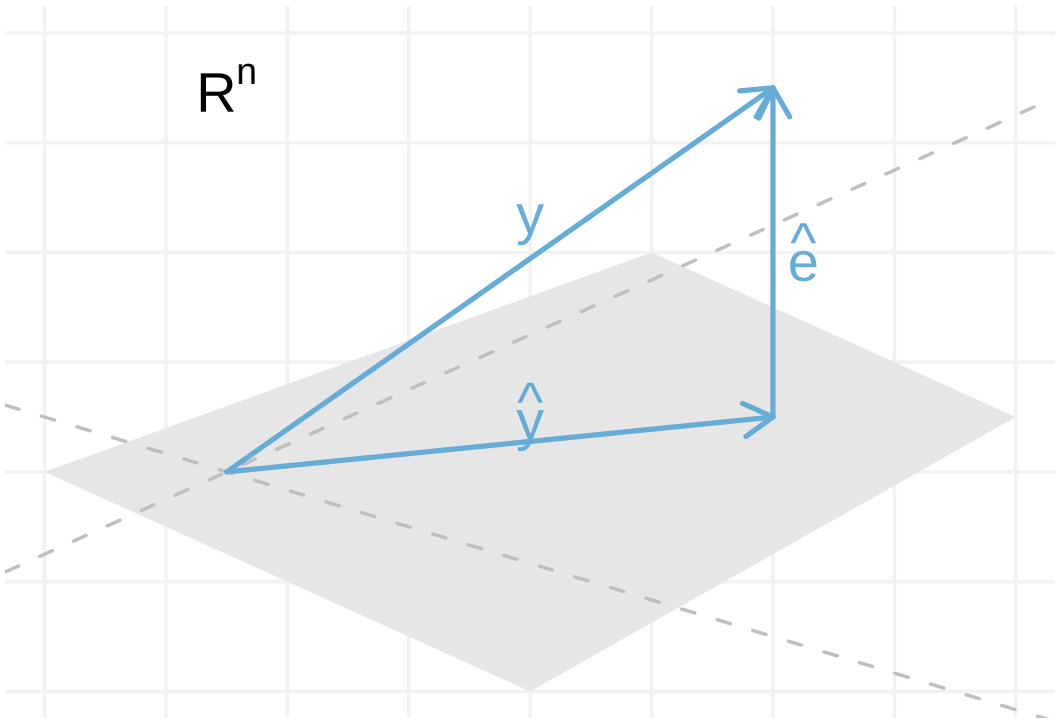
Lineare Modelle

Eine kleine Einführung



Reinhard Meister, Beuth Hochschule Berlin

1 Lineare Modelle zur Analyse von Messgrößen

Lineare Modelle kennen Sie schon als *Regressionsgerade*. In dieser Lerneinheit werden Sie erfahren, wie man diese einfache Modell verallgemeinern kann und damit viele Zusammenhänge zwischen einer metrischen Zielgröße und den verschiedensten Einflussgrößen leicht und dennoch ausreichend komplex beschreiben zu können. Ziel ist es, die erforderlichen mathematischen Grundlagen soweit zu vermitteln, dass Sie in der Lage sind, selbst Analysen zu konzipieren und durchzuführen. Die Inhalte dieses Kapitels sind die Grundlage für die weiteren Lerneinheiten, in denen die Modellierung von 0-1 Zielgrößen und von Ereigniszeiten dargestellt wird. Die Theorie wird mit Hilfe der **R**-Software umgesetzt und an zahlreichen Beispielen erläutert.

Lernziele und Überblick

Lernziele

Nach dem Durcharbeiten dieses Kapitels sollten Sie

- Lineare Modelle kennen und interpretieren können.
- das Prinzip der Kleinst-Quadrate Methode zur Parameterschätzung erläutern und anwenden können.
- Analysen im Linearen Modell mit Hilfe der R-Software selbständig durchführen können.

Gliederung der Lerneinheit

1. Datenbeispiele und Modellierung
2. Parameterschätzungen in Linearen Modellen
3. Lineare Modelle und Matrizen
4. Inferenz bei der Linearen Regression
5. Lineare Regression in **R** und Anwendungsbeispiele

1.1 Einführung

Die Modellierung von Daten gehört zu den wichtigsten Aufgaben der angewandten Statistik. In der medizinischen Forschung dient die Modellierung verschiedenen Zielen. Stets hat man dabei eine Zielgröße im Auge, deren Erwartungswert – der theoretische Mittelwert – für unterschiedliche Bedingungen, erfasst mit Hilfe so genannter Einflussgrößen, beschrieben werden soll. Mögliche Fragestellungen sind dabei fast unerschöpflich. Beispiele für Fragestellungen sind:

- Wie hängt das Geburtsgewicht von der Schwangerschaftswoche ab?
- Wie steigt die Herzaktivität mit der Belastung am Ergometer?
- Wie hängt eine Blutdrucksenkung von der Dosis eines ACE-Hemmers ab? Wird die Blutdrucksenkung dabei von anderen Vorbedingungen wie z.B. Body-Mass-Index der Patienten, dem Geschlecht etc. beeinflusst?

In diesem Kapitel werden wir uns mit metrischen Zielgrößen wie Blutdruck, Körpergewicht, beschäftigen. Später werden wir sehen, dass die Modellierung auch bei dichotomen Zielgrößen wie Heilerfolg, Auftreten von Nebenwirkungen unter einer Therapie, etc. möglich ist. Das Kapitel über Ereigniszeiten wird dann sogar die Modellierung der Zeit bis zur Remission eines Tumors erlauben.

1.2 Datenbeispiele und Aufgaben

Damit Sie eine Vorstellung möglicher Fragestellungen erhalten und in der Folge die theoretischen Konzepte gleich in der Praxis ausprobieren können, haben wir einige Beispiele, nicht nur aus der Medizin, zusammengestellt. Angegeben sind Kontext und Fragestellung, sowie die Daten selbst. Wir werden Teilaspekte dieser Beispiele immer wieder in der Beschreibung der Methoden aufgreifen.
Das Durcharbeiten der Beispiele ist für Sie mit folgenden Aufgaben verbunden:

Übung: Daten und Modelle
Studieren Sie die folgenden Datenbeispiele und beantworten Sie dann die Fragen.

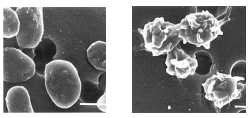
1. Was sind Ziel-, was Einflussgrößen?
2. Welchen Zusammenhang erwarten Sie zwischen Ziel- und Einflussgröße(n)?
3. Können Sie Aussagen über die Datengewinnung machen? Handelt es sich um Beobachtungsstudien oder randomisierte Studien?
4. Nennen Sie eigene Fragestellungen, und stellen sie soweit möglich eigene oder Ihnen zugängliche Daten für eine spätere Analyse zusammen.

Optimale Fahrweise

Ein Fahrdienst hat Durchschnittsgeschwindigkeiten und Verbrauchswerte einiger Fahrzeuge seiner Flotte notiert. Es wurden jeweils zwei Fahrzeuge mit gleicher Durchschnittsgeschwindigkeit gewählt. Dabei muss bei geringeren Durchschnittsgeschwindigkeiten von mehr Brems- und Beschleunigungsvorgängen ausgegangen werden.

Geschwindigkeit	Verbrauch
56.3	10.7
56.3	11.8
64.4	8.4
64.4	7.6
72.4	6.4
72.4	6.2
80.4	5.7
80.4	6.0
88.5	6.9
88.5	6.4
96.5	8.7
96.5	7.8

Bypass Patienten



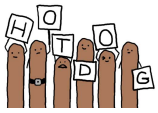
Thrombozyten in Ruhe und aktiviert. Fotos Universität Heidelberg.

In einer randomisierten Studie wurden 22 Patienten drei unterschiedlichen Beatmungsregimen unterworfen. Gruppe1: Stickoxid und Sauerstoff 50%:50% über 24 Stunden. Gruppe2: Mischung wie Gruppe 1 aber nur während der OP, Gruppe 3: Kein Stickoxid, aber 35-50% Sauerstoff über 24 Stunden. Es wurde die Thrombozytenzahl (Folate) nach 24 Stunden bestimmt. Ziel ist es, einen möglichst niedrigen Wert zu erzielen. Der Datensatz `redcell` entstammt dem Buch von Rabe-Hesketh, Everit (2001).

Group	Folate	Group	Folate
g1	243	g2	249
g1	251	g2	255
g1	275	g2	273
g1	291	g2	285
g1	347	g2	295
g1	354	g2	309
g1	380	g3	241
g1	392	g3	258
g2	206	g3	270
g2	210	g3	293
g2	226	g3	328

Datensatz `redcell`

Kaloriengehalt verschiedener Hotdogsorten



Im folgenden Datensatz sind die Kaloriengehalte verschieder Würstchensorten angegeben. Referenz: Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics. Original source: Consumer Reports, June 1986, pp. 366-367.

Beef	186	181	176	149	184	190	158	139	175
	148	152	111	141	153	190	157	131	149
	135	132							
Meat	173	191	182	190	172	147	146	139	175
	136	179	153	107	195	135	140	138	
Poultry	129	132	102	106	94	102	87	99	107
	113	135	142	86	143	152	146	144	

Quecksilber in Forellenbarschen

Die Kontamination von Speisefischen stellt eine ernsthafte Gesundheitsgefährdung dar. In einer Untersuchung über den Quecksilbergehalt in Forellenbarschen (kommen häufig in nordamerikanischen Gewässern vor) wurde die Wasserzusammensetzung von 38 Seen Floridas sowie der Quecksilbergehalt in dreijährigen Fischen aus den jeweiligen Seen bestimmt.
In der Chemie gibt es viele Gesetzmäßigkeiten, die auf proportionale Zusammenhänge hinweisen. Deshalb ist es nicht erstaunlich, wenn Sie zwischen den Konzentrationen von Quecksilber und von Alkalischem sowie Kalzium keinen linearen Zusammenhang finden. Es empfiehlt sich hier eine logarithmische Transformation. Der pH-Wert ist übrigens schon eine Angabe in Logarithmen.

Mercury	Alkalin	Calcium	pH	Mercury	Alkalin	Calcium	pH
1330	2.5	2.9	4.6	250	67.0	58.6	7.8
250	19.6	4.5	7.3	410	28.8	10.2	7.4
450	5.2	2.8	5.4	160	119.1	38.4	7.9
160	71.4	55.2	8.1	160	25.4	8.8	7.1
720	26.4	9.2	5.8	230	106.5	90.7	6.8
810	4.8	4.6	6.4	560	8.5	2.5	7.0
710	6.6	2.7	5.4	890	87.6	85.5	7.5
510	16.5	13.8	7.2	180	114.0	72.6	7.0
1000	7.1	5.2	5.8	190	97.5	45.5	6.8
150	83.7	66.5	8.2	440	11.8	24.2	5.9
190	108.5	35.6	8.7	160	66.5	26.0	8.3
1020	6.4	4.0	5.8	670	16.0	41.2	6.7
450	7.5	2.0	4.4	550	5.0	23.6	6.2
590	17.3	10.7	6.7	580	25.6	12.6	6.2
810	7.0	6.3	6.9	980	1.2	2.1	4.3
420	10.5	6.3	5.5	310	34.0	13.1	7.0
530	30.0	13.9	6.9	430	15.5	5.2	6.9
310	55.4	15.9	7.3	280	17.3	3.0	5.2
470	6.3	3.3	5.8	250	71.8	20.5	7.9

Therapien bei Anorexie

Familientherapie

Eine Studie zum Vergleich verschiedener Therapien bei Anorexie lieferte die folgenden Ergebnisse. Angegeben ist das Körpergewicht (kg) vor und nach der Behandlung. Die Behandlungen sind Standardtherapie (g1), Kognitive Verhaltenstherapie (g2) und Familientherapie (g3). Quelle: Rabe-Hesketh, Everitt 2001.

Group	Before	After	Group	Before	After	Group	Before	After	Group	Before	After
g1	36.6	37.4	g1	39.9	40.5	g2	34.1	39.4	g2	40.5	35.8
g1	38.6	38.9	g1	38.3	38.1	g2	36.6	33.4	g3	38.1	43.3
g1	37.0	37.0	g1	39.3	37.6	g2	35.6	38.5	g3	37.9	42.9
g1	37.5	37.2	g1	34.8	34.4	g2	35.3	35.2	g3	39.1	41.6
g1	36.3	34.7	g1	36.5	37.5	g2	40.3	36.1	g3	37.5	41.8
g1	40.3	47.1	g1	39.9	45.6	g2	37.0	40.7	g3	39.4	45.6
g1	43.1	44.7	g1	37.9	38.7	g2	35.5	37.0	g3	36.2	34.9
g1	34.7	42.5	g1	36.2	38.0	g2	32.0	37.2	g3	35.0	34.9
g1	36.8	33.4	g1	38.4	38.5	g2	35.1	35.1	g3	42.8	46.2
g1	36.6	37.3	g1	36.7	43.7	g2	38.7	38.3	g3	33.4	43.1
g1	38.6	44.0	g1	39.7	39.4	g2	39.1	34.3	g3	36.6	34.2
g1	40.5	43.3	g2	36.7	36.5	g2	38.2	36.1	g3	37.1	35.4
g1	37.0	37.5	g2	40.6	36.4	g2	36.2	33.2	g3	37.3	43.4
g1	34.8	33.0	g2	41.7	39.3	g2	38.9	40.1	g3	35.3	41.2
g1	31.8	41.3	g2	33.6	39.2	g2	38.4	38.5	g3	38.0	42.0
g1	36.5	32.4	g2	35.5	34.6	g2	36.2	37.0	g3	40.9	42.6
g1	37.9	38.8	g2	40.1	35.5	g2	35.2	36.9	g3	39.1	41.7
g1	37.7	37.1	g2	39.7	34.1	g2	32.9	40.1	g3	39.7	44.5

Hinweis, neue Beispiele

Sie können die Datenbeispiele gerne ergänzen. Falls Sie eigene Daten zur Verfügung stellen wollen, oder ein interessantes Beispiel gefunden haben, beschreiben Sie bitte die Daten ähnlich wie in den vorherigen Beispielen unter Angabe der Quelle.

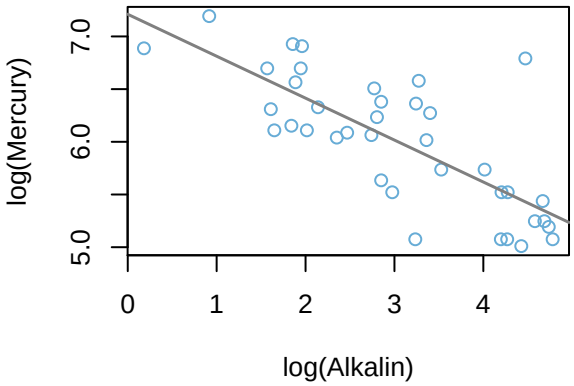
1.3 Lineare Modelle in Beispielen

Bis jetzt haben Sie gesehen, dass man vielen Fragestellungen findet, bei denen eine metrische Zielgröße mit Hilfe von Einflussgrößen erklärt werden soll. Die Linearen Modelle beschränken sich dabei auf Funktionen der bekannten Einflussgrößen, die linear in unbekannten Parametern sind. Diese Beschränkung auf lineare Funktionen hat, wie wir später sehen werden, den großen Vorteil, dass die unbekannten Parameter mit Hilfe einfacher Berechnungen bestimmt werden können. In den folgenden Beispielen geht es darum, dass Sie einen Überblick über die Modellierungsmöglichkeiten erhalten und dass Sie erkennen, was ein lineare Modell ausmacht.

Die Regressionsgerade

Sie kennen die Regressionsgerade schon aus den einführenden Statistik Kursen.

Modell: $y = \alpha + \beta x + e$

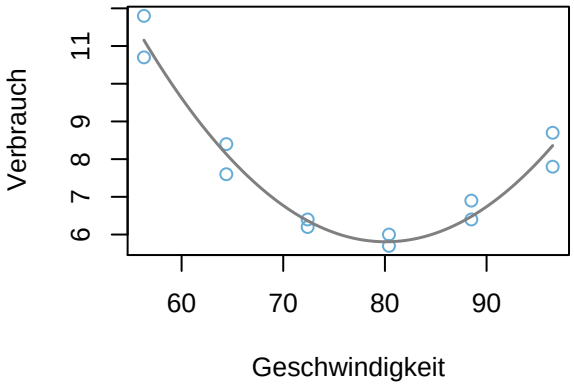


Die Graphik zeigt die Anpassung einer Regressionsgeraden an die logarithmierten Daten. Es handelt sich um die Daten aus unserem Quecksilber Beispiel. Mit y wird der logarithmierte Quecksilbergehalt, mit x die logarithmierte Alkalinität bezeichnet. Die Parameter α und β gehen linear in das Modell ein, der Fehlerterm e soll uns daran erinnern, dass die Messwerte y nicht nur systematisch variieren sondern auch eine Zufallskomponente enthalten.

Die Regressionsparabel

Häufig ist eine Regressionsgerade kein geeignetes Modell, um den Zusammenhang zwischen x und y darzustellen. Im Beispiel des Kraftstoffverbrauchs wird das besonders deutlich.

Modell: $y = \alpha + \beta x + \gamma x^2 + e$



Die Graphik zeigt die Abhängigkeit des Verbrauchs von der Fahrgeschwindigkeit.
Obwohl eine Regressionsparabel überhaupt nicht linear aussieht, ist das Modell dennoch linear

in den Koeffizienten α, β und γ. Der Trick besteht darin, statt einer Einflussgröße x auch deren Quadrat x^2 als weitere Einflussgröße zu betrachten. In Wirklichkeit haben wir es sogar mit drei Einflussgrößen zu tun: die Konstante α können wir nämlich getrost mit einer 1 multiplizieren, ohne dass sich etwas ändert und erhalten als Modellgleichung

$$y = \alpha\, x_0 + \beta\, x_1 + \gamma\, x_2 + e,$$

wenn wir die Bezeichnungen

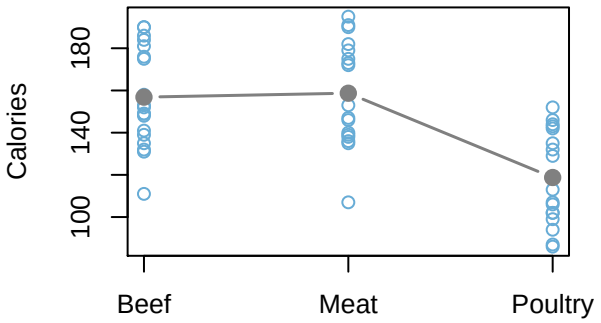
$$x_0 = 1, \quad x_1 = x, \quad x_2 = x^2$$

verwenden.
Wir sehen: im Prinzip hat jedes Lineare Modell genau so viele Einflussgrößen wie Parameter, diese werden paarweise miteinander multipliziert und aufaddiert.

Mehrere Gruppen

Häufig habe wir es mit nicht metrischen Einflussgrößen, wie Geschlecht, Herkunft, etc. zu tun. Um dennoch ein Lineares Modell aufstellen zu können, greifen wir wieder zu dem Trick, aus einer Einflussgröße mehrere zu machen.

Modell: $y_{ij} = \mu_i + e_{ij},\ i = 1, \ldots, k;\ \ j = i, \ldots, n_i$



Die Graphik zeigt, dass die Gruppenmittel des Kaloriengehalts von Hotdogs Unterschiede zwischen den verwendeten Fleischsorten erkennen lassen. Zur Beschreibung der Daten benötigen wir drei Parameter: die erwarteten Gruppenmittel. Verwendet man für das Modell eine etwas umständliche Schreibweise:

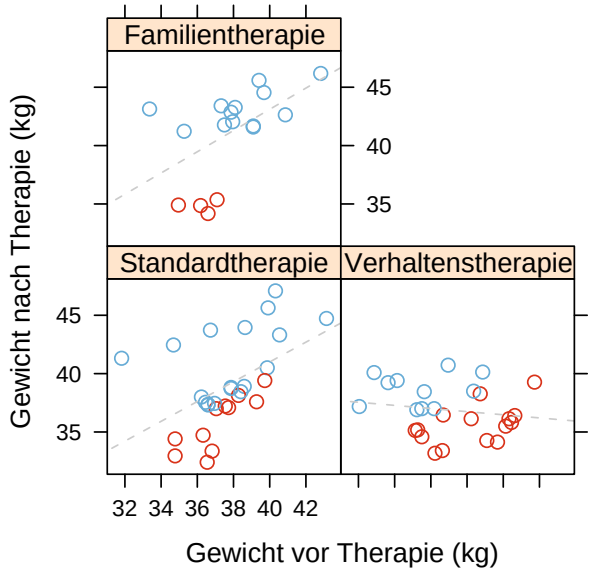
$$y = \mu_1 \mathbf{1}_{group=1} + \mu_2 \mathbf{1}_{group=2} + \mu_3 \mathbf{1}_{group=3} + e$$

erkennt man, dass wieder eine lineare Abhängigkeit von den Parametern besteht. Die Funktionen $\mathbf{1}_{group=i}$ sind so genannte Indikatorfunktionen, sie nehmen den Wert 1 an, wenn die angegebene Beziehung erfüllt ist, ansonsten sind sie gleich Null. Da wir nur Hotdogs aus jeweils einer Fleischsorte betrachten, ist von den Indikatorvariablen auch stets nur eine gleich 1. Im Detail: Für unser Datenbeispiel Hotdog bezeichnet μ_1 den erwarteten Kaloriengehalt für *Beef-Hotdogs*, und die neue Variable $\mathbf{1}_{group=1}$ hat an den ersten 20 Stellen eine Eins und ansonsten Null als Wert. Bei $\mathbf{1}_{group=2}$ erhalten wir zunächst 20 mal Null, dann 17 mal Eins, das sind die 17 *Meat-Hotdogs* und dann wieder 17 mal Null, usw. Man nennt diese Null-Eins Variablen häufig auch Dummy-Variable, weil sie als Platzhalter für die Gruppenzugehörigkeit herhalten müssen.

Mehrere Regressionsgeraden

Hat man es gleichzeitig mit metrischen und nichtmetrischen Einflussgrößen zu tun, kann man die Modellierung natürlich mischen. Wieder wird gegebenenfalls eine Einflussgröße in mehrere Variable verwandelt, man sagt auch kodiert. So lassen sich schon relativ komplex Modellgleichungen aufstellen.
Im folgenden Beispiel werden die Therapieform und das Ausgangsgewicht als Einflussgrößen, das Gewicht nach Therapie als Zielgröße verwendet. Nimmt man einen linearen Zusammenhang zwischen Ausgangs- und Endgewicht je Therapieform an, kommt man zur folgenden Darstellung.

Modell: $y_{ij} = \mu_i + \beta_i x_{ij} + e_{ij},\ i = 1, \dots, k;\ \ j = i, \dots, n_i$



Die Graphik zeigt den Zusammenhang zwischen Gewicht vor und nach Therapie, getrennt für die verschiedenen Therapieformen. Der Therapieerfolg bei Anorexia junger Mädchen ist anscheinend zwischen den Therapieformen unterschiedlich. Das zeigen die rot eingefärbten Misserfolge, bei denen das Gewicht nach Therapie kleiner als das Anfangsgewicht ist.
Die Analyse der Abhängigkeit einer Zielgröße vom Wert vor Therapie ist eine gängige Praxis bei medizinischen Studien. Zum einen ist bei unterschiedlichen Ausgangsbedingungen von vornherein ein Unterschied der Werte der Zielgröße zu erwarten, zum anderen dient diese Vorgehen dazu, den Vergleich der Therapien um Unterschiede in den Vorwerten zu bereinigen, man sagt auch, um die Daten auf die Vorwerte zu *adjustieren*. Genaueres werden noch später sehen.

Übung: Modell für das Anorexie-Beispiel

1. Was sind x und y im Modell?
2. Welche Werte haben k und n_i, $\quad i = 1, \dots, k$
3. Welche Bedeutung haben die Parameter μ_i und β_i?
4. Versuchen Sie, analog zur Modellformulierung im Beispiel *Mehrere Gruppen* das Modell mit Hilfe von Indikatorvariablen aufzuschreiben und so die lineare Abhängigkeit von den Parametern deutlich zu machen.

Lösung: 1. Gewicht vor und nach Therapie. 2. $k = 3$ Therapieformen, n_i: Auszählen. 3. Parameter: Achsenabschnitt und Steigung je Therapie. 4. Formel:

$$y = \mu_1 \mathbf{1}_{g=1} + \mu_2 \mathbf{1}_{g=2} + \mu_3 \mathbf{1}_{g=3} + \beta_1 \mathbf{1}_{g=1}\, x + \beta_2 \mathbf{1}_{g=2}\, x + \beta_3 \mathbf{1}_{g=3}\, x + e$$

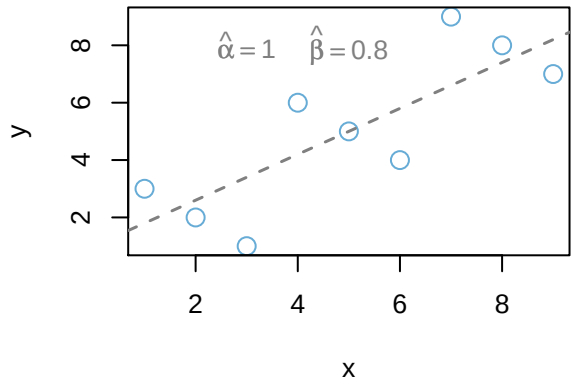
Einschub: Weshalb sind multiple Regressionsmodelle erforderlich?

Im Prinzip ist es zunächst verwunderlich, dass für eine scheinbar einfache Sache, das Anpassen von Regressionsgeraden oder einfachen Mittelwerten eine ganze Theorie herhalten soll. Kann man bei komplizierteren Fragestellungen nicht einfach Schritt für Schritt vorgehen? Man analysiert die Beziehung zur Zielgröße einzeln für jede der in Frage kommenden Einflussgrößen. Was soll daran falsch sein?

Genau dieser Frage gehen wir im folgenden nach. An einem künstlichen Datensatz wird erläutert was geschieht, wenn man beim Modellieren von Daten wichtige Einflussgrößen vergisst. Wir betrachten folgende Daten:

x	1	2	3	4	5	6	7	8	9
y	3	2	1	6	5	4	9	8	7

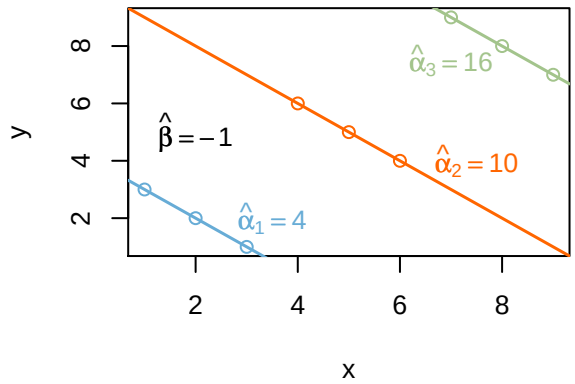
Für die lineare Regression würde man ein Modell der Form $y = \alpha + \beta x + e$ annehmen. Stellt man die Abhängigkeit der y-Werte von x samt Regressionsgerade graphisch dar erhält man den Eindruck eines positiven Zusammenhangs.



Allerdings fällt ein gewisses Muster in den Daten auf. Tatsächlich lässt sich diese Muster leicht erklären, wenn wir über Informationen für eine zusätzliche Variable z verfügen:

x	1	2	3	4	5	6	7	8	9
y	3	2	1	6	5	4	9	8	7
z	1	1	1	2	2	2	3	3	3

Schätzt man nämlich je Werte von z gesonderte Regressionsgeraden, Modell: $y = \alpha_z + \beta\, x + e$ ergibt sich ein völlig anderes Bild.



Die y-Werte nehmen mit wachsendem x ab und mit wachsendem z zu. Der positive Zusam-

menhang bei der einfachen linearen Regression kam nur dadurch zustande, dass die zusätzliche Heterogenität wegen der Kombination der x- und z-Werte zu einer Verfälschung führte.
Fazit: Einfache lineare Regression kann irreführende Ergebnisse liefern, falls weitere Einflussgrößen vorhanden sind. Deshalb sind Methoden der multiplen Regression, wo mehrere Einflussgrößen simultan untersucht werden, für die Anwendung unerlässlich.

Diese einfache Tatsache wird leider von vielen Anwendern gerne übersehen oder ignoriert. Dadurch erhöht sich bestenfalls nur die Variabilität der Ergebnisse. Bei Unausgewogenheit in der Zusammensetzung von Vergleichsgruppen kann es aber zu erheblichen Verzerrungen der Ergebnisse kommen, die sich bei simultaner Analyse vermeiden lassen.

1.4 Parameterschätzung in Linearen Modellen

Da Sie anhand der Theorie der Linearen Modelle ein grundlegendes Verständnis für das Modellieren von Daten erlangen können, werden wir die Analyse relativ ausführlich behandeln. In der Praxis werden Sie später fertige Software verwenden, mit deren Hilfe sich alle notwendigen Berechnungen fast von selbst erledigen, aber Sie sollten schon versuchen, den Hintergrund zu verstehen. Zunächst sollen Sie sich an der folgenden Einführung in die Kleinstquadrateschätzung klarmachen, dass auch bei Modellen mit mehr als einer Einflussgröße keine grundsätzlich neuen Gesichtspunkte auftauchen. Der Rechengang wird lediglich etwas länger. Danach werden wir sehen, wie sich mit Hilfe der Matrix-Notation die Darstellung noch weiter vereinfachen lässt. Zuletzt verwenden wir geometrisch Argumente, um die Methode er kleinsten Quadrate noch besser einschätzen zu können. Wenn das geschafft ist, werden wir anhand der Analyse unserer Beispieldatensätze sehen, wie wir die Datenanalyse mit Hilfe der R-Software durchführen können.

Die Methode der Kleinsten Quadrate: Einführung

Die Methode der Kleinsten Quadrate wurde von Gauss und Legendre vor etwa 200 Jahren eingeführt und publiziert. Die wichtigste Idee war dabei der Ansatz, Beobachtungen möglichst gut mit Hilfe einer linearen Funktion mehrerer bekannter Einflussgrößen anzunähern. Als Gütekriterium wird dabei die Summe der Abweichungsquadrate zwischen Beobachtung und Näherung betrachtet. In Formeln lautet die Aufgabe wie folgt: Gesucht werden die Parameterwerte $\beta_0, \beta_1, \ldots, \beta_p$ für die

$$S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\})^2$$

minimal wird. Dabei werden die Datenpunkte $y_i, x_{i1}, \ldots, x_{ip}; i = 1, \ldots, n$ als bekannt aufgefasst. Sucht man das Minimum der Zielfunktion $S(\beta_0, \beta_1, \ldots, \beta_p)$ hat man es im mathematischen Sinne mit einer Extremwertaufgabe zu tun. Die Lösung dieser Aufgabe kann man durch Nullsetzen der Ableitung der Zielfunktion – hier der partiellen, analog zur Extremwertbestimmung für Funktionen einer Variablen – nach den verschiedenen Parametern bestimmen. Wir werden zunächst das Prinzip an einigen Beispielen studieren.

Der Mittelwert

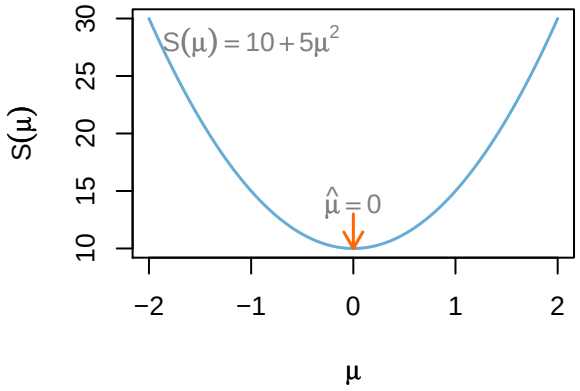
Wie gut lassen sich Daten durch eine Konstante μ approximieren? Dazu betrachten wir:

$$S(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2$$

Die Ableitung lautet:

$$\partial S(\mu)/\partial \mu = -2 \sum (y_i - \mu)$$

Nullsetzen ergibt $\widehat{\mu} = \overline{y}$, also minimiert das arithmetische Mittel die Summe der Abweichungsquadrate.
Zahlenbeispiel: Für $y = (-2, -1, 0, 1, 2)$ erhalten wir als Summe der Abweichungsquadrate $S(\mu) = (-2-\mu)^2 + (-1-\mu)^2 + \mu^2 + (1-\mu)^2 + (2-\mu^2) = 10 + 5\mu^2$ Natürlich sehen wir auch ohne Ableitung, dass das Minimum der Abweichungsquadrate bei $\widehat{\mu} = 0$ liegt, was auch die graphische Darstellung der Zielfunktion in Abhängigkeit von μ zeigt.



Die Regressiongerade
Wie gut lassen sich Daten durch eine Gerade $\alpha + \beta x$ approximieren? Dazu betrachten wir:

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

Ableitungen:

$$\partial S(\alpha,\beta)/\partial\alpha = -2\sum(y_i - \{\alpha + \beta x_i\}) \quad \text{und} \quad \partial S(\alpha,\beta)/\partial\beta = -2\sum x_i(y_i - \{\alpha + \beta x_i\})$$

Nullsetzen und Umformen dieser beiden Gleichungen ergibt zur Bestimmung der sogenannten Kleinstquadrate-Lösung folgendes Gleichungssystem:

$$\begin{aligned} n\,\widehat{\alpha} + \left(\sum x_i\right)\,\widehat{\beta} &= \sum y_i \\ \left(\sum x_i\right)\widehat{\alpha} + \left(\sum x_i^2\right)\widehat{\beta} &= \sum x_i y_i \end{aligned}$$

Zahlenbeispiel: Für $x = (1, 2, 3, 4)$ und $y = (4.4, 7.9, 12.1, 16.6)$ ergibt sich: $n = 4$, $\sum y_i = 41$, $\sum x_i = 10$, $\sum x_i y_i = 122.9$ und $\sum x_i^2 = 30$. Setzen wir diese Werte in das obige Gleichungssystem ein, ergibt sich

$$\begin{aligned} 4\,\widehat{\alpha} + 10\,\widehat{\beta} &= 41 \\ 10\,\widehat{\alpha} + 30\,\widehat{\beta} &= 122.9 \end{aligned}$$

Die Lösung diese Gleichungssystem ergibt übrigens $\widehat{\alpha} = 0.05$ und $\widehat{\beta} = 4.08$, was Sie sicher leicht nachrechnen können. Die Lösung des Gleichungssystems lässt sich auch explizit symbolisch (mit Buchstaben) bestimmen und führt dann zu den schon bekannten Formeln für die Regressionsgerade.
Es gibt übrigens einen interessanten Spezialfall. Wie man sofort erkennt, ergeben sich im Fall $\sum x_i = 0$ zwei Gleichungen mit jeweils nur einer Unbekannten, und die Lösung lautet für diesen Spezialfall $\widehat{\alpha} = \overline{y}$ und $\widehat{\beta} = \sum x_i y_i / \sum x_i^2$.

Die Regressionsparabel
Wie gut lassen sich Daten durch eine Parabel $\alpha + \beta x + \gamma x^2$ approximieren?

Das Aufstellen der Bestimmungsgleichungen für die Parameter der Parabel funktioniert genau wie bei der Geraden. Das zu lösende Gleichungssystem sieht dann folgendermaßen aus:

$$\begin{aligned} n\,\widehat{\alpha} + \left(\sum x_i\right)\,\widehat{\beta} + \left(\sum x_i^2\right)\,\widehat{\gamma} &= \sum y_i \\ \left(\sum x_i\right)\widehat{\alpha} + \left(\sum x_i^2\right)\widehat{\beta} + \left(\sum x_i^3\right)\,\widehat{\gamma} &= \sum x_i y_i \\ \left(\sum x_i^2\right)\widehat{\alpha} + \left(\sum x_i^3\right)\widehat{\beta} + \left(\sum x_i^4\right)\,\widehat{\gamma} &= \sum x_i^2 y_i \end{aligned}$$

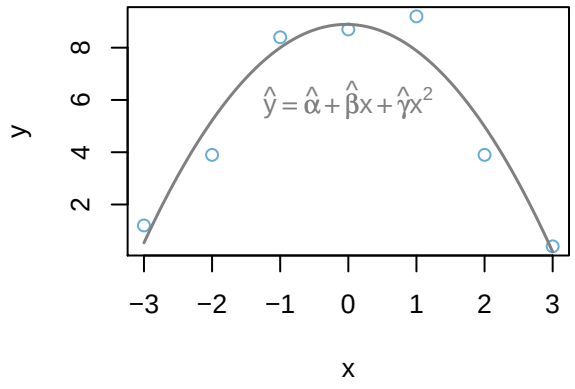
Zahlenbeispiel: Für $x = (-3, -2, -1, 0, 1, 2, 3)$ und $y = (1.2, 3.9, 8.4, 8.7, 9.2, 3.9, 0.4)$ ergibt sich aus dem folgenden Rechenschema

	y	x	x^2	x^3	x^4	xy	x^2y
1	1.2	−3	9	−27	81	−3.6	10.8
2	3.9	−2	4	−8	16	−7.8	15.6
3	8.4	−1	1	−1	1	−8.4	8.4
4	8.7	0	0	0	0	0.0	0.0
5	9.2	1	1	1	1	9.2	9.2
6	3.9	2	4	8	16	7.8	15.6
7	0.4	3	9	27	81	1.2	3.6
Summe	35.7	0	28	0	196	−1.6	63.2

das Gleichungssystem zur Bestimmung der Parabelparameter:

$$\begin{aligned} 7\,\widehat{\alpha} + 0\,\widehat{\beta} + 28\,\widehat{\gamma} &= 35.7 \\ 0\,\widehat{\alpha} + 28\,\widehat{\beta} + 0\,\widehat{\gamma} &= -1.6 \\ 28\,\widehat{\alpha} + 0\,\widehat{\beta} + 196\,\widehat{\gamma} &= 63.2 \end{aligned}$$

Mit etwas Mühe kann man nun die optimalen Parameterwerte berechnen.
Ergebnis: $(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = (8.89, -0.06, -0.95)$. Zur Belohnung schauen wir uns das Ergebnis in einer Graphik an:



1.5 Lineare Modelle und Matrizen

Ein Modell für eine Zielgröße y, die von Einflussgrößen $x_1, x_2, \ldots, x_p$ linear abhängt, ist von der Form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e\ .$$

Dabei haben die Koeffizienten $\beta_j, j = 0 \ldots p$ folgende Bedeutung: β_0 gibt eine allgemeine Konstante an und sorgt somit für ein allgemeines Niveau der y-Werte. Ist einer der anderen Koeffizienten gleich Null, hat die zugehörige Variable keinen Einfluss auf die Zielgröße, bei positiven

Werte erhöht sich der Wert von y bei einer Steigerung von x bei negativen erniedrigt sich y entsprechend. Modelle sind nie perfekt, deshalb wird der Fehlerterm e eingeführt, der sich bei biologischen Daten in der Regel aus der natürlichen biologischen Variabilität und einem meist unvermeidbaren Messfehler zusammensetzt.

Modellgleichung in Matrixform:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Kurzschreibweise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Normalgleichungen:

$$\begin{pmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & & \sum_i x_{ip} \\ & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \dots & \sum_i x_{i1}x_{ip} \\ & & \sum_i x_{i2}^2 & \sum_i x_{il}x_{i2} & \\ & & & \ddots & \vdots \\ & * & & & \sum_i x_{ip}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1}y_i \\ \sum_i x_{i2}y_i \\ \vdots \\ \sum_i x_{ip}y_i \end{pmatrix}$$

*: Die Matrix ist symmetrisch.
Kurzschreibweise:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$$

Lösung:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Prognose:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Residuen:

$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$$

Zur Veranschaulichung haben wir die Berechnungen für die Zahlenbeispiele unter Verwendung von Matrizen in R durchgeführt. Das Ergebnis sehen Sie im folgenden Abschnitt. Sie werden erkennen, dass sich die Formeln sämtlich eins zu eins in R umsetzen lassen.

Zahlenbeispiele in R mit Matrizenkalkül
als separates Window anlegen

Berechnung der numerischen Lösungen der Zahlenbeispiele mit Matrizenschreibweise in R. Sie werden natürlich genau die in den Rechenbeispielen angegebenen Zahlen wiederfinden. Nur, dass die Rechnungen mit R etwas einfacher sind als von Hand.

Regressionsgerade

```
> x <- 1:4
> y <- c(4.4, 7.9, 12.1, 16.6)
> X <- cbind(1, x)

> print("Designmatrix")

[1] "Designmatrix"
```

```
> X

       x
[1,] 1 1
[2,] 1 2
[3,] 1 3
[4,] 1 4

> print("Rechengroessen")

[1] "Rechengroessen"

> t(X) %*% X

      x
   4 10
x 10 30

> t(X) %*% y

   [,1]
   41.0
x 122.9

> print("beta.hat als Loesung des linearen Gleichungssystems")

[1] "beta.hat als Loesung des linearen Gleichungssystems"

> solve(t(X) %*% X, t(X) %*% y)

  [,1]
  0.05
x 4.08
```

Regressionsparabel

```
> x <- c(-3, -2, -1, 0, 1, 2, 3)
> y <- c(1.2, 3.9, 8.4, 8.7, 9.2, 3.9, 0.4)
> X <- cbind(1, x, x^2)

> print("Designmatrix")

[1] "Designmatrix"

> X

        x
[1,] 1 -3 9
[2,] 1 -2 4
[3,] 1 -1 1
[4,] 1  0 0
[5,] 1  1 1
[6,] 1  2 4
[7,] 1  3 9
```

```
> print("Rechengroessen")

[1] "Rechengroessen"

> t(X) %*% X

      x
   7  0  28
x  0 28   0
  28  0 196

> t(X) %*% y

  [,1]
  35.7
x -1.6
  63.2

> print("beta.hat als Loesung des linearen Gleichungssystems")

[1] "beta.hat als Loesung des linearen Gleichungssystems"

> beta.hat <- solve(t(X) %*% X, t(X) %*% y)
> beta.hat

         [,1]
   8.89047619
x -0.05714286
  -0.94761905

> print("Prognosen und Residuen")

[1] "Prognosen und Residuen"

> y.hat <- as.vector(X %*% beta.hat)
> e.hat <- as.vector(y - y.hat)
> round(rbind(y, y.hat, e.hat), 2)

      [,1]  [,2] [,3]  [,4] [,5]  [,6] [,7]
y     1.20  3.90  8.4  8.70 9.20  3.90 0.40
y.hat 0.53  5.21  8.0  8.89 7.89  4.99 0.19
e.hat 0.67 -1.31  0.4 -0.19 1.31 -1.09 0.21
```

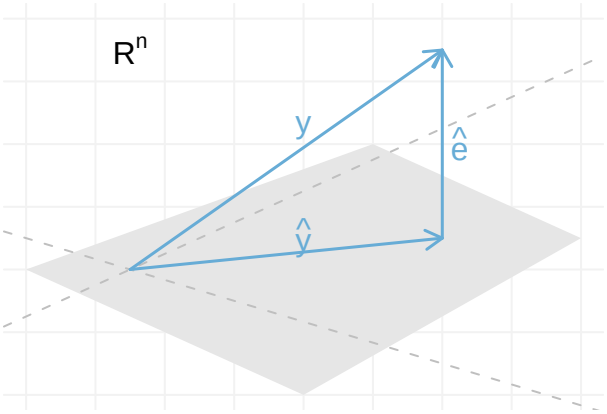
Tatsächlich lass sich Lineare Modelle noch viel einfacher mit der R Software berechnen, aber das werden wir nach noch etwas mehr Theorie an praktischen Beispielen erkunden.

1.6 Inferenz bei der linearen Regression

In diesem Abschnitt werden Sie einige Konzepte kennen lernen, die es ermöglichen, Aussagen über ein an die Daten angepasstes Modell zu machen. Dazu wird zunächst eine geometrische Deutung der Schätzung vorgenommen. Sodann werden Verfahren zur Konstruktion von Konfidenzintervallen und Tests im Linearen Modell vorgestellt.

Kleinstquadrateschätzung: Geometrische Deutung

Auch wenn es Ihnen zunächst fremd vorkommen wird, wir können unsere Daten als Vektoren im n – dimensionalen Raum auffassen. Insbesondere können wir die Zerlegung unseres Datenvektors $\mathbf{y}$ in Schätzung $\hat{\mathbf{y}}$ und Residuenvektor $\hat{\mathbf{e}}$ betrachten. Dazu dient die folgende Graphik:



Ein Vektor $\mathbf{v}$ *wird im Text als fett formatierter Buchstabe notiert, Matrizen wie* $\mathbf{M}$ *als fette Großbuchstaben. Für die Länge eines Vektors* $\mathbf{v}$ *hat sich die Bezeichnung* $||\mathbf{v}||$ *durchgesetzt und es gilt* $||\mathbf{v}||^2 = \sum v_i^2$.

Die Graphik zeigt den Beobachtungsvektor im $\mathbf{R}^n$. Grau unterlegt ist ein Ausschnitt des Modellraums. Die gestrichelten Geraden symbolisieren die Spalten der Modellmatrix, welche diesen Modellraum aufspannen. Der Modellraum entspricht einer $(p+1)$ – dimensionalen Hyperebene. Der Residuenvektor liegt somit in einem $(n-p-1)$ – dimensionalen, zum Modellraum orthogonalen Unterraum.

Im Modell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

erhalten wir die Schätzung $\hat{\mathbf{y}}$ durch die Orthogonalprojektion von $\mathbf{y}$ auf den von der Matrix $\mathbf{X}$ aufgespannten Raum. Somit erhalten wir ein rechtwinkliges Dreieck und es gilt der Satz des Pythagoras: $||\mathbf{y}||^2 = ||\hat{\mathbf{y}}||^2 + ||\hat{\mathbf{e}}||^2$. Die Projektion $\hat{\mathbf{y}}$ können wir mit Hilfe des geschätzten Parametervektors $\widehat{\boldsymbol{\beta}}$, der sich aus der Lösung der Normalgleichungen $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ ergibt, bestimmen. Einsetzen liefert dann $\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

Übung: Geometrische Deutung

1. Überlegen Sie sich, weshalb die Summe der Residuenquadrate $\sum \hat{e}_i^2$ gerade dem quadrierten Abstand von $\mathbf{y}$ und $\hat{\mathbf{y}}$ entspricht.

2. Aus der Elementargeometrie sollten Sie wissen, dass der kürzeste Abstand eines Punktes von einer Ebene im Fußpunkt des Lotes auf die Ebene realisiert wird. Diese Überlegung führt zur so genannten Orthogonalprojektion des Beobachtungsvektors auf die von den Spalter der Designmatrix aufgespannten Ebene.

3. Durch die Vektoren $\mathbf{y}$, $\hat{\mathbf{y}}$ und $\hat{\mathbf{e}}$ wird auch im $\mathbf{R}^n$ ein ebenes (*ganz normales*) Dreieck gebildet. Machen Sie sich klar, dass die Quadratsummenzerlegung tatsächlich auf dem Satz des Pythagoras beruht, da die drei Vektoren ein *rechtwinkliges* Dreieck bilden.

Tests und Konfidenzintervalle

Für die Durchführung von Hypothesentests müssen wir noch Annahmen über die Verteilung der Fehlerterme (Residuen) unseres Modell machen: wir gehen also im Folgenden davon aus, dass die Fehlerterme unabhängig und identisch normalverteilt sind: $e_i \sim N(0, \sigma^2), i = 1, \ldots, n$. Diese Voraussetzungen sind die Standardannahmen. Wir werden später noch sehen, wie sich dies Annahmen überprüfen lassen. Ein derartiger *Check der Voraussetzungen* gehört auf jeden Fall zur regelgerechten Durchführung einer Analyse im Linearen Modell. Wir werden in

diesem Abschnitt auf Herleitungen weitestgehend verzichten. Sie finden derartige Betrachtungen in der Regel in Büchern zu Linearen Modellen. Wichtig in diesem Zusammenhang sind einige Prüfverteilungen – die t - , die χ^2 - sowie die F - Verteilung – die wir benötigen werden, um Tests anzugeben.

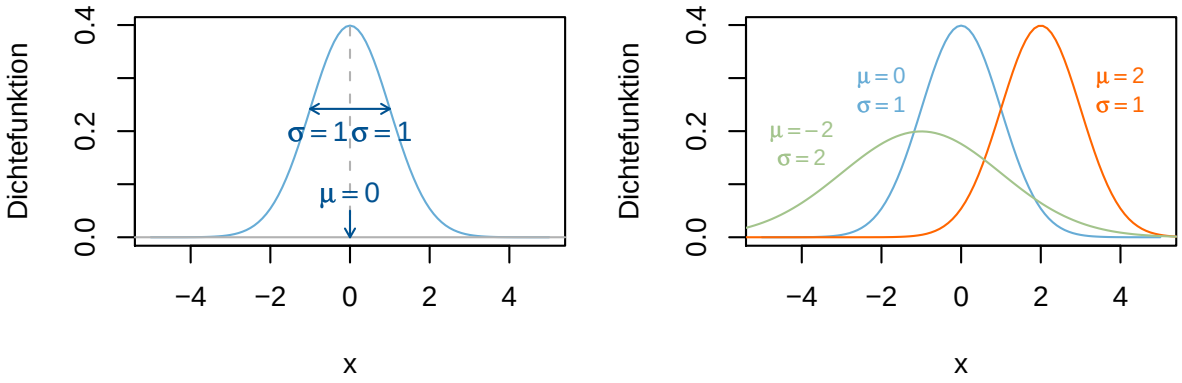
Einschub: Normalverteilung und abgeleitete Prüfverteilungen

Definition 1: Eine Zufallsvariable X mit Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

heißt *standardnormal-verteilt*. Die abgekürzte Schreibweise lautet $X \sim N(0,1)$. Es gilt $\mathbf{E}X = 0$ und $\mathbf{Var}(X) = 1$.
Die Dichtefunktion einer Normalverteilung mit $\mathbf{E}X = \mu$ und $\mathbf{Var}(X) = \sigma^2$ ist gegeben durch

$$f(x) = \frac{1}{(\sqrt{2\pi})\sigma} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)$$



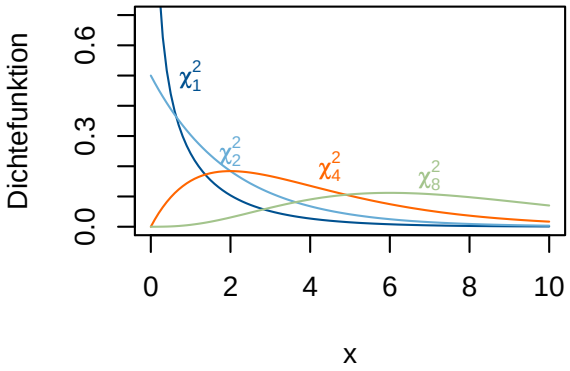
Definition 2. Seien $X_1, \ldots, X_p$ unabhängige $\mathrm{N}(0,1)$ –verteilte Zufallsvariable und sei

$$Y = X_1^2 + \cdots + X_p^2$$

dann heißt Y *chiquadrat–verteilt* mit p Freiheitsgraden. Die abgekürzte Schreibweise lautet $Y \sim \chi^2_p$. Für eine Zufallsvariable $Y \sim \chi^2_p$ gilt:

$$\mathbf{E}\ Y = p; \quad \mathbf{Var}(Y) = 2p.$$

Anstatt die Dichtefunktionen als Formeln anzugeben, betrachten wir eine graphische Darstellung für unterschiedliche Werte von p.



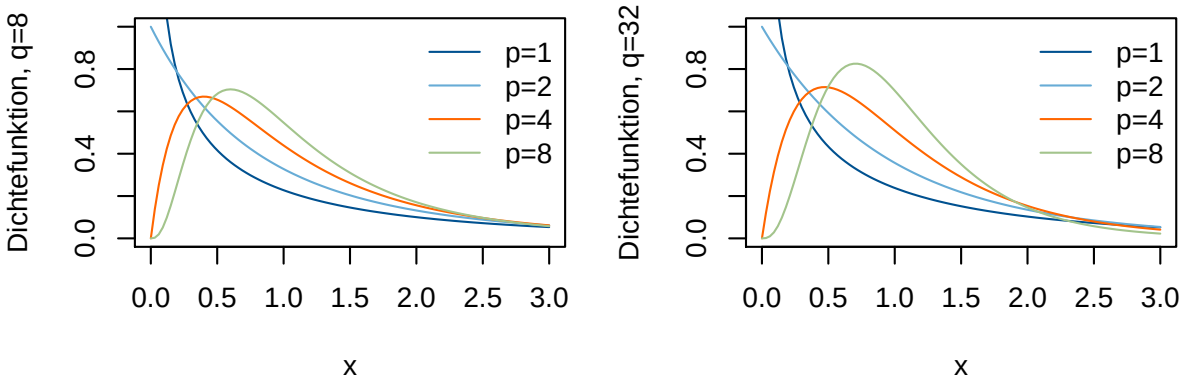
Die χ^2-Verteilungen haben also etwas mit Summen von Quadraten zu tun. Mit Quadratsummen haben wir es immer wieder bei Linearen Modellen zu tun.
Satz 1. Sei $U \sim \chi^2_p$ unabhängig von $V \sim \chi^2_q$. Dann gilt für die Zufallsvariable $W = U + V$: $W \sim \chi^2_{p+q}$.

Dieser Satz hilft uns vor allem, die Verteilung von Zerlegungen von Quadratsummen zu beschreiben.
Definition 3. Sei $U \sim \chi^2_p$ unabhängig von $V \sim \chi^2_q$. Dann heißt die Zufallsvariable

$$Z = \frac{U/p}{V/q}$$

F–verteilt mit p und q Freiheitsgraden. Wir schreiben kurz: $Z \sim F_{p,q}$.

Auch hier wollen wir wieder die Dichtefunktionen graphisch darstellen. Es gilt zu beachten, dass sowohl die Zählerfreiheitsgrade (p) als auch die Nennerfreiheitsgrade (q) variabel sind.

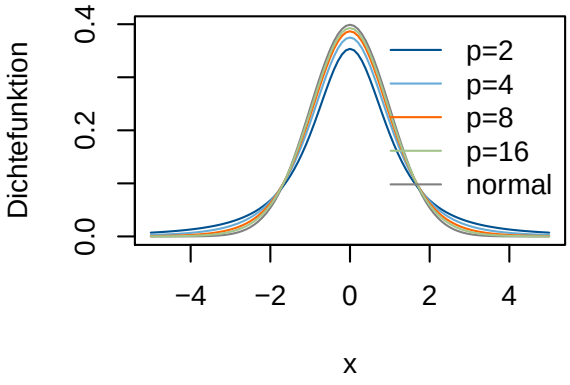


Definition 4. Sei $X \sim \mathrm{N}(0,1)$ unabhängig von $U \sim \chi^2_p$. Dann heißt die Zufallsvariable

$$t = \frac{X}{\sqrt{U/p}}$$

t–verteilt mit p Freiheitsgraden. Wir schreiben kurz: $t \sim t_p$. Offensichtlich gilt für $t \sim t_p$: $t^2 \sim F_{1,p}$.

Für die t-Verteilung erkennt man an der folgenden Graphik, wie sich die Dichtefunktionen mit wachsendem p immer mehr der Dichte der Standardnormalverteilung annähern.



Mehr zu den genannten Verteilungen findet man in den meisten Lehrbüchern der Statistik. Alle Verteilungen sind in R implementiert.

Inferenz für einzelne Parameter

Die Parameterschätzer ergeben sich wie wir schon gesehen haben als Lösung der Normalgleichungen. Die Lösungen sind Linearkombinationen der Beochbachtungen $\mathbf{y}$. Als Konsequenz lassen sich die Varianzen und Kovarianzen der Komponenten des geschätzten Parametervektors explizit angeben. Außerdem ist die Verteilung des Schätzers bei Normalverteilung der Beobachtungen auch eine multivariate Normalverteilung.

Eigenschaften von $\widehat{\beta}$

Im Linearen Modell mit unabhängigen identisch normalverteilten Restfehlern

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N(0, \sigma^2\mathbf{I})$$

gilt:

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Für die einzelnen Komponenten folgt damit

$$\widehat{\beta}_j \sim N(\beta_j, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}), \;\; j = 0, \ldots, p$$

Mit diesen Eigenschaften lassen sich leicht Tests und Konfidenzintervalle angeben. Zur Konstruktion greifen wir wie stets bei normalverteilten Größen auf die Technik der Standardisierung zurück. Für die Restvarianz σ^2 wird dabei die Schätzung

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{e}_i^2$$

verwendet. Mit diesen Schätzwerten erhält man als $1-\alpha$ Konfidenzintervall für eine einzelne Komponente

$$KI(\beta_j, 1-\alpha) = (\widehat{\beta}_j \pm \widehat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}t_{n-p-1,1-\alpha/2}),$$

wobei $t_{n-p-1,1-\alpha/2}$ das $1-\alpha/2$ Quantil der t-Verteilung mit $(n-p-1)$ Freiheitsgraden bezeichnet. Sie erinnern sich, Konfidenzintervalle sind zufällige Bereiche, die mit vorgegebener Wahrscheinlichkeit den wahren aber unbekannten Parameterwert überdecken.
Von den meisten Computerprogrammen zur Analyse Linearer Modelle werden als Standard Tests der Hypothesen $\beta_j = 0$ berechnet. Als Teststatistik dienen die standardisierten t-Werte. Unter der Annahme $H_0 : \beta_j = 0$ gilt

$$T_j = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)} \sim t_{(n-p-1)},$$

die T-Werte sind t-verteilt und so kann eine Testentscheidung mit Hilfe der kritischen Werte bzw. der p-Werte vorgenommen werden. Der geschätzte Standardfehler der Schätzer $SE(\widehat{\beta}_j)$ ist gleich $\widehat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}$ und taucht schon in der Formel für das Konfidenzintervall auf. Je nach Fragestellung werden ein- oder zweiseitige Tests betrachtet. Für zweiseitige Fragestellungen ist der Absolutwert der Teststatistik heranzuziehen.

Simultane und multiple Inferenz

Häufig soll nicht nur über eine Komponente des Parametervektors eine Aussage getroffen werden, vielmehr interessiert man sich gleichzeitig für mehrere Parameter. Wie soll man in einem solchen Falle vorgehen? Man könnte als eine Möglichkeit einfach mehrere Tests, einen je Komponente durchführen. Dabei tritt aber das Problem auf, das nun das ansonsten garantierte Irrtumsrisiko α nicht mehr eingehalten wird. Jeder Test erhöht das Irrtumsrisiko schlimmstenfalls um den Betrag α, also zwei Tests auf dem 5% Niveau ergeben so womöglich ein Irrtumsrisiko von 10%, usw.

Multiple Tests

Eine ganz einfache Variante der so genannten *multiple* Tests besteht darin, das Irrtumsrisiko α durch die Anzahl der Tests zu dividieren, also z.B. bei 5 Tests jeden einzelnen auf dem Niveau $\alpha/5$ statt auf dem Niveau α durchzuführen. So bleibt insgesamt das Irrtumsrisiko α gewahrt. Das Vorgehen wird als *Bonferroni*-Korrektur bezeichnet. Ein großer Nachteil multipler Tests ist der unvermeidbare Verlust an Power. Je mehr Einzelhypothesen geprüft werden, um so größer ist die Hürde, die jeder Test nehmen muss und so können dann nur sehr große Effekte erkannt werden. Es gibt viele Verbesserungen der elementaren Bonferroni-Korrektur, der Powerverlust wird aber lediglich verringert und ist grundsätzlich nicht vermeidbar. Methoden zur Durchführung von multiplen Tests füllen mittlerweile ganze Lehrbücher. In R gibt es ein spezielles Package `multtest`, das die wichtigsten Methoden implementiert hat.

Literaturhinweis:

Katherine S. Pollard, Houston N. Gilbert, Yongchao Ge, Sandra Taylor and Sandrine Dudoit (2009). multtest: Resampling-based multiple hypothesis testing. R package version 2.1.1. http://CRAN.R-project.org/package=multtest

Simultane Tests

Ein ganz andere Herangehensweise besteht darin, gleichzeitig mehrere Einzelhypothesen zu betrachten. Wir wollen diese Problem an unserem Hotdog Beispiel erläutern. Wenn man die verschiedenen Hotdog-Sorten vergleicht, stellt sich doch zunächst die Frage, ob überhaupt ein Unterschied im Kaloriengehalt der einzelnen Sorten (Beef (Rind), Meat (Schwein), Poultry (Geflügel)) besteht, oder ob die Erwartungswerte μ_1, μ_1, μ_3 alle gleich groß sind. Wie kann man die Nullhypothese

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

mit Hilfe eines einzelnen Tests überprüfen? Damit beschäftigt sich das Simultane Testen. Die Antwort auf die Frage nach einem geeigneten Test fällt überraschend einfach aus. Wir wissen inzwischen wie die Reststreuung in einem Linearen Modell für die Kaloriengehalte der Sorten mit individuellen Erwartungswerten je Sorte berechnet wird. Wenn nun die Nullhypothese zutreffen würde, also die mittleren Gehalte alle gleich groß wären, dürften sich die geschätzten Mittelwerte auch nur zufällig voneinander unterscheiden. Würde man also die Restvarianz unter Vernachlässigung der Gruppenzugehörigkeit bestimmen, dürfte sich kein wesentlich anderer Wert ergeben. Genau dieser Ansatz führt so genannten *Extra Sum of Squares* Prinzip. Bevor wir eine exakte Beschreibung liefern, betrachten wir noch numerische Ergebnisse für das Beispiel. Die Mittelwerte der Kalorien lauten:

	Beef	Meat	Poultry
mean	156.8	158.7	118.8
sd	22.6	25.2	17.0
n	20	17	17

Offensichtlich haben Geflügelwürstchen einen geringeren Kaloriengehalt als Rinds- oder Schweinswürstchen. Was zeigt uns die Berechnung der erforderlichen Quadratsummen? Im vollen Modell berechnen wir die Fehlerquadratsumme aus

$$SSE_{voll} = \sum (y_{ij} - \overline{y}_{i.})^2 = 28067.1$$

der Summer der quadrierten Abweichungen der Einzelwerte von den Mittelwerten (das sind gerade die Schätzwerte für y_{ij} im Linearen Modell). Wenn die Nullhypothese zutrifft, dann genügt es die Abweichungen vom Gesamtmittel, den Schätzwerten für y_{ij} im reduzierten Modell zu betrachten, und wir erhalten

$$SSE_{red} = \sum (y_{ij} - \overline{y}_{..})^2 = 45759.3 \quad .$$

Ob dieser Unterschied wesentlich ist können wir mit Hilfe eines F-Tests untersuchen. Wir benötigen lediglich noch die zugehörigen Freiheitsgrade. Im vollen Modell bleiben für die Fehlerquadratsumme $df_{voll} = n - 3 = 51$, im reduzierten $df_{red} = n - 1 = 53$ Freiheitsgrade. Unter der Annahme, dass die Nullhypothese zutrifft gilt nämlich:

$$F = \frac{(SSE_{red} - SSE_{voll})/(df_{red} - df_{voll})}{SSE_{voll}/df_{voll}} \quad \sim \quad F_{(df_{red}-df_{voll}),df_{voll}}$$

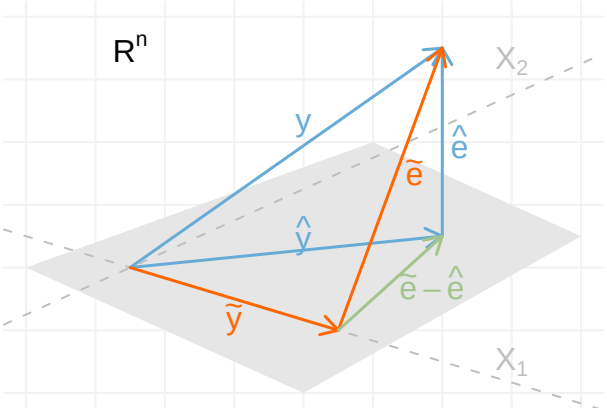
Der berechnete F-Wert gibt also Auskunft darüber, ob die im vollen Modell erreichte Verbesserung der Anpassung wesentlich ist. Die F-Verteilung wurde im vorangehenden schon besprochen, der Zählerfreiheitsgrad ergibt sich aus der Zahl der Parameter, die beim Übergang vom vollen auf das reduzierte Modell gleich Null gesetzt werden, der Nennerfreiheitsgrad entspricht der Stichprobengröße, vermindert um die Zahl der Parameter im vollen Modell. In unserem Fall erhalten wir

$$F_{beob} = \frac{(45759.3 - 28067.1)/(53 - 51)}{28067.1/51} = 16.074$$

ein Wert, der klar den beobachteten Effekt als signifikanten Unterschied nachweist. Der kritische Wert für ein Irrtumsrisiko von $\alpha = 0.05$ beträgt $F_{2,51,0.95} = 3.18$. Damit überschreitet der beobachtete F-Wert den kritischen Wert erheblich und die Nullhypothese *kein Unterschied* wird abgelehnt.

F-Test: Geometrische Deutung

Sie können sich das Vorgehen auch an der folgenden Skizze verdeutlichen. Mit $\hat{\mathbf{e}}$ und $\tilde{\mathbf{e}}$ sind die Residuen im vollen und reduzierten Modell gekennzeichnet. Sie erkennen wieder rechte Winkel und so haben wir schon wieder eine äußert nützliche Anwendung des Satzes von Pythagoras.



Darstellung von vollem und reduzierten Modell (orange). Beachten Sie, dass der Modellraum (grau unterlegt) im vollen Modell von den Teilräumen $\mathbf{X}_1$ und $\mathbf{X}_2$ aufgespannt wird. Wie man sieht gilt

$$||\tilde{\mathbf{e}}||^2 = ||\hat{\mathbf{e}}||^2 + ||\tilde{\mathbf{e}} - \hat{\mathbf{e}}||^2$$

Die Graphik zeigt die orthogonale Zerlegung des Residuenvektors des reduzierten Modells $\tilde{\mathbf{e}}$ in die Komponenten $\hat{\mathbf{e}}$ und $\tilde{\mathbf{e}} - \hat{\mathbf{e}}$. Das volle Modell lautet

$$\mathbf{y} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \mathbf{e},$$

das reduzierte Modell hat die Gestalt

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}.$$

Unter der Nullhypothese $\beta_2 = 0$ sind $||\tilde{\mathbf{e}}||^2$ und $||\hat{\mathbf{e}}||^2$ jeweils χ^2-verteilt und wegen der Orthogonalität damit auch $||\tilde{\mathbf{e}} - \hat{\mathbf{e}}||^2 = ||\tilde{\mathbf{e}}||^2 - ||\hat{\mathbf{e}}||^2$. Der F-Test ergibt sich dann aus den im Einschub Prüfverteilungen angegebenen Regeln. Die zugehörigen Freiheitsgrade ergeben sich wie folgt: Mit n bezeichnen wir den Stichprobenumfang. Der Parametervektor im vollen Modell habe die Länge $p + 1$, wobei der zu überprüfende Teilvektor β_2 gerade r Komponenten besitze. Dann gilt $SSE_{voll} = ||\hat{\mathbf{e}}||^2 \sim \sigma^2 \chi^2_{(n-p-1)}$, und unter der Nullhypothese auch

$SSE_{red} - SSE_{voll} = ||\widetilde{\mathbf{e}}||^2 - ||\widehat{\mathbf{e}}||^2 \sim \sigma^2 \chi^2_{(r)}$ und wegen der Orthogonalität kann die Beziehung

$$\frac{(SSE_{red} - SSE_{voll})/r}{SSE_{voll}/(n-p-1)} \sim F_{r,(n-p-1)}$$

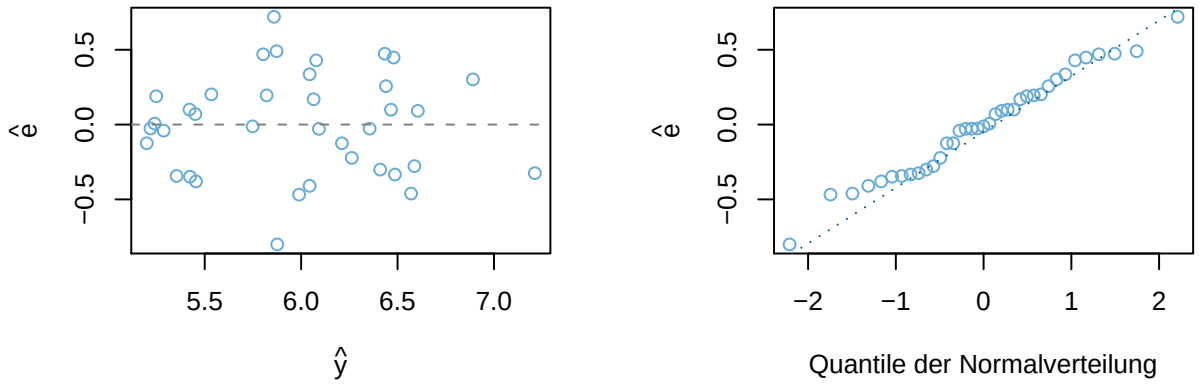
zum Test der Hypothese $\beta_2 = 0$ verwendet werden.
Standardmäßig führen viele Software-Pakete für Lineare Modelle übrigens den so genannten Global-Test für das gesamte Modell durch. Dabei werden in einem Modell mit konstantem Term alle übrigen p Koeffizienten simultan auf Null getestet, es wird also überprüft, ob mindestens einer der Modellkoeffizienten signifikant von Null verschieden ist. Im Beispiel unserer Hotdogs, liegt genau so ein Fall vor. Es gilt nämlich $n = 54$ sowie $p + 1 = 3$ und somit ergibt sich $n - p - 1 = 51$ und $r = 2$. Man sieht, dass wir genau diese Größen zur Berechnung des F-Wertes verwendet haben.

Check der Voraussetzungen

Für die Gültigkeit der Schlüsse, die wir aus Analysen im Linearen Modell ziehen, müssen wir uns darauf verlassen können, dass die Voraussetzungen erfüllt sind. Die sind im Wesentlichen wie folgt gegeben:

- Das Modell ist *passend*.
- Die Fehlervarianz ist konstant.
- Die Fehler sind normalverteilt.

Nun könnte man auf die Idee kommen, man müsse zuerst die Voraussetzungen überprüfen, bevor man ein Lineares Modell anpassen darf. Diese Idee hat man heute verworfen und schaut statt dessen im Nachhinein ob die Voraussetzungen einigermaßen gültig erscheinen. Ohne eine Anpassung eines Modells an die Daten können die Angaben nämlich nicht überprüft werden. Für praktische Anwendungen werden überwiegend graphische Darstellungen verwendet. Stets bilden die geschätzten Restfehler, die *Residuen* den Ausgangspunkt der Untersuchungen. Wir betrachten hier zwei Darstellungen, *residual vs fit* – die Gegenüberstellung von geschätzten Restfehlern und geschätzten Werten , sowie *qq-plot residuals* – die Gegenüberstellung von empirischen Quantilen der Residuen und theoretischen Quantilen der Normalverteilung. Wir erläutern die Interpretation der beiden Darstellungen an einem Beispiel.



Auf der linken Darstellung sehen wir, dass die Residuen einigermaßen symmetrisch um Null schwanken. Es ist kein Trend erkennbar und die Abweichungen scheinen auch nicht systematische mit dem Wert von $\widehat{\mathbf{y}}$ zu variieren. Wir haben also anscheinend ein *passendes* Modell gefunden, bei dem die Fehler näherungsweise konstante Varianz haben. Die beiden extremen Abweichungen sind auch in der zweiten Darstellung erkennbar. Bei Q-Q Plots liegen die Punkte idealerweise auf einer Geraden. Es sind zwar Abweichungen erkennbar, aber die scheinen im wesentlichen durch die beiden extremen Werte verursacht. In der Praxis würde man noch

keine Veranlassung sehen, an der Gültigkeit der abgeleiteten Tests oder Konfidenzintervalle zu zweifeln.

1.7 Lineare Regression in R

Die Anpassung Linearer Modelle an Daten wird in R durch die Funktion `lm` erledigt. Sie sollten sich unbedingt den Hilfetext zu dieser Funktion anschauen `?lm`. Die Funktion erlaubt das Spezifizieren von Modellen mit Hilfe einer sehr mächtigen Syntax, auf die wir hier nicht im Detail eingehen. Die wichtigsten Punkte für die Anwendung geben wir in der folgenden Zusammenstellung. Da der wichtigste Schritt die Darstellung des gewünschten Analysemodells in der Form `y~modell` ist, werden auch einige Befehle zur Spezifikation von Modellen in **R** angegeben.

`formula`	Formel des gewünschten Modells (siehe Modellspezifikation). Beispiel: `y ~ x`
`data`	(optional) Hier kann der Data-Frame angegeben werden, in dem sich die Beobachtungen befinden. Dies hat den Vorteil, dass bei der Modellspezifikation lediglich die entsprechenden Spaltennamen anstelle der Beobachtungsvektoren angegeben werden müssen. Beispiel: `data=Kniebeugen`
`subset`	(optional) Möglichkeit zur Auswahl einer Teilmenge der Beobachtungen
`+`	Hinzunahme einer Variable.
`.`	Hinzunahme aller Variablen aus dem angegebenen Data-Frame, die nicht auf der linken Seite der Tilde auftauchen.
`-`	Entfernen einer Variable. `-1` eliminiert den Achsenabschnitt aus dem Modell.
`I()`	Innerhalb von `I()` behalten mathematische Operatoren ihre ursprüngliche Bedeutung, so dass innerhalb der Modellspezifikation auch gerechnet werden kann.
`:`	Wechselwirkung zwischen zwei Kovariablen.
`*`	Fügt Kovariablen und sämtliche Wechselwirkungen hinzu. Beispiel: `x1*x2` entspricht `x1+x2+x1:x2`.
`^`	Sämtliche Wechselwirkungen bis zum angegeben Grad werden hinzugefügt. Beispiel: `(x1+x2+x3)^2` entspricht `x1+x2+x3+x1:x2+x1:x3+x2:x3`.

Bei der Anwendung empfiehlt es sich, die Ergebnisse der Modellanpassung zu speichern, etwa so: `fit<-lm(y~x)`. Das Objekt `fit` ist eine Liste und kann mit den üblichen **R**-Funktionen weiter bearbeitet werden. Einige sind besonders nützlich. Sie sollte alle diese Funktionen konkret an Beispielen anschauen.

`summary()`	liefert eine Zusammenfassung der Ergebnisse.
`coef()`	extrahiert die geschätzten Modell-Koeffizienten.
`fitted()`	liefert die geschätzten Werte $\widehat{\mathbf{y}}$.
`residuals()`	liefert entsprechend die Residuen
`model.matrix()`	erzeugt die in den Berechnungen verwendete Designmatrix
`predict()`	liefert Vorhersagen, auch für eine neue Werte der verwendeten Einflussgrößen
`plot()`	erzeugt Graphiken zum Check der Modellannahmen

1.8 Beispiele: Analyse Linearer Modelle mit R

Dei folgenden Beispiele zeigen Ihnen den R-Code zur Analyse der Eingangsbeispiele. Am besten, Sie vollziehen die einzelnen Programme nochmals nach. Änderungen und Ergänzungen können Sie dabei natürlich nach Belieben ausprobieren.

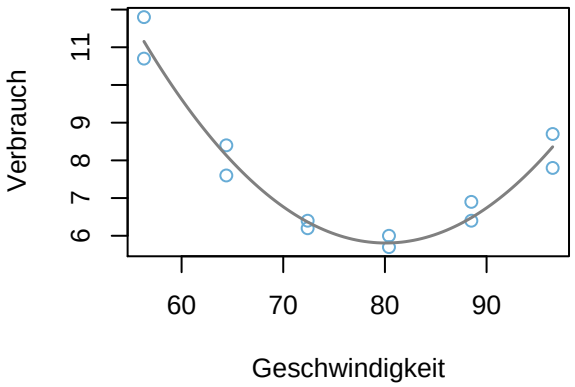
Optimale Fahrweise

Wir erzeugen den `data.frame` für das Beispiel wie folgt:

```
d <- data.frame(
       Geschwindigkeit = rep(c(56.3, 64.4, 72.4, 80.4, 88.5, 96.5), each = 2),
       Verbrauch = c(10.7, 11.8, 8.4, 7.6, 6.4, 6.2, 5.7, 6, 6.9, 6.4, 8.7, 7.8)
                )
```

Die Graphik mit Daten und angepasstem Modell lässt sich leicht generieren. Die Feinbearbeitung für die Darstellung unserer Lerneinheiten haben wir der Lesbarkeit halber weggelassen.

```
# Datenpunkte plot(Verbrauch~Geschwindigkeit, data=d, col="blue",
      xlab = "Geschwindigkeit", ylab = "Verbrauch")
# Modellanpassung
abc<-coef(fit<-lm(Verbrauch~Geschwindigkeit+I(Geschwindigkeit^2),data=d))
# Modellfunktion
curve(abc[1]+abc[2]*x+abc[3]*x^2,col=gray(0.5), lwd=1.5,add=T)
```



Die folgenden Tabelle erzeugen wir mit Hilfe von

```
> summary(fit)

Call:
lm(formula = Verbrauch ~ Geschwindigkeit + I(Geschwindigkeit^2),
    data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55857 -0.23800 -0.02226  0.28585  0.64484

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         66.4528325  4.4080788   15.07 1.08e-07 ***
Geschwindigkeit     -1.5146427  0.1180480  -12.83 4.35e-07 ***
I(Geschwindigkeit^2)  0.0094573  0.0007702   12.28 6.33e-07 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4304 on 9 degrees of freedom
Multiple R-squared: 0.9588,     Adjusted R-squared: 0.9497
F-statistic: 104.7 on 2 and 9 DF,  p-value: 5.843e-07
```

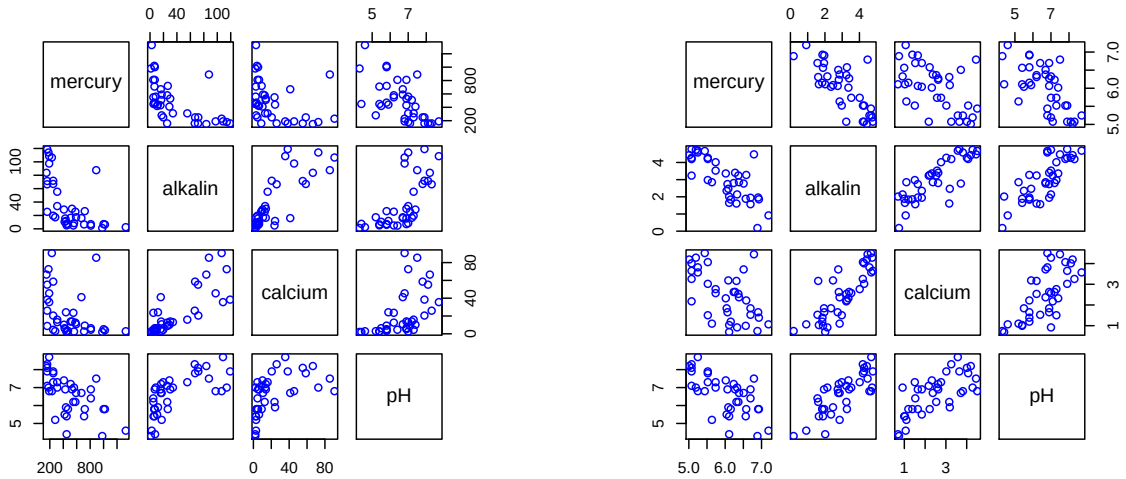
Sie können mit Hilfe der Koeffizienten die Geschwindigkeit mit dem geringsten Verbrauch ermitteln. Überlegen Sie dazu, welchen Wert die Ableitung der Modellfunktion im Minimum hat, und wie man die berechnet.

Analyse des Quecksilbergehalts in Abhängigkeit von der Alkalinität

Zunächst erzeugen wir ein `data.frame Quecksilber` mit den Daten. Wir verwenden die Konsole als Input (`file=stdin()`) und können so die Daten direkt einlesen.

```
Quecksilber<-read.table(file=stdin(), header=T)
   mercury alkalin calcium  pH
1     1330     2.5     2.9 4.6
2      250    19.6     4.5 7.3
3      450     5.2     2.8 5.4
4      160    71.4    55.2 8.1
5      720    26.4     9.2 5.8
6      810     4.8     4.6 6.4
7      710     6.6     2.7 5.4
8      510    16.5    13.8 7.2
9     1000     7.1     5.2 5.8
10     150    83.7    66.5 8.2
11     190   108.5    35.6 8.7
12    1020     6.4     4.0 5.8
13     450     7.5     2.0 4.4
14     590    17.3    10.7 6.7
15     810     7.0     6.3 6.9
16     420    10.5     6.3 5.5
17     530    30.0    13.9 6.9
18     310    55.4    15.9 7.3
19     470     6.3     3.3 5.8
20     250    67.0    58.6 7.8
21     410    28.8    10.2 7.4
22     160   119.1    38.4 7.9
23     160    25.4     8.8 7.1
24     230   106.5    90.7 6.8
25     560     8.5     2.5 7.0
26     890    87.6    85.5 7.5
27     180   114.0    72.6 7.0
28     190    97.5    45.5 6.8
29     440    11.8    24.2 5.9
30     160    66.5    26.0 8.3
31     670    16.0    41.2 6.7
32     550     5.0    23.6 6.2
33     580    25.6    12.6 6.2
34     980     1.2     2.1 4.3
35     310    34.0    13.1 7.0
36     430    15.5     5.2 6.9
37     280    17.3     3.0 5.2
38     250    71.8    20.5 7.9
```

Beim Quecksilberdatensatz geht es darum, die Belastung der Seen mit Quecksilber durch vorhandene Angaben zu pH Wert, Kalziumgehalt bzw. Alkalinität abzuschätzen. Die folgende Abbildung zeigt, dass eine logarithmische Transformation der Variablen einen näherungsweise lineare Zusammenhänge ergibt.

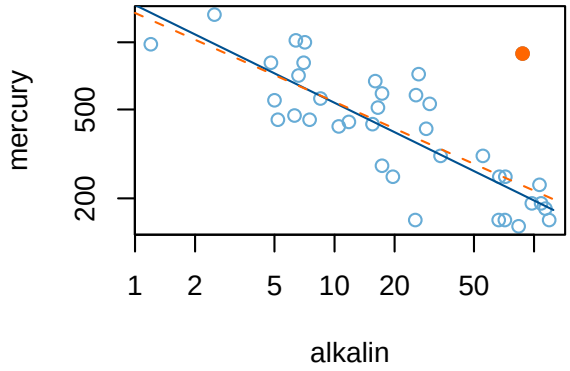


Variablen des Datensatzes links untransformiert, rechts transformiert, pH-Wert unverändert

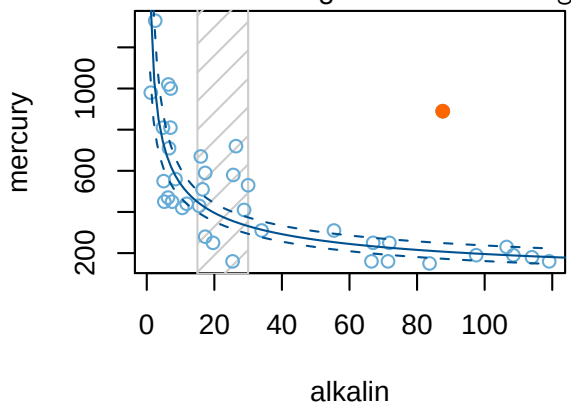
Die Abbildung der transformierten Werte zeigt weiter, dass zwischen den Logarithmen von Quecksilbergehalt und Alkalinität eine relativ enge lineare Beziehung besteht, die anderen Variablen weniger stark mit dem Quecksilbergehalt korrelieren, aber eng mit der Alkalinität zusammenhängen. Deshalb beschränken wir uns zunächst auf die Beziehung zwischen der Quecksilberkonzentration `mercury` und der Alkalinität `alkalin`.
Der Datensatz Nummer 26 stellt offensichtlich einen Ausreißer dar, der für die weiteren Analysen nicht berücksichtigt wird.
Zunächst werden die Ausgangsdaten logarithmiert. Wir erzeugen nun Anpassungen mit und ohne den Ausreißer.

```
Quecklog<-Quecksilber
Quecklog[,-4]<-log(Quecklog)
fit0<-lm(mercury~alkalin,data=Quecklog)
fit<-lm(mercury~alkalin,data=Quecklog[-26,])
ab0<-coef(fit0)
ab<-coef(fit)


# Bildbereich
opar<-par(mar=c(4, 4, 1, 1) + 0.1)
#
plot(mercury~alkalin,log="xy",data=Quecksilber,col=hellblau,pch=21)
curve(exp(ab[1]+ab[2]*log(x)),1,125,add=TRUE,col=blau)
points(87.6,890,col=aronge,pch=19)
curve(exp(ab0[1]+ab0[2]*log(x)),1,125,add=TRUE,lty=2, col=aronge)
par(opar)
```



Die Abbildung zeigt die angepasste Gerade für die Originaldaten in Orange und die bereinigten Daten in Blau. Die Beschriftung der Achsen haben wir in logarithmischem Maßstab vorgenommen. Die verwendeten Farben kann man in R selbst definieren, aber diese Details lassen wir hier weg.
Zum Abschluss zeigen wir Daten und angepasstes Model nochmals ohne die Achsen zu transformieren. Dazu wird lediglich der Befehl `log="xy"` weggelassen.



Im Bereich um den Wert `alkalin = 20` beobachten wir eine etwas größere Streuung der Quecksilberkonzentrationen.

Eine Zusammenfassung der Ergebnisse könnte man wie folgt formulieren:
Die beobachteten Quecksilberkonzentrationen stehen in einem engen Zusammenhang mit der Alkalinität. Die Anpassung einer Regressiongeraden für die logarithmierten Werte ergibt einen starken Abfall der Konzentrationen für Alkalinitat zwischen 0 und 20. Danach nehmen die Werte nur noch langsam ab. Bei den Messungen ist ein Ausreißer klar erkennbar.

1.9 Aufgaben

1. Beispieldatensätze:
   Bearbeiten Sie sämtliche Beispieldatensätze. Passen Sie jeweils geeignete Lineare Modelle an und interpretieren Sie die Ergebnisse. Verwenden Sie dazu geeignete Graphiken.

2. Bei den Anorexie-Daten fällt auf, dass die Verhaltenstherapie anscheinend keine Erfolge bringt. Versuchen Sie eine Anpassung mit Gruppe und Gewicht vor Therapie als Einflussgrößen für den Vergleich von Standard- und Familientherapie.
   Tip: textttfit<-lm(After Before+Group, data=Anorexie[Group!="g3,"])

3. Führen Sie in den bearbeiteten Beispielen den *Check* der Voraussetzungen durch.

4. Versuchen Sie eine Analyse des Kniebeugenexperiments. Untersuchen Sie dabei die Abhängigkeit des Pulswerts nach Belastung von Ruhepuls und Belastung. Versuchen

Sie herauszufinden, ob es genügt, den Einfluss der Belastung linear zu modellieren, oder ob eine quadratische Abhängigkeit nachweisbar ist.

1.10 Zusammenfassung

- Lineare Kombinationen von Einflussgrößen ergeben Lineare Modelle.
- Schätzungen in Linearen Modellen beruhen auf dem Kleinst-Quadrate Kriterium.
- Die Quadratsummen-Zerlegung beruht auf dem Satz des Pythagoras.
- Die Parameterschätzungen ergeben sich aus der Lösungen der Normalgleichungen, einem linearen Gleichungssystem
- Unter den Standardvoraussetzungen lassen sich Tests und Konfidenzintervalle für die Parameter angeben.
- Simultane Hypothesen werden mit Hilfe des *extra-sum-of squares*-Prinzips durch F-Tests überprüft.
- Graphische Darstellungen der Residuen dienen dem Check der Voraussetzungen.
- Die **R**-Funktion `lm()` gestattet eine einfache Modellspezifikation, die Analyse und die Ergebnisdarstellung für Lineare Modelle.