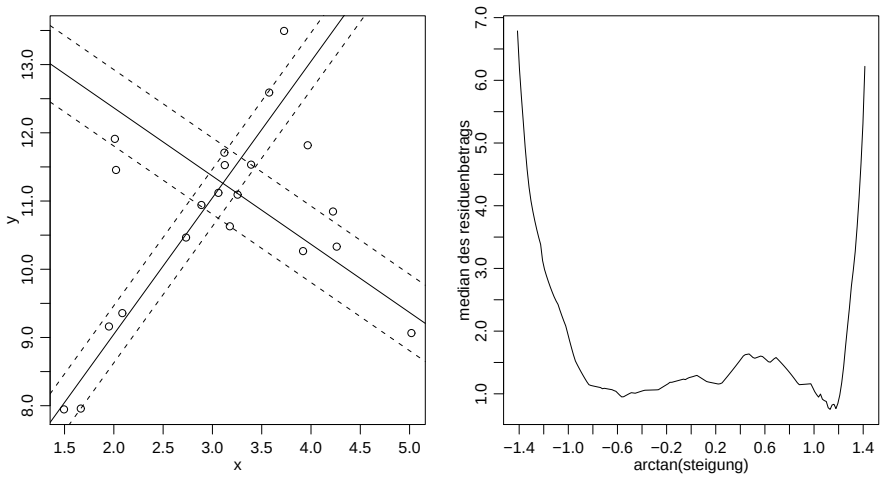

Abbildung 2.2: Least Median of Squares bei einfacher Regression. Links sind zwei Geraden dargestellt zusammen mit je einem Band, das 50% der Beobachtungen enthält. Der Achsenabschnitt wurde so gewählt, dass die Bänder minimalen Durchmesser in y-Richtung haben. Rechts ist der Durchmesser des Bands dargestellt in Funktion der Steigung.

Intuitiv ist es einleuchtend (und kann auch bewiesen werden), dass dieses Verfahren ungefähr 50% Ausreisser toleriert ohne zu divergieren. Die Berechnung des Schätzers stellt jedoch ein grosses Problem dar, weil die Zielfunktion $\operatorname{median}((y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2)$ im Allgemeinen

viele lokale Minima hat, vergleiche Abbildung 2.2, rechts. Man muss daher den ganzen Raum absuchen, um das globale Minimum zu finden, und das wird in hohen Dimensionen sehr rasch zu aufwändig. Üblicherweise verwendet man stochastische Algorithmen, bei denen man $p+1$ Datenpunkte zufällig auswählt, eine Ebene durch diese $p+1$ Punkte legt und den Wert des Kriteriums für das zugehörige θ berechnet.

Ein weiterer Nachteil ist die schlechte Effizienz im Fall, wo die Fehler normalverteilt sind: Der Schätzer konvergiert dann nur mit Rate $n^{-1/3}$. Eine bessere Konvergenzgeschwindigkeit erhält man, indem man z.B. den Median ersetzt durch ein gestutztes Mittel der $((y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2)$, wobei man einen Bruchteil αn (mit $\alpha < 0.5$) der grössten Residuenquadrate weglässt. Meist macht man ausgehend von einem solchen Schätzer dann noch eine Newton-Iteration zur Lösung der Schätzgleichungen für einen M-Schätzer mit einer ψ-Funktion, die auf null zurückgeht. Dies wird als MM-Schätzer bezeichnet.

Die Entwicklung von robusten Regressionsschätzern mit guten statistischen und algorithmischen Eigenschaften ist immer noch ein Thema der Forschung.

2.2 Nichtlineare Kleinste Quadrate

In diesem Kapitel besprechen wir Verfahren für die Schätzung von $\boldsymbol{\theta}$ in Modellen der Form

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i.$$

Dabei ist f eine bekannte Funktion der Versuchsbedingungen und der Parameter. Entscheidend ist, dass f nichtlinear in den Parametern $\boldsymbol{\theta}$ ist und wir dies nicht durch Transformationen auf ein lineares Modell zurückführen können (oder wollen). Die Dimension p des Parameters muss nicht mehr gleich der Dimension der erklärenden Variablen sein. Der Vektor der Fehler $\boldsymbol{\varepsilon}$ soll die gleichen Voraussetzungen erfüllen wie im linearen Modell, d.h. $\mathbf{E}[\boldsymbol{\varepsilon}] = 0$ und $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I$.

Viele Probleme in den Anwendungen sind von diesem Typ, und meistens folgt die Form von f aus einer Theorie der Substanzwissenschaften. Zur Beschreibung des kumulierten Sauerstoffverbrauchs y von Mikroorganismen in Flusswasserproben in Abhängigkeit der Inkubationszeit x verwendet man üblicherweise das Modell

$$f(x, \boldsymbol{\theta}) = \theta_1(1 - \exp(-\theta_2 x)).$$

Der Parameter θ_1 ist also die Sättigungsgrenze und $\theta_1 \cdot \theta_2$ der Anstieg bei $x = 0$. Ein Beispiel von Daten und einer möglichen Regressionsfunktion sind in der Abbildung 2.3 dargestellt.

Ein anderes, ähnliches Beispiel ist das sogenannte Michaelis-Menten Modell, das die Abhängigkeit der Reaktionsgeschwindigkeit y von einer Substratkonzentration x beschreibt. Dort ist

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_2 + x}.$$

Mit der Transformation $y \to 1/y$, $x \to 1/x$ erhält man ein lineares Regressionsmodell, aber bei den Daten, die in Abbildung 2.4 gezeigt werden, ist die Fehlervarianz nach der Transformation nicht mehr konstant und man erhält mit nichtlinearen Kleinsten Quadraten eine wesentlich bessere Anpassung.

In vielen Anwendungen sind die $\mathbf{x}_i$ Zeiten oder Orte, an denen eine Grösse beobachtet wurde, deren Entwicklung einer gewöhnlichen oder partiellen Differentialgleichung genügt.

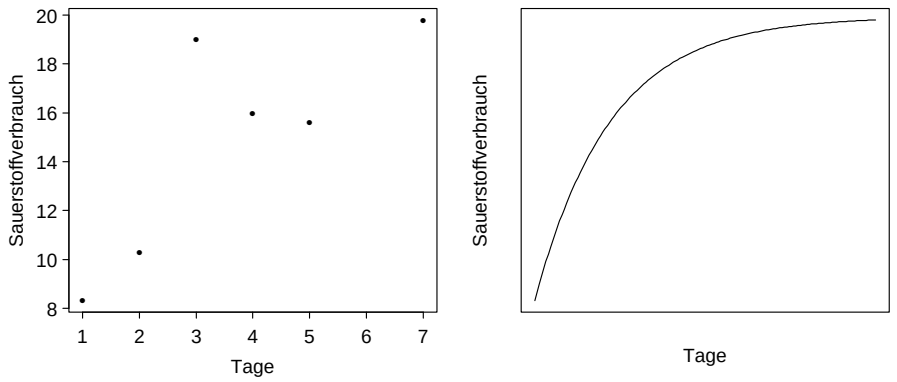

Abbildung 2.3: Daten zum Sauerstoffverbrauch in Abhängigkeit von der Inkubationszeit (links) und eine typische Regressionsfunktion (rechts).

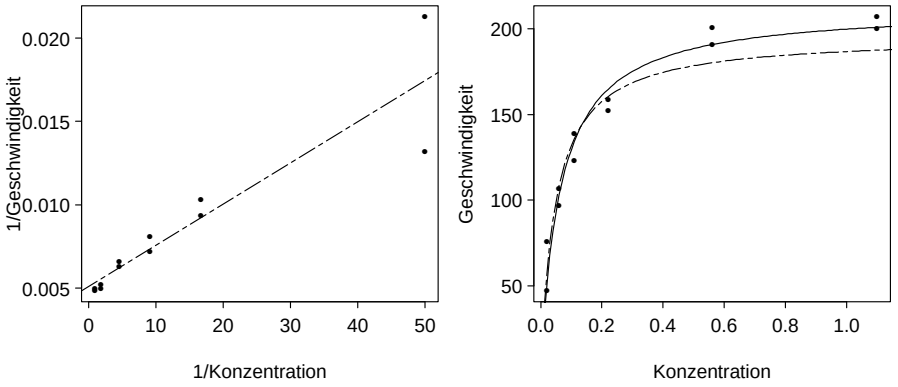

Abbildung 2.4: Reaktionsgeschwindigkeit in Abhängigkeit vom Substrat. Links Anpassung einer Geraden nach einer linearisierenden Transformation, und rechts angepasste Funktionen in der ursprünglichen Skala mit Parametern, welche in der ursrpünglichen Skala (ausgezogen) bzw. in der transformierten Skala (–·–) geschätzt wurden.

Die Parameter $\boldsymbol{\theta}$ sind dann die Parameter in dieser Differentialgleichung (plus eventuell die Anfangsbedingungen), und $f(\mathbf{x}, \boldsymbol{\theta})$ bezeichnet die Lösung der Differentialgleichung mit Parameter $\boldsymbol{\theta}$ an der Stelle $\mathbf{x}$.

Die Kleinste Quadrate Schätzung ist definiert als

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}),$$

wobei

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2.$$

Das kann man geometrisch interpretieren: Variiert man $\boldsymbol{\theta}$, so beschreiben die Punkte $(f(\mathbf{x}_1, \boldsymbol{\theta}), \ldots, f(\mathbf{x}_n, \boldsymbol{\theta}))^T$ eine p-dimensionale, gekrümmte Fläche im $\mathbb{R}^n$, die sogenannte Wirkungsfläche. Wir suchen nun denjenigen Punkt auf der Wirkungsfläche, der am nächsten beim Beobachtungspunkt $(y_1, \ldots, y_n)^T$ liegt.

Die Lösung ist im Allgemeinen nicht geschlossen darstellbar und man verwendet iterative Methoden (Gauss-Newton, bzw. Levenberg-Marquardt). In der Praxis ist es meist entscheidend, gute Startwerte zur Verfügung zu haben.

Die Fehlervarianz σ^2 wird analog zur linearen Regression geschätzt durch

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}))^2.$$

2.2.1 Asymptotische Vertrauensintervalle und Tests

In nichtlinearen Modellen gibt es leider keine exakten Tests und Vertrauensintervalle mehr, selbst wenn man normalverteilte Fehler annimmt. Man stützt sich daher auf die Asymptotik ab.

Die asymptotischen Eigenschaften von $\widehat{\boldsymbol{\theta}}$ erhält man mit Hilfe einer Taylorapproximation um den wahren Parameter $\boldsymbol{\theta}_0$. Man approximiert also die Wirkungsfläche in der Umgebung des Punktes, der dem wahren Parameter entspricht, durch eine Hyperebene:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) \approx f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \mathbf{a}(\boldsymbol{\theta}_0)_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

wobei

$$\mathbf{a}(\boldsymbol{\theta})_i = \left(\frac{\partial}{\partial \boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta}); j = 1, \ldots p\right)^T.$$

Wir haben also in der Umgebung des wahren Parameters genähert ein lineares Modell mit erklärenden Variablen $\mathbf{a}(\boldsymbol{\theta}_0)_i$. Man kann auch zeigen, dass (unter technischen Bedingungen) die Verteilung von $\widehat{\boldsymbol{\theta}}$ asymptotisch gleich ist wie im approximierenden Modell, d.h.

$$\widehat{\boldsymbol{\theta}} \overset{\text{asymptotisch}}{\sim} \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2(A(\boldsymbol{\theta}_0)^T A(\boldsymbol{\theta}_0))^{-1}).$$

Die Matrix $A(\boldsymbol{\theta})$ besteht dabei aus den Zeilen $\mathbf{a}(\boldsymbol{\theta})_i^T$.

Aus diesen genäherten Verteilungen kann man nicht direkt Tests und Vertrauensintervalle konstruieren, weil sowohl σ als auch $A(\boldsymbol{\theta}_0)$ unbekannt sind. Wir können jedoch stattdessen $\widehat{\sigma}$ und $A(\widehat{\boldsymbol{\theta}})$ einsetzen. Analog zum linearen Modell verwendet man im Allgemeinen noch die t-Verteilung statt der Normalverteilung, bzw. F-Verteilungen statt der Chiquadrat-Verteilung. Das Vertrauensintervall für θ_k lautet demnach

$$\widehat{\theta}_k \pm t_{n-p;1-\alpha/2}\, se(\widehat{\theta}_k), \qquad se(\widehat{\theta}_k) = \widehat{\sigma}\sqrt{((A(\widehat{\boldsymbol{\theta}})^T A(\widehat{\boldsymbol{\theta}}))^{-1})_{kk}}.$$

2.2.2 Genauere Tests und Vertrauensintervalle

Analog zum F-Test bei der linearen Regression können wir auch zwei geschachtelte Modelle mit Hilfe des Unterschieds der Quadratsumme der Abweichungen testen. Um so die Nullhypothese $B\boldsymbol{\theta} = \mathbf{b}$ zu testen, brauchen wir noch die Kleinste Quadrate Schätzung unter der Nullhypothese:

$$\widehat{\boldsymbol{\theta}}_0 = \arg\min_{\boldsymbol{\theta}; B\boldsymbol{\theta}=\mathbf{b}} S(\boldsymbol{\theta}).$$

Dann bilden wir die Testgrösse

$$T = \frac{(S(\widehat{\boldsymbol{\theta}}_0) - S(\widehat{\boldsymbol{\theta}}))/q}{S(\widehat{\boldsymbol{\theta}})/(n-p)},$$

wobei q der Rang von B ist. Im linearen Modell war diese Testgrösse identisch mit der Testgrösse, die sich aus der gemeinsamen Normalverteilung von $\widehat{\boldsymbol{\theta}}$ ergibt, und sie hatte unter der Nullhypothese eine $F_{q,n-p}$-Verteilung. Im nichtlinearen Fall sind die beiden

Testgrössen verschieden, und die F-Verteilung stimmt auch nur genähert. Oft ist diese Approximation aber wesentlich besser als die Normalapproximation für $\widehat{\boldsymbol{\theta}}$.

Insbesondere können wir damit die Nullhypothese $\theta_k = \theta_k^*$ testen, wobei θ_k^* ein beliebiger aber fester Wert ist. Die Teststatistik lautet dann

$$T_k(\theta_k^*) = \frac{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})}{S(\widehat{\boldsymbol{\theta}})/(n-p)} = \frac{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})}{\widehat{\sigma}^2},$$

wobei

$$\widehat{\boldsymbol{\theta}}^{(-k)} = \widehat{\boldsymbol{\theta}}^{(-k)}(\theta_k^*) = \arg \min_{\boldsymbol{\theta};\theta_k=\theta_k^*} S(\boldsymbol{\theta}).$$

($\widehat{\boldsymbol{\theta}}^{(-k)}$ ist also die Kleinste Quadrate Schätzung unter der Nullhypothese, d.h. gleich $\widehat{\boldsymbol{\theta}}_0$ in der obigen Notation). Da eine F-Verteilung mit einem Freiheitsgrad im Zähler nichts anderes ist als die Verteilung des Quadrats einer t-verteilten Grösse, können wir auch die Teststatistik

$$\tau_k(\theta_k^*) = \frac{\operatorname{sign}(\theta_k^* - \widehat{\theta}_k)}{\widehat{\sigma}} \sqrt{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})}$$

betrachten. Unter der Nullhypothese hat diese eine t-Verteilung mit $n-p$ Freiheitsgraden. Durch Umkehrung des Tests erhalten wir das folgende Vertrauensintervall für θ_k:

$$\left\{ \theta_k^* \mid \sqrt{S(\widehat{\boldsymbol{\theta}}^{(-k)}) - S(\widehat{\boldsymbol{\theta}})} \le t_{n-p;1-\alpha/2}\, \widehat{\sigma} \right\}.$$

Dieses Vertrauensintervall hat den Vorteil, dass es sich bei monotonen Transformationen von θ_k einfach mittransformiert. Dies ist nicht der Fall beim Intervall $\widehat{\theta}_k \pm t_{n-p;1-\alpha/2}\, se(\widehat{\theta}_k)$. Unterschiede zwischen beiden Intervallen zeigen an, dass sich die Nichtlinearitäten des Modells auswirken. Daher plottet man häufig τ_k gegen θ_k^*, um anhand der Krümmung einen Eindruck von der Stärke der Nichtlinearität zu erhalten.

Für $p = 2$ kann man natürlich auch die Niveaulinien von $S(\theta_1, \theta_2)$ plotten. Im linearen Fall sind es Ellipsen, so dass man die Nichtlinearität auf Grund der Abweichung von einer Ellipsenform beurteilen kann. Die Grenzen eines simultanen Vertrauensbereichs sind gerade diese Niveaulinien. Je stärker sich diese daher von Kreisen unterscheiden, desto stärker ist die Abhängigkeit zwischen den beiden geschätzten Parametern. In der gleichen Figur kann man noch die beiden sogenannten Profilspuren $\widehat{\theta}_2^{(-1)}(\theta_1^*)$ gegen θ_1^* und analog $\widehat{\theta}_1^{(-2)}(\theta_2^*)$ gegen θ_2^* einzeichnen. Sie kreuzen sich im Punkt $\widehat{\boldsymbol{\theta}}$, und der Schnittwinkel zeigt an, wie stark die Abhängigkeiten zwischen den geschätzten Parametern sind. Die Profilspuren. schneiden die Niveaulinien dort, wo die Tangenten horizontal, bzw. vertikal sind, weil der Gradient senkrecht steht auf den Niveaulinien. Wenn das Modell linear ist, ist S quadratisch und die Profilspuren sind Geraden (vgl. Lemma 1.5.1). Damit sieht das Bild genau aus wie in der Abbildung 1.13 (die beiden Regressionsgeraden dort entsprechen genau den beiden Profilspuren). Für die Daten von Abbildung 2.3 und 2.4 sind die Niveaulinien und Profilspuren in Abbildung 2.5 dargestellt. Wie man sieht, sind die Effekte der Nichtlinearität bei der Reaktionsgeschwindigkeit relativ harmlos, während sie beim Sauerstoffverbrauch ziemlich extrem sind.

Wenn $p > 2$ ist, wird es schwieriger, den Effekt der Nichtlinearitäten und die Abhängigkeiten zwischen den Parametern zu visualisieren. Im Prinzip könnte man die Niveaulinien der sogenannten Profil-Likelihood

$$\min_{\boldsymbol{\theta};\theta_j=\theta_j^*,\theta_k=\theta_k^*} S(\boldsymbol{\theta})$$

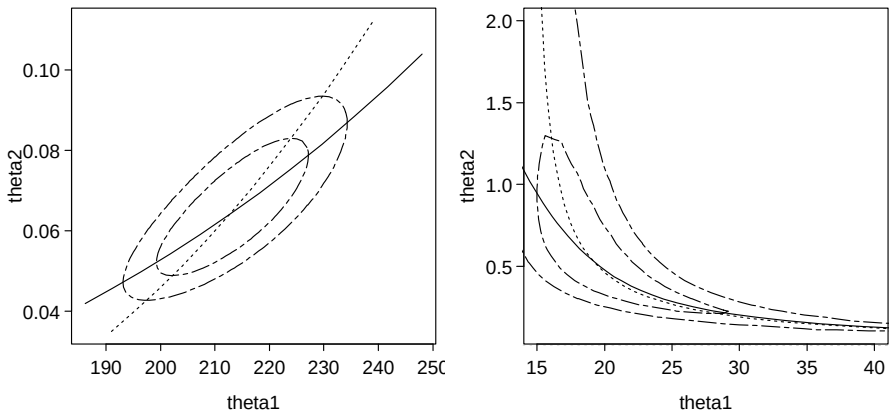

Abbildung 2.5: Niveaulinien und Profilspuren bei den Beispielen Reaktionsgeschwindigkeit (links) und Sauerstoffverbrauch (rechts). Die Daten sind in den Abbildungen 2.4 und 2.3 zu sehen.

berechnen für alle Paare (j, k), doch dies ist oft zu rechenaufwändig. Als Ersatz kann man wenigstens die Paare von Profilspuren zeichnen. Man plottet also $\widehat{\theta}_j^{(-k)}(\theta_k^*)$ gegen θ_k^* und $\widehat{\theta}_k^{(-j)}(\theta_j^*)$ gegen θ_j^* für alle $j < k$. Die Krümmung dieser Profilspuren zeigt die Nichtlinearität an, und der Winkel zwischen den beiden Profilspuren zeigt an, wie gross die Abhängigkeit zwischen den entsprechenden Parameterschätzungen ist. Ausserdem schneiden die Profilspuren auch wieder die Niveaulinien der Profil-Likelihood dort, wo die Tangenten horizontal, bzw. vertikal sind. Daraus erhält man einen Eindruck, wie die Niveaulinien verlaufen müssen.

2.3 Verallgemeinerte Lineare Modelle

2.3.1 Logistische Regression

In vielen Anwendungen ist die Zielvariable Y binär (z.B. in der Medizin geheilt/gestorben) und die Erfolgswahrscheinlichkeit $P[Y = 1]$ hängt von erklärenden Variablen ab. Modelle, die diese Wahrscheinlichkeit als lineare Funktion der erklärenden Variablen darstellen, sind selten sinnvoll, da die Erfolgswahrscheinlichkeit ja zwischen null und eins sein sollte. In der logistischen Regression verwendet man das Modell

$$\log\left(\frac{P_{\boldsymbol{\theta}}[Y_i = 1]}{P_{\boldsymbol{\theta}}[Y_i = 0]}\right) = \sum_{j=1}^{p} x_{ij}\theta_j = \mathbf{x}_i^T\boldsymbol{\theta}$$

(wenn die $\mathbf{x}_i$ auch zufällig sind, dann sind auf der linken Seite bedingte Wahrscheinlichkeiten gegeben $\mathbf{x}_i$ gemeint). Auflösen nach $P_{\boldsymbol{\theta}}[Y_i = 1]$ ergibt

$$P_{\boldsymbol{\theta}}[Y_i = 1] = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta})}{1 + \exp(\mathbf{x}_i^T\boldsymbol{\theta})} = P[U \geq -\mathbf{x}_i^T\boldsymbol{\theta}],$$

wobei U die sogenannte logistische Verteilung hat:

$$P[U \leq u] = P[U \geq -u] = \frac{\exp(u)}{1 + \exp(u)} = \int_{-\infty}^{u} \frac{e^t}{(1 + e^t)^2} dt.$$

Man kann das so interpretieren, dass man latente Variablen Z_i hat, welche einem linearen Modell mit logistischen Fehlern ε_i genügen:

$$Z_i = \mathbf{x}_i^T \boldsymbol{\theta} + \varepsilon_i.$$

Beobachtet wird aber nicht Z_i sondern nur der Indikator $Y_i = 1_{[Z_i \geq 0]}$. Wenn man statt logistischer Fehler normalverteilte nimmt, erhält man das sogenannte Probit-Modell. Die Unterschiede der beiden Modelle sind meist klein, das logistische Modell ist rechnerisch einfacher.

Die Parameterschätzung erfolgt praktisch immer mit Maximum-Likelihood. Man sieht leicht, dass für beliebiges $y \in \{0, 1\}$ gilt

$$P_{\boldsymbol{\theta}}[Y_i = y] = \left(\frac{P_{\boldsymbol{\theta}}[Y_i = 1]}{P_{\boldsymbol{\theta}}[Y_i = 0]} \right)^y P_{\boldsymbol{\theta}}[Y_i = 0] = \exp(y \cdot \mathbf{x}_i^T \boldsymbol{\theta} - \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}))).$$

Also ist die log-Likelihoodfunktion (unter der Annahme, dass die Beobachtungen unabhängig sind)

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i \mathbf{x}_i^T \boldsymbol{\theta} - \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}))).$$

Dies ist eine konkave Funktion und das Maximum ist bestimmt durch

$$\sum_{i=1}^n (y_i - P_{\widehat{\boldsymbol{\theta}}}[Y_i = 1]) \mathbf{x}_i = 0.$$

Diese Gleichungen werden numerisch mit iterativen Methoden gelöst. Wenn man bei Versuchsbedingung $\mathbf{x}_i$ mehrere Beobachtungen $(y_{ij}; 1 \leq j \leq n_i)$ hat, dann hängt $\ell(\boldsymbol{\theta})$ nur von den Summen $y_{i+} = \sum_j y_{ij}$ und den Totalen n_i ab.

Vertrauensintervalle und Tests in diesem Modell beruhen auf der asymptotischen Approximation durch eine Normalverteilung:

$$\widehat{\boldsymbol{\theta}} \overset{\text{asymptotisch}}{\sim} \mathcal{N}(\boldsymbol{\theta}, V(\boldsymbol{\theta})).$$

Die asymptotische Kovarianzmatrix $V(\boldsymbol{\theta})$ von $\widehat{\boldsymbol{\theta}}$ ist dabei die Inverse der *Fisher-Information* (siehe Abschnitt 1.9, bzw. die Mathematische Statistik):

$$V(\boldsymbol{\theta})^{-1} = I(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mathbf{E}\left[(y_i - P_{\boldsymbol{\theta}}[Y_i = 1])^2\right] = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \frac{\exp(\mathbf{x}_i^T \boldsymbol{\theta})}{(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}))^2}.$$

Zum Vergleich zweier geschachtelter Modelle mit den Dimensionen p, bzw. $q < p$ stützt man sich auf das Doppelte des log-Likelihood-Quotienten

$$2(\ell(\widehat{\boldsymbol{\theta}}^{(p)}) - \ell(\widehat{\boldsymbol{\theta}}^{(q)})),$$

welcher asymptotisch Chiquadrat-verteilt ist mit $(p - q)$ Freiheitsgraden.

2.3.2 Allgemeiner Fall

Im allgemeinen Fall hat man Beobachtungen Y_i , welche unabhängig sind und eine Dichte, bzw. Wahrscheinlichkeitsfunktion der Form

$$p_{\beta_i}(y_i) = \exp(y_i \beta_i + c(\beta_i)) h(y_i)$$

haben (sogenannte exponentielle Familie). In einem solchen Modell gilt

$$\mathbf{E}\,[Y_i] = \mu(\beta_i) = -c'(\beta_i).$$

(dies folgt durch Ableiten von $\int p_\beta(y)dy = 1$ nach β und Vertauschen von Integration und Ableitung). Viele häufig verwendete Verteilungen bilden eine solche exponentielle Familie, unter anderem die Normal-, die Binomial- und die Poisson- Verteilung. Im Fall der Normalverteilung ist

$$p(y) = \exp(y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}).$$

Bei bekanntem σ hat man also eine exponentielle Familie mit

$$\beta = \frac{\mu}{\sigma^2}, \quad c(\beta) = -\frac{1}{2}\sigma^2\beta^2.$$

Die Binomial(n, p)-Verteilung lässt sich schreiben als

$$p(y) = \left(\frac{p}{1-p}\right)^y (1-p)^n \binom{n}{y}.$$

Dies ist eine exponentielle Familie mit

$$\beta = \log\left(\frac{p}{1-p}\right), \quad c(\beta) = -n\log(1+e^\beta).$$

Ebenso bilden die Poisson(λ)-Verteilungen eine exponentielle Familie mit

$$\beta = \log\lambda, \quad c(\beta) = -e^\beta.$$

Der Effekt der erklärenden Variablen $\mathbf{x}_i$ auf die Beobachtung Y_i wird nun in einem verallgemeinerten linearen Modell beschrieben durch einen Zusammenhang zwischen $\mathbf{x}_i$ und dem Parameter β_i für die i-te Beobachtung, und zwar soll es eine Funktion g geben, so dass:

$$g(\mu(\beta_i)) = \mathbf{x}_i^T\boldsymbol{\theta},$$

d.h. eine geeignete Transformation von β_i soll linear in den erklärenden Variablen sein. Die Funktion g heisst *Link-Funktion*. Wenn g gerade gleich μ^{-1} ist, dann spricht man von der *kanonischen* Linkfunktion. Das lineare Modell mit normalverteilten Fehlern und die logistische Regression sind also Beispiele verallgemeinerter linearer Modelle mit kanonischer Linkfunktion.

Die Behandlung im allgemeinen Fall geht analog wie im Fall der logistischen Regression, d.h. man verwendet die Maximum-Likelihood-Schätzung und benützt für Tests und Vertrauensintervalle die asymptotische Normalität dieser Schätzung, bzw. die asymptotische Chiquadrat-Verteilung für den Likelihoodquotienten. Wir verweisen für die Details auf die Literatur.

2.4 Cox-Regression

In medizinischen und technischen Anwendungen ist die Zielvariable oft eine Überlebens- oder Ausfallzeit. Im Prinzip könnte man ein verallgemeinertes lineares Modell mit Exponential- oder Gamma-Verteilungen verwenden. In der Praxis hat sich jedoch das Cox-Modell durchgesetzt, das nicht an einen speziellen Verteilungstyp gebunden ist. Wenn F

eine Verteilungsfunktion auf den positiven reellen Zahlen mit einer Dichte f ist, dann ist die Ausfallrate (auch Hazard- oder Risiko-Funktion genannt) definiert als

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} P[t \leq T \leq t + h | T \geq t] = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} \log(1 - F(t)).$$

Die Ausfallrate legt also die Verteilung fest, und zwar gilt

$$F(t) = 1 - \exp(-\int_0^t \lambda(u) du).$$

Das Cox-Modell postuliert nun, dass die i-te Ausfallrate in Abhängigkeit der erklärenden Variablen $\mathbf{x}_i$ die Form hat

$$\lambda_i(t) = \exp(\mathbf{x}_i^T \boldsymbol{\theta}) \lambda_0(t),$$

wobei λ_0 eine nicht näher spezifizierte Basisrate ist. In diesem Modell darf es natürlich keinen Achsenabschnitt geben, weil eine Konstante in λ_0 absorbiert werden kann. Eine Erhöhung um eine Einheit der j-ten Komponente der erklärenden Variablen bringt also eine multiplikative Erhöhung der Ausfallrate um den Faktor $\exp(\boldsymbol{\theta}_j)$ (gleichmässig für alle Zeiten). Deshalb heisst dieses Modell auch "proportionales Hazard-Modell". Man sieht leicht, dass bei einer strikt monotonen und differenzierbaren Transformation der Überlebenszeiten ein Cox-Modell wieder in ein Cox-Modell mit gleichen Parametern, aber anderem λ_0 übergeführt wird. Wenn man λ_0 nicht spezifiziert, heisst das einfach, dass die Wahl der Zeitskala offen bleibt.

In der Likelihoodfunktion kommt natürlich auch die Basisrate λ_0 vor, weshalb man nicht einfach den Maximum Likelihood Schätzer verwenden kann. Zur Schätzung der Parameter $\boldsymbol{\theta}$ maximiert man stattdessen die sogenannte partielle Likelihood, welche definiert ist als

$$\prod_{i=1}^{n} \frac{\exp(\mathbf{x}_i^T \boldsymbol{\theta})}{\sum_{j; t_j \geq t_i} \exp(\mathbf{x}_j^T \boldsymbol{\theta})}.$$

Der i-te Faktor ist die bedingte Wahrscheinlichkeit, dass die i-te Beobachtungseinheit im Intervall $[t_i, t_i + dt)$ ausfällt, gegeben dass eine der Einheiten, die unmittelbar vor t_i noch in Betrieb sind, ausfällt. Von den Ausfallzeiten wird nur die Information über die Reihenfolge des Ausfalls verwendet.

Bei fast allen Daten dieses Typs hat man als zusätzliche Komplikation sogenannte zensierte Beobachtungen, bei denen man von der Ausfallzeit T_i nicht den genauen Wert weiss, sondern nur, dass T_i grösser ist als eine beobachtete Zensierungszeit C_i. Gründe dafür sind, dass die Studie beendet wird, bevor alle Einheiten ausgefallen sind, oder dass Patienten wegziehen oder an andern Krankheiten sterben. Die partielle Likelihood lässt sich in diesen Fällen analog definieren: Man bildet das Produkt über alle unzensierten Beobachtungen, summiert aber im Nenner jeweils über alle unzensierten Beobachtungen mit $t_j \geq t_i$ und alle zensierten Beobachtungen mit $c_j \geq t_i$.

Für Tests und Vertrauensintervalle stützt man sich ebenfalls auf asymptotische Näherungen ab, auf die wir hier nicht eingehen.

2.5 Nichtparametrische Regression

In den letzten 20-30 Jahren haben Verfahren sehr viel Aufmerksamkeit gewonnen, die im Modell

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

keine parametrischen Annahmen über die Form von f machen, sondern nur annehmen, dass f eine glatte Funktion ist. Diese Abschwächung von Voraussetzungen ist natürlich ein grosser Vorteil, vor allem in explorativen Untersuchungen, wo man möglichst gegenüber den Daten offen sein will.

Man unterscheidet den deterministischen Fall, wo die $\mathbf{x}_i$ regelmässig angeordnet sind, und den stochastischen Fall, wo die $\mathbf{x}_i$ Realisierungen von Zufallsvariablen mit einer Verteilung G sind, unabhängig von den Fehlern ε_i. Ein noch allgemeineres Modell nimmt an, dass die $(\mathbf{X}_i, Y_i)$ i.i.d Zufallsvektoren sind mit einer beliebigen gemeinsamen Verteilung und dass man

$$f(\mathbf{x}) = \mathbf{E}\left[Y_i \mid \mathbf{X}_i = \mathbf{x}\right]$$

schätzen möchte. Man kann zeigen, dass für jede beliebige Funktion g gilt

$$\mathbf{E}\left[(Y_i - f(\mathbf{X}_i))^2\right] \leq \mathbf{E}\left[(Y_i - g(\mathbf{X}_i))^2\right]$$

(sofern die zweiten Momente existieren). Das heisst $f(\mathbf{X}_i)$ ist die beste Prognose von Y_i gestützt auf $\mathbf{X}_i$ im Sinne des mittleren quadratischen Fehlers. Die bedingte Fehlervarianz

$$\mathbf{E}\left[(Y_i - f(\mathbf{x}_i))^2 \mid \mathbf{X}_i = \mathbf{x}_i\right],$$

welche $\mathrm{Var}[\varepsilon_i]$ entspricht, hängt dann aber im Allgemeinen von $\mathbf{x}_i$ ab.

Alle nichtparametrischen Schätzungen von $f(\mathbf{x})$ beruhen im Wesentlichen auf einer Mittelung der y_i für diejenigen i, für die $\mathbf{x}_i$ nahe bei $\mathbf{x}$ ist. Es gibt aber grosse Unterschiede darin, wie gemittelt wird, und wie man festlegt, was nahe bei $\mathbf{x}$ heisst.

2.5.1 Einige Verfahren im eindimensionalen Fall

Kernschätzer:

Die Schätzung von $f(x)$ ist das gewichtete Mittel der y_i, wobei das Gewicht von y_i monoton mit dem Abstand $|x - x_i|$ abnimmt. Die Gewichte werden über einen sogenannten Kern K und eine Bandbreite $h > 0$ festgelegt. Ein Kern ist eine bezüglich 0 symmetrische Wahrscheinlichkeitsdichte, die entweder den Träger $[-1, 1]$ hat oder sonst sehr rasch abfällt (wie z.B. die Normalverteilungsdichte). Es gibt zwei Varianten, die sich im stochastischen Fall wesentlich unterscheiden. Die Nadaraya-Watson-Version ist definiert als

$$\widehat{f}(x) = \frac{\sum_{i=1}^n y_i K((x - x_i)/h)}{\sum_{i=1}^n K((x - x_i)/h)}.$$

Für die Gasser-Müller Version nehmen wir an, dass

$$0 \leq x_1 < x_2 < \ldots < x_n \leq 1,$$

und wir setzen $s_0 = -\infty$, $s_i = (x_i + x_{i+1})/2$ für $0 < i < n$ und $s_n = +\infty$. Die geschätzte Funktion ist dann

$$\widehat{f}(x) = \sum_{i=1}^n y_i \int_{s_{i-1}}^{s_i} \frac{1}{h} K((x - u)/h) du.$$

Die Unterschiede zwischen diesen beiden Versionen sind von Bedeutung, falls die x_i sehr unregelmässig verteilt sind. Die Unterschiede sind in Figur 2.6 illustriert. Wir sehen, dass

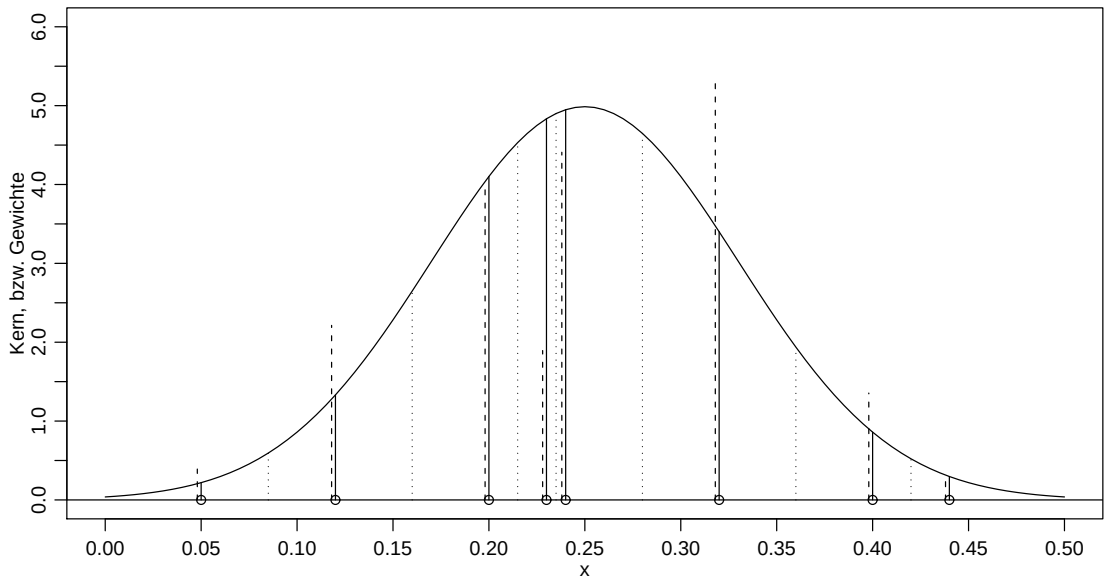

Abbildung 2.6: Gewichte der Nadaraya-Watson (ausgezogen) und der Gasser-Müller Version (gestrichelt). Die Beobachtungsstellen x_i sind durch Kreise auf der x-Achse dargestellt, die Grenzen s_i durch die gepunkteten Linien. Die Gewichte sind mit einem festen Faktor multipliziert.

die Gasser-Müller Version Beobachtungen mehr Gewicht gibt, wenn die zugehörigen x_i isoliert sind.

Bei beiden Versionen reguliert die Bandbreite h die Glattheit von $\widehat{f}$: Je grösser h, desto glatter die Schätzung, aber desto weniger passt sie sich den Daten an.

Wenn die x_i zufällig sind, dann kann die Anzahl Beobachtungen mit wesentlich von null verschiedenem Gewicht sehr stark mit x variieren. Bei der ersten Version kann sie sogar null sein. Um das zu vermeiden, kann man stattdessen eine variable Bandbreite so wählen, dass zum Beispiel die Anzahl Beobachtungen mit $x - h \leq x_i \leq x + h$ konstant bleibt. Man spricht dann von einem Nächste-Nachbar-Schätzer.

Die Kernschätzer haben vor allem Schwierigkeiten mit der Schätzung von f in der Nähe des Randes, d.h. für $x < h$ und $x > 1 - h$. Weil man dort über Beobachtungen mittelt, welche fast alle auf einer Seite von x liegen, ergeben sich systematische Fehler.

Lokale Polynome:

Hier nimmt man nicht mehr an, dass die Funktion f lokal konstant ist, sondern man verwendet lokal ein Polynom vom Grad $p > 0$, welches mit gewichteten Kleinsten Quadraten angepasst wird. Das bedeutet

$$\widehat{f}(x) = \widehat{\theta}_0(x),$$

wobei

$$\widehat{\boldsymbol{\theta}}(x) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} K((x - x_i)/h)(y_i - \sum_{j=0}^{p} \theta_j (x_i - x)^j)^2.$$

Es stellt sich heraus, dass es besser ist, p ungerade zu wählen. In der Praxis benutzt man meist $p = 1$ oder $p = 3$. Die Vorteile sind vor allem am Rand deutlich sichtbar.

Anstatt einer festen Bandbreite kann man wieder die Nächste-Nachbar Variante verwenden, wo man für eine feste Anzahl Beobachtungen in $[x-h, x+h]$ sorgt. Dies ist die Grundidee der Funktion `loess` in den Statistikprogrammen S-Plus und R.

Glättungssplines:

Ein Glättungsspline ist implizit definiert als Lösung des Minimierungsproblems

$$\arg\min_f \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 f''(x)^2 dx.$$

Der erste Term, die Summe über i, gibt an, wie genau sich f an die Beobachtungen anpasst. Der zweite Term, die L_2-Norm der zweiten Ableitung, misst, wie glatt f ist, und der Parameter λ reguliert, wie man den Kompromiss zwischen den gegensätzlichen Zielen festlegt, beide Terme klein zu machen.

Man kann zeigen, dass die Lösung des Minimierungsproblems ein kubischer Spline mit Knoten an den Stellen x_i ist, der ausserdem noch linear auf den Randintervallen $[0, x_1]$ und $[x_n, 1]$ ist. Für $\lambda \to 0$ erhält man den Spline, der die Daten interpoliert, und für $\lambda \to \infty$ die Kleinste Quadrate Gerade. Die Rolle von λ entspricht also der Rolle der Bandbreite h.

Zur Berechnung des Glättungssplines wählt man eine Basis im Vektorraum der Splines mit Knoten x_i. Die Bestimmung der Koeffizienten der Lösung bezüglich dieser Basis führt dann auf die Minimierung einer quadratischen Funktion. Für numerisch stabile und schnelle Verfahren ist die Wahl der Basis entscheidend. Bewährt haben sich die sogenannten B-Splines.

2.5.2 Bias/Varianz-Dilemma

Alle nichtparametrischen Verfahren haben einen Glättungsparameter, der das Verhalten der Schätzung stark bestimmt. Bei den Splines ist dies λ und bei den Verfahren mit Kernen die Bandbreite h. Die Rolle dieses Glättungsparameters wird klarer, wenn man Bias und Varianz der Schätzung betrachtet. Man kann zeigen, dass bei lokalen Polynomen mit ungeradem p

$$\mathbf{E}\left[\widehat{f}(x)\right] - f(x) \sim const(K,p)\, h^{p+1}\, f^{(p+1)}(x)$$

und

$$\mathrm{Var}[\widehat{f}(x)] \sim const(K,p)\, \frac{\sigma_\varepsilon^2}{nh} \left(\frac{1}{nh} \sum K((x-x_i)/h)\right)^{-1}$$

gilt. Dabei werden die $\mathbf{x}_i$ als fest angenommen (bzw. man bedingt auf diese, falls sie zufällig sind), und die Bandbreite $h = h_n$ muss $h_n \to 0$ und $nh_n \to \infty$ erfüllen. Man sieht aus den obigen Formeln, dass man für einen (absolut) kleinen Bias h möglichst klein wählen sollte, während für eine kleine Varianz h möglichst gross sein sollte. Dies erklärt den Titel dieses Abschnittes.

Eine Grösse, die den Bias und die Varianz gleichzeitig berücksichtigt, ist der mittlere quadratische Fehler:

$$\mathbf{E}\left[(\widehat{f}(x) - f(x))^2\right] = \mathrm{Var}[\widehat{f}(x)] + (\mathbf{E}\left[\widehat{f}(x)\right] - f(x))^2 = O(\frac{1}{nh}) + O(h^{2(p+1)}).$$

Dies wird von minimaler Ordnung, wenn beide Terme von gleicher Grössenordnung sind, d.h. $h = O(n^{-1/(2p+3)})$. Eine solche Wahl ergibt einen mittleren quadratischen Fehler der Ordnung $O(n^{-(2p+2)/(2p+3)})$. Von den Grössenordnungen her würde man ein möglichst grosses p wählen, aber in der Praxis ist das nicht ganz richtig, weil für grosses p auch die Konstanten gross sind und man stärkere Annahmen für f braucht. Die Konstanten sind überhaupt das grösste Problem bei der Anwendung dieser Resultate: Sie enthalten unbekannte Terme wie die Ableitungen von f, und das optimale h hängt von der Stelle x ab. Daher ist die datenabhängige, optimale Wahl der Bandbreite ziemlich diffizil.

Man kann auch den optimalen Kern K bestimmen (durch Minimieren von $const(K, p)$). Es stellt sich jedoch heraus, dass die Wahl von K sekundär ist: Fast alle stetigen Kerne sind praktisch gleich gut.

Bei den Nadaraya-Watson-Kernschätzern hat der Bias eine kompliziertere Form, aber die Varianz ist asymptotisch gleich wie bei den lokalen Polynomen mit $p = 1$. Bei der Gasser-Müller-Variante ist der Bias gleich wie bei den lokalen Polynomen mit $p = 1$, dafür ist die Konstante bei der Varianz 1.5 mal so gross wie beim lokalen Polynom vom Grad $p = 1$. Auch für die Glättungssplines kennt man das asymptotische Verhalten von Bias und Varianz: Es ist analog wie bei einem lokalen Polynom vom Grad $p = 3$.

2.5.3 Fluch der Dimension

Im Prinzip können die Kernschätzer und die lokalen Polynome direkt auf mehr als eine Dimension verallgemeinert werden. In der Praxis versagen sie aber meistens schon ab Dimension 3 oder 4. Dies hat damit zu tun, dass der Raum in hohen Dimensionen sehr gross wird und nur sehr schlecht mit einer endlichen Anzahl Punkten überdeckt werden kann. Mit andern Worten: Zwei verschiedene $\mathbf{x}_i$'s liegen fast immer sehr weit auseinander und es ist kein vernünftiger Kompromiss im Bias-Varianz-Dilemma mehr möglich. Dieser "Fluch der Dimension" wird im folgenden Beispiel sehr gut illustriert. Wenn die $\mathbf{x}_i$ gleichverteilt sind im Würfel $[-1, 1]^p$, dann ist der Anteil Punkte, die in der Einheitskugel $\{\mathbf{x}; ||\mathbf{x}|| \leq 1\}$ liegen, ungefähr gleich der Wahrscheinlichkeit, dass ein $\mathbf{x}_i$ in diese Einheitskugel fällt, d.h. gleich dem Volumen der Einheitskugel mal 2^{-p}. Für $p = 2$ ist dies 79%, für $p = 5$ noch 16% und für $p = 10$ nur noch 0.25% ! Das Verhältnis der Durchmesser von Kugel und Würfel ist hingegen $1 : \sqrt{p}$, also für $p = 10$ ungefähr $1 : 3$. Die Annahme, dass f konstant oder linear in der Einheitskugel ist, unterscheidet sich also nicht wesentlich von der Annahme, dass f auf dem ganzen Würfel konstant oder linear ist, aber trotzdem hat man innerhalb der Einheitskugel praktisch immer zu wenige Beobachtungen, um eine konstante oder lineare Funktion gut schätzen zu können.

Anhang A

Resultate aus der Wahrscheinlichkeitstheorie

A.1 Rechenregeln für Momente

Es seien:

- X, Y, Z Zufallsvariablen
- a, b Konstanten

Regeln für den Erwartungswert:

- $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$.
- $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.
- $\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y]$ falls X und Y unabhängig oder zumindest unkorreliert sind.

Regeln für die Varianz und die Kovarianz:

- $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$, woraus folgt: $\sigma[aX + b] = |a|\sigma[X]$
- $\mathrm{Cov}[X, Y] := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] = \mathrm{Cov}[Y, X]$.
- $\mathrm{Var}[X + Y] = \mathrm{Var}\,X + \mathrm{Var}\,Y + 2\,\mathrm{Cov}[X, Y]$
- Speziell: $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ falls X und Y unabhängig oder zumindest unkorreliert sind.
- $\mathrm{Cov}[aX + bY, Z] = a\,\mathrm{Cov}[X, Z] + b\,\mathrm{Cov}[Y, Z]$.

Momente von Zufallsvektoren:

Sei $\mathbf{Y}$ ein $n \times 1$-Zufallsvektor. Wir definieren $\mathbf{E}[\mathbf{Y}]$ als den Vektor mit Komponenten $\mathbf{E}[y_i]$ und die Kovarianzmatrix von $\mathbf{Y}$ als

$$\mathrm{Cov}[\mathbf{Y}] = (\mathrm{Cov}(y_i, y_j)_{ij}) = \begin{pmatrix} \mathrm{Var}[y_1] & \mathrm{Cov}[y_1, y_2] & \dots & \mathrm{Cov}[y_1, y_n] \\ \mathrm{Cov}[y_2, y_1] & \mathrm{Var}[y_2] & \dots & \mathrm{Cov}[y_2, y_n] \\ \dots & \dots & \dots & \dots \\ \mathrm{Cov}[y_n, y_1] & \mathrm{Cov}[y_n, y_2] & \dots & \mathrm{Var}[y_n] \end{pmatrix}$$

Eine Kovarianzmatrix ist also stets symmetrisch.

Es seien A eine feste $m \times n$-Matrix, $\mathbf{b}$ ein fester $m \times 1$-Vektor. Dann gelten analog wie im skalaren Fall:

- $\mathrm{Cov}[\mathbf{Y}] = \mathbf{E}\left[\mathbf{Y}\mathbf{Y}^T\right] - \mathbf{E}[\mathbf{Y}]\,\mathbf{E}[\mathbf{Y}]^T$.
- $\mathbf{E}[A\mathbf{Y} + \mathbf{b}] = A\mathbf{E}[\mathbf{Y}] + \mathbf{b}$,
- $\mathrm{Cov}[A\mathbf{Y} + \mathbf{b}] = A \cdot \mathrm{Cov}[\mathbf{Y}] \cdot A^T$.

Insbesondere erhalten wir aus der letzten Regel für einen beliebigen $n \times 1$ Vektor $\mathbf{a}$:

$$0 \leq \mathrm{Var}[\mathbf{a}^T\mathbf{Y}] = \mathrm{Cov}[\mathbf{a}^T\mathbf{Y}] = \mathbf{a}^T\,\mathrm{Cov}[\mathbf{Y}]\mathbf{a},$$

das heisst jede Kovarianzmatrix ist **positiv semidefinit** (und üblicherweise sogar **positiv definit**).

Wenn Σ eine beliebige, positiv semidefinite Matrix ist, dann existieren Matrizen A mit $AA^T = \Sigma$. Wir nennen jede solche Matrix A eine Wurzel von Σ und schreiben $A = \Sigma^{1/2}$. Man beachte, dass $\Sigma^{1/2}$ also nur bis auf Multiplikation mit einer orthogonalen Matrix bestimmt ist. Die numerisch einfachste Bestimmung von $\Sigma^{1/2}$ benutzt die Choleski-Zerlegung, welche eine untere Dreiecksmatrix liefert. Bei Rechnungen mit $\Sigma^{1/2}$ muss man etwas aufpassen, weil z.B. $(\Sigma^{-1})^{1/2} = (\Sigma^{1/2})^{-T}$.

Wenn $\mathbf{Y}$ ein Zufallsvektor ist mit Kovarianzmatrix Σ und $A = \Sigma^{1/2}$, dann hat auf Grund obiger Regeln der Vektor $\mathbf{X} = A^{-1}\mathbf{Y}$ die Identität als Kovarianzmatrix, d.h. die Komponenten von $\mathbf{X}$ sind unkorreliert und haben Varianz 1.

A.2 Die Normalverteilung

A.2.1 Eindimensionale Normalverteilung

a) **Dichte der "Standard-Normalverteilung"** $\mathcal{N}(0,1)$**:**

$$\boxed{\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}$$

(Dies ist eine Wahrscheinlichkeitsdichte, d.h. das Integral von $\varphi(x)$ über die ganze reelle Achse ist eins). Die kumulative Verteilungsfunktion

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy$$

ist nicht in geschlossener Form beschreibbar, aber es gibt Tabellen.

b) **3 Gründe für die Wichtigkeit der Normalverteilung (als ideales Modell):**

(i) **Einfachheit** (und schöne Theorie).

(ii) **Zentraler Grenzwertsatz** (und **Hypothese der Elementarfehler):** Ein Messfehler ist zusammengesetzt aus verschiedenen kleinen, unabhängigen Elementarfehlern, welche sich **additiv überlagern**, so dass die Summe genähert eine **Normalverteilung** hat. Eine **multiplikative Überlagerung** der Elementarfehler liefert eine **logarithmische Normalverteilung**, welche bei der Logarithmus-Transformation in eine Normalverteilung übergeht.

(iii) **Erfahrung** Viele Datensätze sind genähert normalverteilt (eventuell erst nach einer geeigneten Transformation).

c) **Warum ist in gewissem Sinn die Normalverteilung so einfach ?**

Wenn $f(x)$ eine Dichte auf $\mathbb{R}$ ist mit

$$\frac{d \log f(x)}{dx} = \frac{f'}{f}(x) = ax + b$$

dann folgt daraus $f(x) = e^{\frac{1}{2}ax^2+bx+c}$. Die Grösse f'/f spielt eine wichtige Rolle in der Statistik. In diesem Sinn ist die Normalverteilung tatsächlich die einfachste stetige Verteilung auf der ganzen reellen Achse!

d) **Lineare Transformationen der Standard-Normalverteilung $\mathcal{N}(0,1)$:**

Betrachte: $X \sim \mathcal{N}(0,1)$ und die Transformation: $x \longmapsto y := \mu + \sigma x$, wobei μ beliebig ist und $\sigma > 0$. Die Verteilung von $Y := \mu + \sigma X$ ist die allgemeine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$.

Berechnung der Dichte von Y: Aus

$$f_X(x)dx = f_Y(y)dy \text{ mit } dy = \sigma dx,$$

folgt

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \underbrace{\frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}}_{\text{Dichte von } Y := \mu + \sigma X} dy$$

e) **Die Momente der Normalverteilung:**

Standardnormalverteilung: $X \sim \mathcal{N}(0,1)$.

$$\begin{aligned}
\mathbf{E}[X] &= \int_{-\infty}^{\infty} x\varphi(x)dx = 0 \\
\mathrm{Var}[X] &= \int_{-\infty}^{\infty} x \cdot x\varphi(x)dx = \underbrace{-x\varphi(x) \mid_{-\infty}^{+\infty}}_{0} + \underbrace{\int_{-\infty}^{\infty} \varphi(x)dx}_{1} = 1 \\
\mathbf{E}\left[X^3\right] &= 0 \text{ (Symmetrie)} \\
\mathbf{E}\left[X^4\right] &= \int_{-\infty}^{\infty} x^3 \cdot x\varphi(x)dx = -x^3\varphi(x) \mid_{-\infty}^{+\infty} + \int_{-\infty}^{\infty} 3x^2\varphi(x)dx = 3.
\end{aligned}$$

Allgemeine Normalverteilung: $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Erwartungswert:	$\mathbf{E}[Y]$	$=$	μ
Varianz:	$\mathrm{Var}[Y]$	$=$	σ^2
Schiefe (Standardisiertes 3. Moment):	γ_1	$=$	$\mathbf{E}\left[(Y-\mu)^3\right]/\sigma^3 = 0$
Kurtosis, Exzess (Stand. 4. Moment):	γ_2	$=$	$\mathbf{E}\left[(Y-\mu)^4\right]/\sigma^4 - 3 = 0$

f) **Form der Normalverteilung**

In Abb. A.1 ist die Dichte der Standardnormalverteilung $\mathcal{N}(0,1)$ graphisch dargestellt.

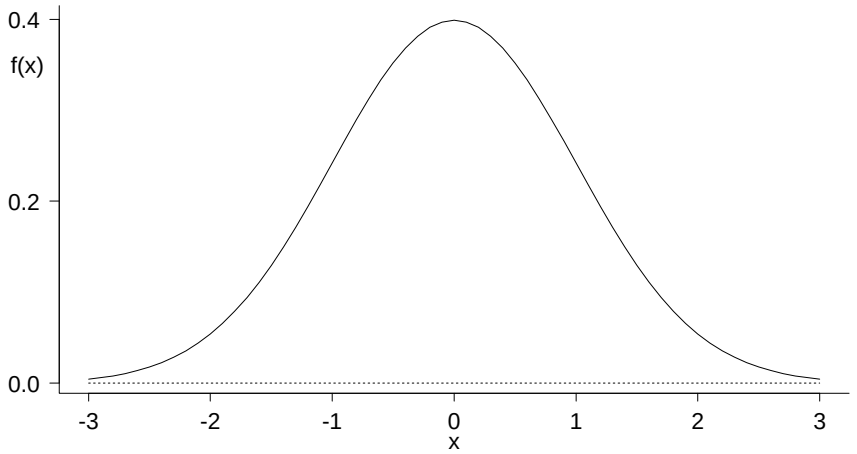

Abbildung A.1: Dichte der Standardnormalverteilung

Einige Werte der kumulativen Verteilungsfunktion:

x	0	0.6745	1	1.64	1.96	2.58	3.3
$\Phi(x)$	$\frac{1}{2}$	$\frac{3}{4}$	84 % $\approx \frac{5}{6}$	95 %	97.5	99.5	99.95 %

Die Dichte der Normalverteilung $f(x)$ geht für $x \longrightarrow \pm\infty$ relativ immer rascher gegen 0 ($|f'|/f \longrightarrow \infty$). Obwohl die Normalverteilung von $-\infty$ bis $+\infty$ eine **positive Dichte** hat, ist sie in der Praxis eine sehr **"kurzschwänzige"** Verteilung, welche ausserhalb von $\mu \pm 3\sigma$ oder $\mu \pm 4\sigma$ praktisch "verschwindet". Dies steht im Gegensatz zu den meisten empirischen Verteilungen von Messfehlern. Deshalb ist auch die **"normale Approximation"** meist nur **in der Mitte** der Verteilung brauchbar, bis ca. $\mu \pm 2\sigma$ oder $\mu \pm 2.5\sigma$.

A.2.2 Mehrdimensionale Normalverteilung

Seien $Y_1, Y_2 \ldots Y_n$ unabhängige standardnormalverteilte Zufallsvariablen . Dann ist ihre gemeinsame Dichte gleich dem Produkt der Einzeldichten, also

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-n/2} \exp(-\tfrac{1}{2}\mathbf{y}^T\mathbf{y}).$$

Dies ist die n-dimensionale Standardnormalverteilung. Sie ist sphärisch-symmetrisch. Die allgemeine n-dimensionale Normalverteilung ist per Definition die Verteilung eines Vektors $\mathbf{X}$, der durch eine lineare Transformation eines standard-normalverteilten n-dimensionalen Vektors entsteht:

$$\mathbf{X} = A\mathbf{Y} + \boldsymbol{\mu}$$

wobei $\boldsymbol{\mu}$ ein $(n \times 1)$-Vektor und A eine $(n \times n)$-Matrix ist. Wenn A singulär ist, nennt man die Verteilung degeneriert. Wir diskutieren zuerst den nicht-degenerierten Fall. Dann hat die Verteilung von $\mathbf{X}$ die Dichte

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= (2\pi)^{-n/2}(|\det A|)^{-1} \exp(-\tfrac{1}{2}(A^{-1}(\mathbf{x}-\boldsymbol{\mu}))^T(A^{-1}(\mathbf{x}-\boldsymbol{\mu}))) \\ &= (2\pi)^{-n/2}(\det \Sigma)^{-1/2} \exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})). \end{aligned}$$

wobei $\Sigma = AA^T$. Gemäss den allgemeinen Regeln gilt

$$\mathbf{E}[\mathbf{X}] = A\mathbf{E}[\mathbf{Y}] + \boldsymbol{\mu} = \boldsymbol{\mu}$$

und

$$\mathrm{Cov}[\mathbf{X}] = \mathbf{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] = \mathbf{E}\left[A\mathbf{Y}\mathbf{Y}^TA^T\right] = A\mathbf{E}\left[\mathbf{Y}\mathbf{Y}^T\right]A^T = AA^T = \Sigma.$$

Wie im univariaten Fall ist die Normalverteilung also bestimmt durch die ersten beiden Momente, und wir schreiben daher $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$. Als Kovarianzmatrix Σ kann eine beliebige positiv definite Matrix auftreten. Die gemeinsame Dichte ist maximal im Punkt $\boldsymbol{\mu}$ und konstant auf ähnlichen Ellipsoiden mit Zentrum $\boldsymbol{\mu}$ (die Hauptachsen der Ellipsoide sind gegeben durch die Eigenvektoren von Σ).

Die beiden folgenden Resultate folgen sofort aus der Definition:

Satz A.2.1. *Unkorrelierte, gemeinsam normalverteilte Zufallsvariablen sind unabhängig. Genauer: Wenn $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ und wenn $\Sigma_{ij} = 0$ für alle $i \in I$ und $j \in J$ für zwei disjunkte Indexmengen $I, J \subset \{1, \ldots, n\}$, dann sind $(X_i, i \in I)$ und $(X_j, j \in J)$ voneinander unabhängig.*

Beweis Die gemeinsame Dichte zerfällt in das Produkt der Einzeldichten.

Satz A.2.2. *Standardnormalverteilte Zufallsvariablen gehen durch orthogonale Transformation wieder in ebensolche über.*

Mit etwas mehr Aufwand kann man auch das Folgende zeigen:

Satz A.2.3. *Linearkombinationen gemeinsam normalverteiler Zufallsvariablen sind wieder gemeinsam normalverteilt.*

Beweis Für Linearkombinationen $A\mathbf{X} + \mathbf{b}$ mit einer $n \times n$ Matrix A ist das unmittelbar klar aus der Definition der n-dimensionalen Normalverteilung. Für den Fall von weniger als n Linearkombinationen nehmen wir ohne Beschränkung der Allgemeinheit an, dass $\mathbf{X}$ standardnormalverteilt ist. Zur Vereinfachung betrachten wir eine einzelne Linearkombination $\mathbf{a}^T\mathbf{X} = \sum_{i=1}^n a_i X_i$ und nehmen an, dass $\mathbf{a}$ Länge eins hat. Dann wählen wir eine orthogonale Matrix A mit erster Zeile gleich $\mathbf{a}^T$, d.h. $\mathbf{a}^T\mathbf{X}$ ist gerade die erste Komponente von $A\mathbf{X}$. Damit ist klar, dass $\mathbf{a}^T\mathbf{X}$ normalverteilt ist, denn die Komponenten von $A\mathbf{X}$ sind unabhängig und normalverteilt.

Damit kann man auch die Struktur der degenerierten Normalverteilung verstehen:

Satz A.2.4. *Wenn $\mathbf{X}$ eine degenerierte n-dimensionale Normalverteilung hat, dann existieren r Komponenten $(X_{i_1}, X_{i_2}, \ldots X_{i_r})$ welche eine nichtdegenerierte r-dimensionale Verteilung haben, während die restlichen $n-r$ Komponenten sich als Linearkombinationen davon darstellen lassen. Dabei ist r der Rang von Σ.*

Ferner folgt aus Satz A.2.3 insbesondere:

Korollar A.2.1. *Die n Randverteilungen einer n-dimensional normalverteilten Zufallsvariablen sind alle auch Normalverteilungen.*

Allerdings gilt die Umkehrung dieses Korollars nicht. Falls alle Randverteilungen einer Zufallsvariablen normalverteilt sind, so ist die gemeinsame Verteilung nicht notwendigerweise auch eine Normalverteilung!

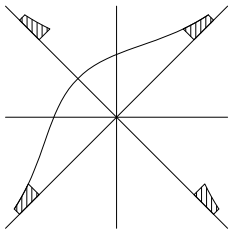

Abbildung A.2: Trotz normalverteilter Randverteilungen braucht die gemeinsame Verteilung keine Normalverteilung zu sein.

Betrachte dazu das folgende Beispiel:
Sei U eine eindimensionale standardnormalverteilte Zufallsvariable. Wir setzen $X = Y = U$, d.h. die gemeinsame Verteilung von (X, Y) ist auf der Diagonalen konzentriert, und die Randverteilungen sind sicher normal. Schneidet man nun (wie in der Abb. A.2 gezeichnet) die gemeinsame Verteilung irgendwo symmetrisch ab und setzt das abgeschnittene Stück auf der anderen Diagonalen wieder hinzu, so haben wir **keine gemeinsame Normalverteilung** mehr, jedoch die Projektionen, welche die Randverteilungen darstellen, sind immer noch die gleichen, also normalverteilt!

Man beachte ferner, dass

$$\begin{array}{ll} \text{Abschneidepunkt} \longrightarrow \infty & \Longrightarrow \text{Korrelation } +1 \\ \text{Abschneidepunkt} \longrightarrow 0 & \Longrightarrow \text{Korrelation } -1 \end{array}$$

dazwischen ist die Korrelation stetig variierend, also ist sie irgendwo Null. Die beiden Variablen sind jedoch stets abhängig. Also sieht man hier auch, dass ohne gemeinsame Normalverteilung Unkorreliertheit nicht Unabhängigkeit impliziert.

A.2.3 Chiquadrat-, t- und F-Verteilung

Diese Verteilungen sind aus der Normalverteilung abgeleitet und spielen eine wichtige Rolle bei verschiedenen Tests in der Regression. Es seien:

$$X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n \text{ unabhängig } \sim \mathcal{N}(0,1).$$

Dann heisst die Verteilung von

$$Z_m = \sum_{i=1}^{m} X_i^2$$

Chiquadrat-Verteilung mit m Freiheitsgraden (in Formeln χ_m^2). Insbesondere gilt $\mathbf{E}\left[Z_m\right] = m$, $Var[Z_m] = 2m$ und

$$\mathcal{L}\left(\frac{Z_m - m}{\sqrt{2m}}\right) \stackrel{m\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

($\mathcal{L}$ steht für law, (Verteilungs)–Gesetz, und Konvergenz ist im Sinne der schwachen Konvergenz).

Die Verteilung von

$$V_n = \frac{X_1}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}$$

heisst t-**Verteilung** mit n Freiheitsgraden (in Formel t_n). Insbesondere ist t_1 die (Standard)-"Cauchy-Verteilung", und es gilt

$$\mathcal{L}(V_n) \stackrel{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

Die Verteilung von

$$W_{m,n} = \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{i=1}^n Y_i^2}$$

heisst F-**Verteilung** mit m Freiheitsgraden im Zähler und n Freiheitsgraden im Nenner (in Formeln $F_{m,n}$). Insbesondere gilt

$$\mathcal{L}(W_{m,n}) \to \frac{1}{m}\chi_m^2 \quad (m \text{ fest}, n \to \infty)$$

und

$$\mathcal{L}(W_{m,n}) \to 1, \quad (m \to \infty, n \to \infty).$$

Zur Berechnung der Dichten dieser Verteilungen
Wir beginnen mit χ_1^2, d.h. gegeben sei $X \sim \mathcal{N}(0,1)$ und gesucht ist $\mathcal{L}(X^2)$. Lösung:

$$P[X^2 \le c] = P[|X| \le \sqrt{c}] = P[-\sqrt{c} \le X \le \sqrt{c}] = \Phi(\sqrt{c}) - \Phi(-\sqrt{c})$$

Durch Ableiten folgt die Dichte.

Die Formel für die χ_m^2-Dichte folgt durch wiederholte Anwendung der Faltungsformel (Dichte der Summe von unabhängigen Zufallsvariablen).

Die Formel für die t- und F-Verteilungen beruht auf der folgenden Überlegung: Seien U und $V > 0$ zwei unabhängige Zufallsvariablen mit Dichten f_U und f_V. Dann gilt

$$P[\frac{U}{V} \le x] = \int\int_{\{u \le xv\}} f_U(u) f_V(v) du dv = \int_0^\infty f_V(v) F_U(xv) dv.$$

Durch Ableiten der rechten Seite nach x erhält man dann die Dichte von U/V:

$$f_{U/V}(x) = \int_0^\infty f_V(v) f_U(xv) v dv.$$

Die konkrete Berechnung der Integrale ist langwierig und bringt keine neue Erkenntnis, daher verzichten wir darauf.

Zum Schluss erwähnen wir noch ein Resultat, das manchmal nützlich ist:

Lemma A.2.1. *Wenn ein Zufallsvektor* $\mathbf{X}$ *eine* $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$*-Verteilung hat, dann ist* $(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})$ *chiquadrat-verteilt mit* n *Freiheitsgraden.*

Beweis Man schreibt $\mathbf{X}$ als $\boldsymbol{\mu} + A\mathbf{Y}$ mit $\mathbf{Y}$ $\mathcal{N}_n(0, 1_n)$-verteilt und $AA^T = \Sigma$. Dann ist die quadratische Form in $\mathbf{X}$ gerade gleich $\mathbf{Y}^T\mathbf{Y} = \sum Y_i^2$, und damit folgt die Behauptung aus der Definition der Chiquadrat-Verteilung.

Anhang B

Literatur

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, N. Y. *Verallgemeinerte Lineare Modelle.*

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*, Wiley N. Y. *Standardwerk zur nichtlinearen Regression.*

Bowman, A. W. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis*, Oxford University Press. *Nichtparametrische Regression.*

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole. *Diskussion von Regressions- und andern Methoden im Hinblick auf deren Implementation in S und R.*

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, Wiley, N. Y. *Graphische und exploratorische Methoden für Regressionsmodelle.*

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London. *Überlebenszeiten, Cox-Regression.*

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N. Y. *Ein Klassiker für Datenanalyse und Anwendungen.*

Dobson, A. J. (1991). *An Introduction to Generalized Linear Models*, Chapman and Hall, London. *Kurze anschauliche Einführung in lineare und verallgemeinerte lineare Modelle.*

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N. Y. *Spezialliteratur zum Thema "Errors in Variables".*

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. *Robustheit allgemein, mit 2 Kapiteln über robuste Regression.*

Lindsey, J. K. (1997). *Applying Generalized Linear Models*, Springer, N. Y. *Verallgemeinerte lineare Modelle, mit komplexeren Anwendungen.*

McCullagh, P. and Nelder, J. A. (1989, 1997). *Generalized Linear Models*, 2nd edn, Chapman-Hall, London. *Standardwerk über verallgemeinerte lineare Modelle.*

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N. Y. *Robuste Regression, einfacher zu lesen als Hampel et al.*

Ryan, T. P. (1997). *Modern Regression Methods*, Wiley, N. Y. *vor allem als Nachschlagewerk.*

Seber, G. A. F. (1980). *The Linear Hypothesis: A General Theory*, 2nd edn, Charles Griffin, London. *Mathematisch elegante, knappe Darstellung.*

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*, Wiley, N. Y. *Etwas mathematischer als Bates and Watts, umfangreich.*

Sen, A. and Srivastava, M. (1990). *Regression Analysis; Theory, Methods, and Applications*, Springer-Verlag, N. Y. *Lehr- und Nachschlagewerk, ähnlich wie die Vorlesung.*

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer Series in Statistics, Springer-Verlag, N. Y. *Nichtparametrische Regression.*

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, 3rd edn, Springer-Verlag, N. Y. *Überblick über alle gängigen statistischen Verfahren und deren Implementation in S und R.*

Weisberg, S. (1990). *Applied Linear Regression*, 2nd edn, Wiley, N. Y. *Lehrbuch, betont Anwendungen.*