Scriptum - Fortgeschrittene Statistik

Robert Rein

Invalid Date

Table of contents

Vorwort

Dies ist das Skriptum für den Master-Statistikkurse Fortgeschrittene Statistik und ist die Vorlage für die Kurse LTC4 und SBG4. Es werden in den Kursen nicht alle Themen des Skriptums behandelt. Das Skriptum befindet sich derzeit noch in einem frühen Stadium, so dass die Inhalte noch nicht vollständig ausgearbeitet sind.

Part I

Statistik

Die erste Frage die sich im Umgang mit der Anwendung von Verfahren der Statistik stellt ist: Wofür benötigen wir Statistik überhaupt?

Beispielsweise wurden ein Datensatz gesammelt, bei dem zwei Gruppen miteinander verglichen werden, eine Treatmentgruppe (TRT) und eine Kontrollgruppe (CON). In beiden Gruppen wurden jeweils $N_i = 20$ Personen untersucht. Es wurde das folgende Ergebnis erhalten (siehe Figure 1).

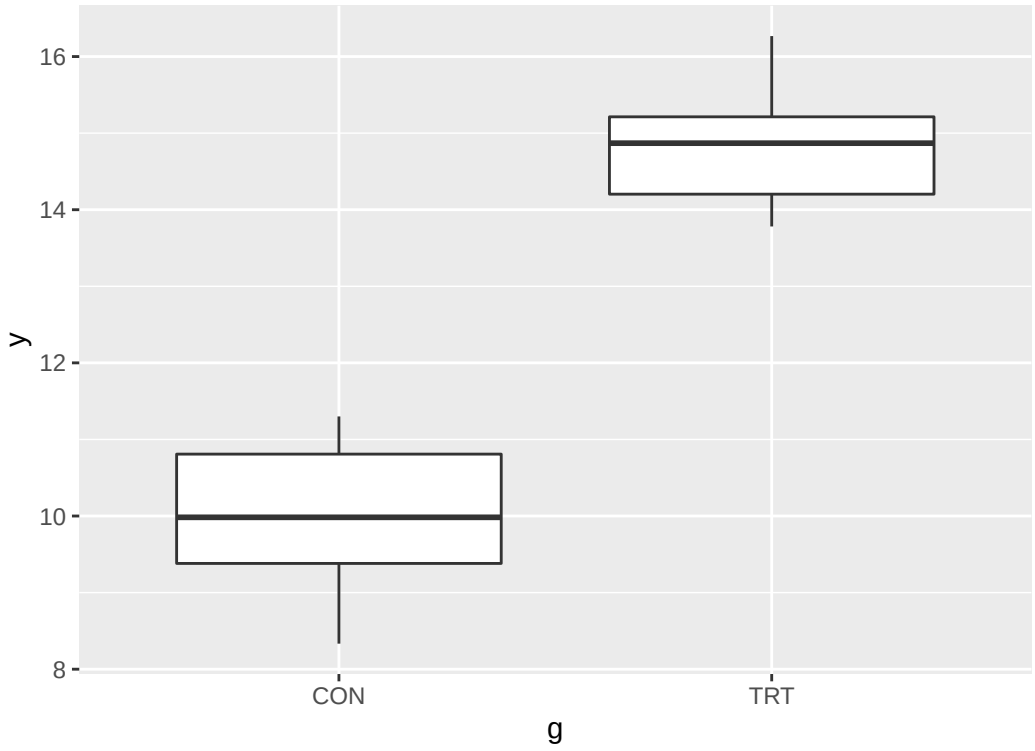


Figure 1: Boxplot der Kontroll- und der Treatmentgruppe bezüglich einer abhängigen Variable

Offensichtlich sind die Werte in der Treatmentgruppe deutlich höher als diejenigen in der Kontrollgruppe. Warum ist es nicht ausreichend das offensichtliche zu dokumentieren? Warum ist eine statistische Analyse der Daten notwendig?

Diese Fragestellung wird in dem folgenden Abschnitt untersucht. Gleichzeitig werden die notwendigen Werkzeuge entwickelt um die verschiendenen Schritte die einer statistische Analyse von Daten zugrundeliegenen zu verstehen und anwenden zu können.

1 Eine kleine Welt der Unsicherheit

Beginnen wir mit eine einfachen Modell. Dazu nehmen wir eine kleine Welt die nur aus 20 Personen besteht. In Figure 1.1 können wir alle Personen einzeln sehen. Die Gesamtheit aller Personen (allgmeine Objekte) über die wir eine Aussage treffen woll bezeichnen wir als eine Population.

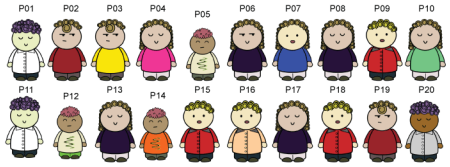


Figure 1.1: Eine kleine Welt

Definition 1.1 (Population)**.** Eine Population oder auch die Grundgesamtheit ist Gesamtheit aller Objekte/Dinge/Personen über die eine Aussage getroffen werden soll.

1.1 Ein Experiment

Wir wollen nun eine Krafttrainingsstudie durchführen um die Beinkraft zu erhöhen. Wir haben allerdings nur sehr wenige Ressourcen (bzw. wir sind faul) und können insgesamt nur sechs Messungen durchführen. Aus einem kürzlich durchgeführten Census haben wir aber die Kraftwerte der ganzen Population. Wir stellen die Kraftwerte zunächst mittels einer Tabelle dar (siehe Table 1.1)

Table 1.1: Kraftwerte (in Newton) der Lummerländer an der einbeinigen Beinpresse

ID	Kraft[N]	ID	Kraft[N]
P01	2414	P11	2243
P02	2462	P12	2497
P03	2178	P13	1800
P04	2013	P14	2152
P05	2194	P15	2089

ID	Kraft[N]	ID	Kraft[N]
P06	2425	P16	2090
P07	2305	P17	3200
P08	2117	P18	2196
P09	2298	P19	2485
P10	2228	P20	2440

Selbst bei 20 Werten ist diese Darstellung wenig übersichtlich. Wir könnten zwar Zeile für Zeile durchgehen und nach etwas notieren und suchen würden wir sehen das der Maximalwert bei 3200N für P17 und der Minimalwert von Person P13 bei 1800N liegt. Aber wirklich einfach ist diese Darstellung nicht. Für solche univariaten Daten (uni = eins) kann eine übersichtlichere Darstellung mittels eines sogenannten Dotplots erreicht werden (siehe Figure 1.2).

Figure 1.2: Dotplot der Lummerlandkraftdaten

Hier kann deutlich schneller abgelesen werden was das Minimum und das Maximum der Daten ist, sowie es kann auch direkt abgeschätzt werden in welchem Bereich sich der Großteil der Daten befindet. Allerdings wird durch diese Art der Darstellung die Information über welche Person die jeweiligen Werte besitzt nicht mehr dargestellt. Dies stellt in den meisten Fällen allerdings kein Problem dar, da wir in den meisten Fällen aussagen über die Gruppe und weniger über einzelne Personen machen wollen.

Gehen wir jetzt von der folgenden Fragestellung aus. Wir wollen den Gesundheitsstatus unserer Lummerländer verbessern und wollen dazu ein Krafttraining durchführen. Da evidenzbasiert arbeiten wollen, möchten wir überprüfen ob wirklich ein Verbesserung der Kraft durch das Training stattgefunden hat. Da es sich aber gleichzeitig um unsere selbst geschaffene Welt handelt führen wir natürlich ein perfektes Krafttraining, eine perfekte Intervention, durch. D.h wir stellen uns immer wieder als unwissend da und geben vor das wir gar nicht wissen, das das Training perfekt effektiv ist.

D.h. wir führen gleichzeitig ein Gedankenexperiment durch. Wir führen ein Krafttraining für die Beine durch. Das Training ist perfekt und verbessert die Kraftleistung um genau +100N. Dieser Kraftzuwachs unabhängig davon welche Person aus unserer Population das Training durchführt (Warum ist das keine realistische Annahme?). Wir wollen zwei Gruppen

miteinander vergleichen eine Interventionsgruppe und eine Kontrollgruppe. In beiden Gruppen sollen jeweils $n_{\text{TRT}} = n_{\text{CON}} = 3$ TeilnehmerInnen bzw. Teilnehmer einbezogen werden da wir nicht mehr Ressourcen für mehr ProbandInnen haben.

Die erste Frage die sich nun stellt ist wie wählen wir die sechs Personen aus unserer Population aus und wie teilen wir die sechs Personen in die beiden Gruppen? Nach etwas überlegen kommen wir darauf, dass wir am besten eine zufällige Stichprobe ziehen sollten (Warum?).

Definition 1.2 (Stichprobe)**.** Eine Stichprobe ist eine Teilmenge der Objekte aus der Population.

Definition 1.3 (Zufallsstichprobe)**.** Eine Zufallsstichprobe ist eine Teilmenge der Objekte aus der Population die *zufällig* ausgewählt wurde.

Diese sechs Personen, unsere Stichprobe, wird dann wiederum zufällig auf die beiden Gruppen aufgeteilt.

Ein Zufallszahlengenerator hat die Zahlen $i = \{3, 7, 8, 9, 10, 20\}$ gezogen. Die entsprechenden Personen werden aus der Population ausgewählt und wiederum zufällig in die beiden Gruppen aufgeteilt (siehe Table 1.2).

Table 1.2: Zufällig ausgewählte Stichprobe der Kontrollgruppe (CON) und der Interventionsgruppe (TRT).

ID	Kraft[N]	Gruppe
P08	2117	CON
P09	2298	CON
P03	2178	CON
P07	2305	TRT
P10	2228	TRT
P20	2440	TRT

Mit diesen sechs Personen führen wir jetzt unser Experiment durch. Die drei Personen aus der Kontrollgruppe, unterlaufen im Interventionszeitraum nur ein Stretchtraining während die Interventionsgruppe zweimal die Woche für 12 Wochen unser perfektes Krafttraining durchführt. Nach diesem Zeitraum messen wir alle Personen aus beiden Gruppen und erhalten das folgende Ergebnis (siehe Table 1.3).

Table 1.3: Ergebnis der Intervention in Experiment 1 für die Kontroll- und die Interventionsgruppe.

(a) Kontrollgruppe

ID	Kraft[N]
P08	2117
P09	2298
P03	2178
$\bar{K}$	2198

(b) Interventionsgruppe

ID	Kraft[N]
P07	2405
P10	2328
P20	2540
$\bar{K}$	2424

Für beide Gruppen ist jeweils der Mittelwert berechnet worden, um die Wert miteinander vergleichen zu können. Später werden wir noch weitere Maße kennenlernen die es ermöglichen zwei Mengen von Werten miteinander zu vergleichen.

Definition 1.4 (Mittelwert)**.** Der Mittelwert über n Werte berechnet sich nach der Formel:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

Der Mittelwert wird mit einem Strich über der Variable dargestellt.

Damit lernen wir direkt auch ein neues Konzept kennen. Nämlich das der Statistik. Ein Wert der auf der erhobenen Stichprobe berechnet wird, wird als Statistik bezeichnet.

Definition 1.5 (Statistik)**.** Ein auf einer Stichprobe berechnet Wert, wird als Statistik bezeichnet.

Um jetzt Unterschied zwischen den beiden Gruppen zu untersuchen berechnen wir die Differenz D zwischen den beiden Mittelwerten $D = \bar{K}_{\text{TRT}} - \bar{K}_{\text{CON}}$. Die Differenz kann natürlich auch in die andere Richtung berechnet werden und es würde sich das Vorzeichen ändern. Hier gibt es keine Vorgaben, sondern die Richtung kann frei bestimmt werden. Wenn bekannt ist in welcher Richtung der Unterschied berechnet wird, dann stellt dies keine Problem dar. Im vorliegenden Fall ziehen wir die Interventionsgruppe von der Kontrollgruppe ab, da wir davon ausgehen, dass die Intervention zu einer Krafterhöhung führt und wir dadurch einen positiven Unterschied erhalten (vgl. Equation 1.2)

$$D = 2424N - 2198N = 226N \tag{1.2}$$

Da der Wert D, wiederum auf den Daten der Stichprobe berechnet wird, handelt es sich ebenfalls um eine Statistik.

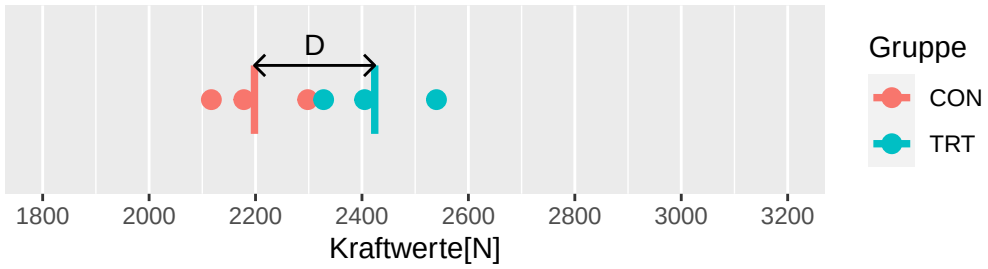


Figure 1.3: Dotplot der beiden Stichproben. Senkrechte Striche zeigen die jeweiligen Mittelwerte an.

In Figure 1.3 sind die Werte der beiden Gruppen, deren Mittelwerte $\bar{K}_{\text{CON}}$ und $\bar{K}_{\text{TRT}}$ und der Unterschied D zwischen diesen abgebildet. Wie erwartet zeigt die Interventionsgruppen den höheren Kraftwert im Vergleich zu der Kontrollgruppe. Allerdings ist der Wert mit $D = 226$ größer als der tatsächliche Zuwachs von $\Delta_{\text{Training}} = 100$ (Warum ist das so?).

Der Unterschied zwischen den beiden Gruppen ist natürlich auch zum Teil auf die Unterschiede die zwischen den beiden Gruppen vor der Intervention bestanden haben zurück zu führen. Was wäre denn passiert, wenn wir eine andere Stichprobe gezogen hätten?

Sei $i = \{12, 2, 19, 4, 8, 16\}$ eine zweite Stichprobe. Dies würde zu den folgenden Werten führen nach der Intervention führen.

Table 1.4: Ergebnis der Intervention in Experiment 2 für die Kontroll- und die Interventionsgruppe.

ID	Kraft[N]	Gruppe
P08	2117	CON
P09	2298	CON
P03	2178	CON
P07	2405	TRT
P10	2328	TRT
P20	2540	TRT

In Figure 1.4 sind wiederum die Datenpunkte, Mittelwerte und der Unterschied abgetragen. In diesem Fall ist allerdings die Differenz zwischen den beiden Gruppen genau in der anderen Richtung $D = -308$, so dass die Interpretation des Ergebnisses genau in der anderen Richtung wäre. Nämlich, nicht nur hat das Krafttraining zu keiner Verbesserung in der Kraftfähigkeit geführt, sondern zu einer Verschlechterung!

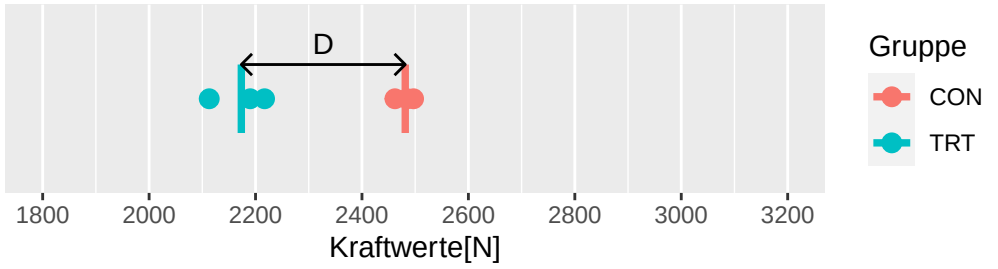


Figure 1.4: Dotplot der beiden Stichproben in Experiment 2. Senkrechte Striche zeigen die jeweiligen Mittelwerte an.

Es hätte aber auch sein können, das wir noch eine andere Stichprobe gezogen hätten, z.B. $i = \{6, 5, 7, 20, 14, 16\}$. Dies würde zu dem folgenden Ergebnis führen (siehe Table 1.5).

Table 1.5: Mittelwertsdaten aus Experiment 3 und der Unterschied D zwischen den beiden Gruppenmittelwerten

Gruppe	Kraft[N]
CON	2308
TRT	2327
D	19

In diesem Fall haben wird zwar wieder einen positiven Unterschied zwischen den beiden Gruppen in der zu erwartenden Richtung gefunden. Der Unterschied von $D = 19$ ist allerdings deutlich kleiner als das tatsächlichen $\Delta = 100$. Daher würden wir möglicherweise das Ergebnis so interpretieren, führen, dass wir das Krafttraining als ineffektiv bewerten würden und keine Empfehlung ausprechen.

Zusammengenommen, ist keines der Ergebnisse 100% korrekt. Entweder der Unterschied zwischen den beiden Gruppen ist deutlich zu groß, oder in der anderen Richtung oder deutlich zu klein. Das Ergebnis des Experiments hängt ursächlich damit zusammen, welche Stichprobe gezogen wird. Diese Einsicht gilt in jedem Fall generell für jedes Ergebnis eines Experiments.

Das Phänomen, das der Wert der berechneten Statistik zwischen Wiederholungen des Experiments schwankt wird als Stichprobenvariabilität bezeichnet.

Definition 1.6 (Stichprobenvariabilität)**.** Durch die Anwendung von Zufallsstichproben, variiert eine auf den Daten berechnete Statistik. Die Variabilität wird als Stichprobenvariabilität bezeichnet.

Streng genommen, führt die Stichprobenvariabilität für sich genommen noch nicht dazu, das sich die Statistik zwischen Wiederholungen des Experiments verändert, sondern die zu untersuchenden Werte in der Population müssen selbst auch noch eine Streuung aufweisen. Wenn wir eine Population untersuchen würden, bei der alle Personen die gleiche Beinkraft hätten, würden unterschiedliche Stichproben immer den gleichen Mittelwert haben und wiederholte Durchführung des Experiment würden immer wieder zu dem selben Ergebnis führen. Dieser Fall ist in der Realität aber praktisch nie gegeben und sämtlich Parameter für die wir uns hier interessieren zeigen immer eine natürlich Streuung in der Population. Diese Streuung in der Population führt daher zu dem besagten Ergebnis, das das gleiche Experiment mehrmals wiederholt zu unterschiedlichen Zufallsstichproben führt und dementsprechend immer zu unterschiedlichen Ergebnissen führt.

Daher ist eine der zentrale Aufgabe der Statistik mit dieser Variabilität umzugehen und die Forscherin trotzdem in die Lage zu versetzen rationale Entscheidungen zu treffen. Eine implizite Kernannahme dabei ist, das wir mit Hilfe von Daten überhaupt etwas über die Welt lernen können. D.h. das uns die Erhebung von Daten überhaupt auch in die Lage versetzt rationale Entscheidungen zu treffen. Entscheidungen wie ein spezialisiertes Krafttraining mit einer klinischen Population durchzführen oder eine bestimmte taktische Variante mit meiner Mannschaft zu trainieren um die Gegner besser auszuspielen. Alle diese Entscheidungen sollten rational vor dem Hintergrund von Variabilität getroffen werden und auch möglichst oft korrekte Entscheidungen zu treffen. Wie wir sehen werden, kann uns die Statistik leider nicht garantieren immer die korrekte Entscheidungen zu treffen. Nochmal auf den Punkt gebracht nach Wild and Seber (2000, 28)

> The subject matter of statistics is the process of finding out more about the real world by collecting and then making sense of data.

Untersuchen wir jedoch zunächst unsere Einsicht, das Wiederholungen des gleichen Experiments zu unterschiedlichen Ergebnissen führt, weiter. In unserem Beispiel aus Lummerland haben wir nämlich den Vorteil, das uns die Wahrheit bekannt ist. In Figure 1.5 ist die Verteilung unsere bisheringen drei Ds abgetragen.

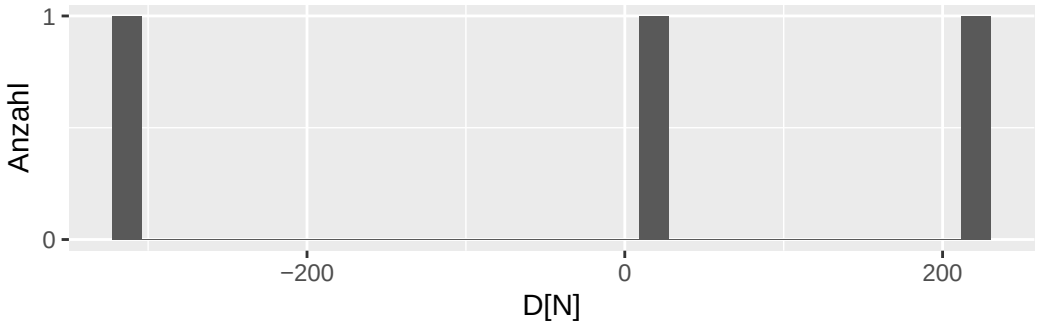


Figure 1.5: Bisherige Verteilung der Unterschiede D

Die drei Werte liegen ja relativ weiter auseiander. Eien Anschlussfrage könnte jetzt sein: "*Welche weiteren Werte sind denn überhaupt möglich mit der vorliegenden Population?*".

1.2 Die Stichprobenverteilung

Wir können jetzt ja einfach mal das Experiment anfangen zu wiederholen. In Figure 1.6 sind mal 15 verschiedene Stichproben abgetragen. Wir haben in jeder Zeile jeweils sechs TeilnehmerInnen gezogen. Drei für die Kontrollgruppe und drei für die Inervationsgruppe. Für jede dieser Zeilen können wir jeweils den Gruppenmittelwert berechnen und den Unterschied D bestimmen.

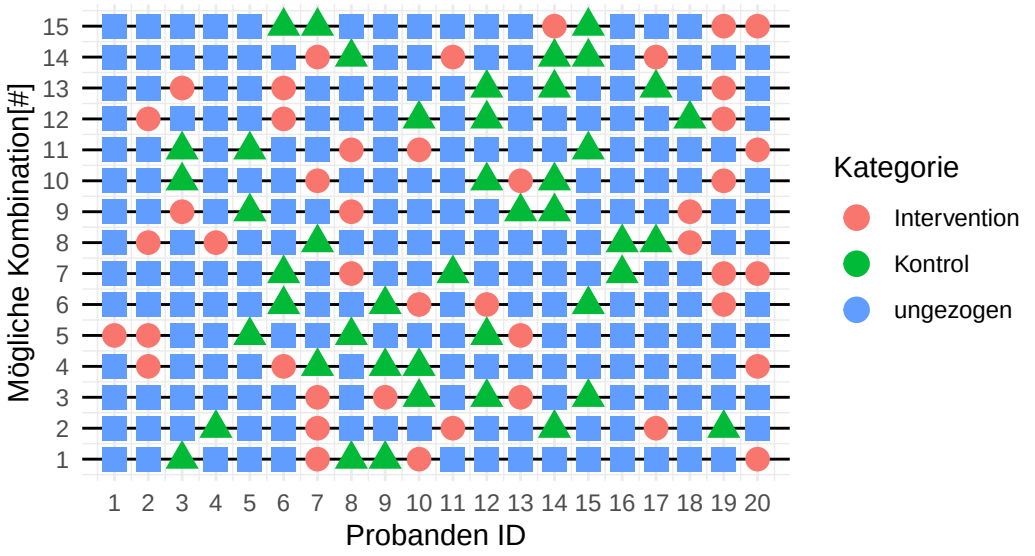


Figure 1.6: Beispiele für verschiedene Möglichkeiten zwei Stichproben mit jeweils $n_i = 3$ aus der Population zu ziehen

Warum eigentlich bei 15 aufhören. Wir haben ja den Vorteil, das unsere Population relativ übersichtlich ist. Vielleicht können wir uns ja noch aus unserer Schulezeit an Kombinatorik erinnern. Da haben wir den Binomialkoeffizienten kennengelernt. Die Anzahl der möglichken Kombination von k Elementen aus einer Menge von n Elementen berechnet sich nach:

$$\text{Anzahl} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.3}$$

In unserem Fall wollen wir zunächst sechs Elemente aus $N = 20$ auswählen und dann drei Elemente aus den sechs gezogenen Elementen auswählen um diese entweder der Intervention-

sgruppe oder der Kontrollgruppe zu zuweisen (Warum brauchen wir uns nur eine Gruppe anzuschauen?). Die Anzahl der möglichen Stichprobenkombinationen ist folglich:

$$\text{Anzahl} = \binom{20}{6}\binom{6}{3} = 7.752 \times 10^5 \tag{1.4}$$

Das sind jetzt natürlich selbst bei dieser kleinen Population ein große Menge von einzelnen Experimenten, aber dafür sind Computer da, die können alle diese Experiment in kurzer Zeit durchführen. In Figure 1.7 ist die Verteilung aller möglichen Experimentausgänge, d.h. alle Differenzen D zwischen der Interventions- und der Kontrollgruppe, abgebildet.

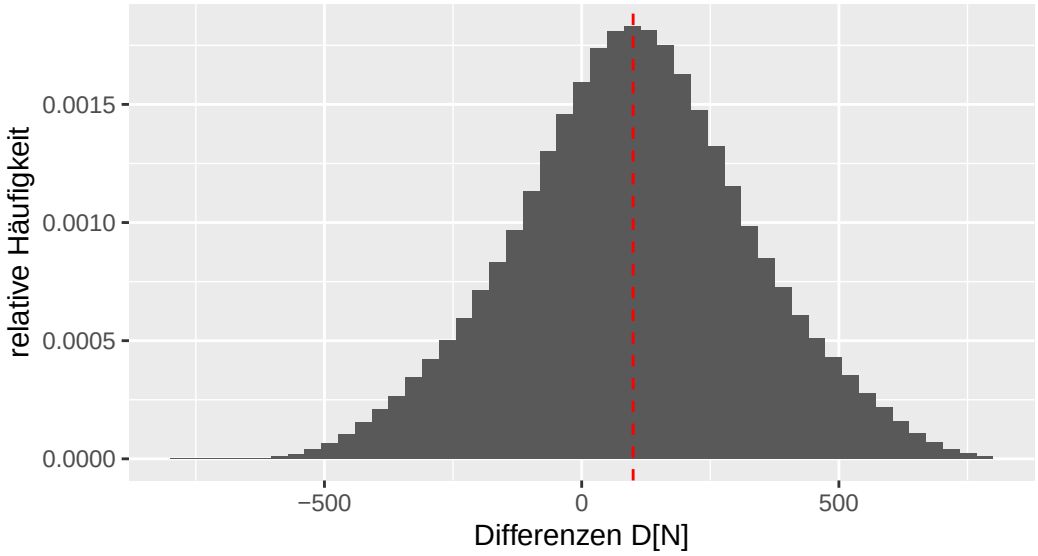


Figure 1.7: Verteilung aller möglichen Differenzen zwischen Kontroll- und Interventionsgruppe bei einer Intervention mit $\Delta = 100$ (im Graphen mittels der roten Linie angezeigt).

Auf der x-Achse sind die möglichen Differenzen D abgetragen, während auf der y-Achse die relative Häufigkeit, d.h. die Häufigkeit für einen bestimmten D-Wert geteilt durch die Anzahl 7.752×10^5 aller möglichen Werte. Die Verteilung der D's wird als Stichprobenverteilung bezeichnet.

Definition 1.7. Die Stichprobenverteilung kennzeichnet die Verteilung der beobachteten Statistik.

Die Figure 1.7 zeigt, dass die überwiegende Anzahl der Ausgänge tatsächlich auch im Bereich von $\Delta = 100$ liegen. Noch präziser das Maximum der Verteilung, also die höchste relative

Häufigkeit liegt genau auf der roten Linie. Dies sollte uns etwas beruhigen, denn es zeigt, das unsere Art der Herangehensweise mittels zweier Stichproben auch tatsächlich in den meisten Fällen einen nahezu korrekten Wert ermittelt. Allerdings zeigt die Stichprobenverteilung auch das Werte am rechten Ende die deutlich zu hoch sind wie auch Werte am linken Ende der Verteilung die deutlich in der falschen Richtung möglich sind. Das bedeutet, wenn wir das Experiment nur einmal durchführen wir uns eigentlich nie sich sein können, welches dieser vielen Experimente wir durchgeführt haben. Es ist zwar warscheinlicher, dass wir eins aus der Mitte der Verteilung durchgeführt haben, einfach da die Anzahl größer ist, aber wir haben keine 100% Versicherung, das wir nicht *Pech* gehabt haben und das Experiment ganz links mit $D = -500$ oder aber das Experiment ganz rechts mit $D = 700$ durchgeführt haben. Diese Unsicherheit wird leider keine Art von Experiment vollständig auflösen können. Eine weitere Eigenschaft der Verteilung ist ihre Symmetrie bezüglich des Maximums mit abnehmenden relativen Häufigkeiten umso weiter von Maximum D entfernt ist (Warum macht das heuristisch Sinn?).

Die Darstellungsform von Figure 1.7 wird als Histogramm bezeichnet und eignet sich vor allem dazu die Verteilung einer Variablen z.B. x darzustellen. Dazu wird der Wertebereich von x zwischen dem Minimalwert x_{min} und dem Maximalwert x_{max} in k gleich große Intervalle unterteilt und die Anzahl der Werte innerhalb jedes Intervalls wird abgezählt und durch die Anzahl der Gesamtwerte geteilt um die relative Häufigkeit zu erhalten.

Zum Beispiel für die Werte:

$$x_i \in \{1, 1.5, 1.8, 2.1, 2.2, 2.7, 2.8, 3.5, 4\}$$

könnte das Histogram ermittelt werden, indem der Bereich von $x_{\text{min}} = 1$ bis $x_{\text{max}} = 4$ in vier Intervalle unterteilt wird und dann die Anzahl der Werte in den jeweiligen Intervallen ermittelt wird (siehe Figure 1.8). Die ermittelte Anzahl würde dann noch durch die Gesamtanzahl 9 der Elemente geteilt um die relative Häufigkeit zu berechnen.

Die Form des Histogramms hängt davon ab wie viele Intervalle verwendet werden, so wird die Auflösung mit mehr Intervallen besser, aber es die Anzahl wird geringer und andersherum wird die Auflösung mit weniger Intervallen geringer aber die Anzahl der Elemente pro Intervall wird größer und somit stabiler. Daher sollte in den meisten praktischen Fällen die Anzahl variiert werden um sicher zu gehen, das nicht nur zufällig eine spezielle Darstellung verwendet wird.

Zurück zu unserer Verteilung von D unter $\Delta = 100\text{N}$ in Figure 1.7. Wie schon besprochen sind alle Werte zwischen etwa $D = -500N$ und $D = 700\text{N}$ plausibel bzw. möglich. Schauen wir uns doch einmal an, was passiert wenn das Training überhaupt nichts bringen würde und es keine Verbesserung gibt, also $\Delta = 0$.

Die Verteilung in Figure 1.9 sieht praktisch genau gleich aus, wie diejenige für $\Delta = 100$. Der einzige Unterschied ist lediglich das sie nach links verschoben ist und zwar scheinbar genau um die 100N Unterschied zwischen den beiden Δs. Dies ist letztendlich auch nicht weiter verwunderlich, bei der Berechnung des Unterschied D zwischen den beiden Gruppen

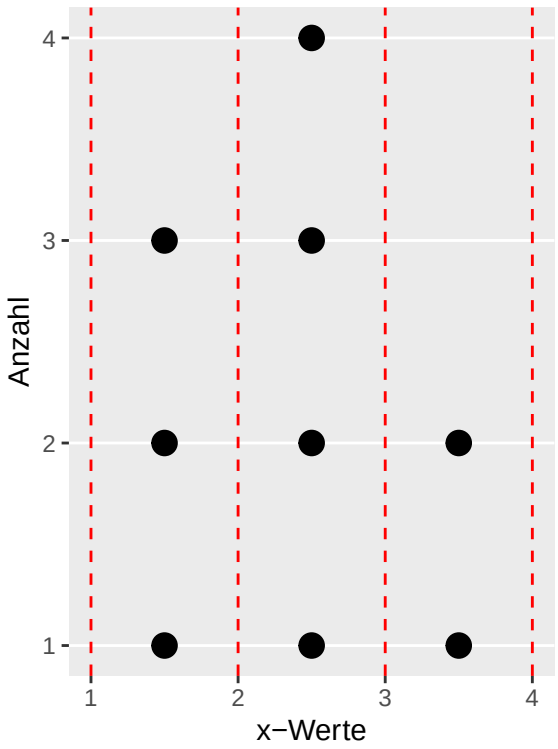


Figure 1.8: Beispiel für die Darstellung eines Histogramms für die Daten x_i.

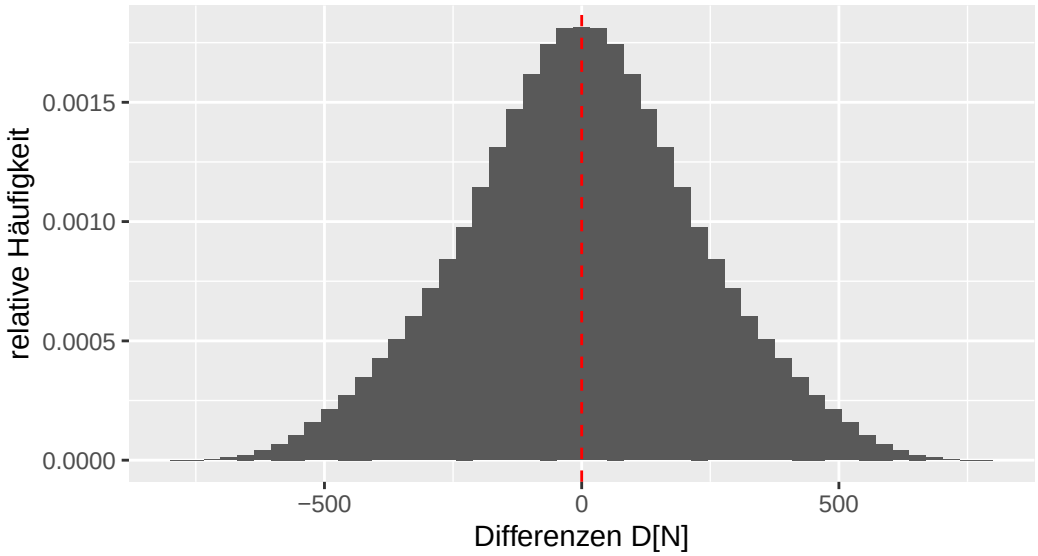


Figure 1.9: Verteilung aller möglichen Differenzen zwischen Kontroll- und Interventionsgruppe wenn $\Delta = 0$ (rote Linie).

kommen in beiden Fällen genau die gleichen Kombination vor. Bei $\Delta = 100$ wird aber zu der Interventionsgruppe das Δ dazuaddiert bevor die Differenz der Mittelwerte berechnet wird. Da aber gilt:

$$D = \frac{1}{3}\sum_{i=1}^{3} x_{\text{KON}i} - \frac{1}{3}\sum_{j=1}^{3}(x_{\text{TRT}j} + \Delta) = \bar{x}_{\text{KON}} - \bar{x}_{\text{TRT}} + \Delta$$

Daher bleibt die Form der Verteilung immer genau gleich und wird lediglich um den Wert Δ im Vergleich zur Nullintervention verschoben. Wobei mit Nullintervention Umgangssprachlich die Intervention bezeichnet, bei der nichts passiert also $\Delta = 0$ gilt.

1.3 Unsicherheit in Lummerland

Das führt jetzt aber zu einem Problem für uns. Gehen wir jetzt nämlich von diesen beiden Annahmen aus, das entweder die Intervention effektiv ist $\Delta = 100$ gilt oder das die Intervention nichts bringt also $\Delta = 0$ gilt. Wenn wir diese beiden Verteilungen übereinander legen erhalten wir Figure 1.10. Wir haben die Darstellung jetzt etwas verändert und eine Kurve durch die relativen Häufigkeiten gelegt. Dieser Graphen wird jetzt nicht mehr als Histogramm sondern als Dichtegraph bezeichnet.

In Figure 1.10 ist klar zu sehen, dass die beiden Graphen zu großen Teilen überlappen und dazu noch in einem Bereich wo beide Ergebnisse ihrer höchsten relativen Häufigkeiten, also auch die größte Wahrscheinlichkeit haben unter den jeweiligen Annahmen aufzutreten. Unser Problem besteht jetzt darin, dass wir in der Realität gar nicht diese Information haben welchen Effekt unser Training auf die Stichprobe ausführt. Wenn wir dies wüssten, dann müssten wir das Experiment ja gar nicht durchführen. Wir haben im Normalfall nur ein einziges Ergebnis, nämlich den Ausgang unseres einen Experiments.

Wenn wir jetzt unser Experiment einmal durchgeführt haben und ein einziges Ergebnis für D erhalten haben, sei zum Beispiel $D = 50$ dann haben wir ein Zuweisungsproblem (siehe Figure 1.11). Wie weisen wir unser Ergebnis jetzt den beiden möglichen Realität zu? Einmal kann es sein, das das Krafttraining aber auch gar nichts gebracht hat und wir haben lediglich eine der vielen möglichen Stichprobenkombination beobachtet haben die zu einem positiven Wert für D führt. Oder aber das Krafttraining ist effektiv gewesen und hat zu einer Verbesserung von $\Delta = 100$N geführt und wir haben lediglich ein Stichprobenkombination aus den vielen möglichen Stichprobenkombination gezogen die zu einem Ergebnis von $D = 50$ führt. Noch mal, in der Realität wissen wir nicht welche der beiden Annahmen korrekt ist und können es auch nie vollständig wissen. Denn egal wie viele Experimente wir machen, wir können immer den zwar unwahrscheinlichen aber nicht unmöglichen Fall haben, das wir nur Werte beispielsweise aus dem linken Teil der Verteilung beobachten. Das heißt wir haben immer mit einer Ungewissheit zu kämpfen. Wir können nicht im Sinne eines Beweises zeigen, das das Training effektiv ist.

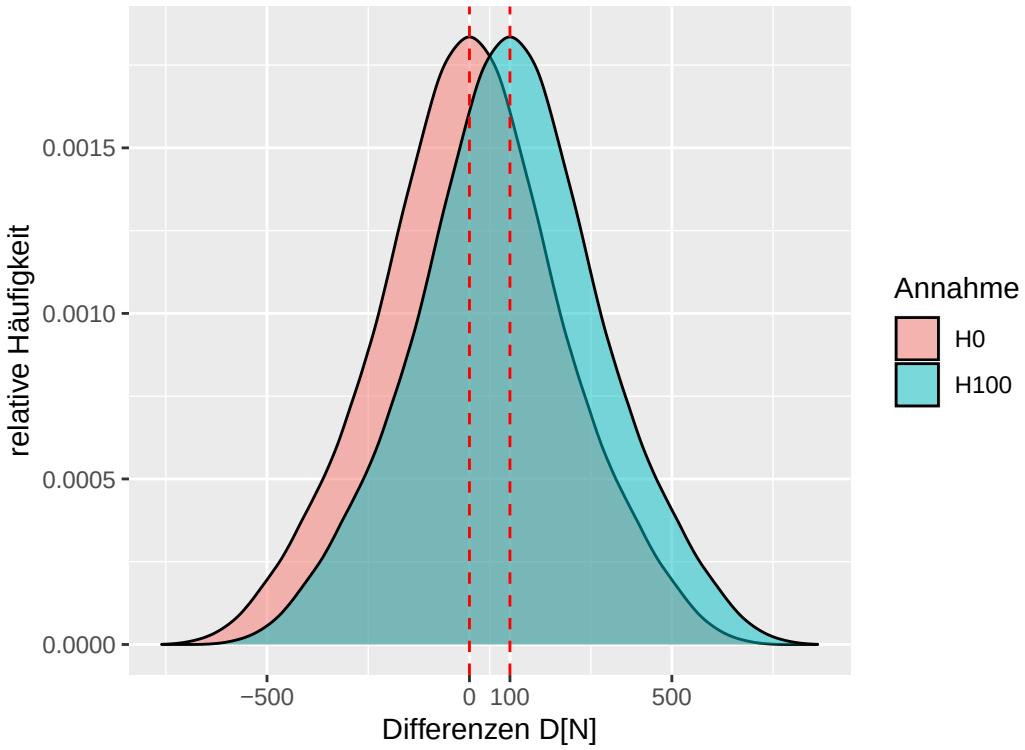


Figure 1.10: Verteilung aller möglichen Differenzen zwischen Kontroll- und Interventionsgruppe wenn $\Delta = 0$ und $\Delta = 100$.

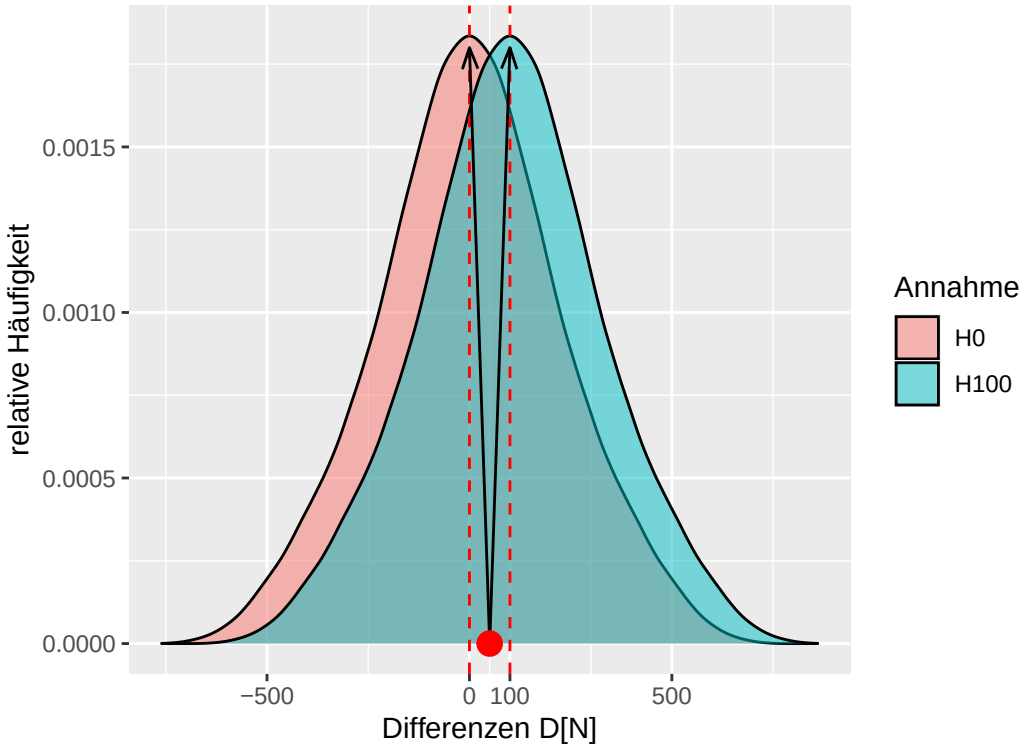


Figure 1.11: Zuweisung eines beobachteten Unterschieds D nach einem Experiment

Table 1.6: Entscheidungsmöglichkeiten wenn entweder H_0 oder H_{1} zutrifft.

	Realität	
	H_0	H_1
H_0	korrekt	β
H_1	α	korrekt

Die Methoden der Statistik liefern uns nun Werkzeuge an die Hand um trotzdem rational zu Entscheiden welche der beiden Annahmen möglicherweise wahrscheinlicher ist. Gleichzeitig ermöglicht uns die Statistik abzuschätzen respektive zu berechnen wie groß die Unsicherheit in dieser Entscheidung ist. Die Statistik sagt dabei immer nur etwas über die beobachteten Daten aus. Die Statistik sagt jedoch nichts über die zugrundeliegenden wissenschaftlichen Theorien aus.

Schauen wir uns jetzt als vorläufig letzten Punkt an welche Entscheidungsmöglichkeiten wir haben.

1.4 Eine Entscheidung treffen

Wir hatten im Beispiel zwei verschiedene Annahmen, einmal das das Training nichts bringt und keine Verbesserung der Kraftfähigkeit folgt $\Delta = 0N$. Andererseits hatten wir das Beispiel gestartet damit, dass die Kraftfähigkeit um $100N$ zunimmt, also $\Delta = 100N$. Wie bezeichnen jetzt diese beiden Annahmen als Hypothesen und bezeichnen $\Delta = 0N$ als die Nullhypothese H_0 und $\Delta = 100N$ als die Alternativhypothese H_1.

Wenn wir jetzt das Experiment durchgeführt haben, können wir uns also entweder für die H_0 oder die H_1 entscheiden. Aus Gründen der Symmetrie ist dies gleichbedeutend wenn wir uns nur auf die H_0 fokussieren und entweder die H_0 annehmen bzw. beibehalten oder verwerfen also uns gegen H_0 entscheiden.

In Table 1.6 sind die verschiedenen Entscheidungsmöglichkeiten abgetragen. In der Realität gehen wir, wie gesagt, von zwei Fällen aus. Entweder trifft die H_0 oder die H_1 zu. Wenn die $H_=$ zutrifft und wir uns für die H_0 entscheiden, dann haben wir eine korrekte Entscheidung getroffen. Wenn H_0 zutrifft und wir allerdings die H_0 ablehnen, also uns für die H_1 entscheiden ist unsere Entscheidung falsch und wir begehen einen Fehler. Dieser Fehler wird als Fehler 1. Art bzw. α-Fehler bezeichnet. Trifft in der Realität dagegen die H_1 zu und wir entscheiden uns gegen die H_0 und für die H_1, dann haben wir wiederum eine korrekte Entscheidung getroffen. Zuletzt, wenn die H_1 zutrifft und wir uns aber für die H_0 entscheiden, also die H_0 beibehalten bzw. uns gegen die H_1 entscheiden, treffen wir wieder eine falsche Entscheidung. Dieser Fehler wird als Fehler 2. Art, bzw. β-Fehler bezeichnet.

Definition 1.8. Wenn eine Entscheidung gegen die H_0 getroffen wird, obwohl die H_0 korrekt ist, wird dies als α-Fehler bezeichnet.

Definition 1.9. Wenn eine Entscheidung gegen die H_1 getroffen wird, obwohl die H_1 korrekt ist, wird dies als β-Fehler bezeichnet.

2 Statistische Signifikanz, p-Wert und Power

Im vorherigen Kapitel haben wir gesehen, wie Unsicherheit ein zentrales Problem bei der Interpretation von Ergebnissen von Experimenten oder Daten allgemein ist. Im nun folgenden Abschnitt wollen wir eine Prozess aufbauen, der es uns vor dem Hintergrund dieser Unsicherheit eine Entscheidung zu treffen.

2.1 Wie treffe ich eine Entscheidung?

In unserem kleine Welt Bespiel waren wir in der komfortablen Position, das wir genau wussten was passiert bzw. welcher Prozess unseren beobachteten Datenpunkt erzeugt hat. D.h wir kannten den datengenerieren Prozesses.

Definition 2.1 (Datengeneriereden Prozess)**.** Der Prozess in der realen Welt der die beobachteten Daten und damit die daraus folgende Statistik erzeugt wird als datengenerierender Prozess (DGP) bezeichnet.

Letztendlich zielt unsere Untersuchung, unser Experiment, darauf ab, Informationen über den DGP zu erhalten, weil diese Information uns erlaubt Aussagen über die reale Welt zu treffen. Dabei muss allerdings beachtet werden, dass dieser Prozess in den allermeisten Fällen ein starke Vereinfachung des tatsächlichen Prozesses in der Realität darstellt. Meistens sind die Abläufe in der Realität zu komplex um sie ins Gänze abzubilden. Somit wird fast immer nur ein Modell verwendet.

Zurück zu unseren Problem, wenn wir ein Experiment durchführen, dann haben wir normalerweise nur eine einzige beobachtete Statistik. In unseren bisherigen Beispiel also den berechneten Unterschied D in der Kraftfähigkeit nach der Intervention zwischen der Kontroll- und der Interventionsgruppe.

In Figure 2.1 sehen wir unseren beobachteten Wert. Dieser sei $D = 50$. Wir wissen ja aber von vorne herein schon, das dieser Wert beeinflusst ist durch die zufällige Wahl der Stichprobe und die daran geknüpfte Streuung der Werte in der Population. Wie können wir den nun überhaupt eine Aussage treffen darüber, ob das Krafttraining was bringt oder vielleicht nur einen kleinen Effekt zeigt oder möglicherweise sogar schädlich ist?

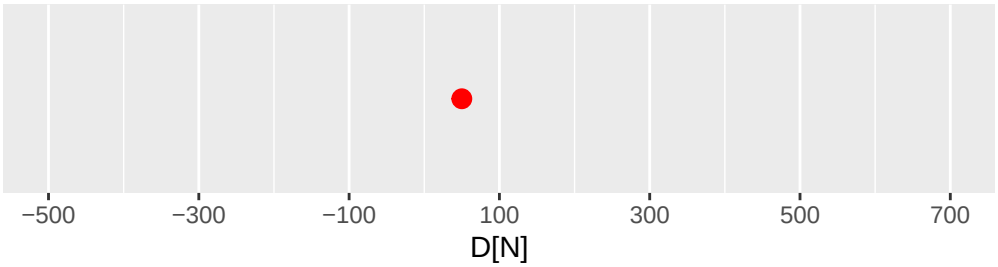


Figure 2.1: Beobachteter Unterschied nach der Durchführung unseres Experiments

Überlegen wir uns zunächst, welche Prozesse unseren beobachteten Wert zustande gebracht haben könnten. Wir haben schon zwei Prozesse kennengelernt, einmal den Prozess mit $\Delta = 100$ und auch den Prozess mit $\Delta = 0$

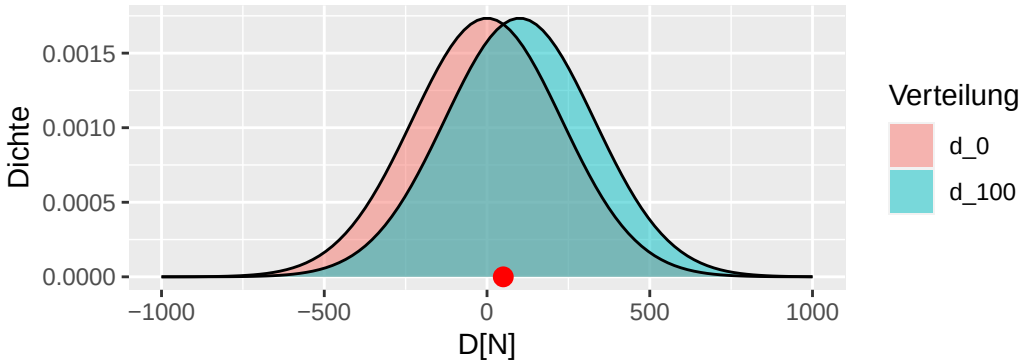


Figure 2.2: Mögliche datengenerierende Prozesse für den beobachteten Unterschied D (rot)

In Figure 2.2 ist wieder unser beobachteter Wert $D = 50$ und die beiden Verteilungen abgetragen. Leider können wir nicht sagen, welche der beiden Verteilungen, bzw. deren zugrundeliegende Prozesse, unseren beobachteten Wert erzeugt haben. Da unser beobachteter Wert D genau zwischen den beiden Maxima der Verteilungn liegt. Etwas motiviertes Starren auf die Abbildung wird uns allerdings auf die Idee bringen, dass der Wert ja nicht nur von diesen beiden Verteilungen erzeugt worden sein kann sondern durchaus noch mehr Verteilungen dafür in Frage kommen.

Figure 2.3 zeigt, dass selbst die Verteilung mit $\Delta = -250N$ und $\Delta = 350N$ nicht unplausibel sind den beobachteten Wert erzeugt zu haben. Warum aber bei diesen fünf Verteilungen aufhören, warum sollte *Delta* nicht -50 oder 127 sein. Und überhaupt, ich bin mir nicht sicher, das die Natur nur ganzzahlige Wert kennt (siehe π). Warum sollte D nicht auch $123.4567N$ sein?

Wenn diese Überlegung weitergeführt wird, dann wird schnell klar, dass letztendlich eine un-

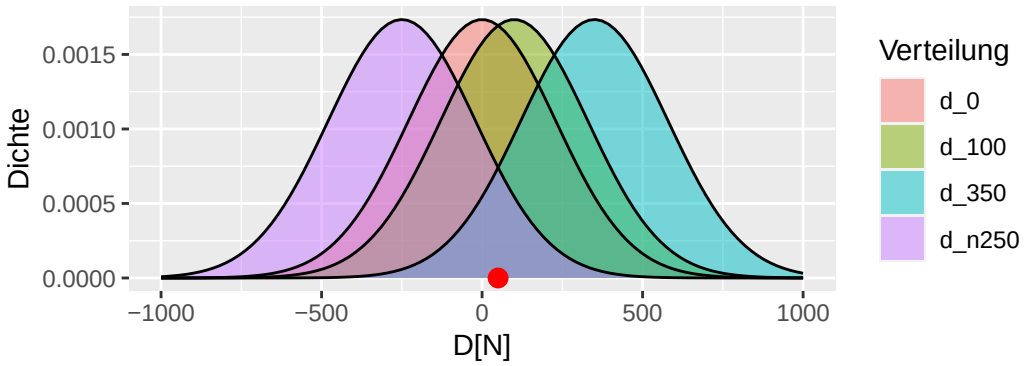


Figure 2.3: Beispiele für weitere mögliche Verteilungen als DGP.

endliche Anzahl von Verteilung in der Lage ist unseren beobachteten Wert plausibel zu generieren. D.h. wir haben ein Experiment durchgeführt und den ganzen Aufwand betrieben und haben wochenlang mit unseren ProbandInnen Krafttraining durchgeführt und sind hinterher eigentlich keinen Schritt weiter da wir immer noch nicht wissen was der datengenerierende Prozess ist. Also können wir selbst nach dem Experiment nicht sagen ob unser Krafttraining tatsächlich nützlich ist.

Zum Glück werden wir sehen und unser Unterfangen ist nicht ganz so aussichtslos. Schauen wir uns zum Beispiel die Verteilung für $\Delta = -350N$ an (Figure 2.4).

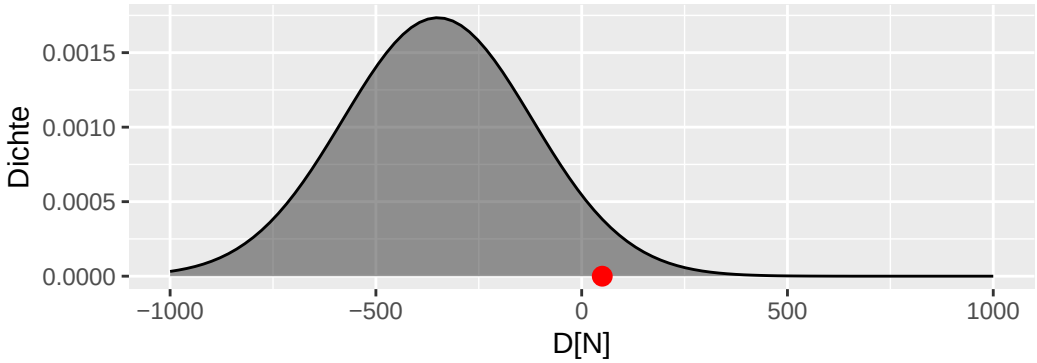


Figure 2.4: Verteilung für $\Delta = -350N$ und der beobachtete Wert D

Unser beobachteter Wert ist jetzt nicht vollkommen unmöglich unter der Annahme das $\Delta = -350N$ ist, aber so richtig *wahrscheinlich* ist er auch nicht. Der Wert liegt relativ weit am Rand der Verteilung. Die Kurve ist dort schon ziemlich nahe bei Null. D.h. der beobachtete Wert ist zwar schon möglich aber es wäre schon überraschend wenn wir bei einer Durchführung des Experiments ausgerechnet so einen Wert beobachten würden.

Wenn wir jetzt dagegen von der Annahme ausgehen, dass dem DGP der Wert $\Delta = 50N$

zugrundeliegen würde, hätten wir die Verteilung in Figure 2.5. Hier ist der beobachtete Wert mitten drin in dem Teil der Verteilung der auch zu erwarten würde. D.h. unser beobachteter Wert ist durchaus plausibel unter der Annahme das $\Delta = 50N$ gilt.

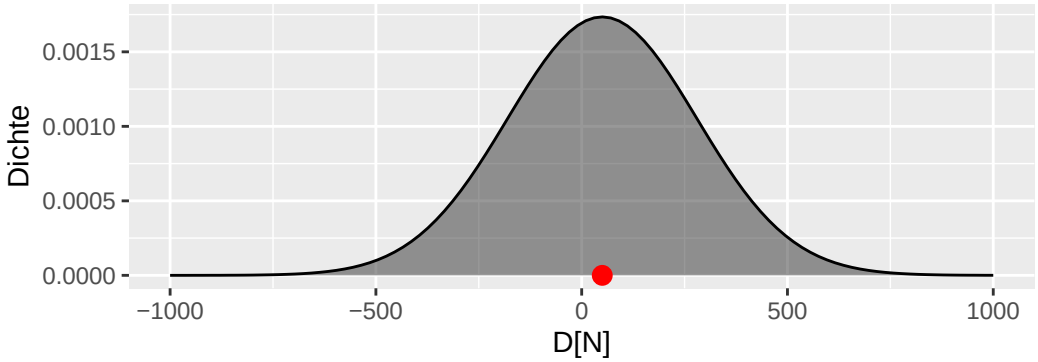


Figure 2.5: Verteilung für $\Delta = 50N$ und der beobachtete Wert D

Diesen Ansatz können wir verwenden um mit Hilfe unseres Experiments doch etwas über den DGP auszusagen. Allerdings müssen wir uns noch einmal etwas eingehender mit Verteilungen auseinandersetzen. D.h. wir müssen uns erst ein mal ein paar neue Konzepte erarbeiten.

2.2 Verteilungen - 1. deep dive

2.3 Eigenschaften von Verteilungen - Mittelwert μ

[1]

2.4 Eigenschaften von Verteilungen - Varianz σ^2

[2]

2.5 Formeln

n := Anzahl der Stichprobenelemente, x_i := Messwerte

[1] auch Lageparameter oder Erwartungswert
[2] auch Skalenparameter

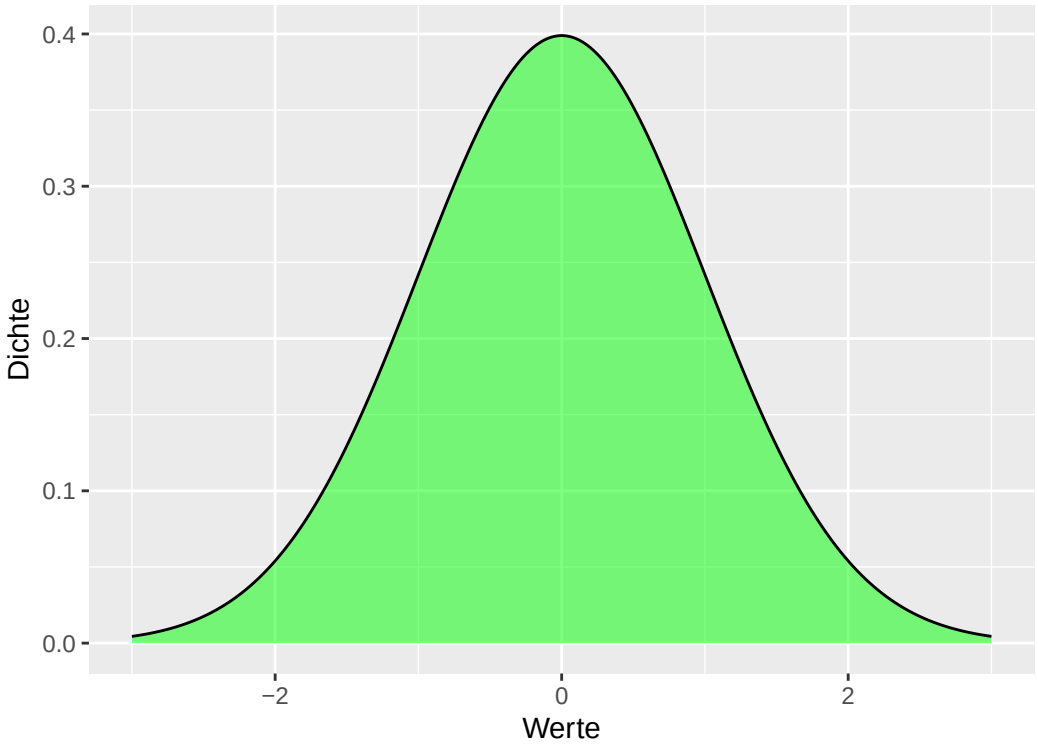


Figure 2.6: Eine Dichtefunktion

Table 2.1: Parameter einer Verteilung und deren Schätzer

Population	Stichprobe
Mittelwert μ	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Varianz σ^2	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
Standardabweichung σ	$s = \sqrt{s^2}$

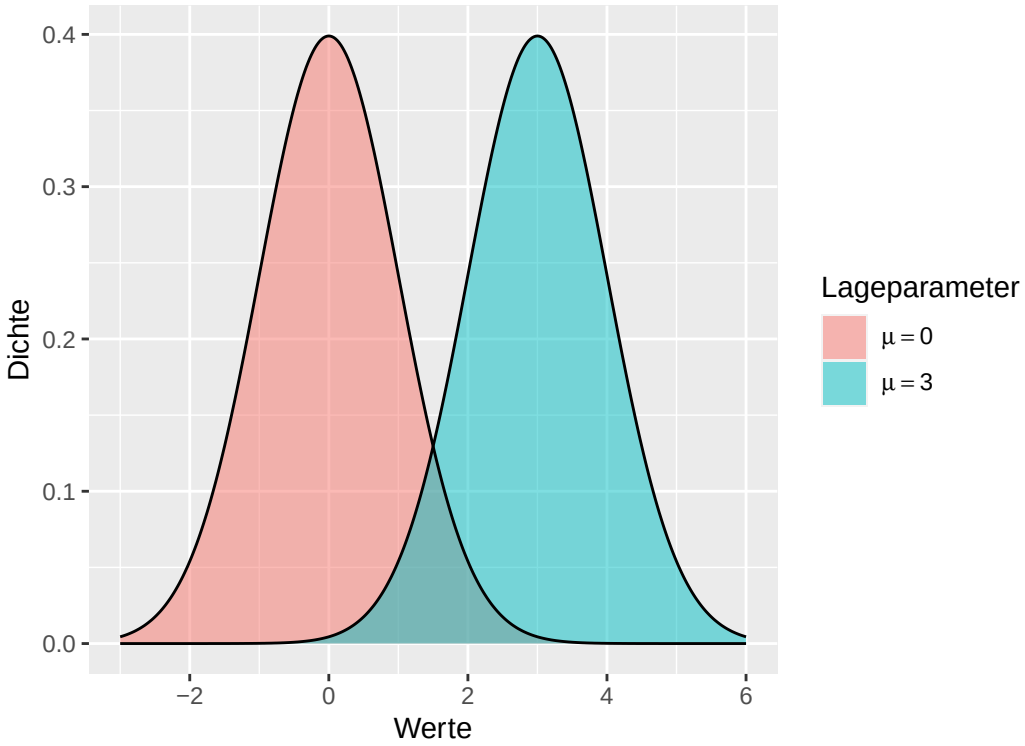


Figure 2.7: Verteilungen mit unterschiedlichen Mittelwerten

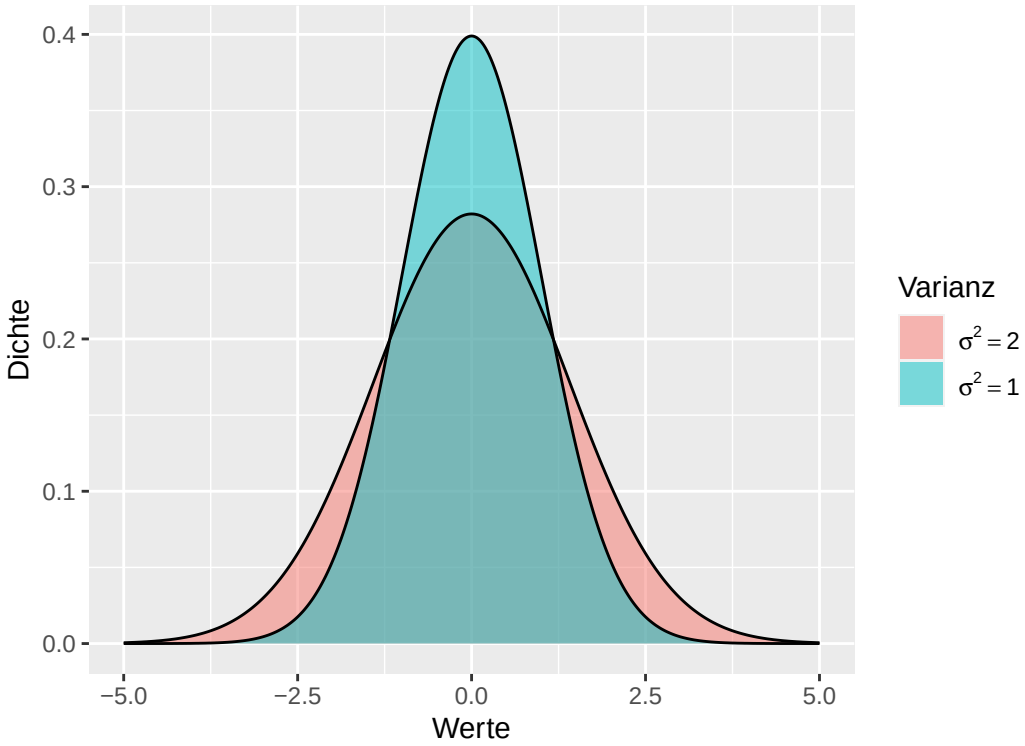


Figure 2.8: Verteilungen mit unterschiedlichen Varianzen

2.6 Nebenbei: Warum der Mittelwert Sinn macht

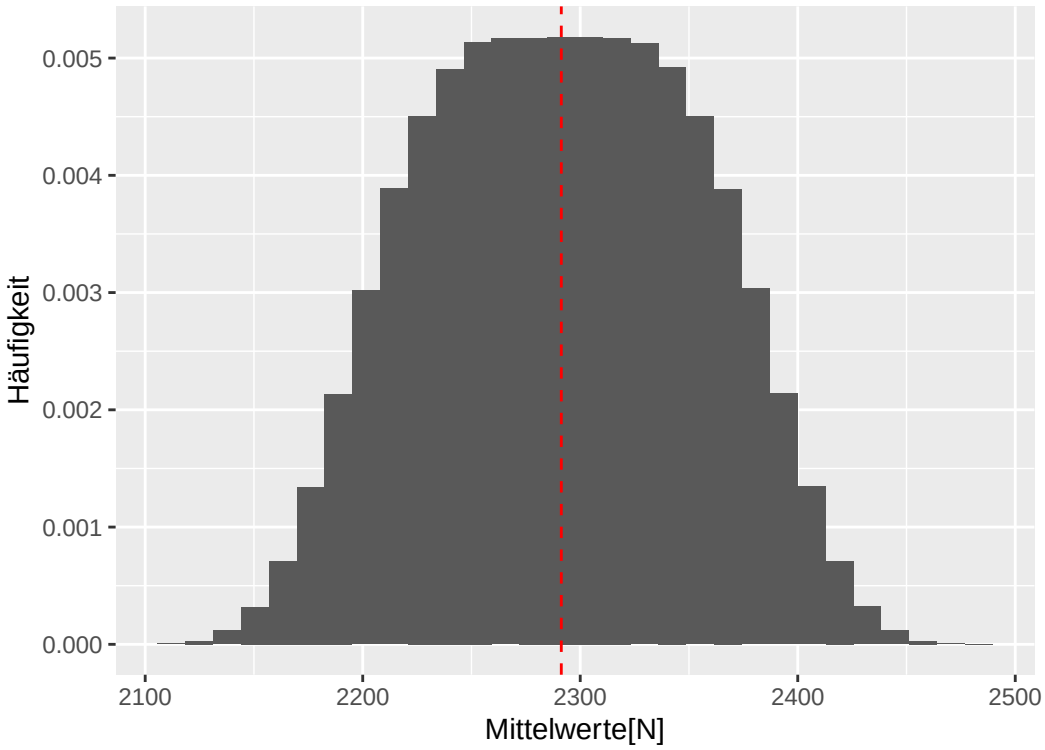


Figure 2.9: Verteilung der Mittelwerte von Stichproben der Größe $n = 10$, Kleine Welt Population μ (rot)

2.7 Mit der Verteilung die annimmt das nichts passiert!

2.8 *Signifikanter* Wert

Wenn der Stichprobenwert der Statistik in der *kritischen* Region auftritt, dann wird von einem **statistisch** signifikanten Effekt gesprochen. *Unter der H_0 bin ich überrascht diesen Wert zu sehen!*

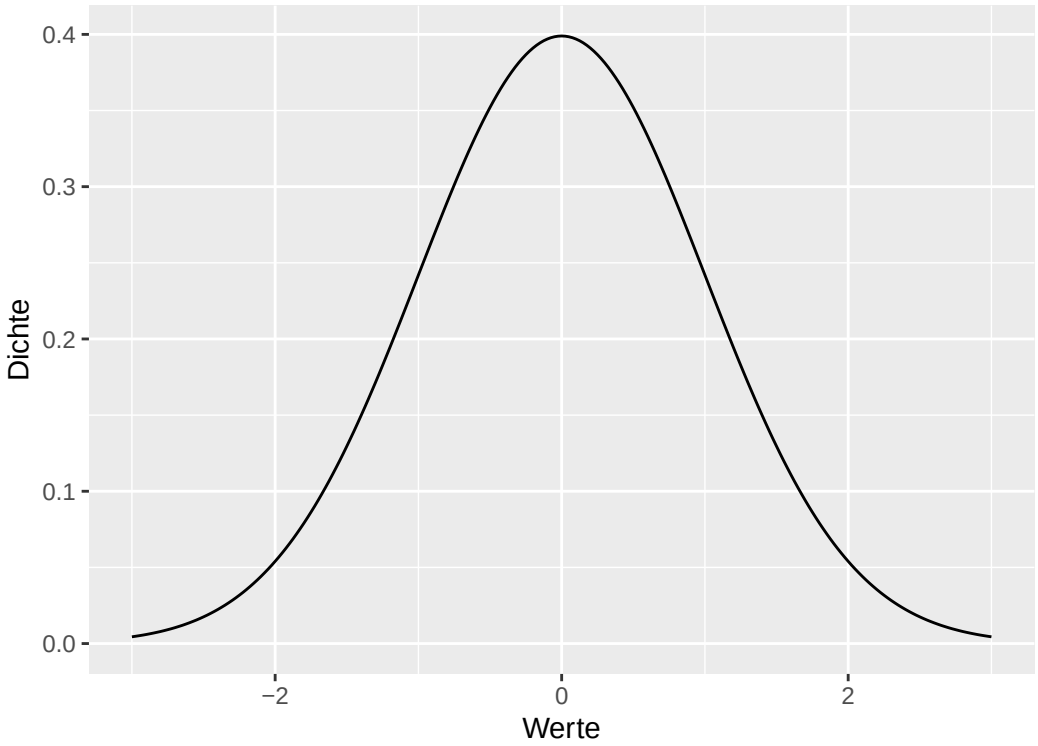


Figure 2.10: Verteilung wenn nichts passiert.

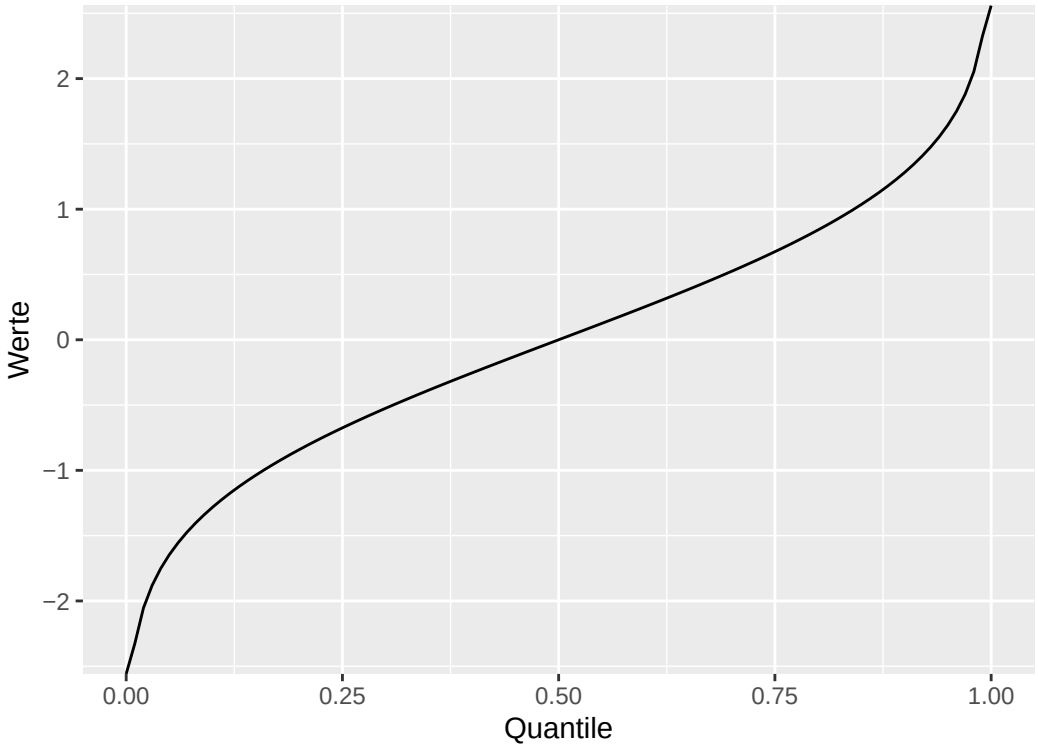


Figure 2.11: Quantilefunktion wenn nichts passiert.

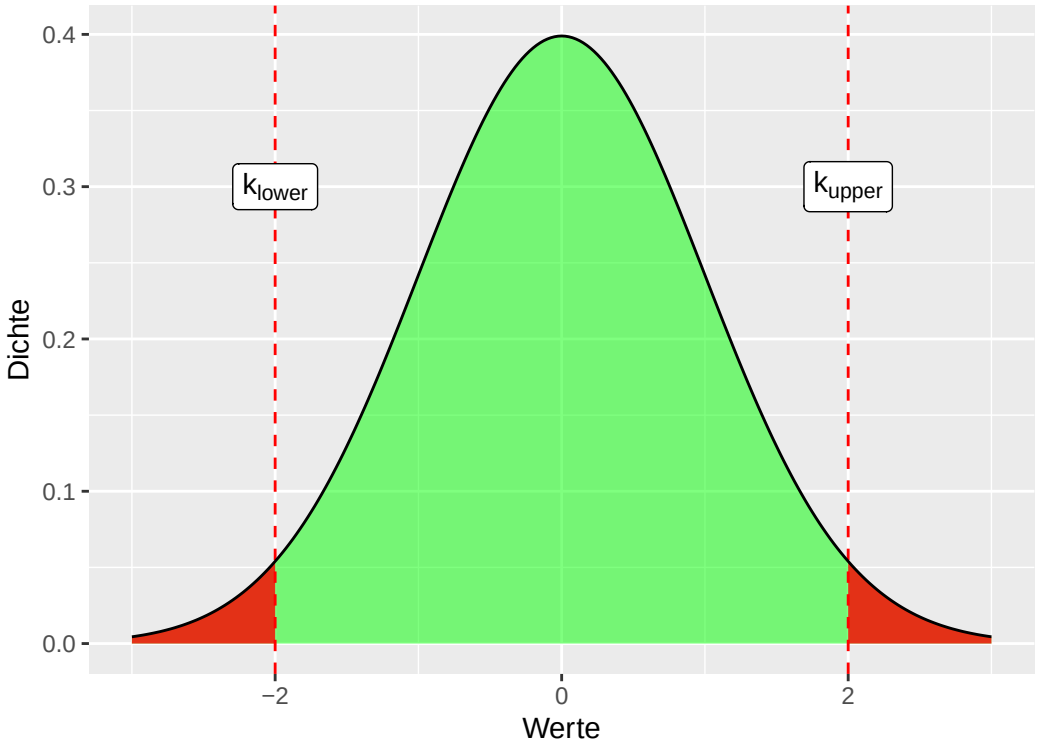


Figure 2.12: Verteilung wenn nichts passiert und kritische Regionen.

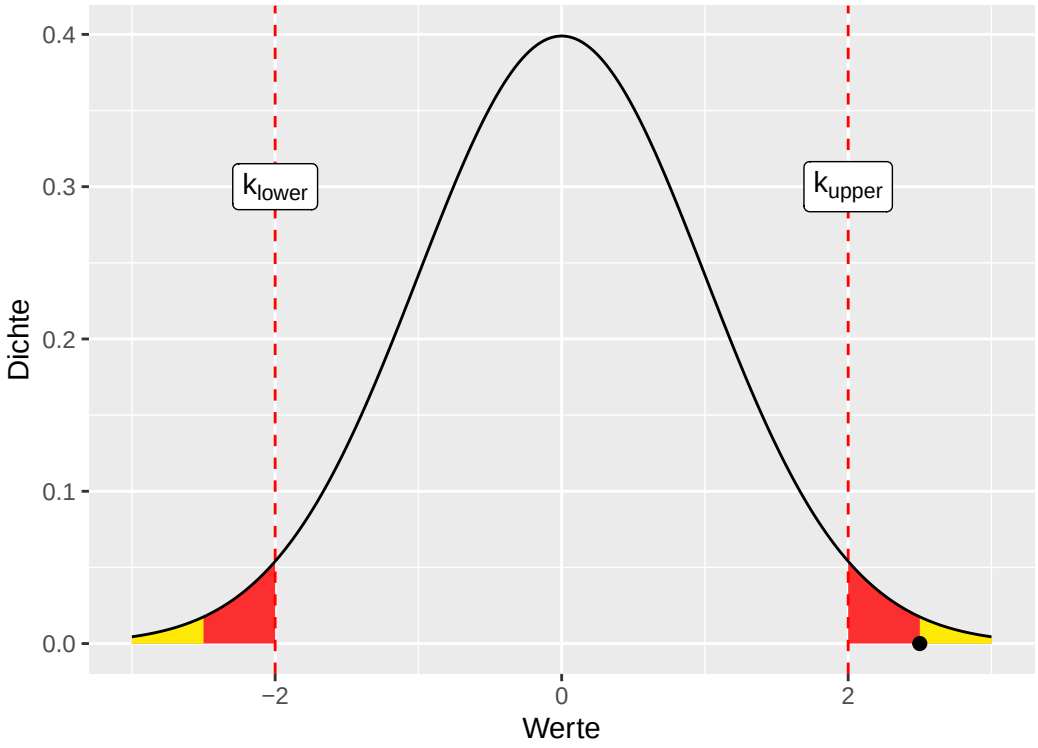


Figure 2.13: Der gelben Flächen zeigen den p-Wert für den Wert der Statistik von d = 2,5 an.

2.9 Der p-Wert

2.10 p-Werte

Der p-Wert gibt die Wahrscheinlichkeit für den gefundenen oder einen noch extremeren Wert unter der H_0 an.

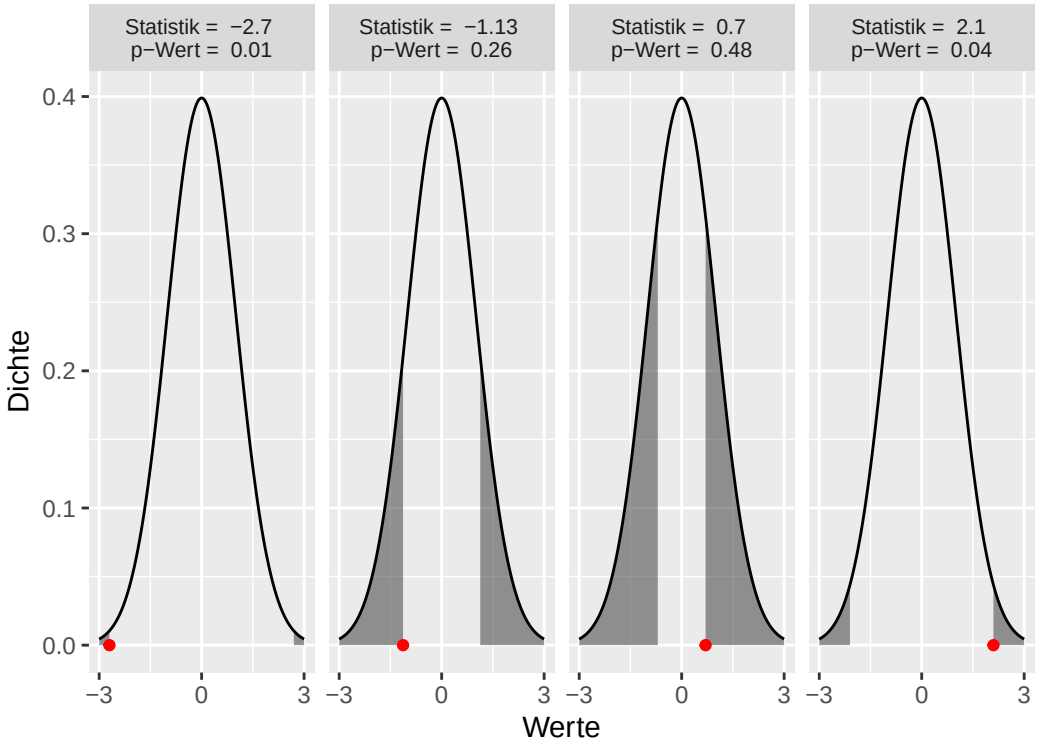


Figure 2.14: Verschiedene P-Werte

2.11 p-Werte

"[A] p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." (Wasserstein and Lazar 2016, 131)

"[T]he P value is the probability of seeing data that are as weird or more weird than those that were actually observed." (Christensen 2018, 38)

2.12 Signifikanter Wert - Das Kleingedruckte

- **Vor** dem Experiment wird für ein H_0 ein α-Level angesetzt (per Konvention $\alpha = 0,05 = 5\%$)
- Anhand des α-Levels können **kritische Werte** (k_{lower}, k_{upper}) bestimmt werden. Diese bestimmen die Grenzen der **kritischen Regionen**.
- Wenn der gemessene Wert w der Statistik in die kritische Region fällt, also $w \leq k_{lower}$ oder $w \geq k_{upper}$ gilt, dann wird von einem **statistisch** signifikanten Wert gesprochen und die dazugehörige Hypothese wird **abgelehnt**. Äquivalent: Der p-Wert ist kleiner als α.
- Da in α-Fällen ein Wert in der kritischen Region auftritt, auch wenn die H_0 zutrifft, wird in α-Fällen ein α-Fehler gemacht.

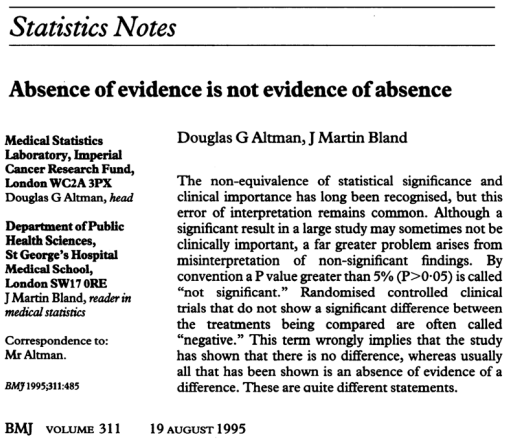
2.13 Signifikanter Wert - Das Kleingedruckte

- Wenn der Wert w der Statistik nicht in den kritischen Regionen liegt, oder gleichwertig der p-Wert größer als α ist, wird die H_0 **beibehalten**. D.h. nicht, dass **kein Effekt** vorliegt, sondern lediglich, dass anhand der Daten keine Evidenz diesbezüglich gefunden werden konnte!
- Die **statistische** Signifikanz sagt nichts über die Wahrscheinlichkeit der Theorie aus!
- Ein p-Wert von $p = 0.0001$ heißt nicht, dass mit 99,99% Wahrscheinlichkeit ein Effekt vorliegt!
- *Statistisch* signifikant heißt nicht automatisch *praktisch* relevant!

2.14 Nochmal, wenn die H_0 nicht abgelehnt wird

2.15 Nochmal p-Wert (Wasserstein and Lazar (2016))

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.



Statistics Notes

Absence of evidence is not evidence of absence

Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX
Douglas G Altman, *head*

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE
J Martin Bland, *reader in medical statistics*

Correspondence to:
Mr Altman.

BMJ 1995;311:485

Douglas G Altman, J Martin Bland

The non-equivalence of statistical significance and clinical importance has long been recognised, but this error of interpretation remains common. Although a significant result in a large study may sometimes not be clinically important, a far greater problem arises from misinterpretation of non-significant findings. By convention a P value greater than 5% (P>0·05) is called "not significant." Randomised controlled clinical trials that do not show a significant difference between the treatments being compared are often called "negative." This term wrongly implies that the study has shown that there is no difference, whereas usually all that has been shown is an absence of evidence of a difference. These are quite different statements.

Figure 2.15: Ausschnitt aus D. G. Altman and Bland (1995)

2.16 Was passiert nun aber wenn die "andere" Hypothese zutrifft?

2.17 Wir machen einen β-Fehler!

2.18 Snap!(1989) - The Power

2.19 Terminologie noch mal

- α: Die Wahrscheinlichkeit sich gegen die H_0 zu entscheiden, wenn die H_0 zutrifft. α-Level wird vor dem Experiment festgelegt um zu kontrollieren welche Fehlerrate toleriert wird.
- β: Die Wahrscheinlichkeit sich gegen die H_1 zu entscheiden, wenn die H_1 zutrifft.
- Power $:= 1 - \beta$: Die Wahrscheinlichkeit sich für die H_1 zu entscheiden, wenn die H_1 zutrifft. Sollte ebenfalls **vor** dem Experiment festgelegt werden.

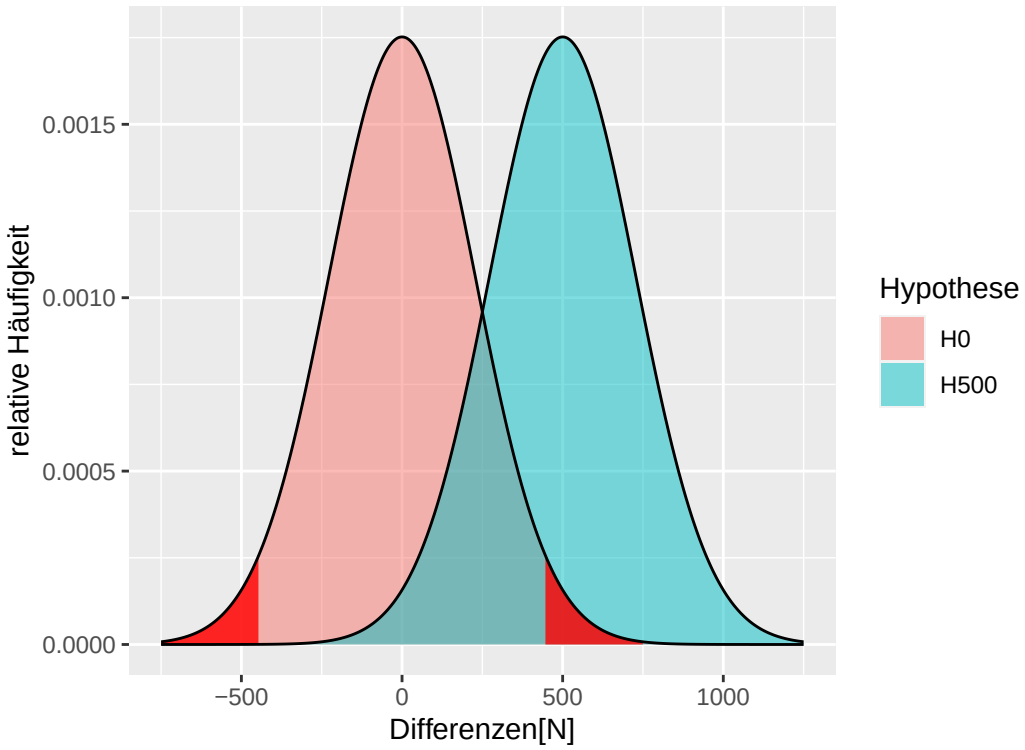


Figure 2.16: Differenzen mit kritischen Regionen (rot) mit einer Wahrscheinlichkeit von α wenn H_0 zutrifft.

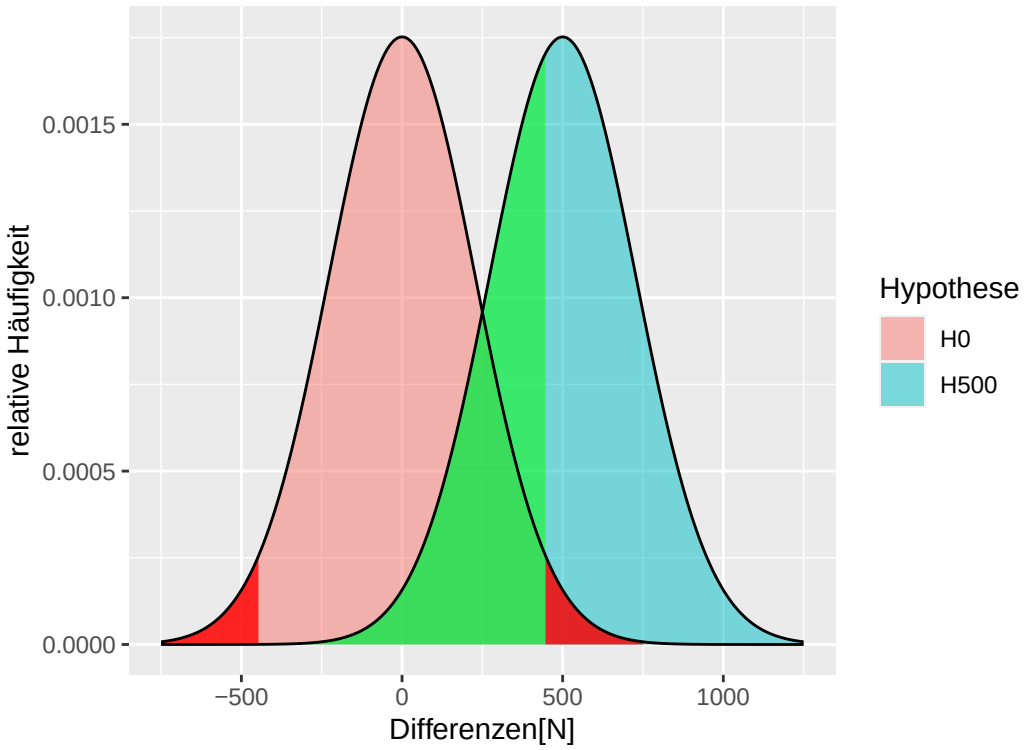


Figure 2.17: Differenzen mit kritischen Regionen (rot) mit einer Wahrscheinlichkeit von α wenn H_0 zutrifft und β (grün) wenn H_1 zutrifft.

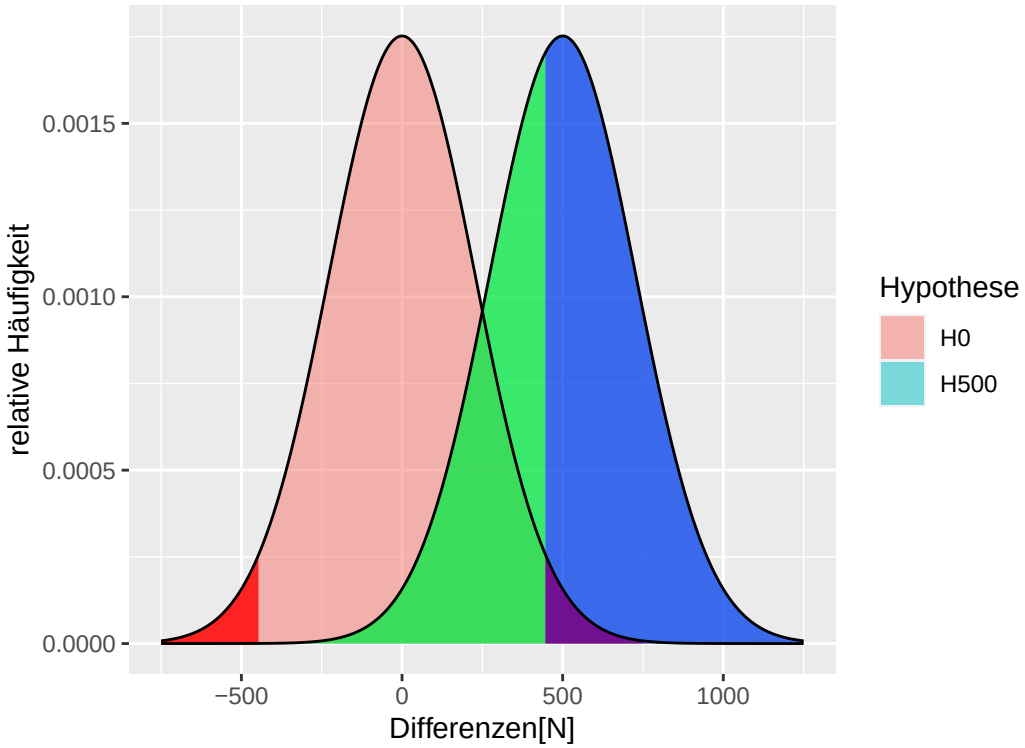


Figure 2.18: $1 - \beta$ = Power des Tests (blaue Fläche).

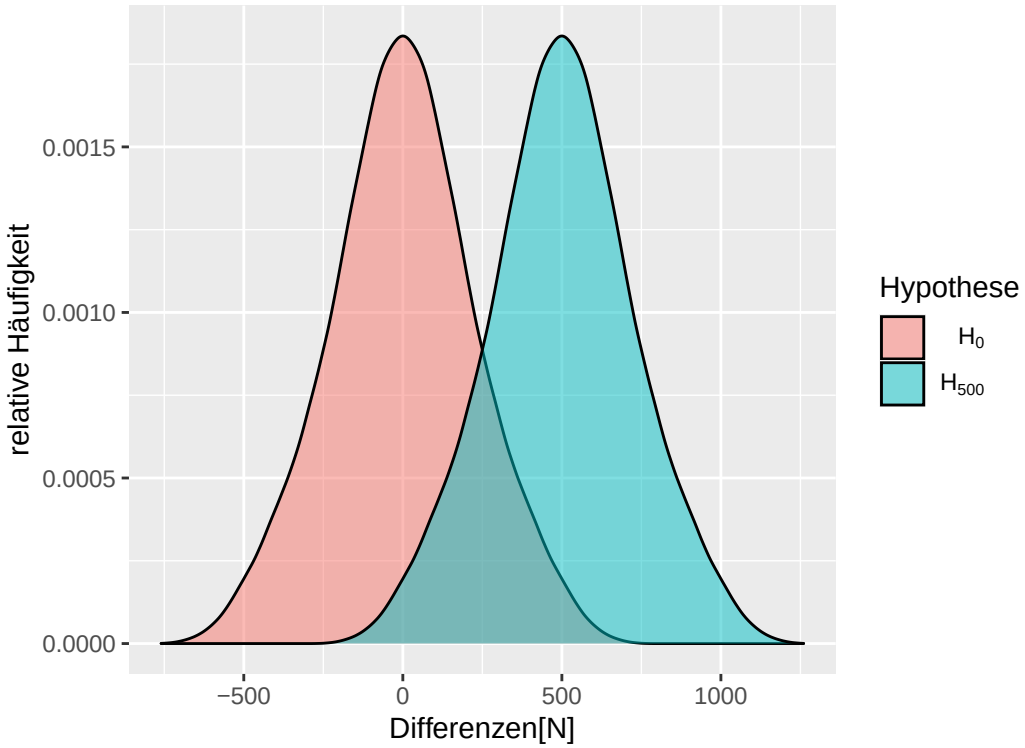


Figure 2.19: Verteilungen wenn δ=500 und δ=0 in unserem kleine Welt Beispiel mit n = 3.

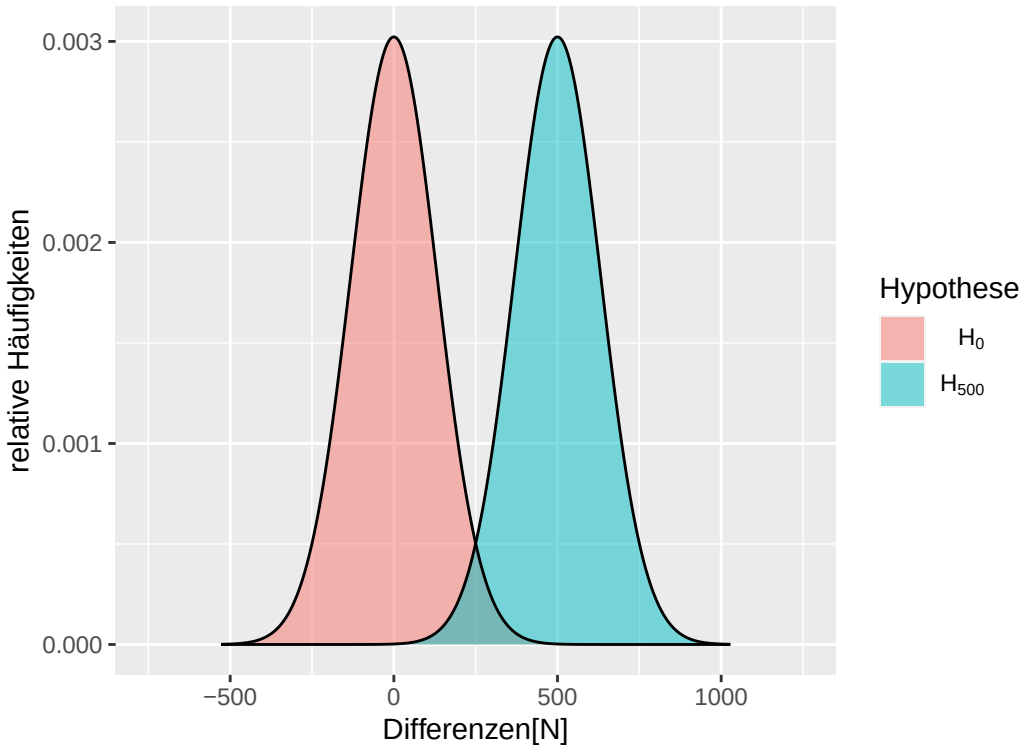


Figure 2.20: Stichprobenverteilungen der Differenz unter H_0 und $H_1 : \delta = 500$N bei einer Stichprobengröße von n = 9

Table 2.2: Standardfehler des Mittelwerts, n = Stichprobengröße

Population	Stichprobe
$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$	$s_e = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$

2.20 Wie können wir die Power erhöhen?

2.21 Stichprobengröße von n = 3 auf n = 9 erhöhen?

2.22 Standardfehler

Die Standardabweichung der Statistik wird als **Standardfehler** s_e bezeichnet[3]. Der Standardfehler ist nicht gleich der Standardabweichung in der Population bzw. der Stichprobe. Es gilt für den Mittelwert:

[3]Der Standardfehler schätzt die Reliabilität der Statistik ab (Cohen (1988))

3 Parameterschätzung

3.1 Problem bei einer dichotomen Betrachtung der Daten

Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Lucky (2008) used two independent groups each of size $N = 22$, and Noluck (2008) used two groups each with $N = 18$. Each study reported the difference between the means for the new treatment and the current treatment.

Lucky (2008) found that the new treatment showed a statistically significant advantage over the current treatment: *M(difference)* = 3.61, *SD(difference)* = 6.97, $t(42) = 2.43$, $p = .02$. The study by Noluck (2008) found no statistically significant difference between the two treatment means: *M(difference)* = 2.23, *SD(difference)* = 7.59, $t(34) = 1.25$, $p = .22$.

Figure 3.1: Auszug aus Cumming (2013, 1)

3.2 Wie groß ist der Effekt?

3.3 Schätzung der Populationsparameter

Kleine Welt: Experiment wird einmal mit n = 9 durchgeführt

3.3.1 Beobachtete Stichprobenkennwerte

$$d = \bar{x}_{treat} - \bar{x}_{con} = 350$$
$$s = 132$$
$$s_e = 44$$

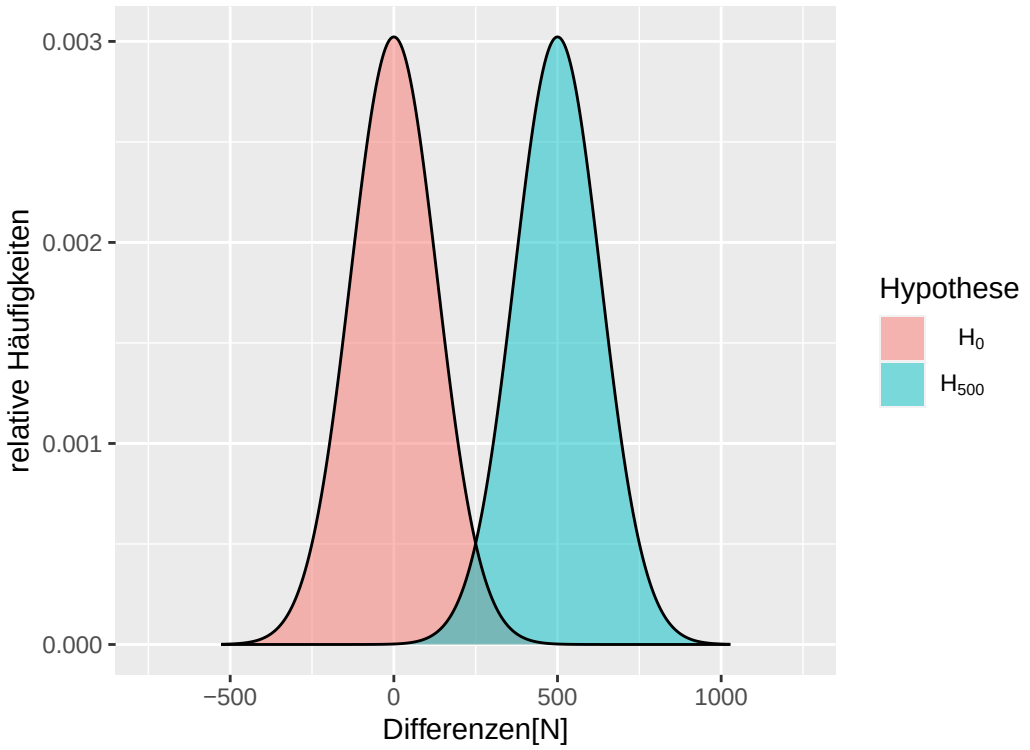


Figure 3.2: Stichprobenverteilungen der Differenz unter H_0 und $H_1 : \delta = 500$N bei einer Stichprobengröße von n = 9

Wie präzise ist meine Schätzung und welche anderen Unterschiedswerte sind anhand der beobachteten Daten noch plausibel?

3.4 Welche δs sind plausibel für $d = 350$?

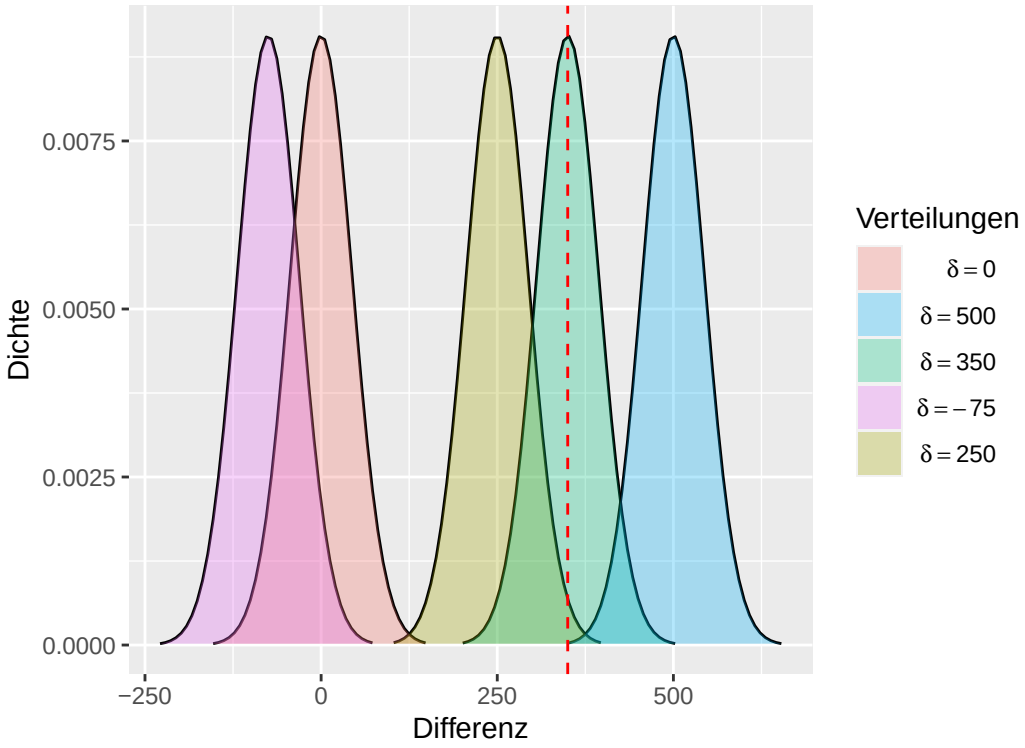


Figure 3.3: Verschiedene Verteilungen von Gruppendifferenzen, beobachteter Unterschied (rot)

Plausibel unter einem gegebenem α**-Level**!

3.5 Alle möglichen δs die plausibel sind

3.6 Was passiert wenn ich das Experiment ganz oft wiederhole?

3.7 Konfidenzintervall - Das Kleingedruckte

- Das Konfidenzintervall für ein gegebenes α-Niveau gibt nicht die Wahrscheinlichkeit an mit der der *wahre* Parameter in dem Intervall liegt.

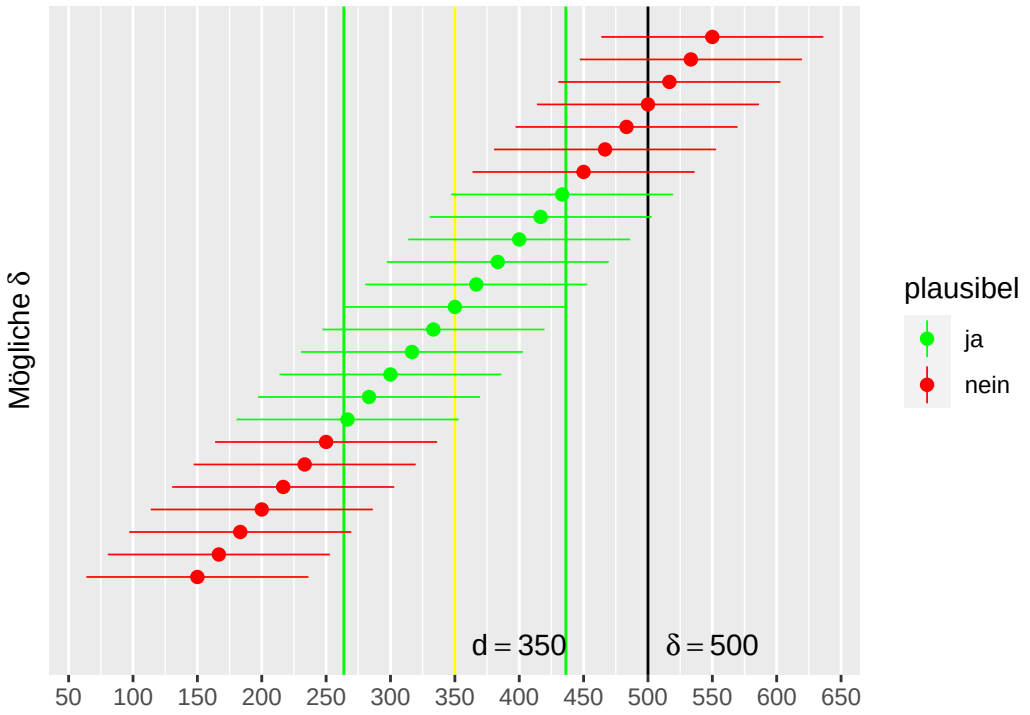


Figure 3.4: Konfidenzintervall (grün), Populationsparameter δ und α-Level für die beobachtete Differenz (gelb).

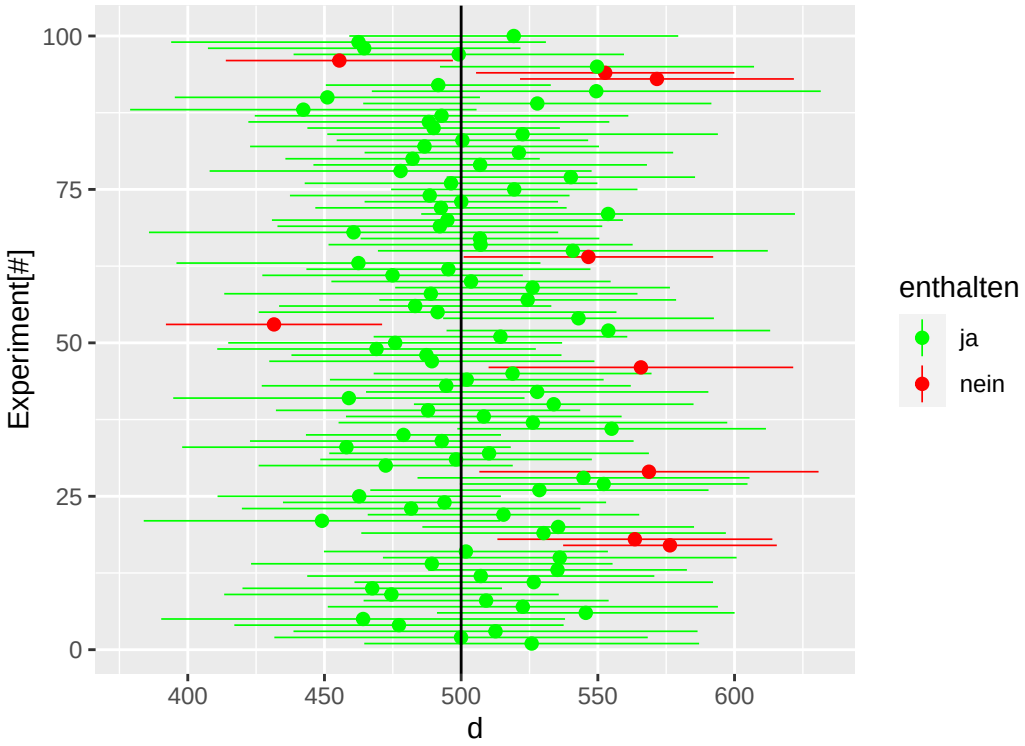


Figure 3.5: Simulation von $n = 100$ Konfidenzintervallen.

- Das Konfidenzintervall gibt alle mit den Daten kompatiblen Populationsparameter an.
- Das α-Niveau des Konfidenzintervalls gibt an bei welchem Anteil von Wiederholungen davon auszugehen ist, das das Konfidenzintervall den wahren Populationsparameter enthält.

3.8 Konfidenzintervall herleiten nach Spiegelhalter (2019, 241)

1. We use probability theory to tell us, for any particular population parameter, an interval in which we expect the observed statistic to lie with 95% probability.
2. Then we observe a particular statistic.
3. Finally (and this is the difficult bit) we work out the range of possible population parameters for which our statistic lies in their 95% intervals. This we call a "95% confidence interval".
4. This resulting confidence interval is given the label "95%" since, with repeated application, 95% of such intervals should contain the true value.[1]

All clear? If it isn't, then please be reassured that you have joined generations of baffled students.

3.9 Konfidenzintervall berechnen (Vorschau)

$$\text{CI}_{1-\alpha} = \bar{x} \pm z_{\alpha/2} \times s_e$$

3.10 Dualität von Signifikanztests und Konfidenzintervall

Wenn das Konfidenzintervall mit Niveau $1 - \alpha\%$ die H_0 nicht beinhaltet, dann wird auch bei einem Signifikanztest die H_0 bei einer Irrtumswahrscheinlichkeit von α abgelehnt.

[1] Strictly speaking, a 95% confidence interval does ***not*** mean there is a 95% probability that this particular interval contains the true value [...]

4 Verteilungen

5 Die Normalverteilung

5.1 Normalverteilung - $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$

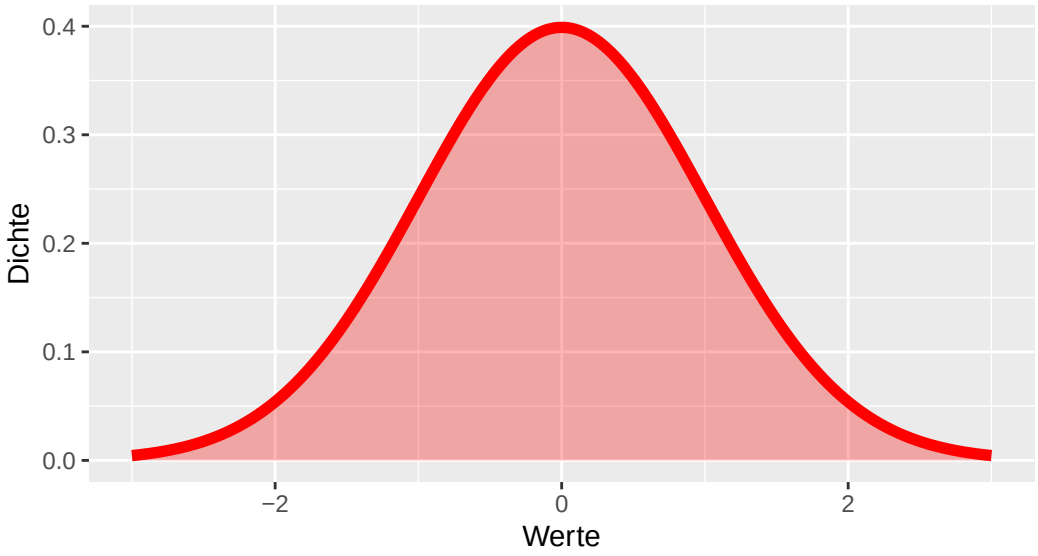


Figure 5.1: Dichtefunktion der Normalverteilung mit Parametern μ und σ.

5.2 Zentraler Grenzwertsatz oder *Warum die Normalverteilung überall auftaucht.*

Seien $X_1, X_2, \ldots, X_n$ n unabhängige, gleichverteilte Zufallsvariablen mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$.

$$\lim_{n\to\infty} \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

5.3 Normalverteilung und Standardabweichung

5.4 Normalverteilung und Standardabweichung

$$P(x \in [\mu - 1.96\sigma, \mu + 1.96\sigma]) = 0.95$$

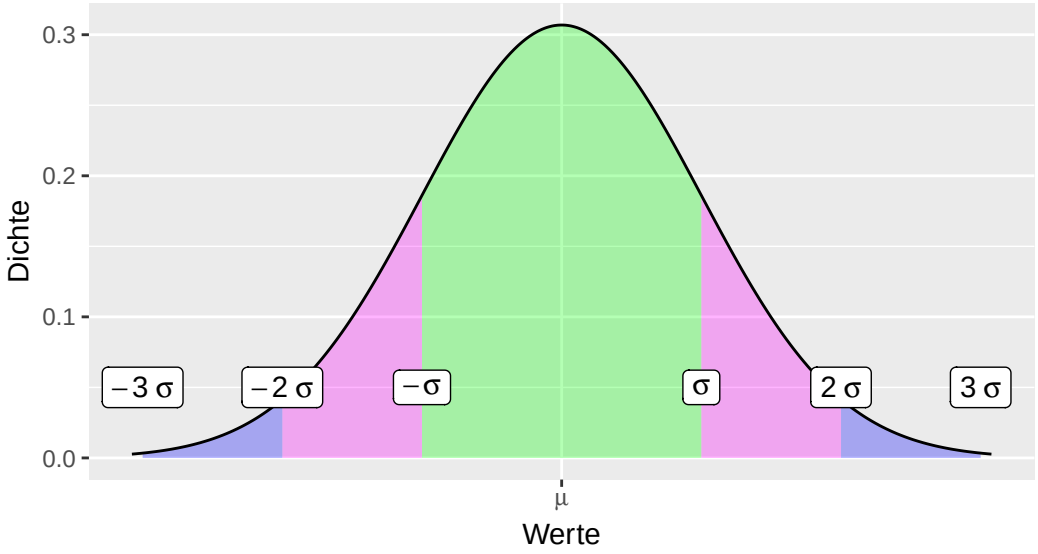


Figure 5.2: Dichtefunktion von $\mathcal{N}(\mu, \sigma^2)$

Table 5.1: Wahrscheinlichkeiten P für verschiedene Bereiche der Normalverteilung.

Bereich	P
$[\mu - \sigma, \mu + \sigma]$	0.682
$[\mu - 2\sigma, \mu + 2\sigma]$	0.955
$[\mu - 3\sigma, \mu + 3\sigma]$	0.997

Table 5.2: z-Transformation

Population	Stichprobe
$z = \frac{x-\mu}{\sigma}$	$z = \frac{x-\bar{x}}{s}$

5.5 Standardnormalverteilung $\phi(x)$

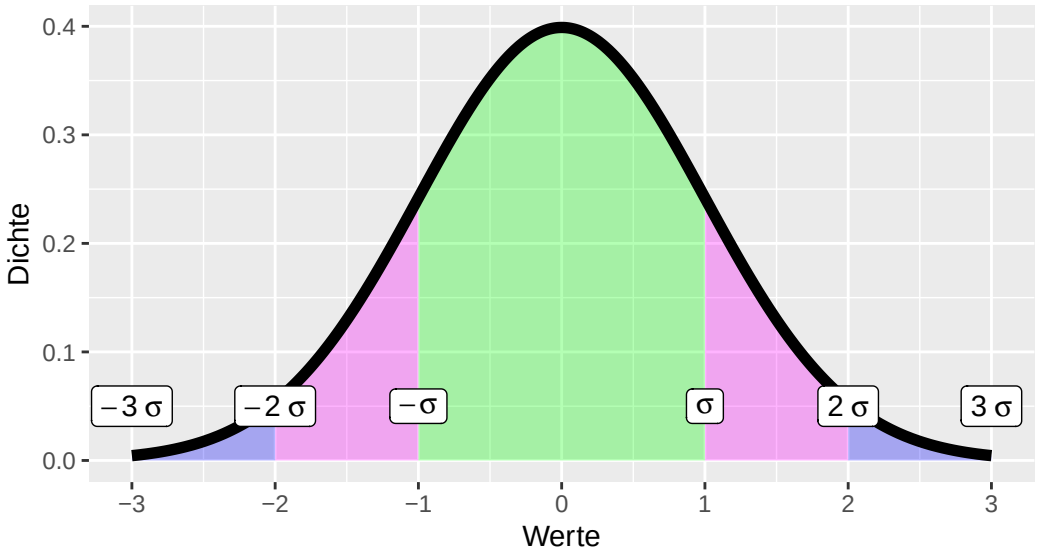


Figure 5.3: Dichtefunktion der Standardnormalverteilung $\phi(x)$ mit $\mu = 0$ und $\sigma^2 = 1$

5.6 Abbildung N(μ,σ) auf N(0,1)

5.7 z-Transformation allgemein bzw. Standardisierung

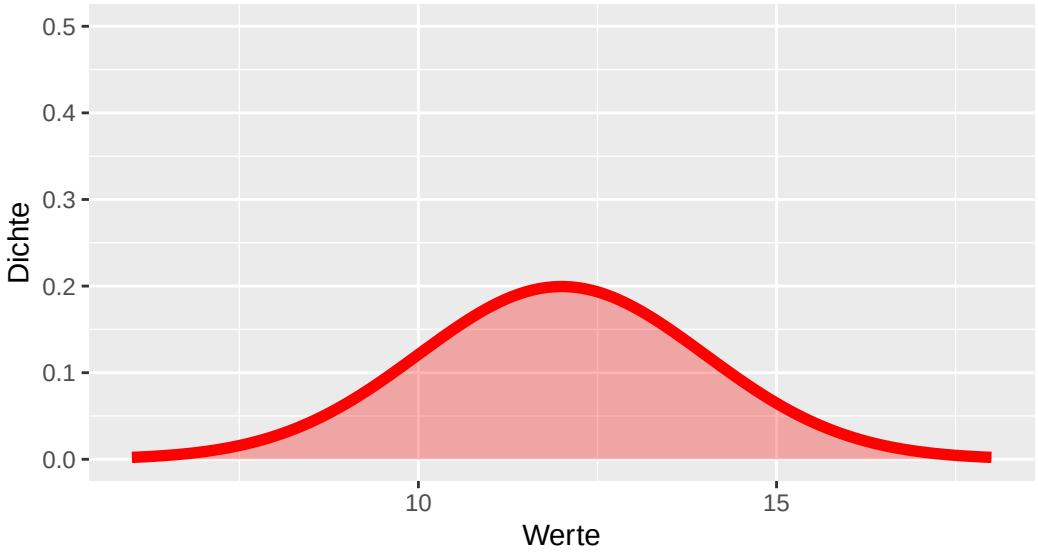


Figure 5.4: Standardnormalverteilung mit $\mu = 12, \sigma^2 = 2$

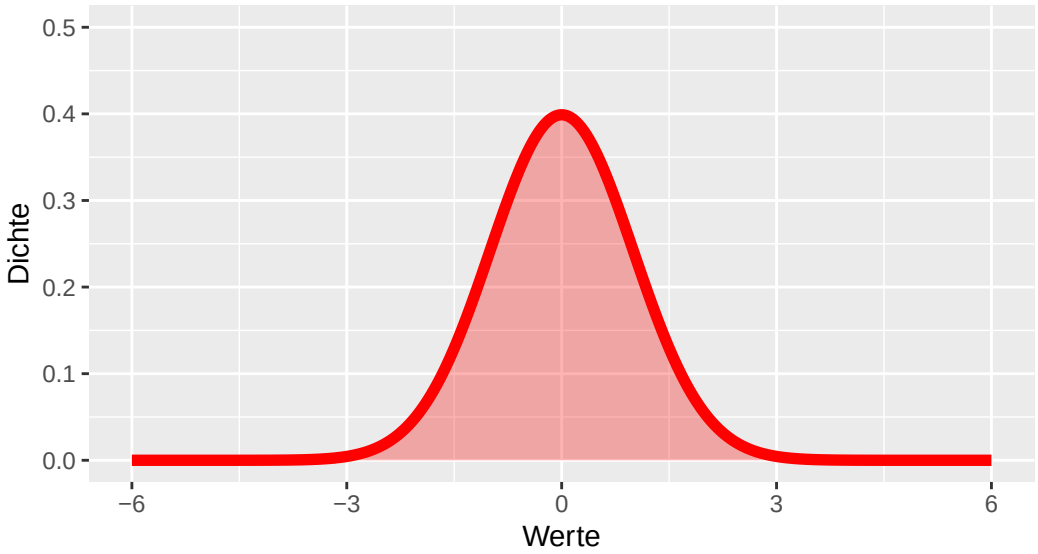


Figure 5.5: Normalverteilung mit $\mu = 0, \sigma = 1$

6 Verteilungszoo

6.1 t-Verteilung

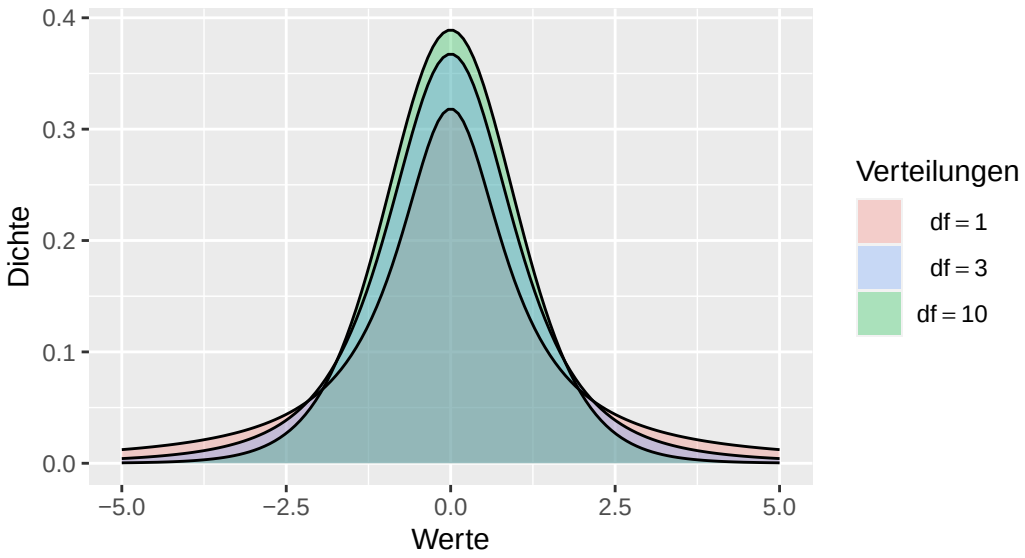


Figure 6.1: Beispiel für verschiedene Dichtefunktionen der t-Verteilung

6.2 χ^2-Verteilung

6.3 F-Verteilung

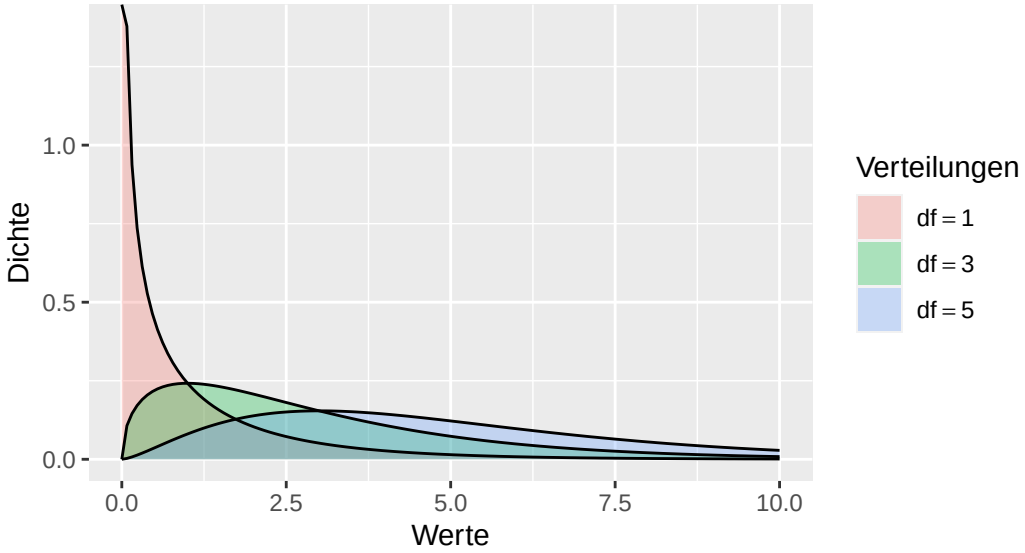


Figure 6.2: Beispiele für verschiedene Dichtefunktion der χ^2-Verteilung.

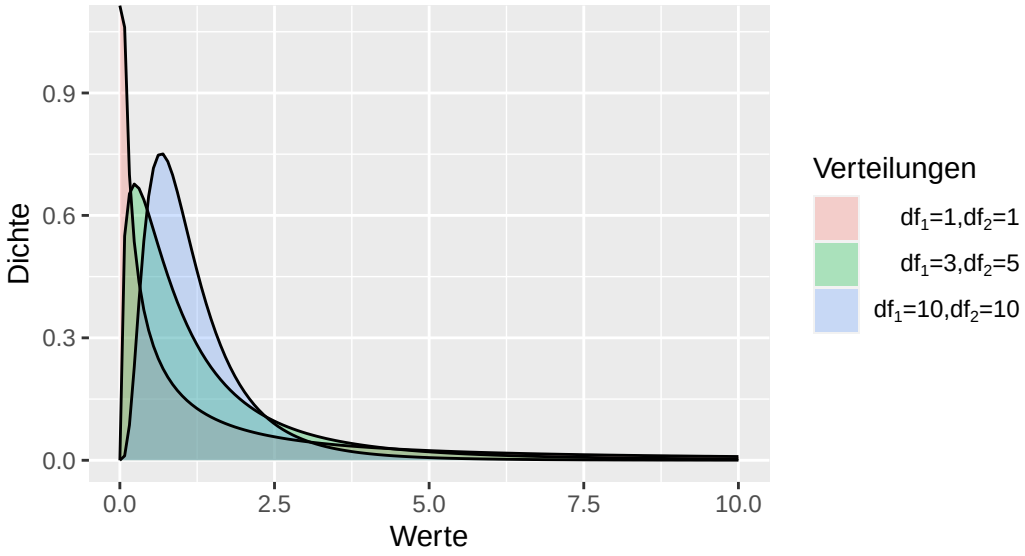


Figure 6.3: Beispiele für verschiedene Dichtefunktion der F-Verteilung.

7 Hypothesen testen

7.1 Wahrscheinlichkeitstheorie

7.2 Schätzer

7.3 Hypothesentestung

Part II

Das einfache Regressionmodell

Wir beginnen nun mit dem einfachen Regressionsmodell. Das Modell knüpft an unsere Vorkenntnisse aus der Schule und bietet die Möglichkeit eine einfaches mentales Template zu erarbeiten auf das wir immer wieder zurück greifen können, da sich bis auf ein paar wenige Konzepte alle wichtige Eigenschaften von linearen Modellen anhand des einfachen Regressionsmodells erklären können. Wenn dann im zweiten Schritt der Übergang auf die multiple Regression durchgeführt wird, sollte dies keine größeren Probleme mehr bereiten, da immer nur ein paar wenige neue Konzepte dazu kommen.

8 Einführung

8.1 Back to school

Wir beginnen mit ein Konzept das wir schon alle kennen. Nämlich die Punkt-Steigungsform aus der Schule (siehe Equation 8.1).

$$y = mx + b \tag{8.1}$$

Wir haben eine abhängige Variable y und eine lineare Formel $mx + b$ die den funktionalen Zusammenhang zwischen den Variablen y und x beschreibt. Um das Ganz einmal konkret zu machen setzen wir $m = 2$ und $b = 3$ fest. Die Formel Equation 8.1 wird dann zu:

$$y = 2x + 3 \tag{8.2}$$

Um ein paar Werte für y zu erhalten setzen wir jetzt verschiedene Wert für x ein indem wir x in Einserschritten zwischen $[0, \dots, 5]$ erhöhen. Um die Werte darzustellen verwenden wir zunächst eine Tabelle (vlg. Table 8.1)

Table 8.1: Tabelle der Daten

x	y
0	3
1	5
2	7
3	9
4	11
5	13

Wenig überraschend nimmt y für den Wert $x = 0$ den Wert 3 an und z.B. für den Wert $x = 3$ nimmt y den Wert $2 \cdot 3 + 3 = 9$ an.

TODO: Einführung eines Index i

Eine andere Darstellungsform ist naturlich eine graphische Darstellung in dem wir die Werte von y gegen x auf einem Graphen abtragen (siehe Figure 8.1).

Wiederum wenig überraschen sehen wir einen linearen Zuwachs der y-Wert mit den größerwerdenden x-Werte. Da in der Definition der Formel Equation 8.2 nirgends festgelt wurde, dass diese nur für ganzzahlige x-Werte gilt, haben wir direkt eine Gerade durch die Punkte gelegt. Hier wird auch die Bedeutung von m und b direkt klar. Die Variable m bestimmt die Steigung der Gleichung während b den y-Achsenabschnitt beschreibt.

Definition 8.1 (y-Achsenabschnitt)**.** Der y-Achsenabschnitt ist der Wert den y einnimmt wenn x den Wert 0 annimmt. Sei y durch eine lineare Gleichung $y = mx + b$ definiert, dann wird der y-Achsenabschnitt durch den Wert b bestimmt.

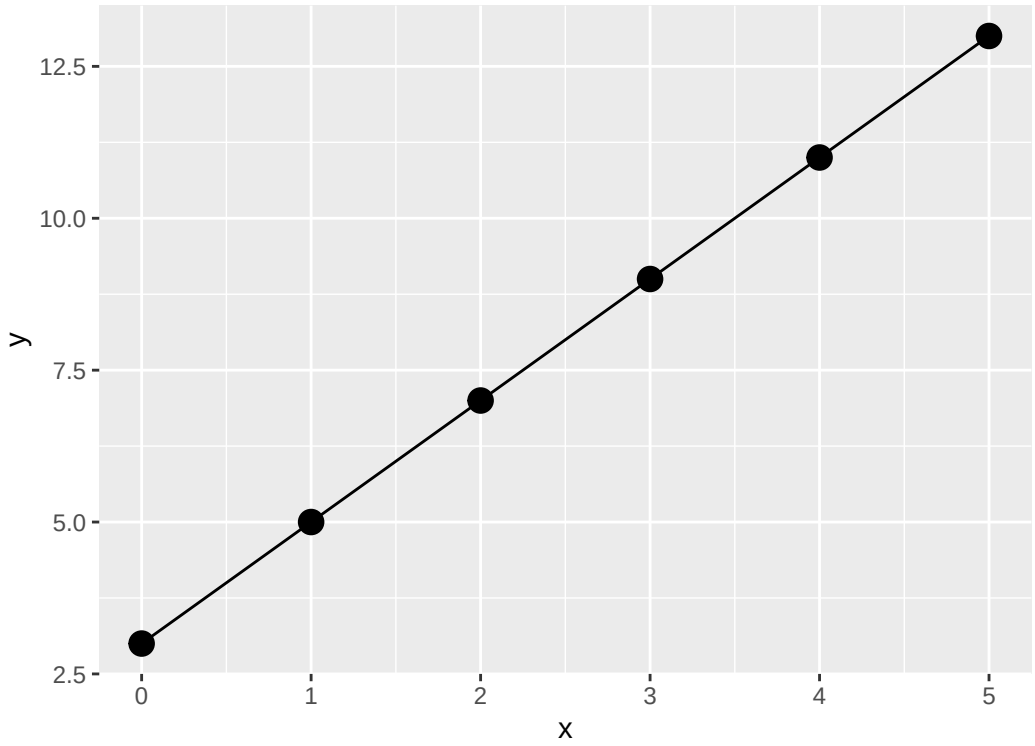


Figure 8.1: Graphische Darstellung der Daten aus Table 8.1

Die Variable m dahingehend bestimmt die Steigung der Gerade.

Definition 8.2. Wenn y durch eine lineare Gleichung $y = mx+b$ definiert ist, dann bestimmt die Variable m die Steiung der dazugehörenden Gerade. D.h. wenn sich die Variable x um einen Einheit vergrößert (verkleinert) wird der Wert von y um m Einheiten größer (kleiner). Gilt $m < 0$ dann umgekehrt.

Diese beiden trivialen Konzepte mit eigenen Definitionen zu versehen erscheint im ersten Moment vielleicht etwas übertrieben. Wie sich allerdings später zeigen wird, sind diese beiden Einsichten immer wieder zentral wenn es um die Interpretation von linearen statistischen Modellen geht.

8.2 Einfaches Beispiel - Daten

Table 8.2: Ausschnitt der Sprungdaten

jump_m	v_ms
4.36	6.13
4.31	6.39
4.56	6.56
4.75	6.44
5.52	7.30
5.63	7.19
5.70	7.30

8.3 Einfaches Beispiel - Grafik

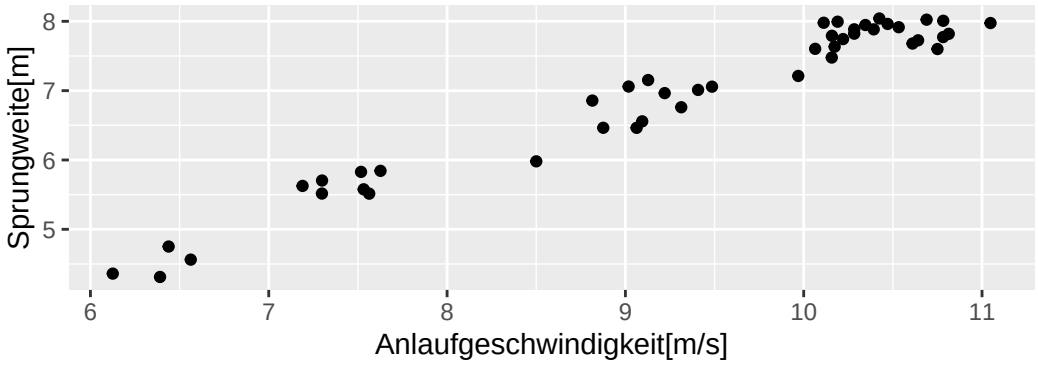


Figure 8.2: Zusammenhang der Anlaufgeschwindigkeit und der Sprungweite

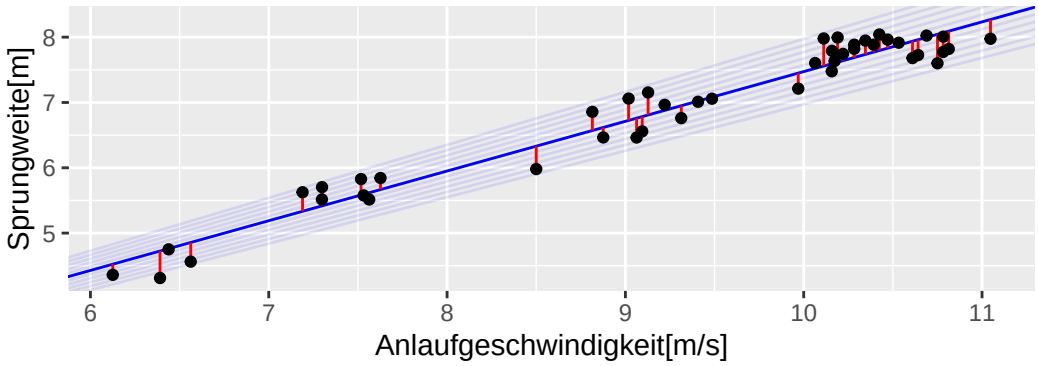


Figure 8.3: Zusammenhang der Anlaufgeschwindigkeit und der Sprungweite

8.4 Einfaches Beispiel - Regressionsgerade

8.5 Loss function

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 v_i))^2$$

$$y = -0.14 + 0.76 \times v$$

8.6 Regression in R

8.6.1 Model fitten mit `lm()`

```
mod <- lm(jump_m ~ v_ms, data = jump)
mod
```

```
Call:
lm(formula = jump_m ~ v_ms, data = jump)

Coefficients:
(Intercept)         v_ms
    -0.1385       0.7611
```

8.7 Formelsyntax in `lm(y ~ x, data)`

Table 8.3: Formelsyntaxbeispiele für `lm()`

Modell	Formel	Erklärung
$y = \beta_0$	`y ~ 1`	y-Ab
$y = \beta_0 + \beta x$	`y ~ x`	y-Ab und StKoef
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	`y ~ x1 + x2`	y-Ab und 2 StKoe

y-Ab = y-Achsenabshnitt, StKoef = Steigungskoeffizient

8.8 `lm()`-fit mit `summary()` inspizieren

```
summary(mod)
```

```
Call:
lm(formula = jump_m ~ v_ms, data = jump)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44314 -0.22564  0.02678  0.19638  0.42148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13854    0.23261  -0.596    0.555
v_ms         0.76110    0.02479  30.702   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2369 on 43 degrees of freedom
Multiple R-squared:  0.9564,    Adjusted R-squared:  0.9554
F-statistic: 942.6 on 1 and 43 DF,  p-value: < 2.2e-16
```

8.9 `lm()` und ein paar friends...

Koeffizienten und Standardschätzfehler

```
coef(mod)
```

```
(Intercept)        v_ms
 -0.1385361   0.7611019
```

```
sigma(mod)
```

```
[1] 0.2369055
```

Residuen

```r
# Nur die ersten beiden
# Residuen
# damit der Ausdruck
# auf das Slide passt.
resid(mod)[1:2]
```

```
         1          2
-0.1626772 -0.4124884
```

9 Inferenz

Nachdem das Modell gefittet wurde stellt sich die Frage ob tatsächlich ein Zusammenhang zwischen der Prädiktorvariable und der abhängigen Variable besteht. Da das einfache lineare Modelle zwei Parameter β_0 und β_1 beinhaltet (streng genommen ist σ^2 ein dritter Parameter) kann diese Frage auf beide Koeffizienten angewendet werden.

Eine kurze Überlegung zeigt, dass wenn zwischen der Prädiktorvariablen und y kein Zusammenhang besteht, dann sollte der Steigungskoeffizient β_1 gleich Null sein bzw. auf Grund von Stichprobenvariabilität in der Nähe von Null sein. Daher ist eine plausible Hypothese die sich statistisch Überprüfung lässt:

$$H_0 : \beta_1 = 0$$

9.1 Inferenz

9.1.1 Modellannahmen

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, N$$
$$\epsilon_i \sim N(0, \sigma^2) \quad \text{identisch, unabhängig verteilt}$$

9.2 Modellannahmen - Verteilung der Werte für gegebene x-Werte

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

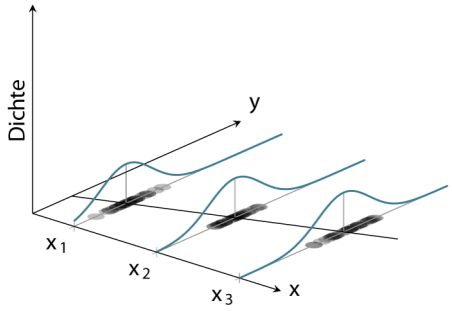


Figure 9.1: Verteilung der Daten für verschiedene x-Werte

9.3 Statistische Hypothesen

9.3.1 Ungerichtet

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

9.3.2 Gerichtet

$$H_0 : \beta_1 \leq 0$$
$$H_1 : \beta_1 > 0$$

9.4 Teststatistik informell herleiten

9.4.1 Simulation unter der H_0

$$N = 45$$
$$x \sim \mathcal{U}(-1, 1)$$
$$y \sim \mathcal{N}(0, \sigma)$$
$$\sigma = 1$$
$$H_0 : \beta_1 = 0$$

9.5 Teststatistik informell herleiten

9.6 Stichprobenverteilung von β_1 unter der Annahme $\beta_1 = 0$

9.7 Verteilung der Statistik unter der H_0

9.7.0.1 Standardfehler von β_1

$$\sigma_{\beta_1} = \sqrt{\frac{\sigma^2}{\sum (X_i - \bar{X})^2}}$$

σ lässt sich abschätzen mit:

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} e_i^2 / (N - K)}$$

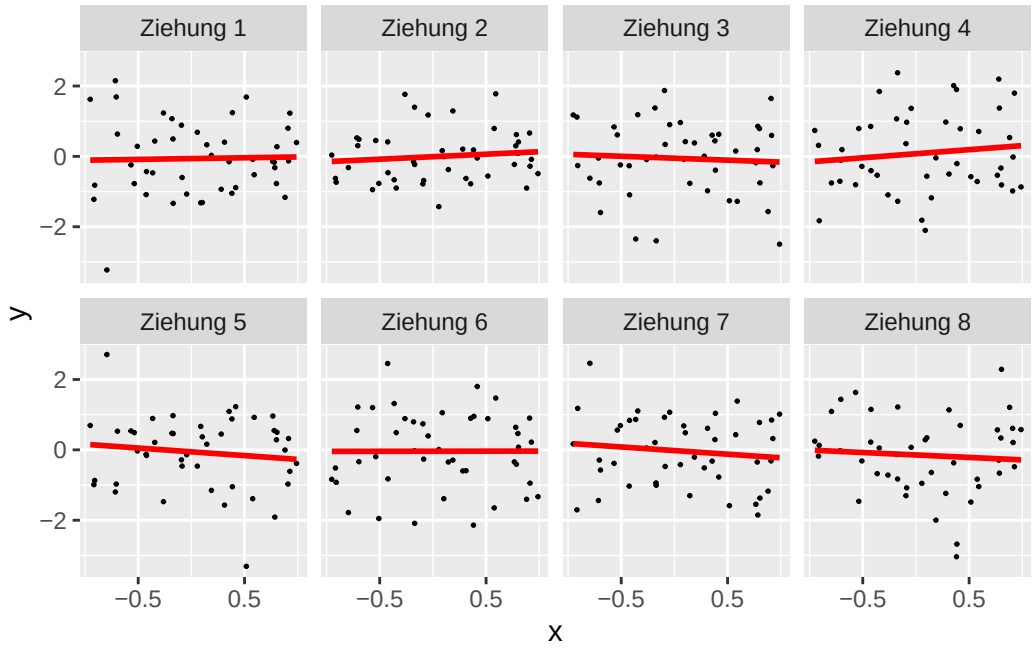


Figure 9.2: Acht Zufallsziehung unter der H_0

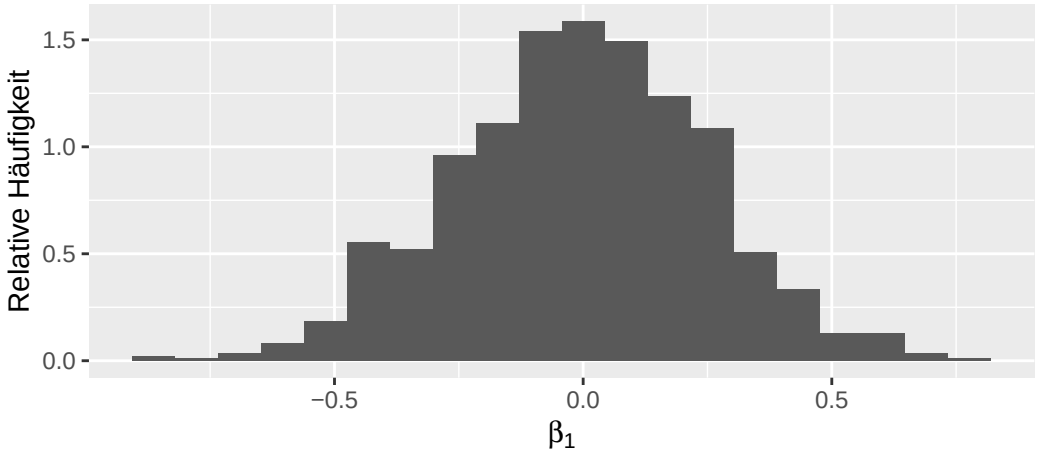


Figure 9.3: Verteilung der β_1s - 1000 Simulationen unter der Annahme der H_0.

9.7.1 in R

```
sigma(mod)
```

```
[1] 0.2369055
```

9.8 Verteilung der Statistik unter der H_0

Unter den Annahmen des Regressionsmodells und der H_0 gilt:

$$\frac{\beta_1}{\sigma_{\beta_1}} \sim t_{N-2}$$

Mittels α lässt sich daher wieder ein kritischer Wert bestimmen ab dem die H_0 verworfen wird.

9.9 Teststatistik

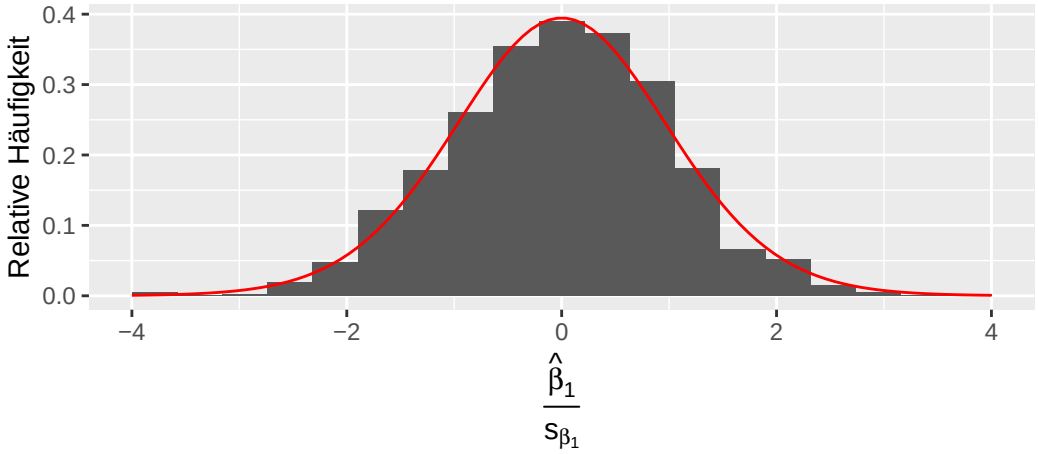


Figure 9.4: Verteilung von $\frac{\beta_1}{s_{\beta_1}}$,Dichtefunktion der t-Verteilung (rot) mit $df = n - 2$

9.10 Verteilung der $\hat{\sigma} = \sqrt{\sum_{i=1}^{N} e_i^2/(N-K)}$

9.11 Nochmal `summary()`

```
summary(mod)
```

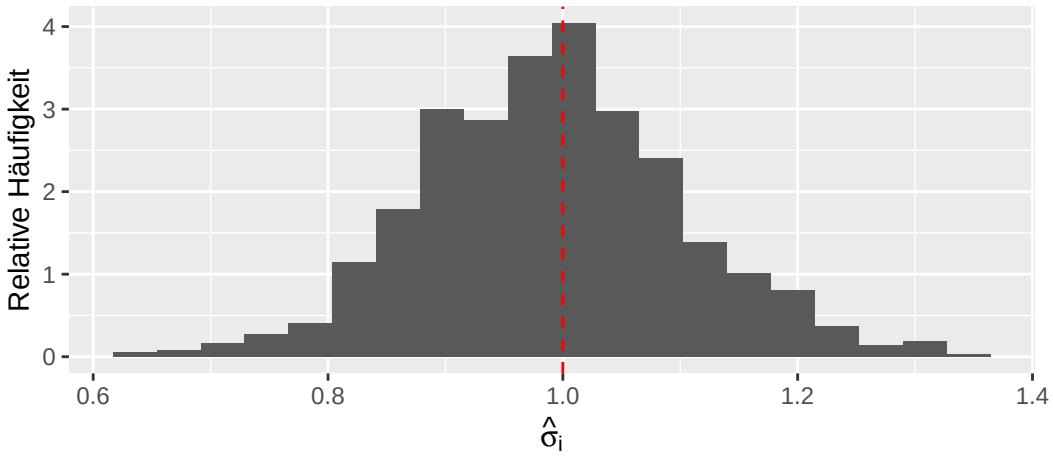


Figure 9.5: Verteilung von $\hat{\sigma}$

```
Call:
lm(formula = jump_m ~ v_ms, data = jump)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44314 -0.22564  0.02678  0.19638  0.42148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13854    0.23261  -0.596    0.555
v_ms         0.76110    0.02479  30.702   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2369 on 43 degrees of freedom
Multiple R-squared:  0.9564,    Adjusted R-squared:  0.9554
F-statistic: 942.6 on 1 and 43 DF,  p-value: < 2.2e-16
```

9.12 Konfidenzintervalle für die Koeffizienten

9.12.1 Formel

$$\hat{\beta_j} \pm q_{t_{\alpha/2, df=N-2}} \times \hat{\sigma}_{\beta_j}$$

9.12.2 In R

```
confint(mod)
```

```
                2.5 %    97.5 %
(Intercept) -0.6076488 0.3305767
v_ms         0.7111082 0.8110957
```

9.13 Zum Nacharbeiten

N. Altman and Krzywinski (2015b) und Kutner et al. (2005, 40–48)

10 Modellfit

10.1 Residuen

10.2 Was sind noch mal Residuen ϵ_i bzw. deren Schätzer $\hat{\epsilon}_i = e_i$

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

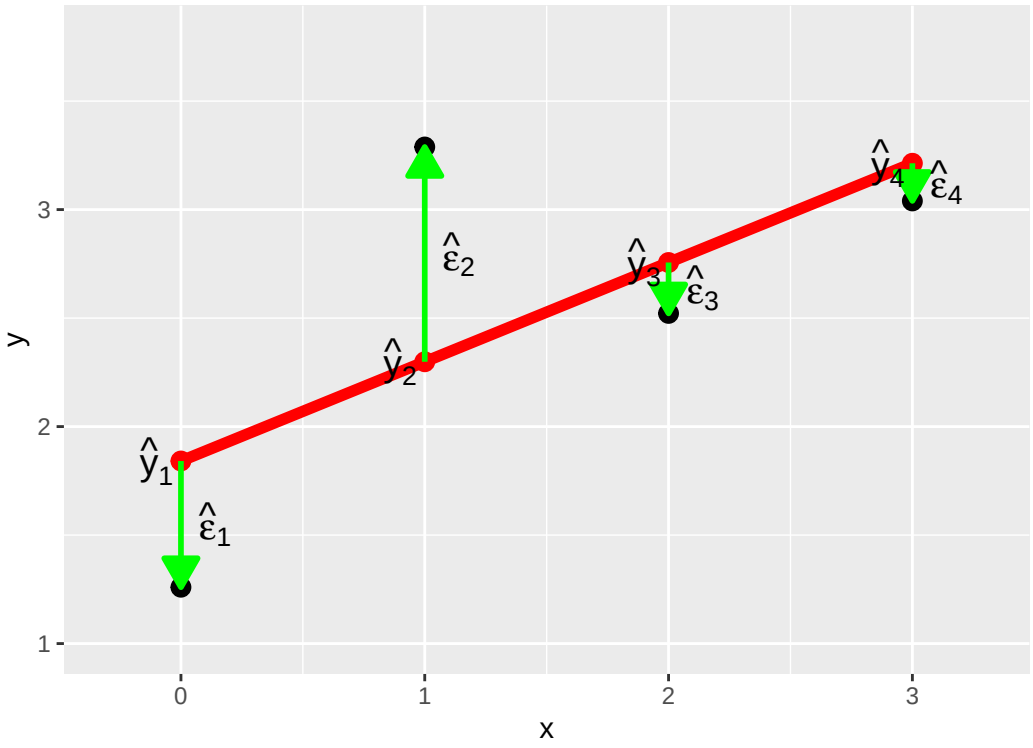


Figure 10.1: Spielzeugbeispiel mit Residuen $\hat{\epsilon}_i = e_i = y_i - \hat{y}_i$

10.3 Annahme: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

10.4 Übersicht Residuen

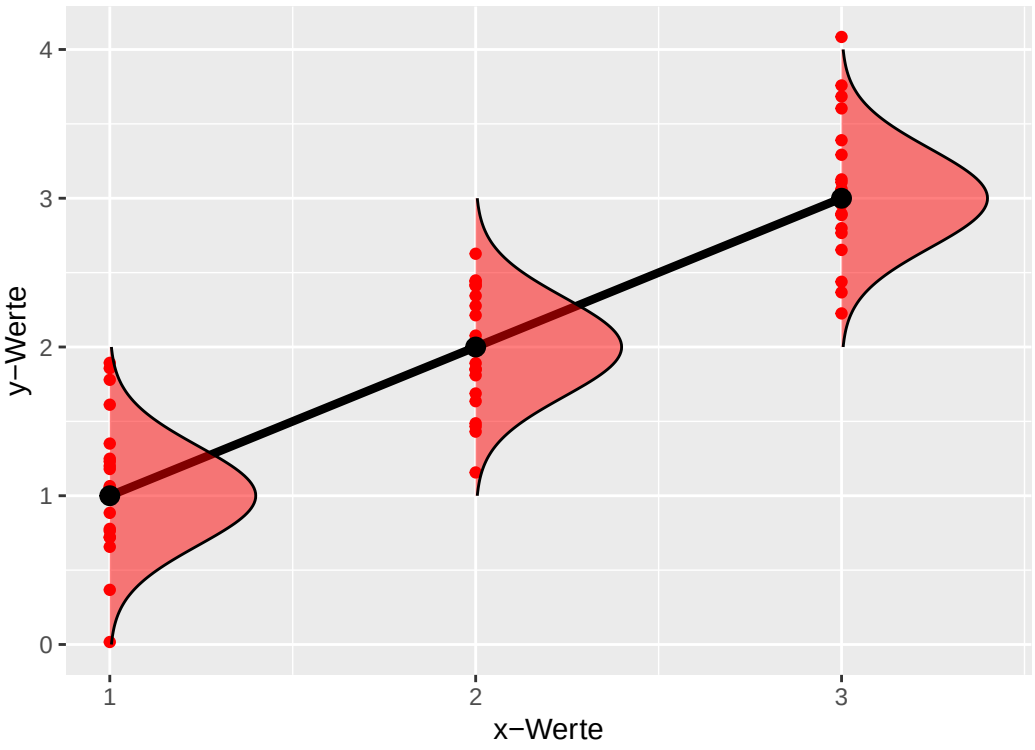


Figure 10.2: Verteilung der Werte für verschiedene x-Werte (rote Punkte) und die resultierende Regressionsgerade mit den Vorhersagewerte $\hat{y}_i$ (schwarze Punkte)

Table 10.1: Übersicht über verschiedene Arten von Residuen[1]

Typ	Berechnung	Ziel
Einfache Residuen	$e_i = y_i - \hat{y}_i$	Verteilungsannahme
Standardisierte Residuen	$e_{Si} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$	Verteilungsannahme
Studentized Residuen	$e_{Ti} = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$	Einfluss auf Modell

10.5 Residuen in R berechnen mit `residuals()` und Freunden

```
residuals(mod)[1:5] # einfache Residuen
```

```
        1         2         3         4         5 
 -9.300928  -9.368288 -11.217658  -5.572108  -6.363565 
```

```
rstandard(mod)[1:5] # standardisierte Residuen
```

```
        1         2         3         4         5 
-1.4592936 -1.4598906 -1.7440573 -0.8724351 -0.9916310 
```

```
rstudent(mod)[1:5] # studentized Residuen
```

```
        1         2         3         4         5 
-1.4814779 -1.4821191 -1.7928881 -0.8697060 -0.9914135 
```

10.6 Residuen in R inspizieren

```
y_hat <- predict(mod)
plot(y_hat, residuals(mod))
plot(y_hat, rstandard(mod))
plot(y_hat, rstudent(mod))
```

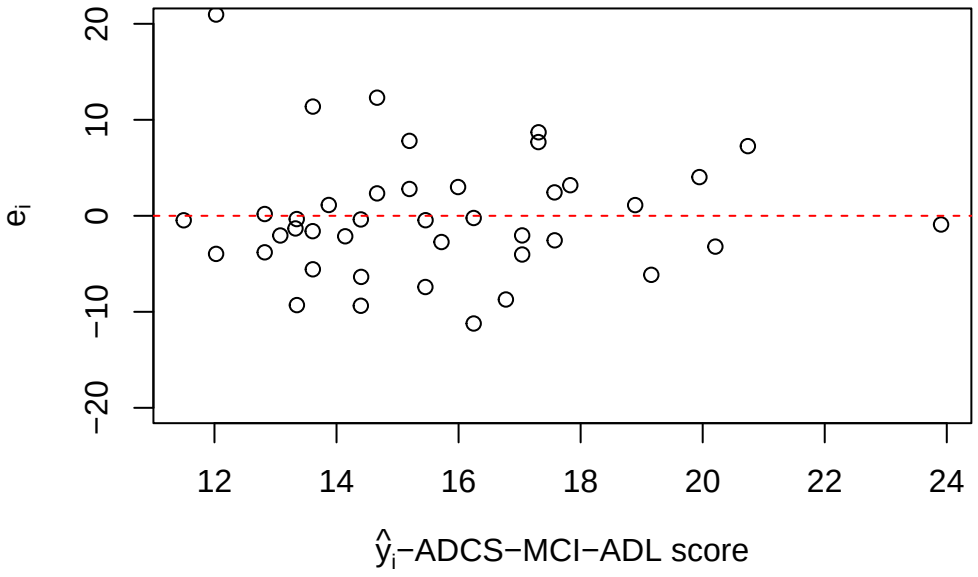


Figure 10.3: Streudiagramm der Residuen $\hat{\epsilon_i}$ gegen die Vorhersagewerte $\hat{y}_i$

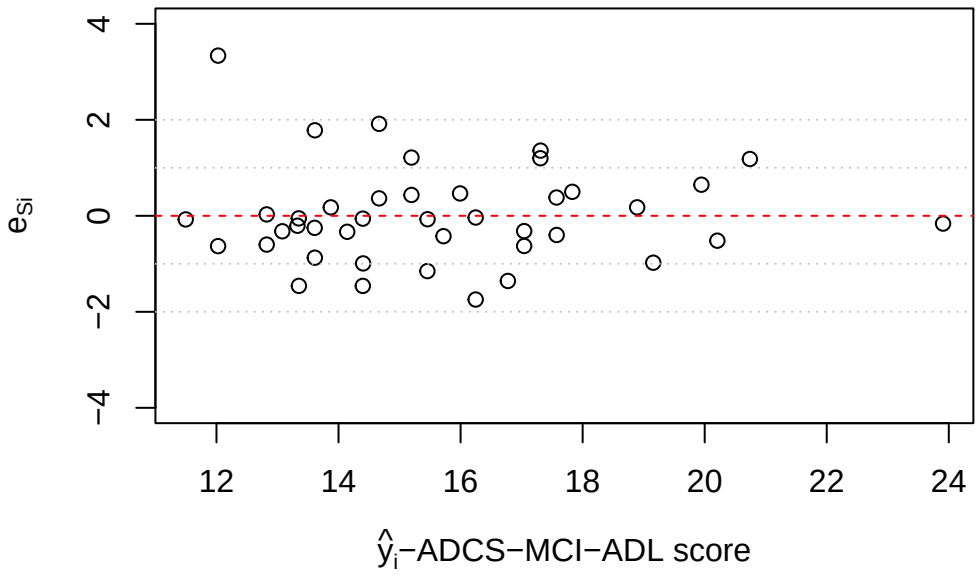


Figure 10.4: Streudiagramm der standardisierten Residuen $\hat{\epsilon}_{Si}$ gegen die Vorhersagewerte $\hat{y}_i$

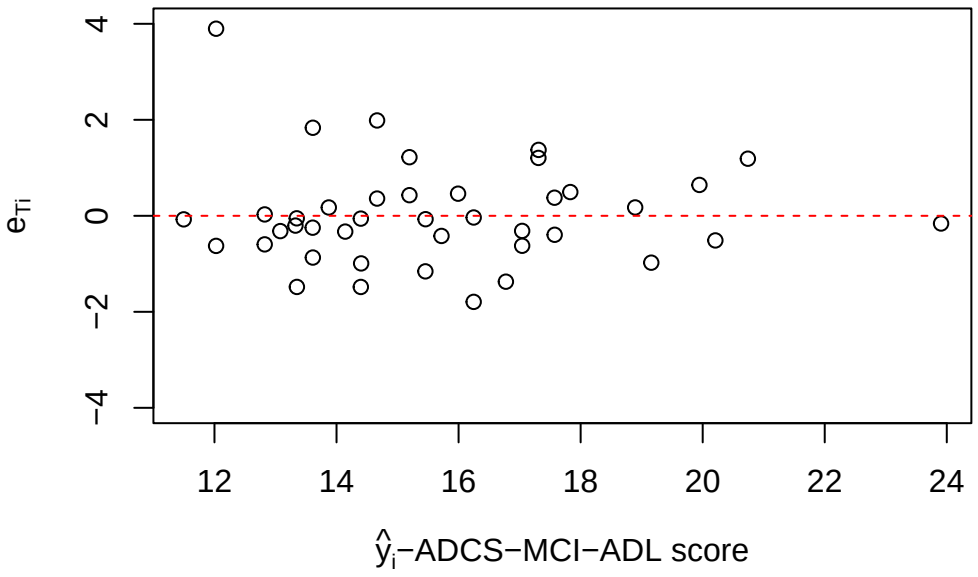


Figure 10.5: Streudiagramm der studentized Residuals $\hat{\epsilon}_{Ti}$ gegen die Vorhersagewerte $\hat{y}_i$

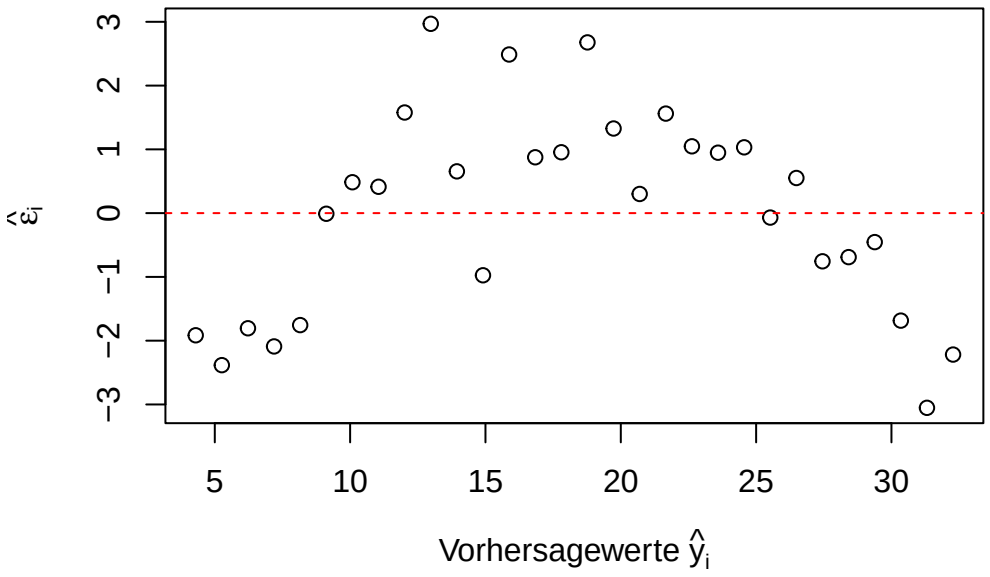


Figure 10.6: Beispielstreudiagramm

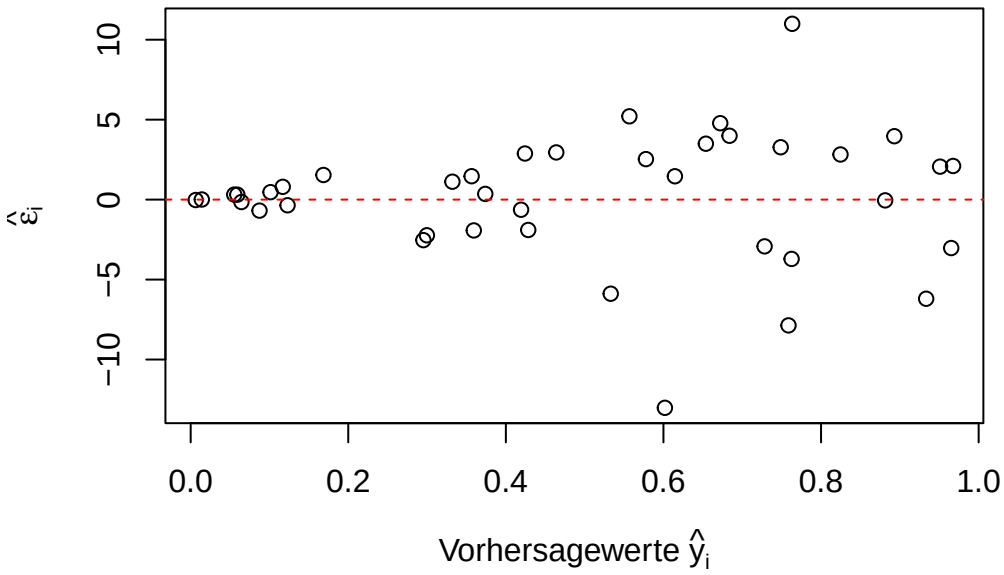


Figure 10.7: Beispielstreudiagramm

10.7 Diagnoseplot - Einfache Residuen $\widehat{\epsilon}_i \sim \widehat{y}_i$

10.8 Diagnoseplot - Standardisierte Residuen $\widehat{\epsilon}_{Si} \sim \widehat{y}_i$

10.9 Diagnoseplot - Studentized Residuen $\widehat{\epsilon}_{Ti} \sim \widehat{y}_i$

10.10 Diagnoseplot - Wie sehen Probleme aus?

10.11 Diagnoseplot - Wie sehen Probleme aus?

10.12 Wie kann die Verteilung der Residuen überprüft werden?

Table 10.2: Spielzeugbeispieldaten mit $n = 5$

y
-2.0
5.0
-1.2
0.1
7.0

[1] h_i = Influenz von Punkt i

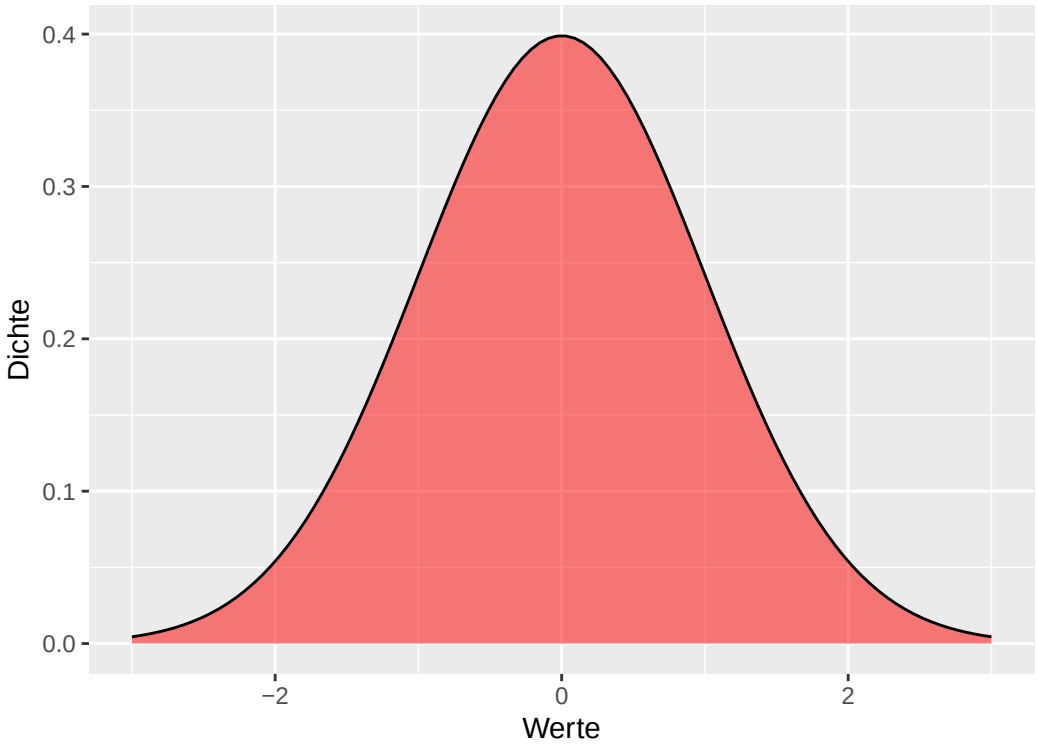


Figure 10.8: Dichtefunktion der Standardnormalverteilung

10.13 Konstruktion eines qq-Graphen

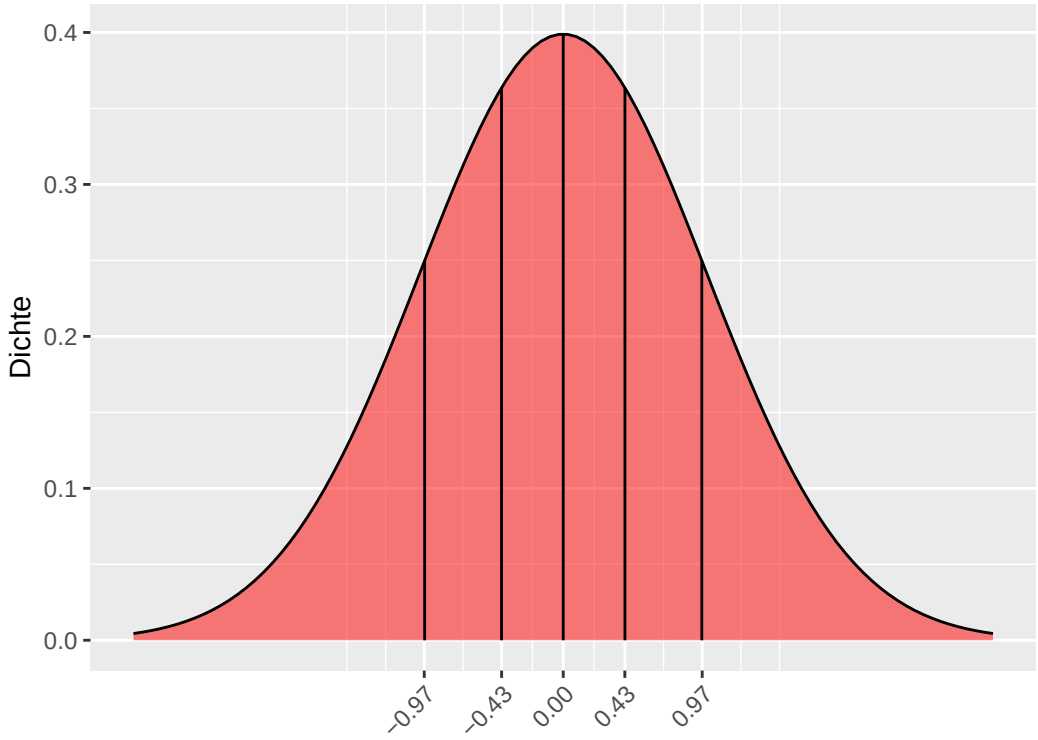


Table 10.3: Sortierte Datenwerte

kleinster	2.kleinster	mittlerer	2.größter	größter
-2	-1.2	0.1	5	7

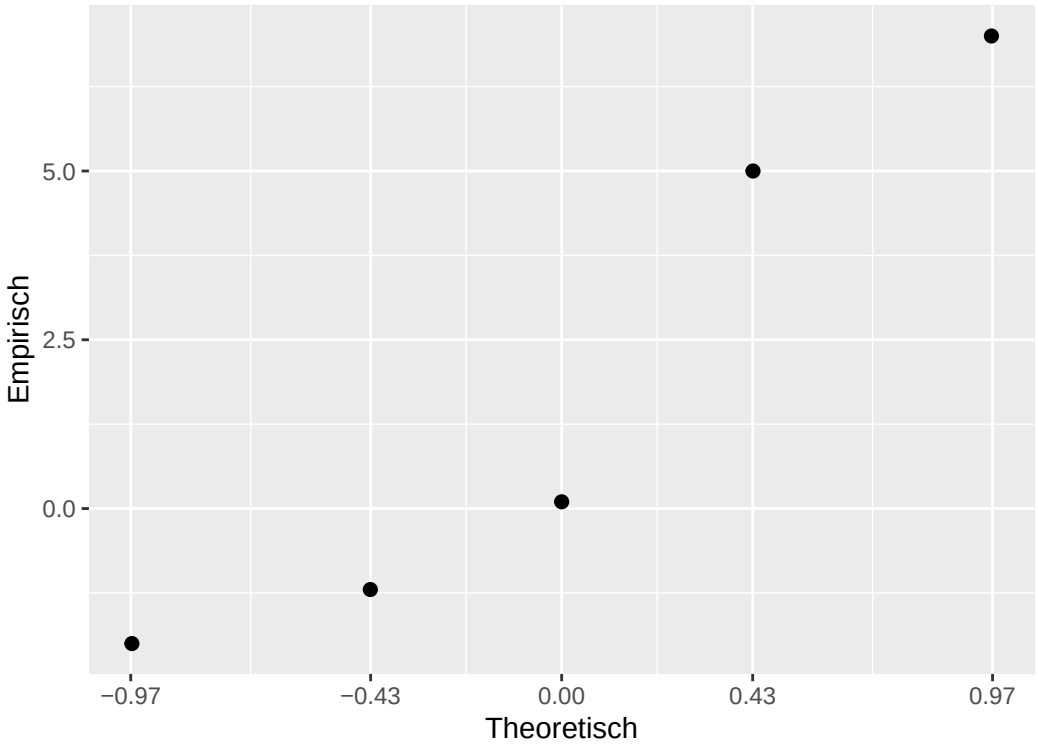
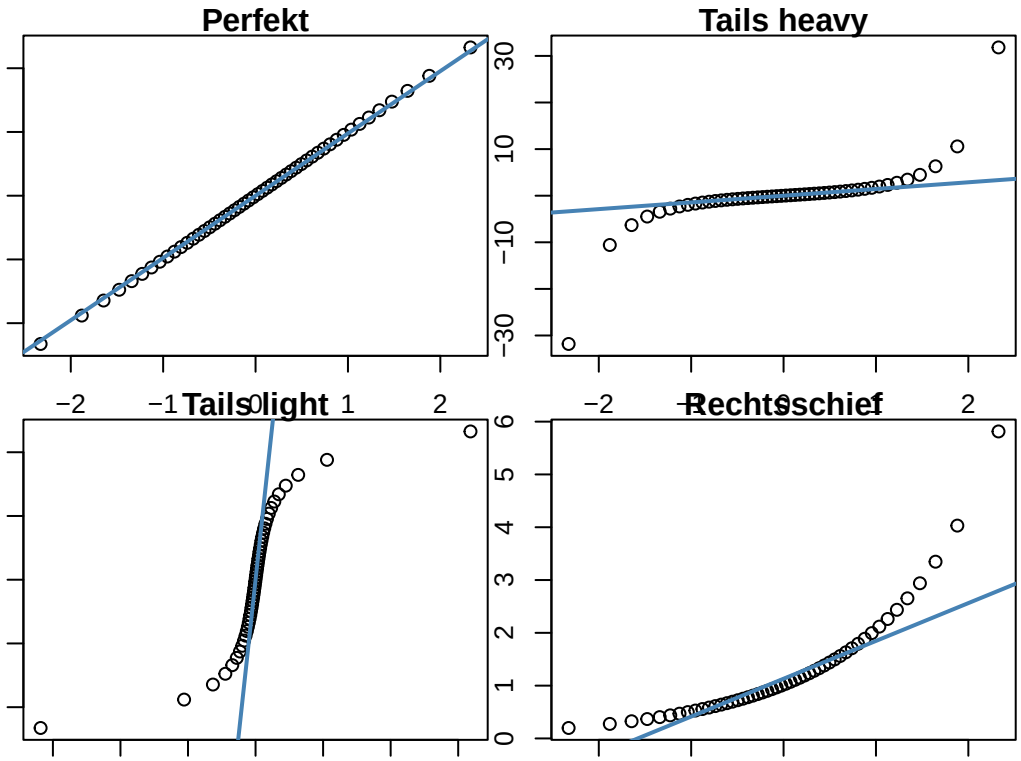


Figure 10.9: Streudiagramm der empirischen Werte gegen die theoretischen Quantilen

10.14 Konstruktion eines qq-Graphen

10.15 Beispiele für qq-Graphen mit `qqnorm()` und `qqline()`



10.16 Diagnoseplot - QQ-Diagramm

[2]

10.17 `summary()`

```
Call:
lm(formula = adcs ~ adas, data = adl)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2177  -3.8033  -0.4663   2.7950  20.9634

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

[2] In R mit `qqnorm(resid(mod))`

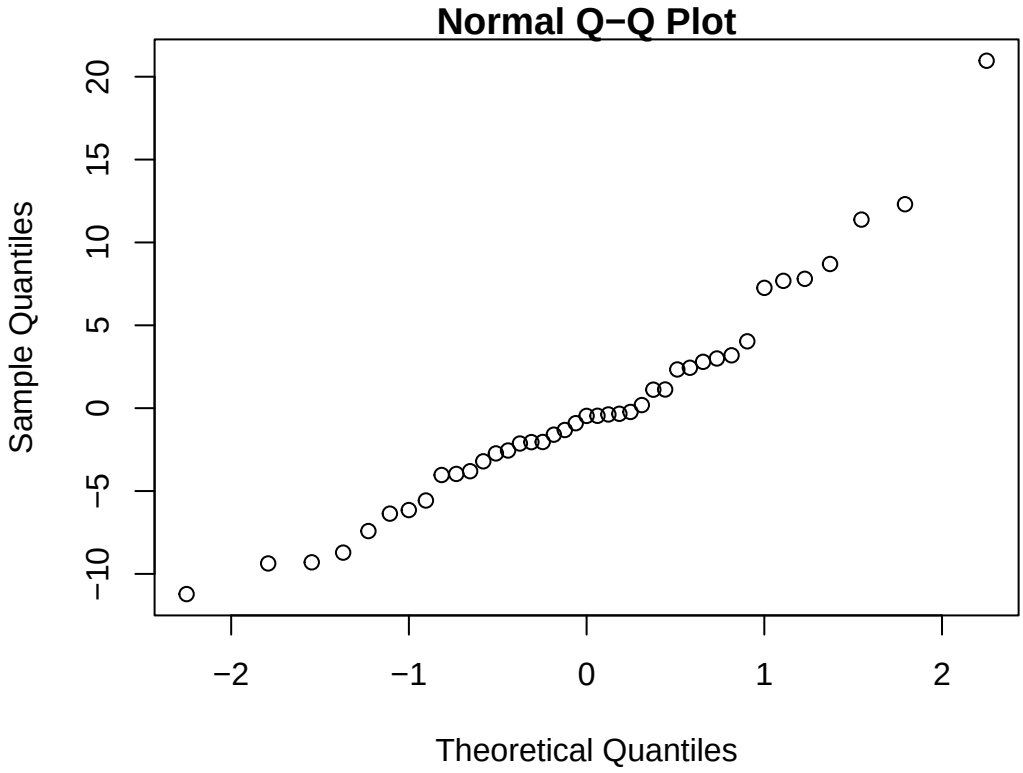


Figure 10.10: QQ-Diagramm der Residuen des ADAS-ADCS-Modells

```
(Intercept)  26.5445     4.3052   6.166 3.05e-07 ***
adas         -0.2638     0.1015  -2.599   0.0131 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.516 on 39 degrees of freedom
Multiple R-squared:  0.1477,    Adjusted R-squared:  0.1258
F-statistic: 6.757 on 1 and 39 DF,  p-value: 0.01312
```

10.18 Neue Idee zu Residuen

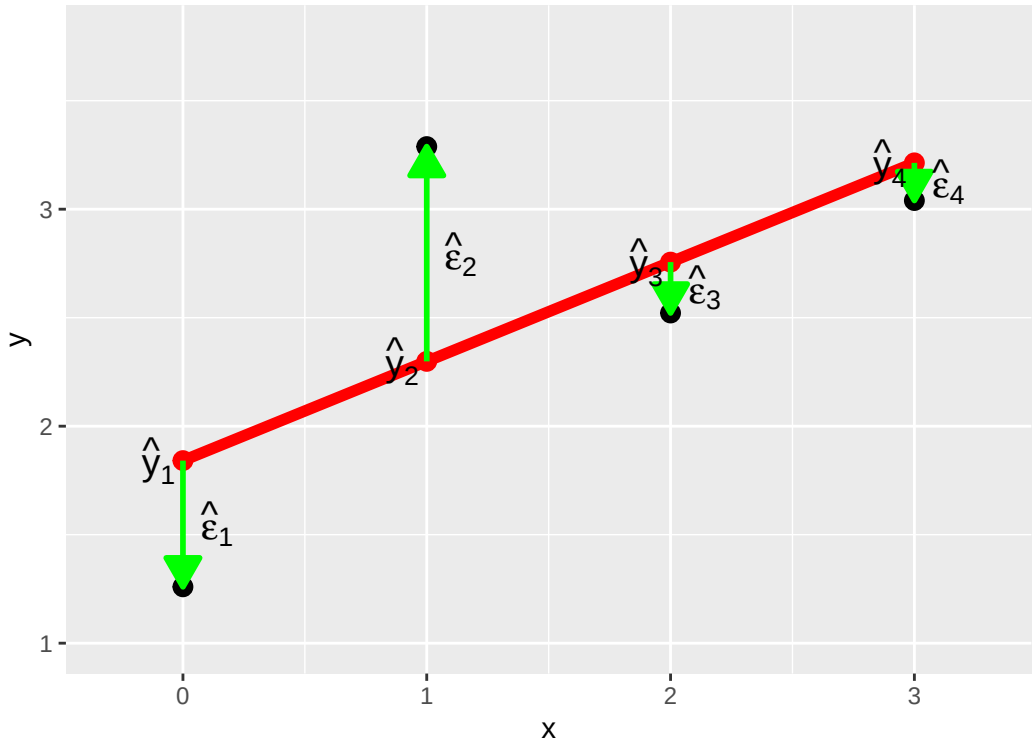


Figure 10.11: Spielzeugbeispiel mit Residuen $\hat{\epsilon}_i = e_i = y_i - \hat{y}_i$

10.19 Zum Nacharbeiten

Kutner et al. (2005, 100–114)
N. Altman and Krzywinski (2016b)
Fox (2011, 285–96)

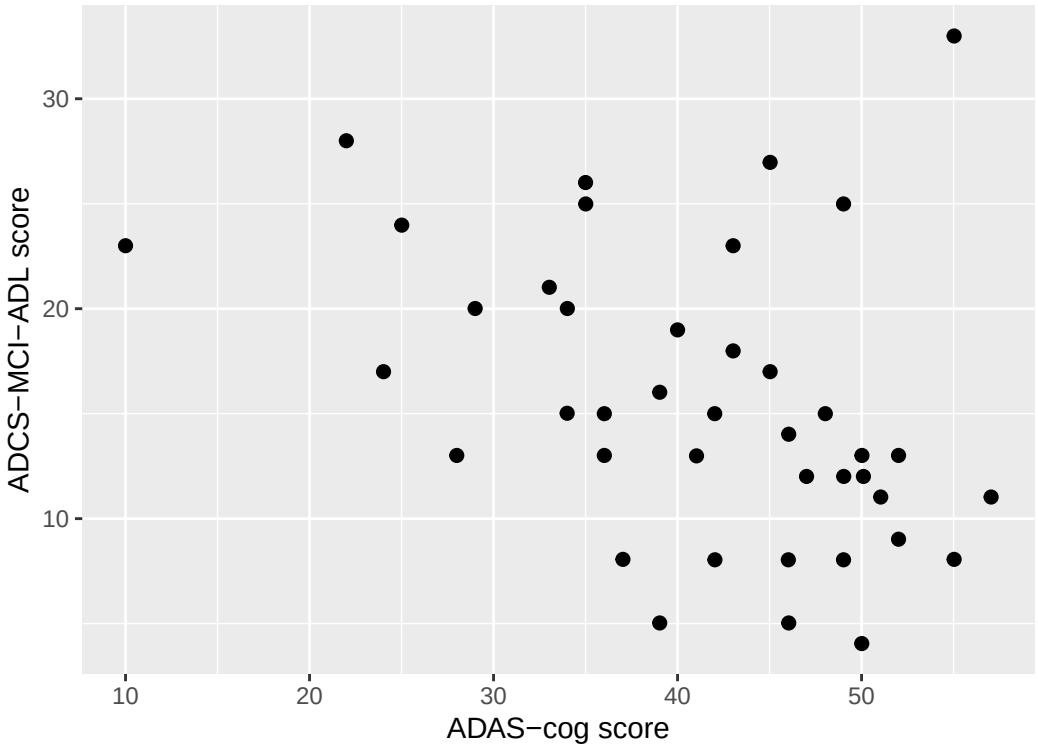


Figure 10.12: Streudiagramm der ADCS-MCI-ADL scores gegen ADAS-cos scores

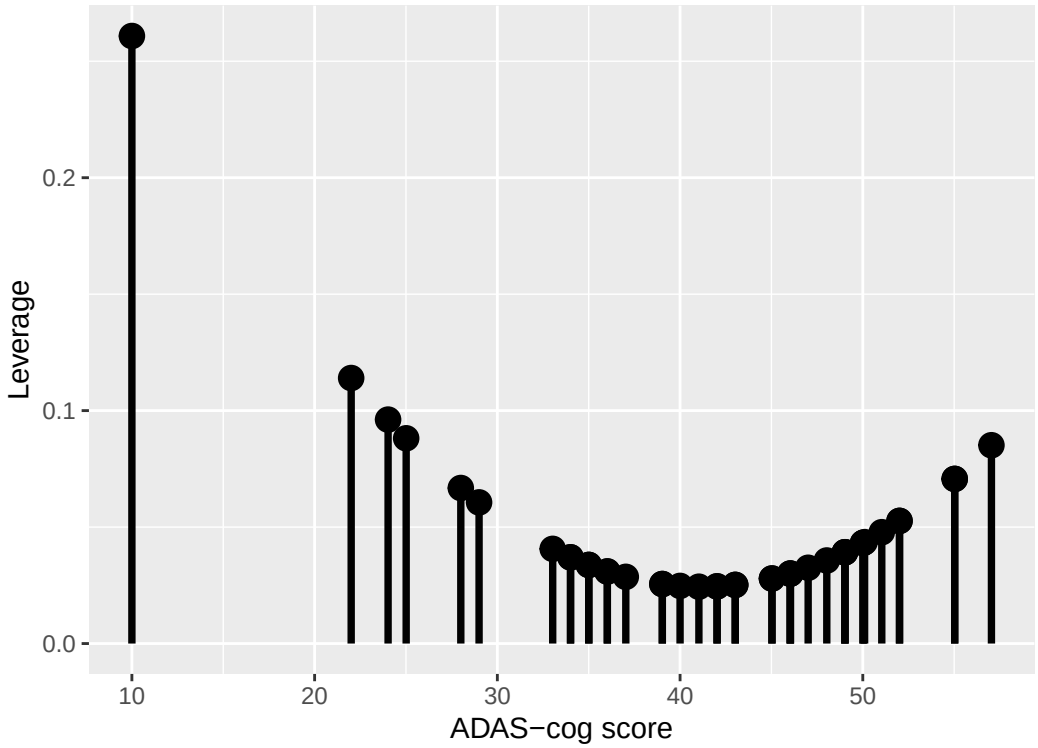


Figure 10.13: Hebelwerte der jeweiligen x_is

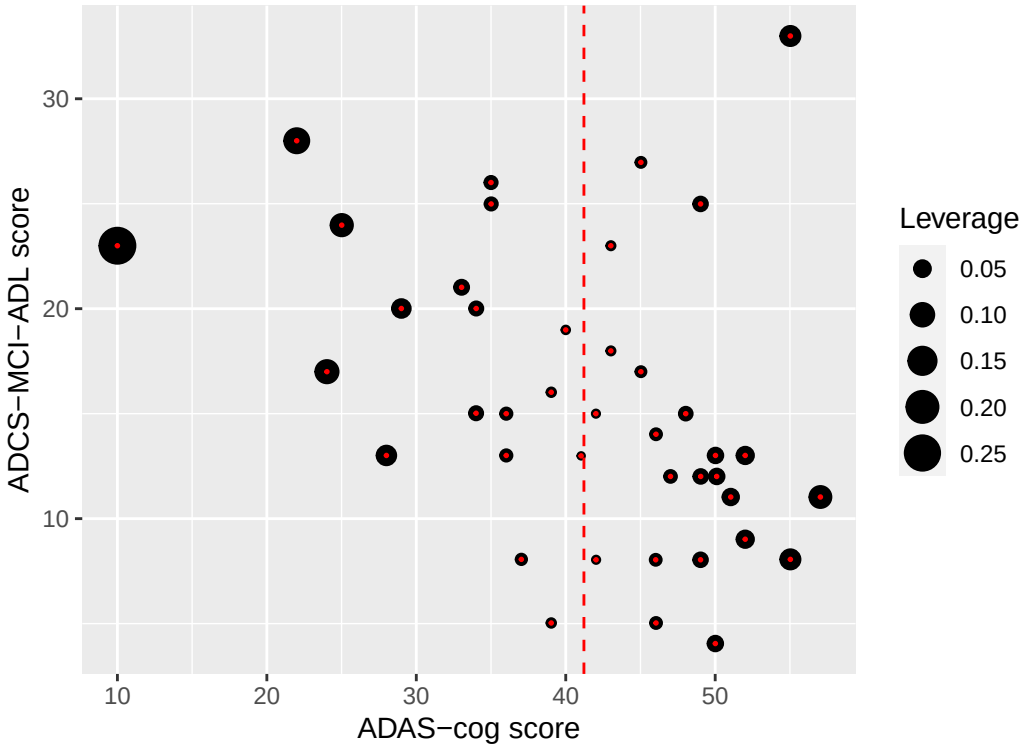


Figure 10.14: Hebelwerte der jeweiligen Datenpunkte

10.20 Hebelwerte

10.21 DFFITS

Mit Hilfe der Hebelwerte lassen sich verschiedene Maße erstellen um den Einfluss von Datenpunkten auf das Modell zu überprüfen. Ein Maß wird als bezeichnet (siehe Equation 10.1)

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{\hat{\sigma}^2 h_i}} \tag{10.1}$$

Im Zähler kommen vin Equation 10.1 zweimal vorhergesagte y-Werte vor. $\hat{y}_i$ ist dabei der ganz normale Vorhersagewert der uns mittlerweile schon mehrfach begegnet ist. Der zweite Wert $\hat{y}_{i(i)}$ bezeichnet den vorhergesagten Wert aus dem Modell aus dem der Wert y_i weggelassen wurde. D.h, dass Modell ist mit einem Wert weniger gefittet worden. Daher misst die Differenz $\hat{y}_i - \hat{y}_{i(i)}$ den Unterschied in den Vorhersagewerte zwischen zwei Modellen bei denen einmal der Wert y_i zum fitten verwendet wurde und einmal wenn y_i nicht zum fitten verwendet wurde. Umso größer der Unterschied zwischen diesen beiden Werte umso größer ist der Einfluss des Wertes y_i auf den Modellfit. Den Nenner von Equation 10.1 lassen wir mal fallen, da es sich dabei nur um einen Normierungswert handelt. Dementsprechend, wird mittels DFFITS für jeden Datenpunkt ein Wert ermittelt und umso größer dieser Wert ist umso größer ist der Einfluss des jeweiligen Datenpunktes auf den Modellfit.

Im idealen Fall sollte alle Datenpunkt ungefähr den gleichen Einfluss haben und einzelne Datenpunkte die einen übermäßig großen Einfluss auf das Modell haben sollten noch einmal genauer inspiziert werden.

> **Tip**
>
> Als Daumenregel, kann für kleine bis mittlere Datensätze ein DFFITS von ≈ 1 auf Probleme hindeuten, während bei großen Datensätzen $\approx 2\sqrt{k/N}$ als Orientierungshilfe verwendet werden kann (k := Anzahl der Prediktoren, N := Stichprobengröße).

> **Warning**
>
> Wenn ein Wert außerhalb der Daumenregel liegt, heißt das nicht, dass er automatisch ausgeschlossen werden muss/soll, sondern lediglich inspiziert werden sollte und das Modell mit und ohne diesen Wert interpretiert werden sollte.

In `R` können die DFFITS werden mittels der `dffits()`-Funktion berechnet werden. Als Parameter erwartet `dffits()` das gefittete `lm()`-Objekt. Ähnlich wie bei den Residuen, werden die DFFITS-Werte gegen die vorhergesagten y_i-Werte graphisch abgetragen um die Wert zu inspizieren und Probleme in der Modellspezifikation zu identifizieren.

```
plot(adl$y_hat, dffits(mod),
     ylim=c(-2,2),
     xlab=expression(hat(y)[i]),
     ylab='DFFIT-Wert')
abline(h=c(-1,1), col='red', lty=2)
```

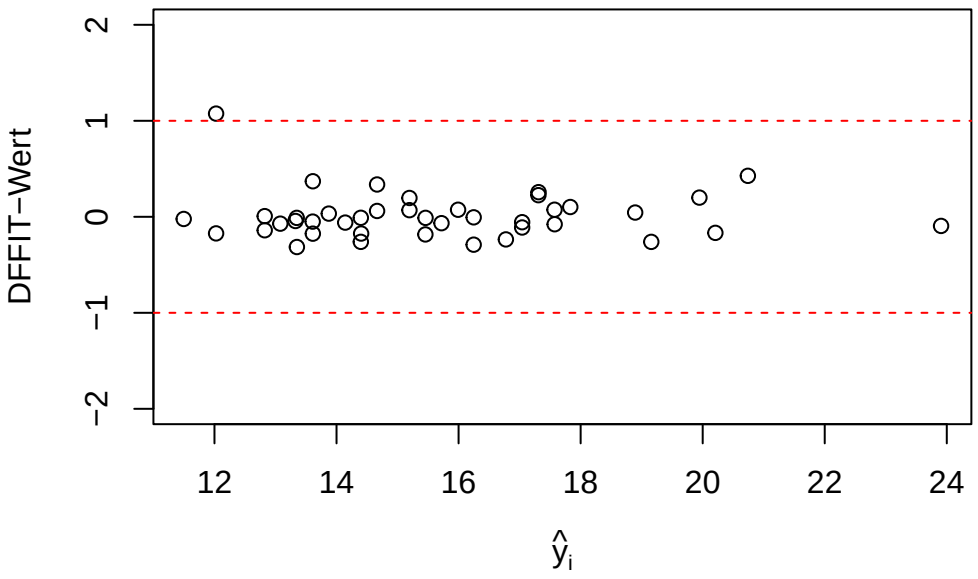


Figure 10.15: Beispiel für DFFITS gegen $\hat{y}_i$

In Figure 10.15 sind die DFFITS-Werte gegen die vorhergesagten Werte $\hat{y}_i$ abgetragen und zusätzlich die Daumenregel ± 1 eingezeichnet. Hier ist ein Wert nur gerade so außerhalb des vorgeschlagenen Bereichs. Hier könnte daher sich dieser Datenpunkt noch einmal genauer angeschaut werden, ob bei Ausschluß des Wertes es zu einer qualitativ anderen Interpretation der Daten kommt oder ob bespielsweise Übertragungsfehler für diesen Wert vorliegen oder sonstige Gründe.

10.22 Cooks-Abstand

Ein Maß um den Einfluss von einzelnen Datenpunkten auf die Vorhersagewerte $\hat{y}_i$ über alle Werte abzuschätzen.

$$D_i = \frac{\sum_{j=1}^{N} (\hat{y_j} - \hat{y}_{j(i)})}{k\hat{\sigma}^2}$$

10.22.1 Daumenregel

$D_i > 1$

10.22.2 In R

```
cooks.distance()
```

10.23 Cooks-Abstand plot

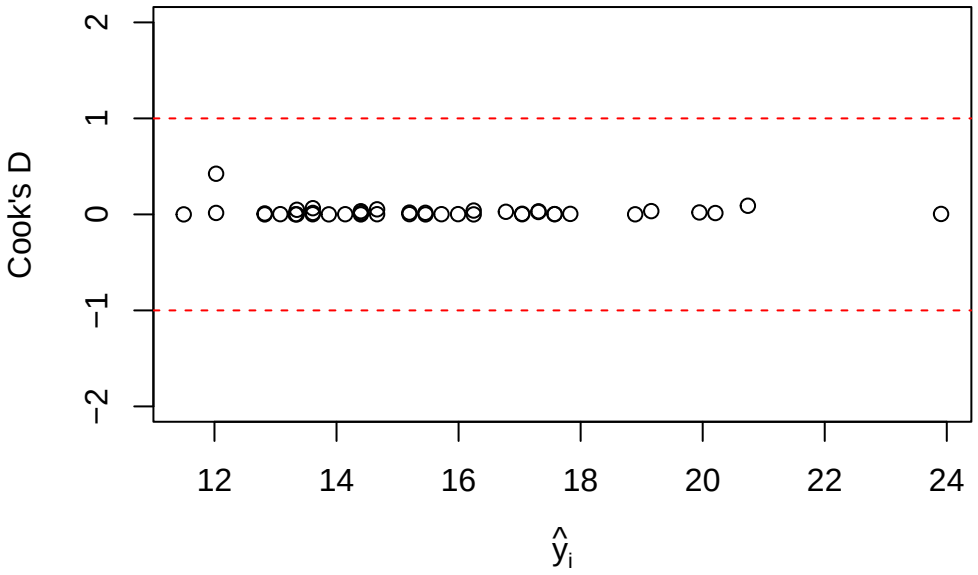


Figure 10.16: Cook's D_i gegen $\hat{y}_i$

10.24 DFBETAS

Ein Maß für die Veränderung der β-Koeffizienten durch einzelne Datenpunkte i.

$$(DFBETAS)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\hat{\sigma}^2 c_{kk}}}$$

10.24.1 Daumenregel

Für kleine bis mittlere Datensätze ≈ 1
Für große Datensätze $\approx 2/\sqrt{N}$

10.24.2 In R

`dfbeta()`[3]

[3]Es wird eine Matrize mit k-Spalten zurückgegeben.

10.25 DFBETAS

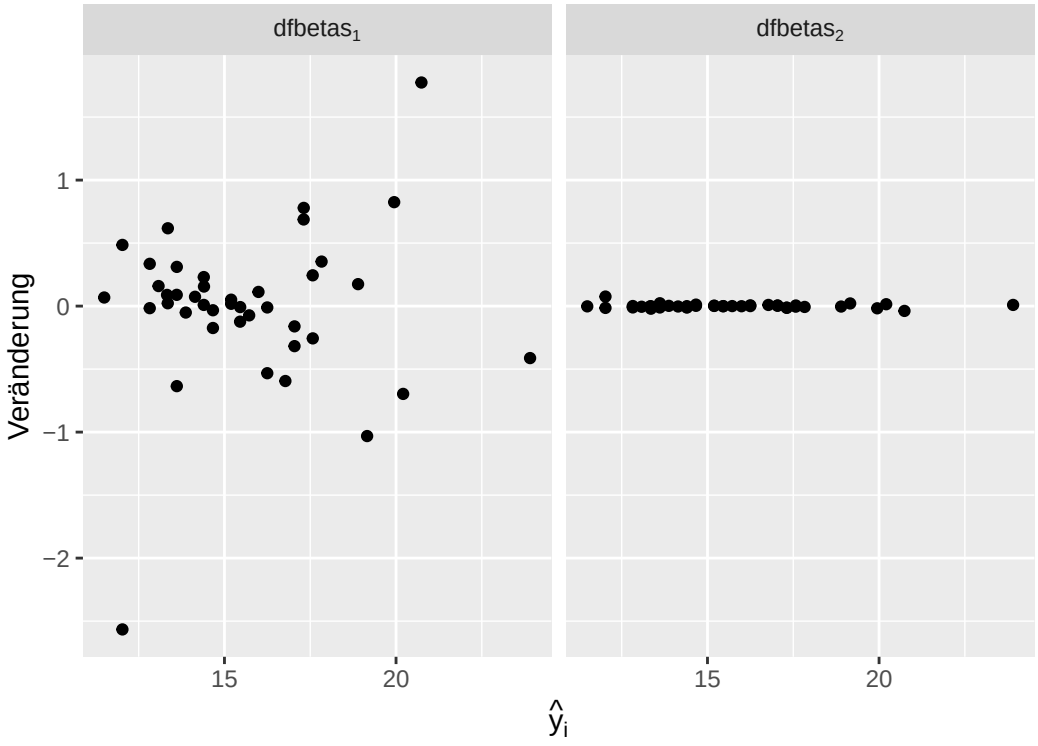


Figure 10.17: DFBETA-Werte für β_0 und β_1 gegen $\hat{y}_i$

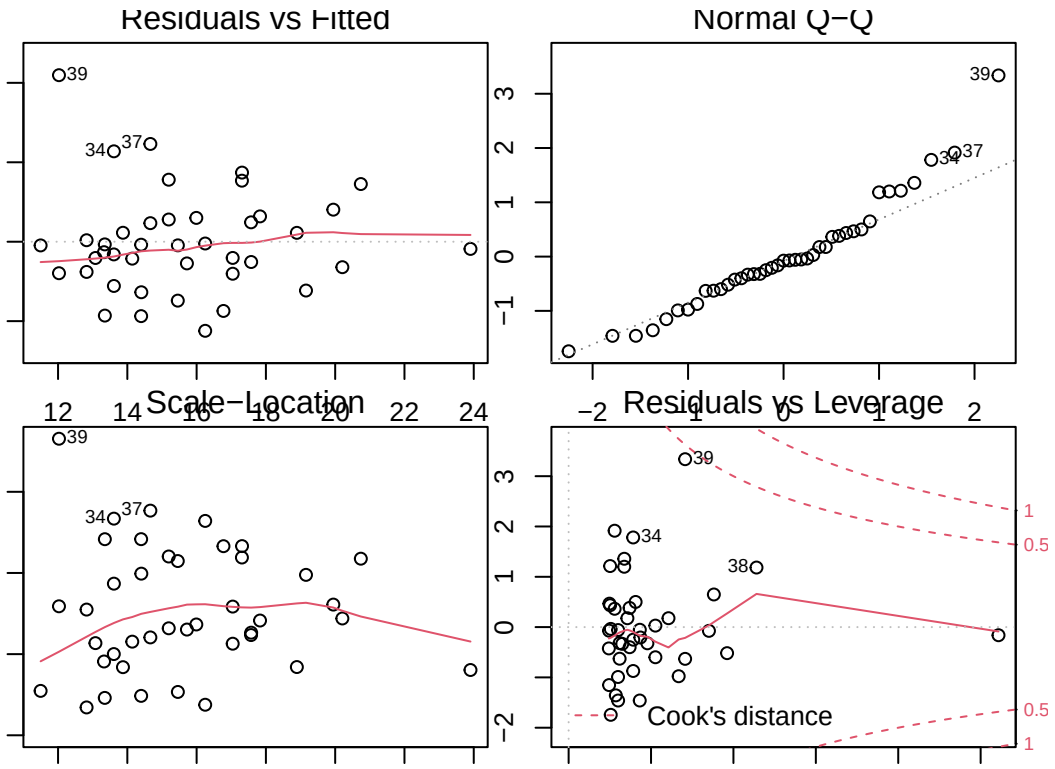
10.26 Zusammenfassung

Table 10.4: Übersicht über die verschiedene Einflussmaße zur Bewertung der Modellgüte

Typ	Veränderung	Daumenregel
$(DFFITS)_i$	Vorhersagewert i	$2\sqrt{k/N}$
Cook	Durchschnittliche Vorhersagewerte	> 1
$(DFBETAS)_{k(i)}$	Koeffizient i	$2\sqrt{N}$
e_{Ti}	Residuum i	t-Verteilung(n-k-2)

10.27 Diagnoseplots in R mit `plot(mod)`

```
plot(mod)
```



10.28 Zum Nacharbeiten

N. Altman and Krzywinski (2016a)
Fox (2011, 294–302)

10.28.1 Weiterführendes

Young (2019)

11 Vorhersage

11.1 Vorhergesagte Werte $\hat{y}_i$

Wenn ein einfaches lineares Modell gefittet wurde ist eine zentrale Frage welche Vorhersagen anhand des Modell getroffen werden können. Die Vorhersagen $\hat{y}_i$ liegen auf der vorhergesagten Regressionsgerade und berechnen sich nach dem Modell für einen gegeben x-Wert.

$$\hat{y} = \hat{\beta_0} + \hat{\beta_0} x$$

Wie schon mehrfach besprochen unterliegt die Regressionsgerade inherent der Unsicherheit bezüglich der geschätzen Modellkoeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$. Diese Unsicherheit überträgt sich auf die geschätzen Werte $\hat{y}_i$ und muss daher bei deren Interpretation berücksichtigt werden.

In Figure 11.1 sind die bereits behandelten Sprungdaten gegen die Anlaufgeschwindigkeiten zusammen mit der Regressionsgeraden und vorhergesagten Werten (rot) abgetragen.

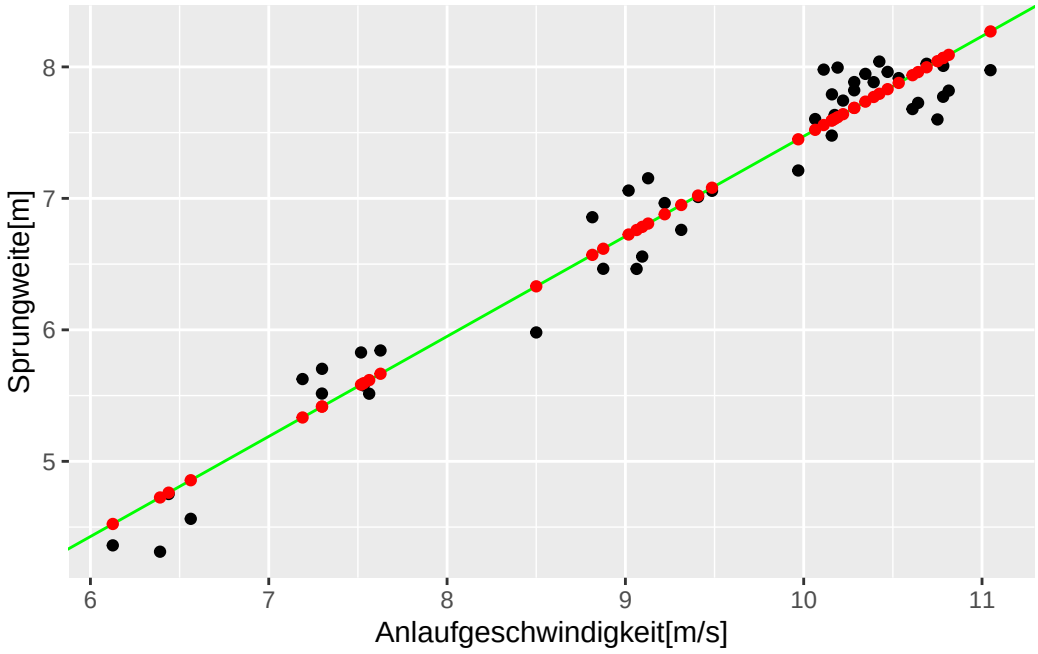


Figure 11.1: Vorhersagewerte $\hat{y}_i$ (rote Punkte) für die Sprungdaten.

In `R` können die vorhergesagten Werte des mittels `lm()` gefitteten Modells mit der Hilfsfunktion `predict()` bestimmt werden. Wenn der Funktion `predict()` keine weiteren Parameter außer dem `lm`-Objekt übergeben werden, berechnet `predict()` die vorhergesagten Werte $\hat{y}_i$ für alle die x-Werte die auch zum fitten des Modells benutzt wurden. Die Reihenfolge der Werte $\hat{y}_i$ enspricht dabei den Werten im Original-`data.frame()`.

```
predict(mod)[1:5]
```

```
       1        2        3        4        5
4.523537 4.725140 4.856256 4.761778 5.416207
```

Wir haben uns hier nur die ersten fünf Werte ausgeben lassen, da nur demonstriert werden soll wie die `predict()`-Funktion angewendet werden kann. Um eine Anwendung zu geben, so können mittels `predict()` die Residuen auch von Hand ohne die `resid()`-Funktion erhalten werden.

```
(jump$jump_m - predict(mod))[1:5]
```

```
          1           2           3           4           5
-0.16267721 -0.41248842 -0.29359256 -0.01047071  0.09927500
```

```
resid(mod)[1:5]
```

```
          1           2           3           4           5
-0.16267721 -0.41248842 -0.29359256 -0.01047071  0.09927500
```

Wiederum nur zur Demonstration die ersten fünf Wert um die Äquivalenz der beiden Methoden zu demonstrieren.

Meistens liegt das Interesse jedoch weniger auf den vorhergesagten Werten $\hat{y}_i$ für die gemessenen Werte, sondern es sollen Werte vorhergesagt werden für x-Werte die nicht im Datensatz enthalten sind. Operational ändert sich nichts, es wird immer noch das gefittete Modell verwendetet und es müssen lediglich neue x-Werte übergeben werden.

In `R` kann dies mittels des zweite Parameter in `predict()` erreicht werden. Soll zum Beispiel die Sprungweite für eine Anlaufgeschwindigkeit von $v = 11.5[m/s]$ berechnen werden, muss zunächst ein neues `tibble()` erstellt werden, welches den gewünschten x-Wert enthält. Dabei muss der Spaltenname in dem neuen `tibble()` demjenigen im Original-`tibble()` entsprechen. Ansonsten funktioniert die Anwendung von `predict()` nicht.

```
df <- tibble(v_ms = 11.5)
df
```

```
# A tibble: 1 x 1
   v_ms
  <dbl>
1  11.5
```

Dieses `tibble()` kann nun zusammen mit dem `lm()`-Objekt an `predict()` übergeben werden.

```
predict(mod, newdata = df)
```

```
       1
8.614136
```

D.h., bei einer Anlaufgeschwindigkeit von $v = 11.5[m/s]$ ist anhand des Modells eine Sprungweite von $8.6m$ zu erwarten.

11.2 Unsicherheit in der Vorhersage

Wie schon angesprochen ist unser Modell natürlich mit Unsicherheiten behaftet. Diese drücken sich in den Standardfehler für die beiden Koeffizienten $\hat{\beta_0}$ und $\hat{\beta_1}$ (siehe Table 11.1).

Table 11.1: Modellparameter und Standardfehler

	Schätzer	s_e
(Intercept)	-0.14	0.23
v_ms	0.76	0.02

Der vorhergesagte Wert $\hat{y}$ ist daher für sich alleine ist noch nicht brauchbar, da auch Informationen über dessen Unsicherheit notwendig sind um die Ergebnisse korrekt zu interpretieren.

Es können zwei unterschiedliche Anwendungsfälle voneinander unterschieden werden.

1. Der mittlere, erwartete Wert $\hat{\bar{y}}_{neu}$
2. Die Vorhersage eines einzelnen Wertes $\bar{y}_{neu}$

Im konkreten Fall werden damit zwei unterschiedliche Fragestellungen beantwortet. Im 1. Fall lautet die Frage, ich habe eine Trainingsgruppe und möchte wissen was der mittlere Wert der Gruppe anhand des Modells ist, wenn alle eine bestimmte Anlaufgeschwindigkeit v_{neu} haben. Im 2. Fall lautet die Frage welche Weite eine einzelne Athletin für die Anlaufgeschwindigkeit v_{neu} springen sollte. In beiden Fällen werden keiner genau den Wert des Regressionsmodells treffen, aber im 1. Fall der Gruppe werden sich Streuungen nach oben bzw. nach unten gegenseitig im Schnitt ausbalancieren während im 2. Fall der einzelnen Athletin dies nicht der Fall ist. Daher hat die Vorhersage im 2. Fall eine höhere Unsicherheit. Diese Unterschied sollte sich dementsprechend in den Varianzen der beiden Vorhersagen wiederspiegeln.

Wie bereits erwähnt, der vorhergesagte Wert $\hat{y}_{neu}$ ist in beiden Fällen gleich und entsprecht der oben beschriebenen Methode anhand des Modell $y_{neu} = \hat{\beta}_0 + \hat{\beta}_1 \times x_{\text{neu}}$.

Für den erwarteten Mittelwert errechnet sich die Varianz nach:

$$Var(\hat{\bar{y}}_{neu}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_{neu} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \hat{\sigma}^2_{\hat{\bar{y}}_{neu}} \tag{11.1}$$

Das dazugehörige Konfidenzintervall errechnet sich danach mittels:

$$\hat{\bar{y}}_{neu} \pm q_{t(1-\alpha/2;n-2)} \times \hat{\sigma}_{\hat{\bar{y}}_{neu}} \tag{11.2}$$

Die Varianz für die Vorhersage eines einzelnen Wertes errechnet sich:

$$Var(\hat{y}_{neu}) = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_{neu} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \hat{\sigma}^2 + \hat{\sigma}^2_{\hat{\bar{y}}_{neu}} = \hat{\sigma}^2_{\hat{y}_{neu}} \tag{11.3}$$

Was wiederum zu dem folgenden Konfidenzintervall führt:

$$\hat{y}_{neu} \pm q_{t(1-\alpha/2;n-2)} \times \hat{\sigma}_{\hat{y}_{neu}} \tag{11.4}$$

In beiden Fällen ist der Term

$$\frac{(x_{neu} - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

enthalten. Anhand des Zählers kann abgeleitet werden, dass die Unsicherheit der Vorhersage mit dem Abstand vom Mittelwert der x-Werte zunimmt. Rein heuristisch macht dies Sinn, da davon ausgegangen werden kann, dass um den Mittelwert der x-Werte auch die meiste Information über y vorhanden ist und dementsprechend umso weiter die Werte sich vom $\bar{x}$ entfernen die Information abnimmt. Im Nenner ist wiederum wie auch beim Standardfehler σ_{β_1} des Steigungskoeffizienten β_1 zu sehen, dass die Varianz abnimmt mit der Streuung der x-Werte. Daher, wenn eine Vorhersage in einem bestimmten Bereich von x-Werten durchgeführt werden soll, dann sollte darauf geachtet werden möglichst diesen Bereich auch zu samplen um die Unsicherheit so klein wie möglich zu halten.

11.3 Vorhersagen in R mit `predict()`

11.3.1 Erwarteter Mittelwert

```
df <- data.frame(v_ms = 11.5) # oder tibble(v_ms = 11.5)
predict(mod, newdata = df, interval = 'confidence')
```

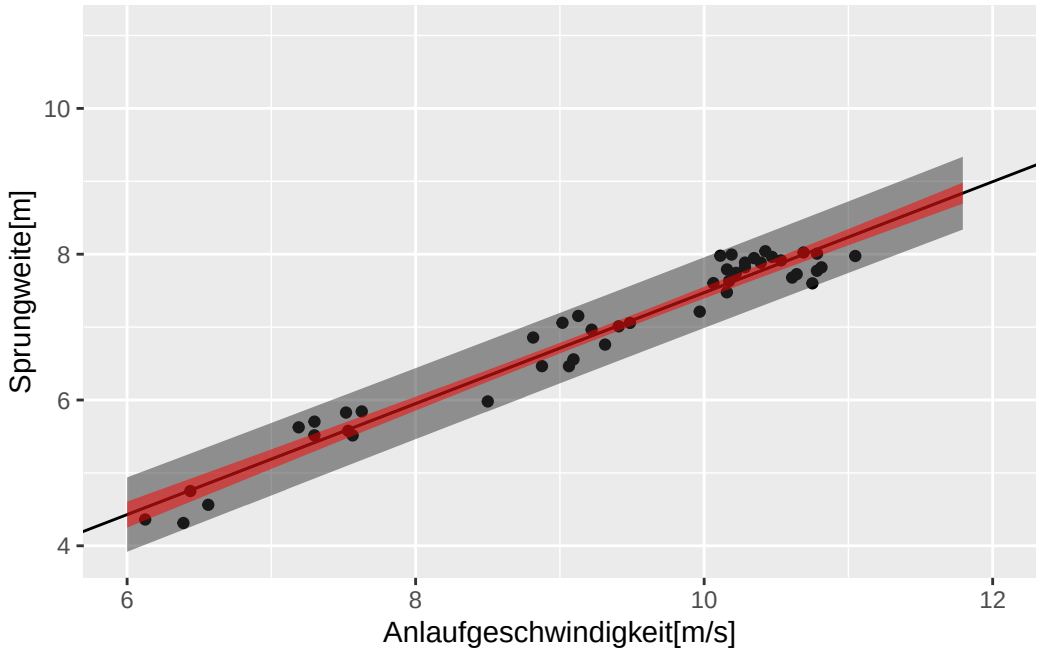
```
       fit      lwr      upr
1 8.614136 8.482039 8.746234
```

11.3.2 Individuelle Werte

```
predict(mod, newdata = df, interval = 'prediction')
```

```
       fit      lwr      upr
1 8.614136 8.118445 9.109827
```

11.4 Konfidenzintervalle graphisch



Weiterführende Literatur sind Kutner et al. (2005)

11.5 R^2 und Root-mean-square

11.6 Einfaches Modell

```
mod0 <- lm(y ~ x, simple)
summary(mod0)
```

```
Call:
lm(formula = y ~ x, data = simple)

Residuals:
      1       2       3       4
-0.5817  0.9898 -0.2345 -0.1736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8414     0.7008   2.628    0.119
x             0.4574     0.3746   1.221    0.346

Residual standard error: 0.8376 on 2 degrees of freedom
Multiple R-squared:  0.4271,    Adjusted R-squared:  0.1406
F-statistic: 1.491 on 1 and 2 DF,  p-value: 0.3465
```

11.7 Nochmal Abweichungen

1. **Gesamtvarianz**:
$$SSTO := \sum_{i=1}^{N} (y_i - \bar{y})^2$$
2. **Regressionsvarianz**:
$$SSR := \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$
3. **Residualvarianz**:
$$SSE := \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

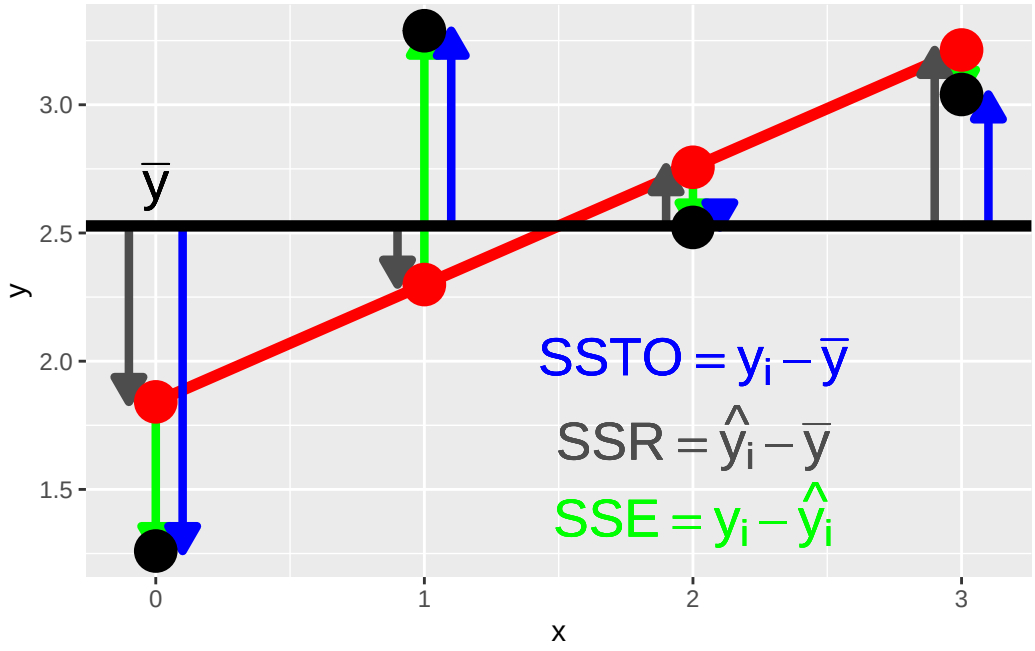


Figure 11.2: Minimalmodell der Abweichungen

11.8 Verhältnis von SSR zu $SSTO$

$$\frac{SSR}{SSTO} = 1$$

$$\frac{SSR}{SSTO} = 0$$

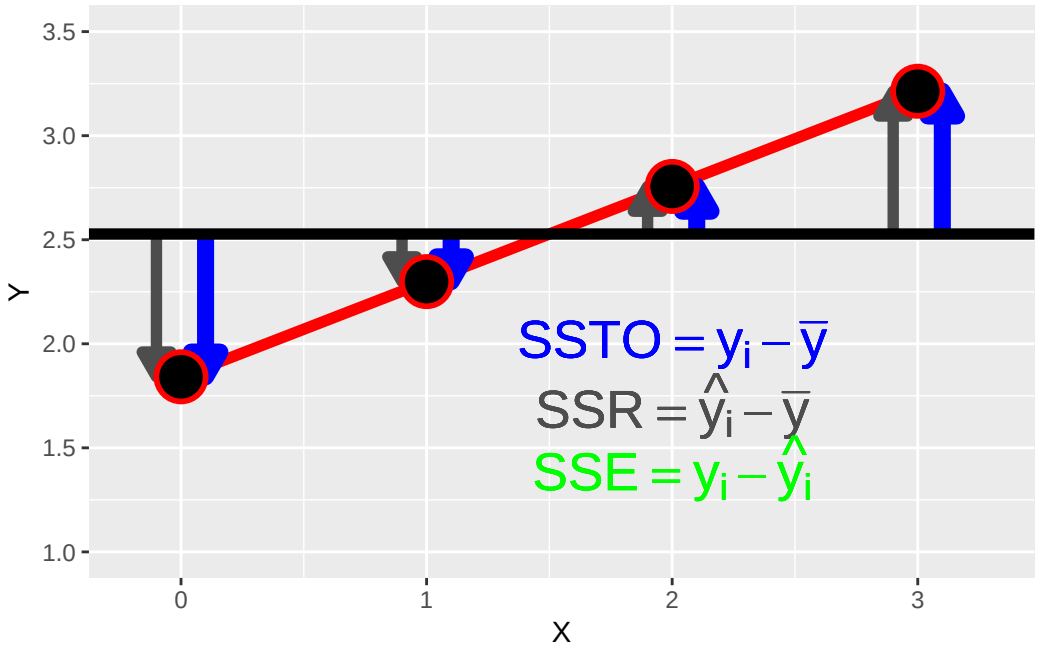


Figure 11.3: Perfekter Zusammenhang

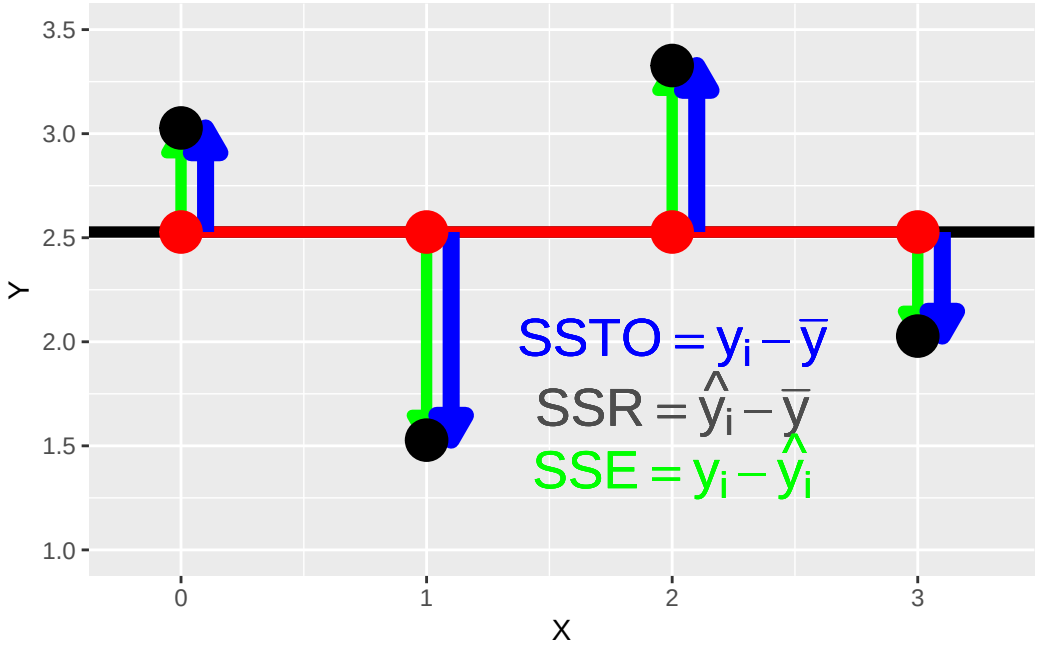


Figure 11.4: Kein Zusammenhang

11.9 Determinationskoeffizient R^2

Es gilt: $SSTO = SSR + SSE$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \in [0, 1]$$

[1]

11.9.1 Korrigierter Determinationskoeffizient R_a^2

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}$$

[1] Bei der einfachen Regression gilt: $r_{xy} = \pm\sqrt{R^2}$

Part III

Multiple Regression

Im folgenden wird das Modell der einfachen linearen Regression erweitert indem zusätzliche Terme in das Modell aufgenommen werden. Die Prinzipien bleiben dabei jedoch weitestgehendst gleich und können direkt auf den komplizierteren Fall der multiplen Regression übertragen werden. Im Laufe der Erweiterung des Modells wird sich dabei wird herausstellen, dass neben mehreren kontinuierlichen Variablen auch nominale Faktoren in das Modell intergriert werden können. Daraus entsteht ein sehr flexibler Modellapparat, der in den verschiedensten Zusammenhängen angewendet werden kann.

12 Einführung

In vielen Fällen in der Praxis liegt selten der einfache Fall vor, dass eine abhängige Variable mitels nur einer einzigen Variable erklärt bzw. vorhergesagt werden soll. Sondern meisten sind mehrere Variablen an dem Prozess der modelliert werden soll beteiligt. Ein einfaches Beispiel aus der Literatur ist der Zusammenhang zwischen der Wurfgeschwindigkeit beim Handball in Abhängigkeit vom Körpergewicht und der Armspannweite. In Table 12.1 ist ein Ausschnitt aus einem möglichen Datensatz abgebildet.

Table 12.1: Datenausschnitt: Wurfgeschwindigkeit, Körpermasse und Armspannweite bei professionellen Handballern (angelehnt an Debanne & Laffaye, 2011).

Velocity[m/s]	body mass[kg]	arm span[cm]
15.8	70.7	189.2
17.2	63.7	182.0
18.3	76.2	192.1
18.4	64.9	171.1
18.4	63.0	181.1

Im Prinzip könnte der isolierte Einfluss der beiden Prädiktorvariablen Körpermasse und Armspannweite auf die Wurfgeschwindigkeit untersucht werden. Allerdings ist den meisten Fällen von größerem Interesse wie sich die beiden Variablen zusammen verhalten und ob durch die Kombination der beiden Variablen ein besseres Modell der Daten erstellt werden kann.

Aus dieser Problemstellung heraus ergibt sich die Notwendigkeit von der einfachen linearen Regression auf eine multiple multiple lineare Regression überzugehen. Formal, geschieht dies einfach dadurch, dass die Formel der einfachen Regression mit dem Prädiktor x um eine zweite Variable erweitert wird.

Dementsprechend wird aus:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{12.1}$$

die Formel für die multiple Regression mit:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i \tag{12.2}$$

Da bei der einfachen Regression nur eine einzige x-Variable in der Formel vorhanden war, ist kein zusätzlicher Index notwendig gewesen, bei der mutliplen Regression mit mehreren Prädiktorvariablen x wird jeder x Variabler ein zusätzlicher Index j angehängt um die Variablen eindeutig zu identifizieren. Per Konvention, wobei diese leider nicht global eingehalten wird, wird die Anzahl der Prädiktorvaiablen mit K bezeichnet. Der y-Achsenabschnitt erhält den Index $j = 0$ und die weiteren Steigungskoeffzienten β_1 bis β_K erhalten den Prädiktorvariablen x_j entsprechden Index.

In welcher Reihenfolge die Prädiktorvariablen mit $j = 1, j = 2, \ldots, j = K$ verteilt werden hat zunächst keine Auswirkung auf das Modell und regelt lediglich die Bezeichnung. In unserem konkreten Fall der Handballwurfdaten wäre zum Beispiel eine mögliche Zuordnung, das x_1 die Körpermasse und x_2 die Armspannweite kodiert.

i	Velocity[m/s]	body mass[kg] $j = 1$	arm span[cm] $j = 2$
1	15.8	70.7	189.2
2	17.2	63.7	182.0

i	Velocity[m/s]	body mass[kg] $j = 1$	arm span[cm] $j = 2$
3	18.3	76.2	192.1
4	18.4	64.9	171.1
5	18.4	63.0	181.1

Rein formal haben wir jetzt schon den Übergang zur multiple Regression vollzogen. Die Frage die sich natürlich direkt anschließt bezieht sich nun auf die Bedeutung der Koeffizienten $\beta_1, \dots, \beta_k$.

12.1 Bedeutung der Koeffizienten bei der multiplen Regression

Um die Bedeutung der Regressionskoeffzienten bei der multiple Regression besser zu verstehen ist es von Vorteil sich noch einmal die Bedeutung der Koeffizienten im einfachen Regressionsmodell zu vergegenwärtigen (siehe Figure 12.1).

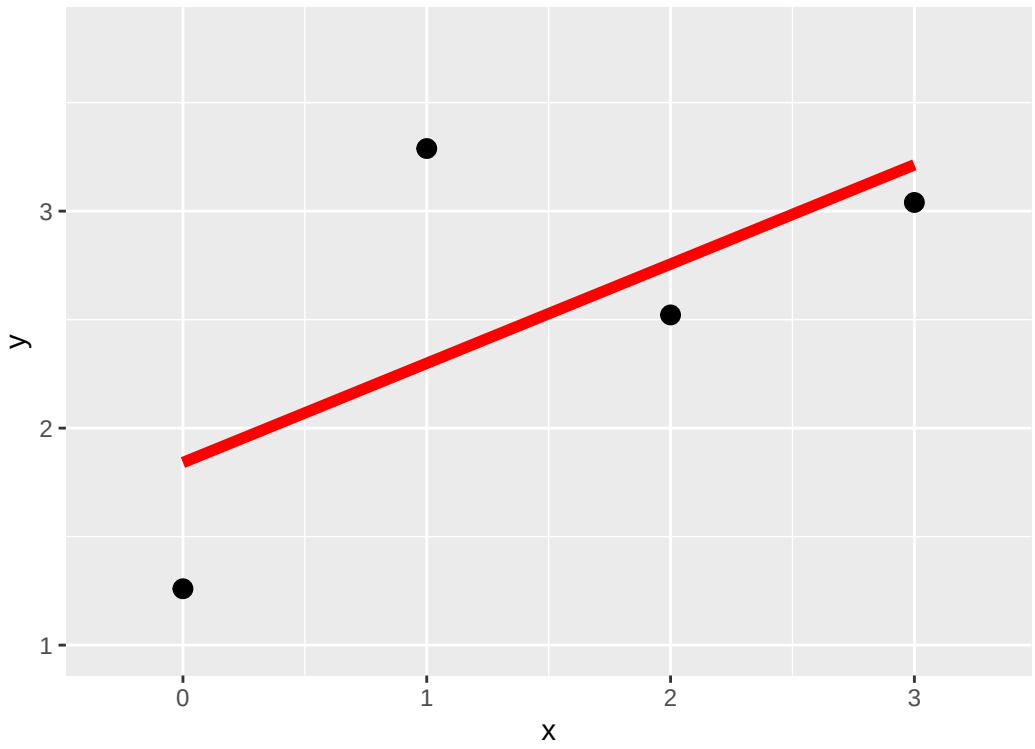


Figure 12.1: Beispiel für eine einfache Regression und der resultierenden Regressiongeraden

Bei der einfachen Regression haben mittels der Methode der kleinsten Quadrate eine Regressiongerade durch unsere Punktwolke gelegt. Dabei haben wir die Regressionsgerade so gewählt, dass die senkrechten Abstände der beobachteten Punkte von der Regressionsgerade minimiert werden bzw. die Abstände zwischen denen auf der Gerade liegenden, vorhergesagten Werte $\hat{y}_i$ und den beobachteten Wert y_i.

Wenn wir nun den Übergang von einer Prädiktorvariablenzum nächstkomplizierteren Fall nehmen mit zwei Prädiktorvariablen x_1 und x_2, dann wäre eine mögliche Darstellungsform der Daten eine Punktwolke im dreidimensionalen Raum (siehe Figure 12.2a).

(a) 3D Punktwolke

(b) 3D Punktwolke mit gefitteter Ebene

Figure 12.2: Punktwolken bei der multiple Regression

Da jetzt eine einzelne Gerade nicht mehr in der Lage ist die Daten zu fitten, ist die nächst Möglichkeit eine Ebene die in die Punktwolke gelegt wird (siehe Figure 12.2b). Dies ermöglicht dann genau die gleiche Herangehensweise wie bei der einfachen linearen Regression anzuwenden. Als Zielgröße wird aus den möglichen Ebenen diejenigen gesucht deren vorhergesagten, auf der Ebene liegenden Punkte $\hat{y}_i$ die geringsten senkrechten Abstand zu den beobachteten Punkten y_i haben. Anders, wir suchen diejenigen Ebene durch die Punktwolke deren Summe der quadrierten Residuen $e_i = y_i - \hat{y}_i$ minimal ist.

Diese Herangehensweise hat den Vorteil, dass sie zum einem die einfache lineare Regression als Spezialfall mit $K = 1$ beinhaltet und sich beliebig erweitern lässt mit der Einschränkung, dass bei $K > 2$ die dreidimenionale Darstellung mittels einer Grafik nicht mehr möglich ist. Das Prinzip der Minimierung der Abweichungen von $\hat{y}_i$ zu y bleibt aber immer erhalten. Zusammenfassend hat dieser Ansatz somit die folgenden Vorteile:

- Die Berechnungen bleiben alle gleich
- Abweichungen $\hat{\epsilon_i}$ sind jetzt nicht mehr Abweichungen von einer Gerade sondern von einer K-dimensionalen Hyperebene. Die Eigenschaften der Residuen bleiben aber alle erhalten.
- Die Modellannahmen bleiben gleich: Unabhängige y_i und $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid
- Inferenz für die Koeffizienten mittels $t_k = \frac{\hat{\beta}_k}{s_k} \sim t(N - K - 1)$ (Konfidenzintervall dito)
- Konzepte für die Vorhersage bleiben erhalten
- Modelldiagnosetools bleiben alle erhalten

Als nächster Schritt versuchen wir nun die Interpretation der Koeffizienten im multiplen Regressionsmodell besser zu verstehen.

12.2 Einfaches Beispiel

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon_i \\
\beta_0 &= 1, \beta_1 = 3, \beta_2 = 0.7 \\
\epsilon_i &\sim N(0, \sigma = 0.5)
\end{aligned}
$$

```
N <- 50 # Anzahl Datenpunkte
beta_0 <- 1
```

```r
beta_1 <- 3
beta_2 <- 0.7
sigma <- 0.5
set.seed(123)
df <- tibble(
  x1 = runif(N, -2, 2),
  x2 = runif(N, -2, 2),
  y = beta_0 + beta_1*x1 + beta_2*x2 +
    rnorm(N, 0, sigma))
```

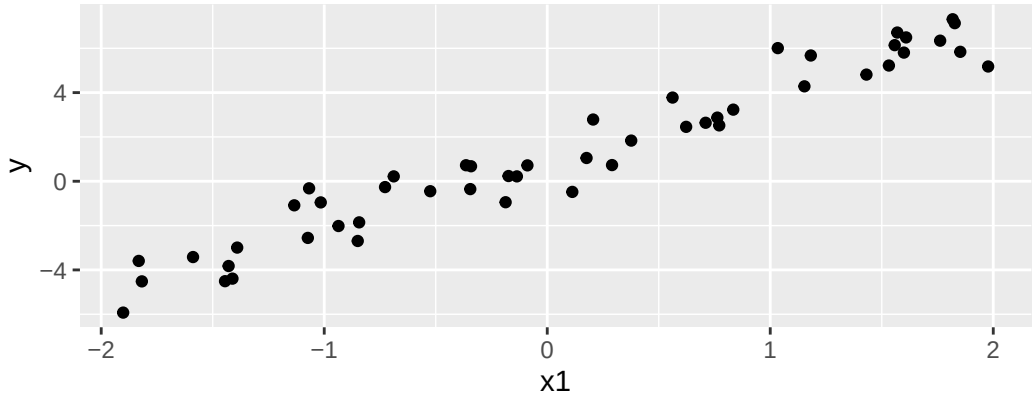


Figure 12.3: Einfacher Zusammenhang y~x1

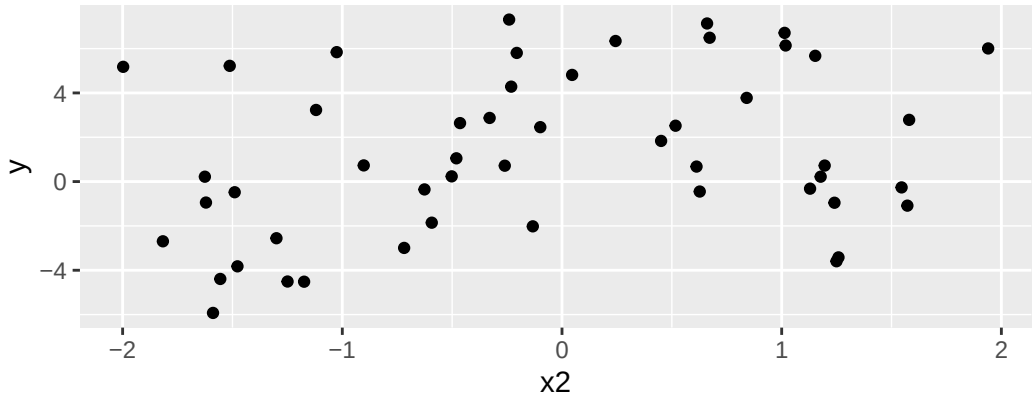


Figure 12.4: Einfacher Zusammenhang y~x2

12.3 Wie sieht der Fit aus?

```
Call:
lm(formula = y ~ x1 + x2, data = df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.20883 -0.26741 -0.00591  0.27315  1.01322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.07674    0.06552   16.43  < 2e-16 ***
x1           2.96537    0.05604   52.91  < 2e-16 ***
x2           0.70815    0.05961   11.88 9.27e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4604 on 47 degrees of freedom
Multiple R-squared:  0.9849,    Adjusted R-squared:  0.9842
F-statistic:  1529 on 2 and 47 DF,  p-value: < 2.2e-16
```

12.4 Was bedeuten die einzelnen Koeffizienten?

Table 12.3: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	1.077	0.066
x1	2.965	0.056
x2	0.708	0.060

Der Unterschied in der abhängigen Variablen, wenn zwei Objekte sich in x_i um eine Einheit unterscheiden und die paarweise gleichen Werte in den verbleibenden $x_j, j \neq i$ annehmen.

12.5 Was bedeuten die Koeffizienten in Kombination?

12.5.1 Full model

Table 12.4: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	1.077	0.066
x1	2.965	0.056
x2	0.708	0.060

12.5.2 um x2 bereinigt

```r
mod_x1_x2 <- lm(x1 ~ x2, df)
res_mod_x1_x2 <- resid(mod_x1_x2)
mod_x1_res <- lm(y ~ res_mod_x1_x2, df)
```

```
             Estimate Std. Error t value
(Intercept)      1.25       0.16    7.61
res_mod_x1_x2    2.97       0.14   20.97
```

12.5.3 um x1 bereinigt

```
mod_x2_x1 <- lm(x2 ~ x1, df)
res_mod_x2_x1 <- resid(mod_x2_x1)
mod_x2_res <- lm(y ~ res_mod_x2_x1, df)
```

```
              Estimate Std. Error t value
(Intercept)       1.25       0.51    2.44
res_mod_x2_x1     0.71       0.47    1.51
```

12.6 Was bedeuten die Koeffizienten in Kombination?

- $\hat{\beta}_1$: Wenn ich x_2 weiß, welche zusätzlichen Informationen bekomme ich durch x_1
- $\hat{\beta}_2$: Wenn ich x_1 weiß, welche zusätzlichen Informationen bekomme ich durch x_2

In Beispiel nicht problematisch, weil nach Konstruktion x_1 und x_2 unabhängig voneinander sind:

```
round(cor(df),3)
```

```
      x1    x2     y
x1 1.000 0.078 0.969
x2 0.078 1.000 0.289
y  0.969 0.289 1.000
```

12.7 Added-variable plots

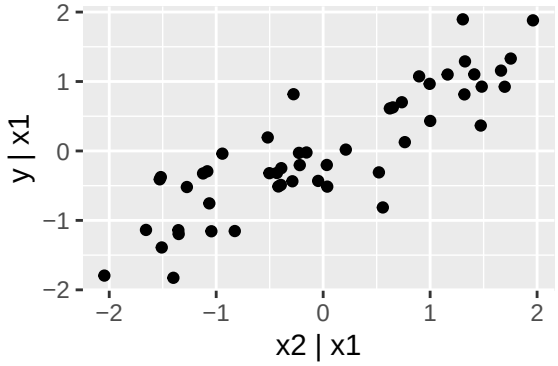
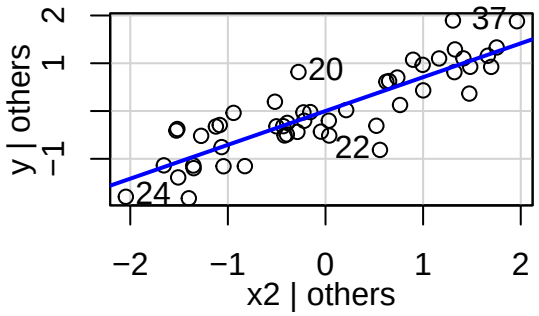


Figure 12.5: Zusammenhang zwischen y und x2 bereinigt um den Einfluß von x1.

12.8 Added-variable plots mit `car::avPlots()`

```
car::avPlots(mod, ~x2)
```



12.9 Was passiert wenn ich einen Prädiktor weg lasse?

Table 12.5: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	1.077	0.066
x1	2.965	0.056
x2	0.708	0.060

```
coef(lm(y ~ x1, df))
```

```
(Intercept)          x1
   1.007466    3.017589
```

```
coef(lm(y ~ x2, df))
```

```
(Intercept)          x2
  1.3377771   0.9555316
```

In unserem Beispiel wieder nicht viel, da die Variablen unabhängig (orthogonal) voneinander sind.

12.10 Was passiert wenn Prädiktoren stark miteinander korrelieren?

Table 12.6: Ausschnitt von Körperfettdaten

triceps	thigh	midarm	body_fat
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7

[1]

12.11 Was passiert wenn Prädiktoren stark miteinander korrelieren?

```
GGally::ggpairs(bodyfat) + theme(text = element_text(size = 10))
```

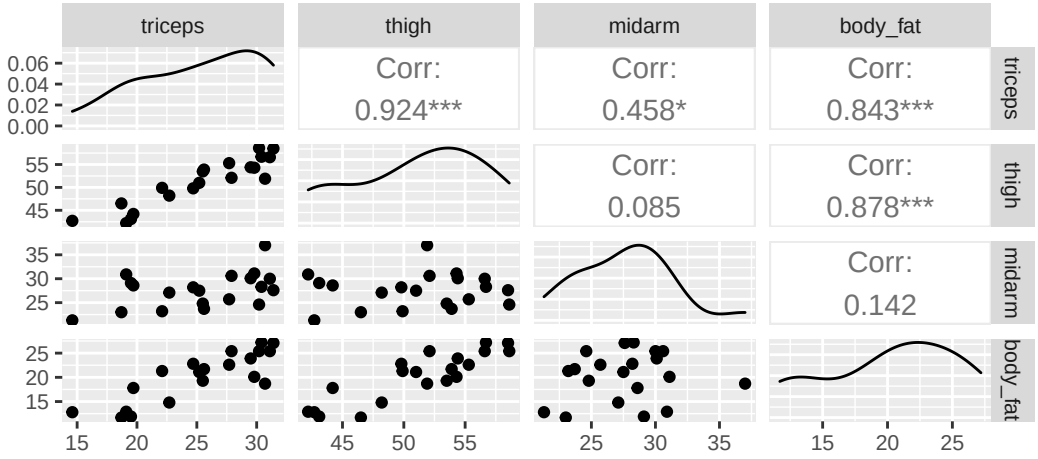


Figure 12.6: Korrelationsmatrize

12.12 Was passiert wenn Prädiktoren stark miteinander korrelieren?

```
# Alle drei Prädiktoren
mod_full <- lm(body_fat ~ triceps + thigh + midarm, bodyfat)
# ohne Arm
mod_wo_midarm <- lm(body_fat ~ triceps + thigh, bodyfat)
```

[1] Beispiel nach Kutner et al. (2005)

```r
# Ohne Oberschenkel
mod_wo_thigh <- lm(body_fat ~ triceps + midarm, bodyfat)
# Ohne Triceps
mod_wo_triceps <- lm(body_fat ~ thigh + midarm, bodyfat)
```

12.13 Was passiert wenn Prädiktoren stark miteinander korrelieren?

Table 12.7: full model

	$\hat{\beta}$	s_e
(Intercept)	117.085	99.782
triceps	4.334	3.016
thigh	-2.857	2.582
midarm	-2.186	1.595

Table 12.8: w/o midarm

	$\hat{\beta}$	s_e
(Intercept)	-19.174	8.361
triceps	0.222	0.303
thigh	0.659	0.291

Table 12.9: w/o thigh

	$\hat{\beta}$	s_e
(Intercept)	6.792	4.488
triceps	1.001	0.128
midarm	-0.431	0.177

Table 12.10: w/o triceps

	$\hat{\beta}$	s_e
(Intercept)	-25.997	6.997
thigh	0.851	0.112
midarm	0.096	0.161

12.14 Multikollinearität[2]

- Große Änderungen in den Koeffizienten wenn Prädiktoren ausgelassen/eingefügt werden
- Koeffizienten haben eine andere Richtung als erwartet
- Hohe (einfache) Korrelationen zwischen Prädiktoren

[2] informell nach Kutner et al. (2005, 407)

- Breite Konfidenzintervalle für "wichtige" Prädiktoren b_j

$$\widehat{\text{Var}}(b_j) = \frac{\hat{\sigma}^2}{(n-1)s_j^2} \frac{1}{1-R_j^2}$$

R_j^2 = Multipler Korrelationskoeffizient der Prädiktoren auf Prädiktorvariable j.

12.15 Variance Inflation Factor (VIF)

$$\text{VIF}_j = \frac{1}{1-R_j^2}$$

> **Tip**
>
> Wenn VIF > 10 ist, dann deutet dies auf hohe Multikollinearität hin.

[3]

12.16 Variance Inflation Factor (VIF)

```
car::vif(mod_full)
```

```
 triceps    thigh   midarm
708.8429 564.3434 104.6060
```

[4]

Üblicherweise wird der größte Wert betrachtet um die Multikollinearität zu bewerten.

12.17 Wenn Prädiktoren sich gegenseitig maskieren[5]

12.18 Wenn Prädiktoren sich gegenseitig maskieren

[3] Manchmal wird auch Tolerance = $\frac{1}{VIF}$ betrachtet.
[4] car::vif berechnet generalized variance inflation factor wenn Prädiktoren Faktoren oder Polynome sind (Fox 2011.)
[5] adaptiert nach McElreath (2016)

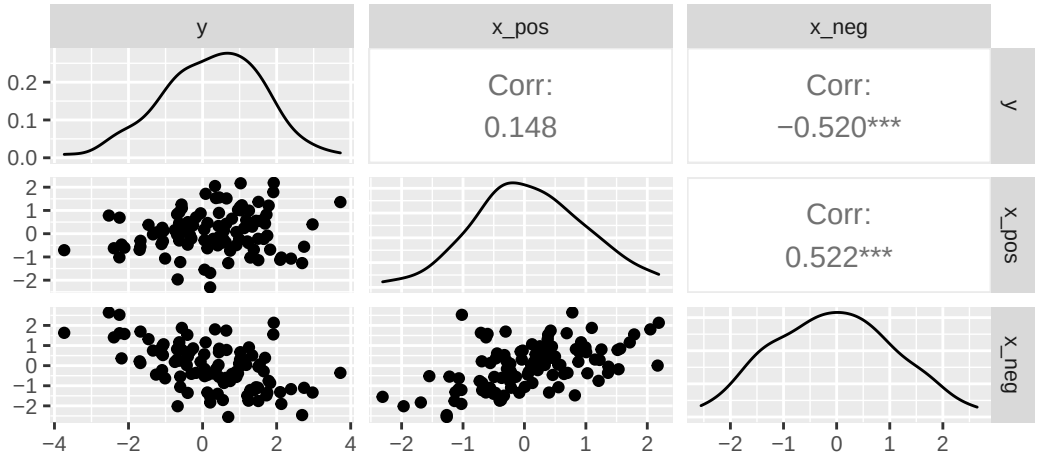


Figure 12.7: x_pos maskiert den Einfluss von x_neg

Table 12.11: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	0.235	0.135
x_pos	0.218	0.147

Table 12.12: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	0.228	0.116
x_neg	-0.618	0.103

Table 12.13: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	0.135	0.096
x_pos	0.850	0.123
x_neg	-0.976	0.099

12.19 Multiple Regression

Aus der einfachen Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

wird

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i$$

mit K Prädiktorvariablen und Multikollinearität.

12.20 Zum Nacharbeiten

N. Altman and Krzywinski (2015a)
Kutner et al. (2005, 278–88)
Fox (2011, 325–27)

13 Interaktionseffekte

13.1 Beispieldaten[1]

Table 13.1: Beispieldaten (synthetisch)

Velocity[m/s]	body mass[kg]	arm span[cm]
185.42	68.71	20.14
184.08	73.85	21.29
200.74	89.43	27.57
170.34	84.97	19.88
176.89	82.40	20.51
200.68	91.57	29.22

13.2 Beispieldaten - Deskriptiv

Table 13.2: Deskriptive Statistik der Handballdaten

	Mean	Std.Dev	Min	Max
arm_span	184.3	7.7	169.4	200.7
body_mass	77.5	10.3	58.0	101.1
vel	21.9	2.3	18.5	29.2

13.3 Beispieldaten

13.4 Beispieldaten - Startmodell

$$Y_i = \beta_0 + \beta_1 \times \text{bm}_i + \beta_2 \times \text{as}_i + \epsilon_i$$

```
mod_1 <- lm(vel ~ body_mass + arm_span, handball)
```

[1] Debanne and Laffaye (2011)

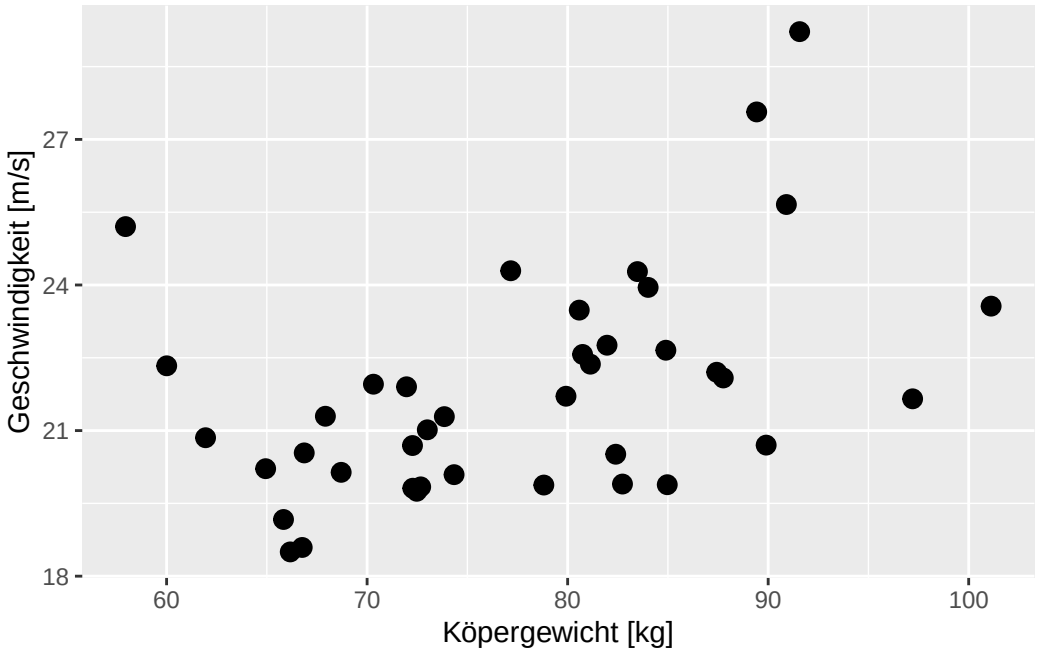


Figure 13.1: Geschwindigkeit gegen Körpergewicht

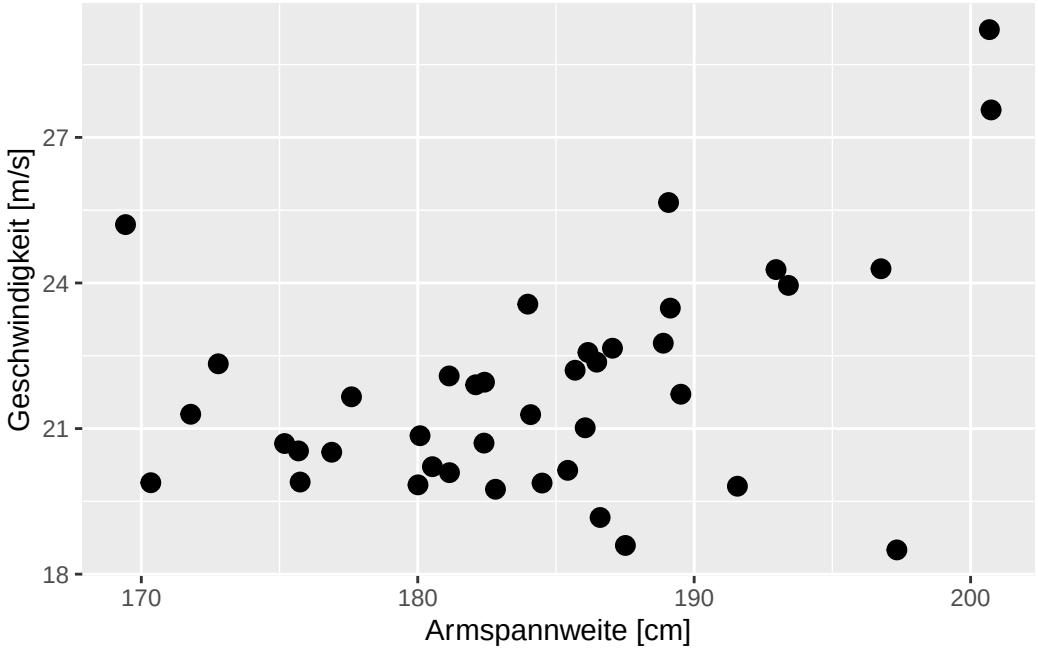


Figure 13.2: Geschwindigkeit gegen Armspannweite

Table 13.3: Modell 1

	$\hat{\beta}$	s_e	t	p
(Intercept)	-1.768	7.632	-0.232	0.818
body_mass	0.077	0.033	2.359	0.024
arm_span	0.096	0.044	2.192	0.035
$\hat{\sigma}$	1.996			

13.5 Modellfit

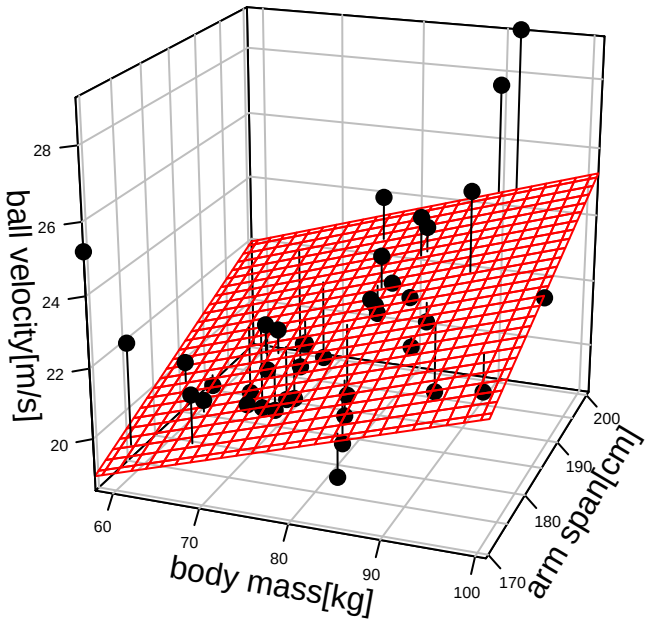


Figure 13.3: 3D Streudiagramm

13.6 Zentrierung

```
handball <- dplyr::mutate(handball,
                          body_mass_c = body_mass - mean(body_mass),
                          arm_span_c = arm_span - mean(arm_span))
```

Table 13.4: Deskriptive Statistik

	Mean	Std.Dev
arm_span	184.29	7.72

	Mean	Std.Dev
arm_span_c	0.00	7.72
body_mass	77.46	10.26
body_mass_c	0.00	10.26
vel	21.85	2.31

13.7 Modell mit zentrierten Variablen

```
mod_2 <- lm(vel ~ body_mass_c + arm_span_c, handball)
```

Table 13.5: Modell 2

	$\hat{\beta}$	s_e	t	p
(Intercept)	21.852	0.316	69.247	<0.001
body_mass_c	0.077	0.033	2.359	0.024
arm_span_c	0.096	0.044	2.192	0.035
$\hat{\sigma}$	1.996			

13.8 Residuen im zentrierten, additiven Modell

13.9 Added-variable plot

13.10 Was passiert wenn die Effekte nicht mehr nur additiv sind?

13.11 Was passiert wenn die Effekte nicht mehr nur additiv sind?

13.11.1 Neues Modell mit Interaktionen:

$$Y_i = \beta_0 + \beta_1 \times \text{bm}_i + \beta_2 \times \text{as}_i + \beta_3 \times \text{bm}_i \times \text{as}_i + \epsilon_i$$

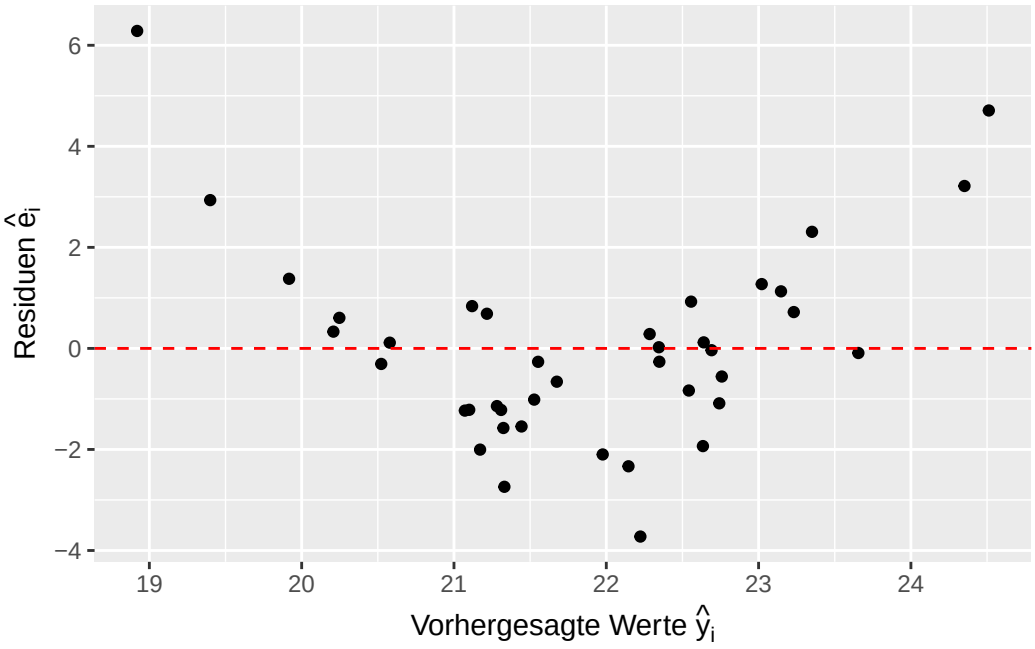


Figure 13.4: Residuenplot

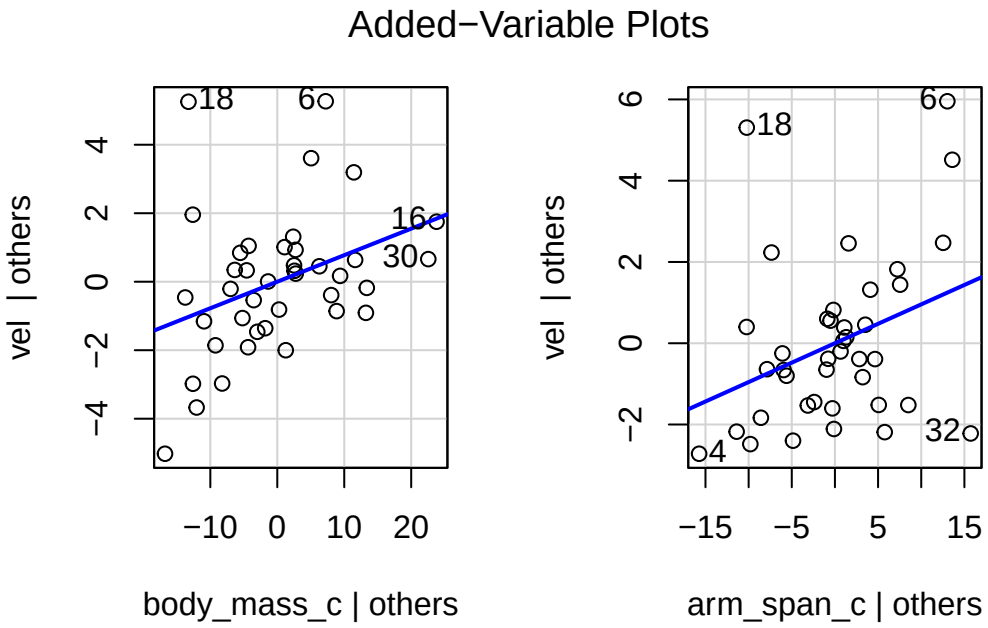


Figure 13.5: Added-variable Graph mit `car::avPlots()`

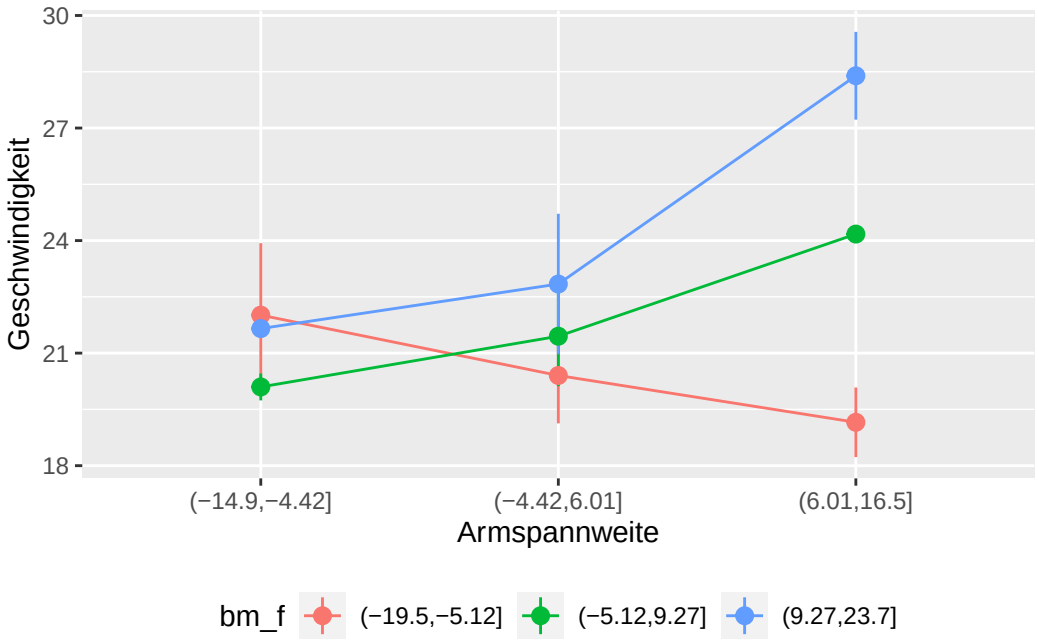


Figure 13.6: Unterteilung von Körpergewicht und Armspannweite in Kategorien

13.12 Modellierung

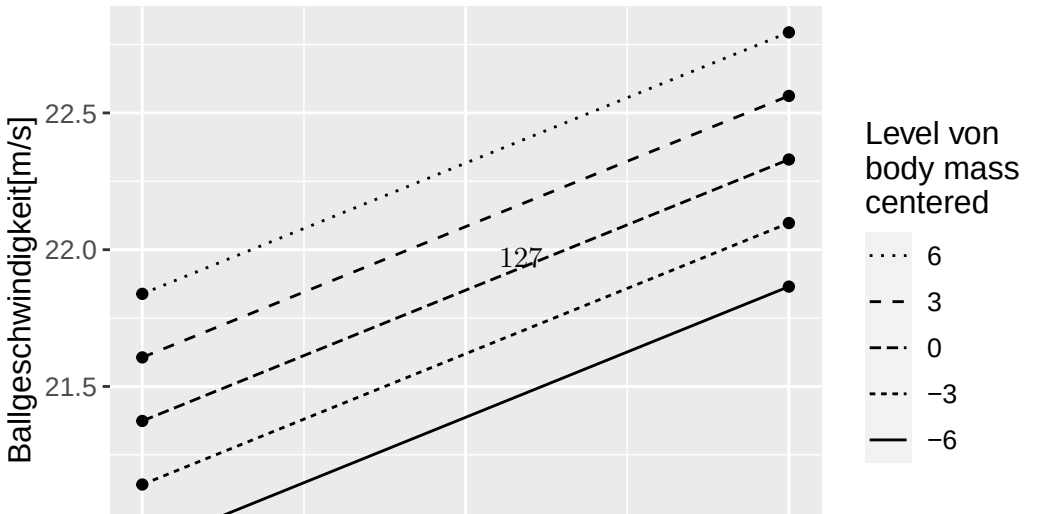
```
mod_3 <- lm(vel ~ body_mass_c * arm_span_c, handball)
```

Table 13.6: Modell 3

	$\hat{\beta}$	s_e	t	p
(Intercept)	21.346	0.143	149.296	<0.001
body_mass_c	0.119	0.015	8.133	<0.001
arm_span_c	0.083	0.019	4.380	<0.001
body_mass_c:arm_span_c	0.021	0.002	12.633	<0.001
$\hat{\sigma}$	0.868			

[2]

13.13 Einfache Steigungen in Vergleich



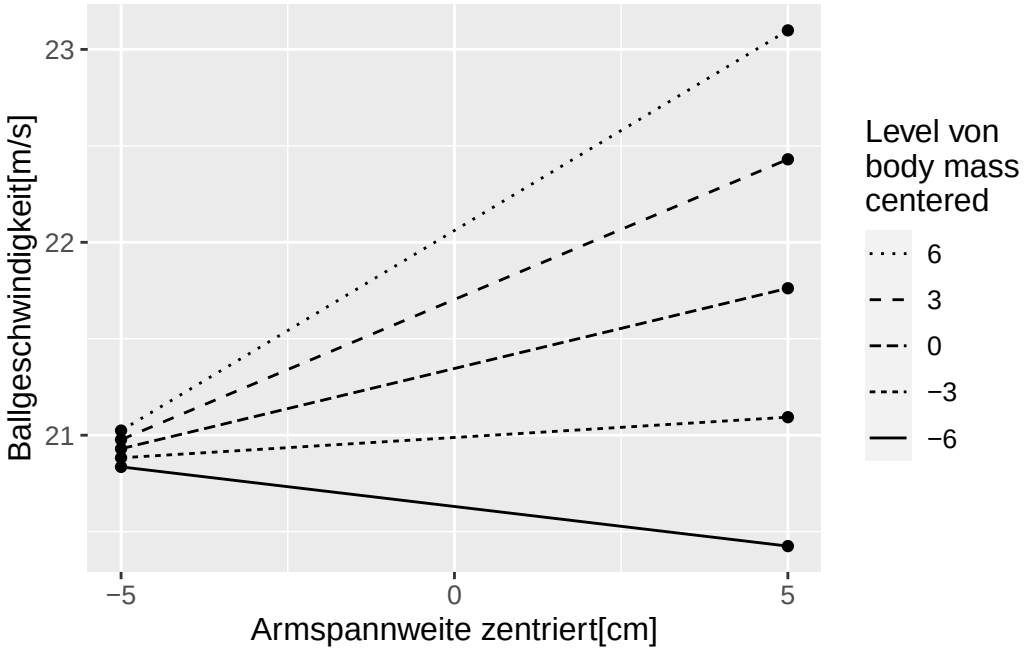


Figure 13.8: Modell mit Interaktionen

13.14 Interaktionen sind symmetrisch

13.15 Warum das Model Sinn macht

Table 13.7: Einfache Steigungen

arm span\centered	β_0	β_1
10	22.18	0.33
0	21.35	0.12
-10	20.51	-0.09

13.16 Warum das Modell Sinn macht

Table 13.8: Einfache Steigungen

arm span\centered	β_0	β_1
10	22.18	0.33
0	21.35	0.12
-10	20.51	-0.09

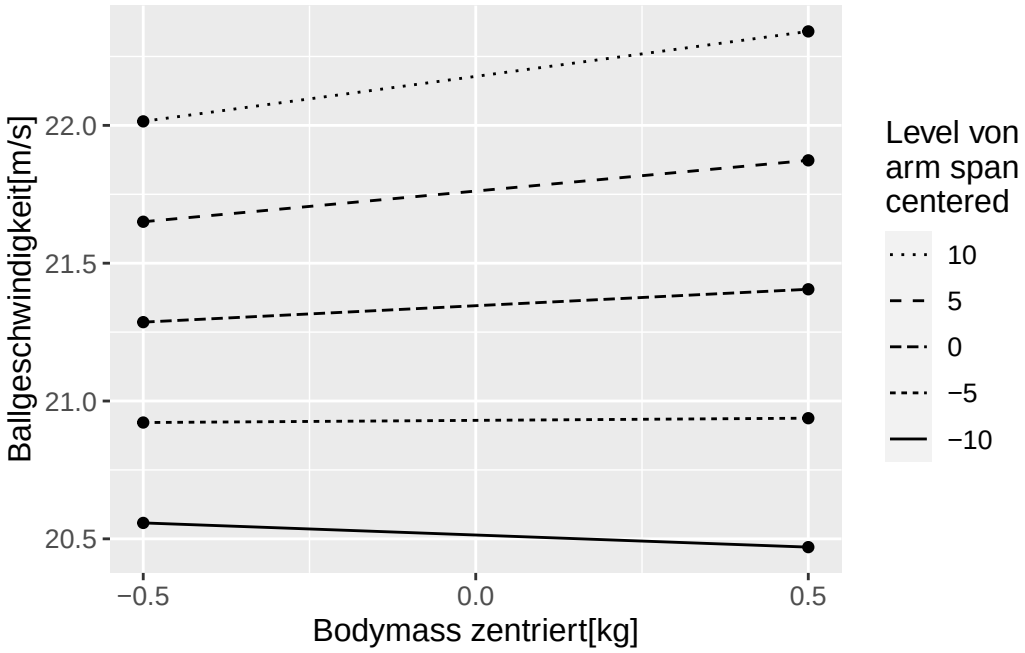


Figure 13.9: Veränderung mit der Körpergewicht

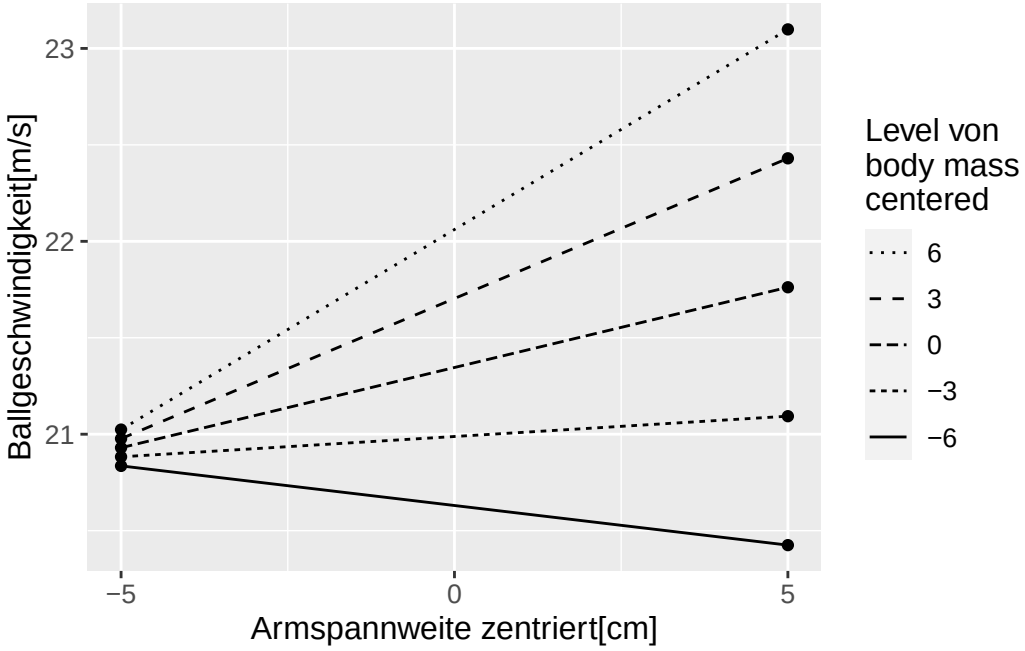


Figure 13.10: Veränderung mit dem Armspannweite

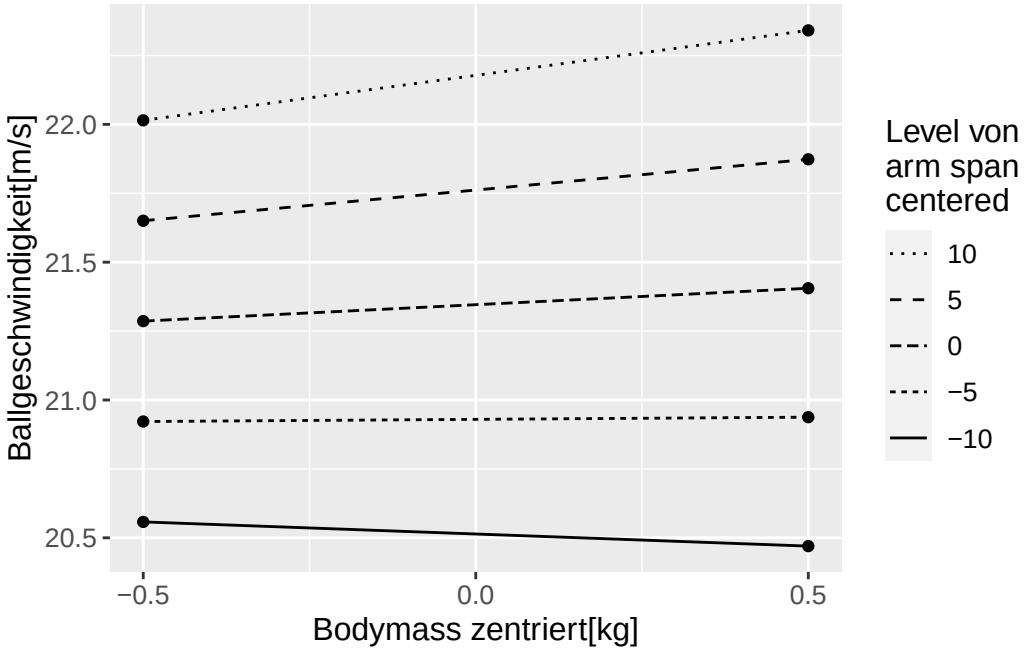


Figure 13.11: Veränderung mit dem Körpergewicht

Table 13.9: Modellkoeffizienten

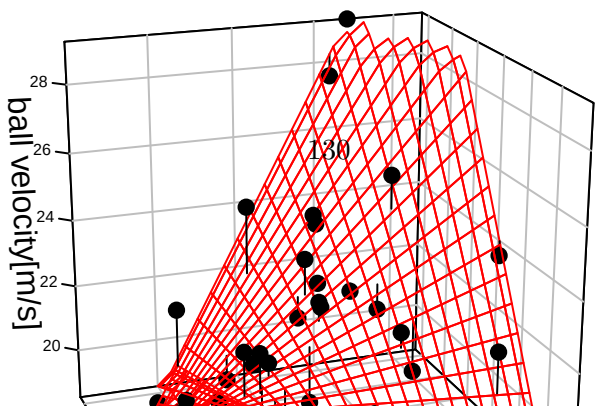
	betas
b0	21.35
bm_c	0.12
as_c	0.08
bm_c:as_c	0.02

13.17 Interpretation der Koeffizienten

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_1 \cdot x_2 + \epsilon_i$$

- b_0: (y-Achsenabschnitt) der Wert von $\hat{Y}$ wenn $x_1 = 0$ und $x_2 = 0$ gilt.
- b_1: Der Unterschied in $\hat{Y}$ wenn zwei Objekte sich in x_1 um eine Einheit unterscheiden und $x_2 = 0$ ist.
- b_2: Der Unterschied in $\hat{Y}$ wenn zwei Objekte sich in x_2 um eine Einheit unterscheiden und $x_1 = 0$ ist.
- b_3: (Interaktionskoeffizient) Die Veränderung des Effekts von x_1 auf $\hat{Y}$ wenn x_2 um eine Einheit größer wird bzw. genau andersherum für x_2.

13.18 Aus der Ebene wird eine gekrümmte Fläche



13.19 Residuenvergleich

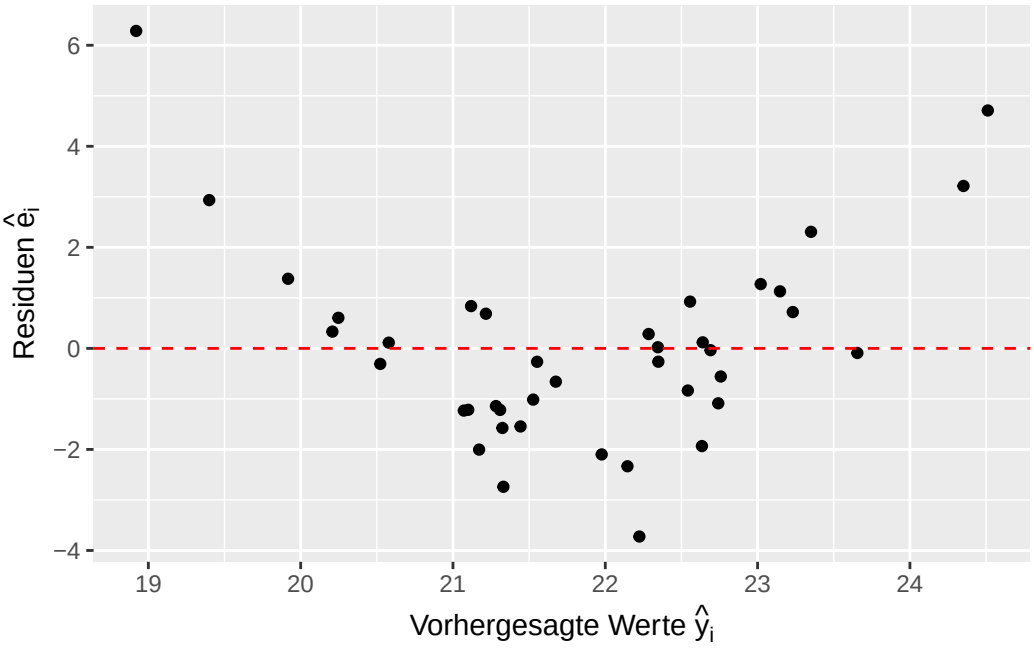


Figure 13.13: Residuen im additiven Modell

13.20 Residuenvergleich - qq-Plot

13.21 Take-away

Interaktionsmodell

- Erhöht die Flexibilität des linearen Modells.
- Bei Interaktionen hängt der Einfluss der einzelnen Variablen immer von den Werten der anderen Variablen ab.
- Achtung: Interpretation der einfachen Haupteffekte nicht mehr möglich bzw. sinnvoll!

13.22 Zuschlag

Was passiert im Interaktionsmodell mit den Koeffizienten wenn die x_{ki}s zentriert werden?

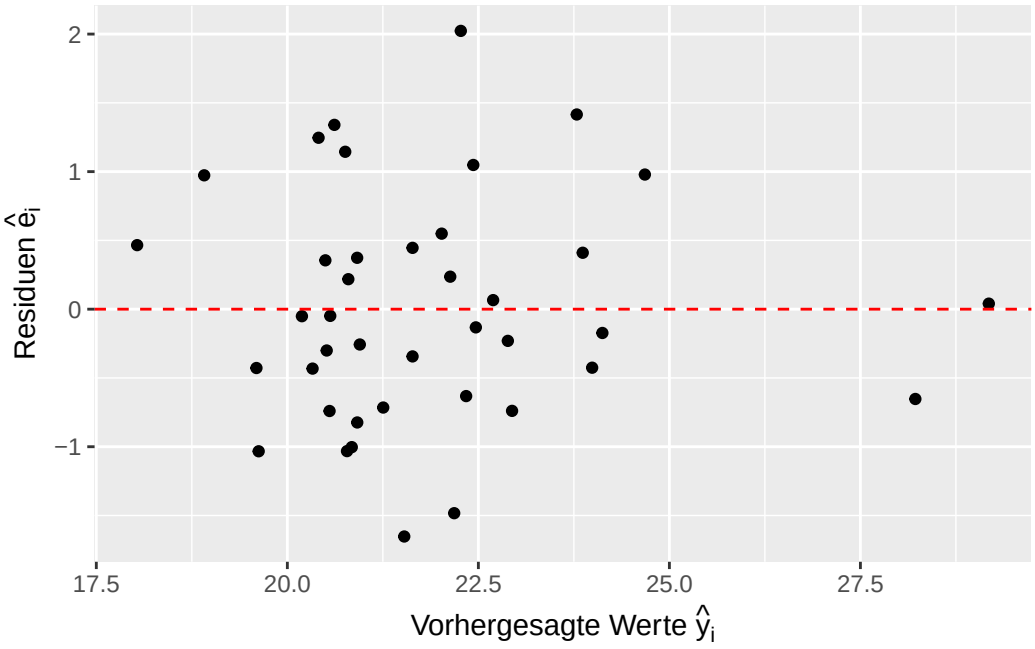


Figure 13.14: Residuen im Interaktionsmodell

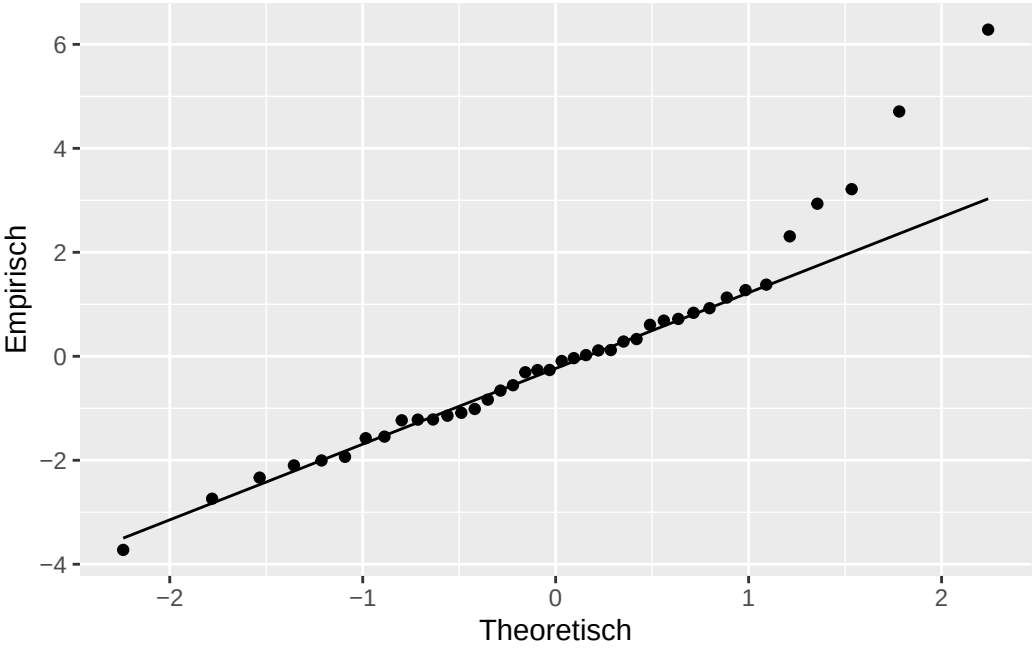


Figure 13.15: additives Modell

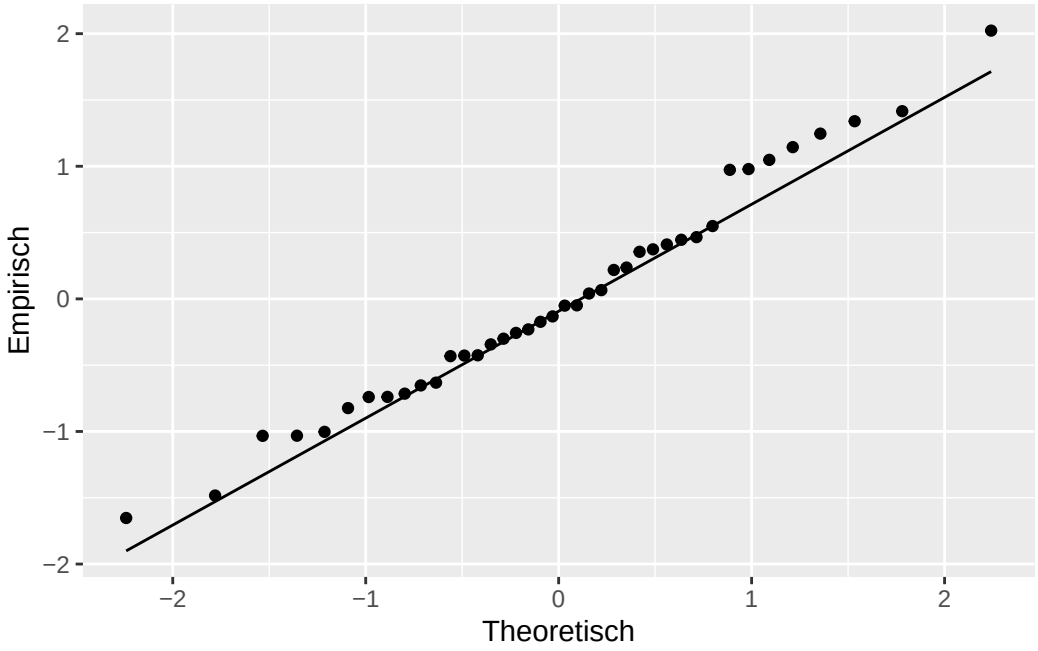


Figure 13.16: Interaktionsmodell

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \\
&= \beta_0 + \beta_1 x_{1i} - \beta_1\bar{x}_1 + \beta_2 x_{2i} - \beta_2\bar{x}_2 + \beta_3 x_{1i}x_{2i} - \beta_3 x_{1i}\bar{x}_2 - \beta_3\bar{x}_1 x_{2i} + \beta_3\bar{x}_1\bar{x}_2 \\
&= \beta_0 - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 + \beta_3\bar{x}_1\bar{x}_2 + \beta_1 x_{1i} - \beta_3\bar{x}_2 x_{1i} + \beta_2 x_{2i} - \beta_3\bar{x}_1 x_{2i} + \beta_3 x_{1i}x_{2i} \\
&= \underbrace{\beta_0 - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 + \beta_3\bar{x}_1\bar{x}_2}_{\beta_0} + \underbrace{(\beta_1 - \beta_3\bar{x}_2)x_{1i}}_{\beta_1 x_{1i}} + \underbrace{(\beta_2 - \beta_3\bar{x}_1)x_{2i}}_{\beta_2 x_{2i}} + \beta_3 x_{1i}x_{2i}
\end{aligned}
$$

13.23 Zum Nacharbeiten

Kutner et al. (2005, 306–13)

14 Integration von nominale Variablen

14.1 Beispiel: Körpergröße bei Frauen und Männern

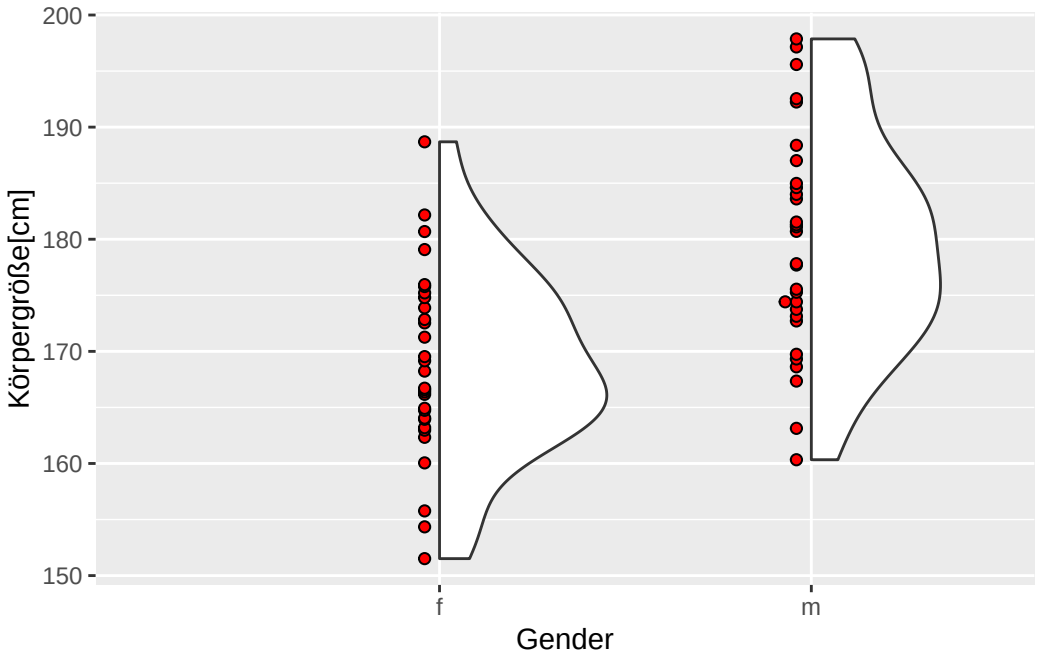


Figure 14.1: Simulierte Daten: Verteilung von Körpergrößen nach Geschlecht

14.2 Datensatz

Table 14.1: Ausschnitt aus den Daten

cm	gender
174.4	m
177.7	m
195.6	m
171.3	f
164.0	f

cm	gender
176.0	f

14.3 Nominale Variablen in R

Nominale Variablen werden in R als `factor()` dargestellt.

```
gender <- factor(c(0,0,1,1),
                 levels = c(0,1),
                 labels = c('m','f'))
gender
```

```
[1] m m f f
Levels: m f
```

[1]

14.4 t-Test in R mit `t.test()`

```
t.test(cm ~ gender, data=height, var.equal=T)
```

```
 Two Sample t-test

data: cm by gender
t = -4.57, df = 58, p-value = <0.001
d = -10.75, s_e = 2.35
95 percent confidence interval
[-15.45, -6.04]
```

[1] Viele Funktionen in R transformieren eine Vektor mit Zeichenketten in einen `factor()` um. z.B. `factor(c('m','m','f','f'))`

Table 14.2: Deskriptive Werte

gender	m	sd
f	168.8	8.4
m	179.5	9.8

14.5 Modellformulierung beim t-Test $(n_w = n_m)$

$$
\begin{aligned}
Y_{if} &= \mu_f + \epsilon_{if}, \quad \epsilon_{if} \sim \mathcal{N}(0, \sigma^2) \\
Y_{im} &= \mu_m + \epsilon_{im}, \quad \epsilon_{im} \sim \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

14.5.1 Hypothesen

$$
\begin{aligned}
H_0 &: \delta = 0 \\
H_1 &: \delta \neq 0
\end{aligned}
$$

14.5.2 Teststatistik

$$
t = \frac{\bar{y}_m - \bar{y}_w}{\sqrt{\frac{s_m^2 + s_w^2}{2}}\sqrt{\frac{2}{n}}}
$$

14.5.3 Referenzverteilung

$$
t \sim t_{df=2n-2}
$$

14.6 Kann ich aus dem t-Test ein lineares Modell machen?

14.6.1 t-Test

$$
\begin{aligned}
Y_{if} &= \mu_f + \epsilon_{if}, \quad \epsilon_{if} \sim \mathcal{N}(0, \sigma^2) \\
Y_{im} &= \mu_m + \epsilon_{im}, \quad \epsilon_{im} \sim \mathcal{N}(0, \sigma^2) \\
t &= \frac{\bar{y}_m - \bar{y}_w}{\sqrt{\frac{s_m^2 + s_w^2}{2}}\sqrt{\frac{2}{n}}} \\
t &\sim t_{df=2n-2}
\end{aligned}
$$

14.6.2 Lineares Modell

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 \times x_i + \epsilon_i \\
\Delta_m &= \mu_m - \mu_f \\
Y_i &= \beta_0 + \beta_1 \times x_{??} + \epsilon_i \\
Y_i &= \mu_f + \Delta_m \times x_{??} + \epsilon_i
\end{aligned}
$$

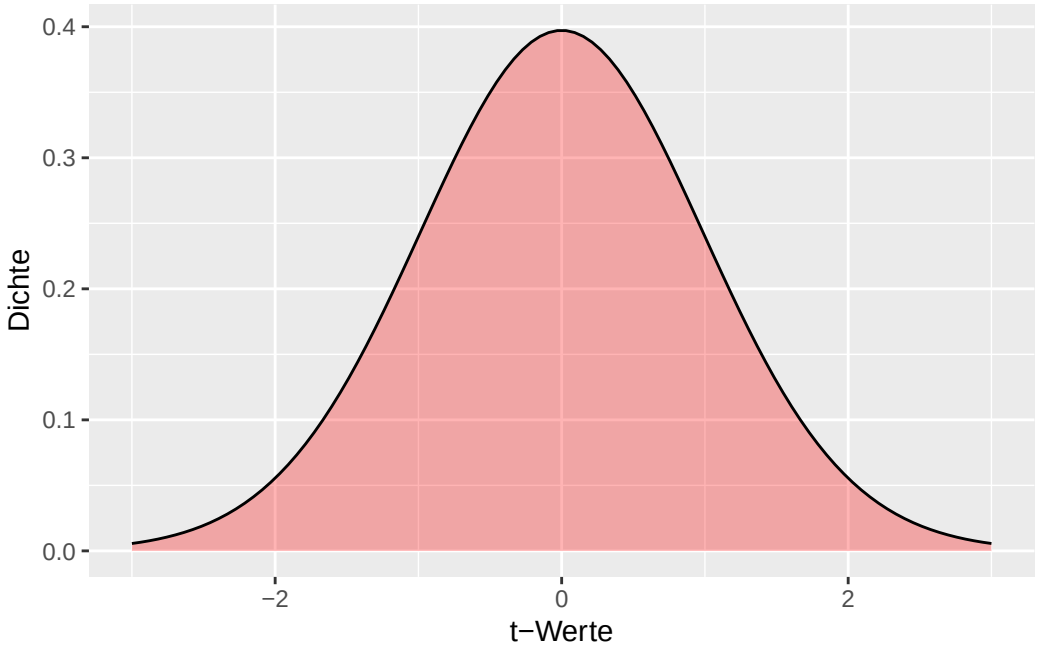


Figure 14.2: t-Verteilung mit $df = 58$

14.7 Dummy- oder Indikatorkodierung

$$
\begin{aligned}
Y_i &= \mu_f + \Delta_m \times x_{1i} + \epsilon_i \\
\Delta_m &= \mu_m - \mu_f \\
x_1 &= \begin{cases} 0 \text{ wenn weiblich} \\ 1 \text{ wenn männlich} \end{cases}
\end{aligned}
$$

Für eine nominale Variable wird eine Indikatorvariablen (Dummyvariable) definiert. Über diese Indikatorvariable kann die Zugehörigkeit eines Messwerts Y_i zu einer Faktorstufe k bestimmt werden. Eine Faktorstufe ist dabei immer die Referenzstufe bei der die Indikatorvariable gleich 0 ist.

14.8 Einfach mal stumpf in `lm()` eingeben

```
mod <- lm(cm ~ gender, height)
```

Table 14.3: Modellfit

	$\hat{\beta}$	s_e	t	p
(Intercept)	168.783	1.663	101.477	<0.001
genderm	10.746	2.352	4.568	<0.001

[2]

14.9 Vergleich der Konfidenzintervalle

14.9.1 Lineares Modell

```
confint(mod)
```

```
               2.5 %    97.5 %
(Intercept) 165.45401 172.11276
genderm       6.03713  15.45403
```

14.9.2 t-Test

```
t.test(cm ~ gender,
       data = height,
       var.equal=T)$conf
```

```
[1] -15.45403  -6.03713
attr(,"conf.level")
[1] 0.95
```

[3]

14.10 Auf welchen Werten wird ein lineares Modell gerechnet???

[2] `R` gibt die Faktorstufe nach dem Namen des Faktors an. Im Beispiel steht `genderm` für Stufe `m` im Faktor `gender`.

[3] Mit `t.test()$conf.int` kann auf das berechnete Konfidenzintervall zugegriffen werden.

Table 14.4: Repräsentation der Faktorvariablen

cm	gender	x_1
174.40	m	1
177.70	m	1
195.59	m	1
160.05	f	0
164.92	f	0
154.35	f	0

14.11 Residuen

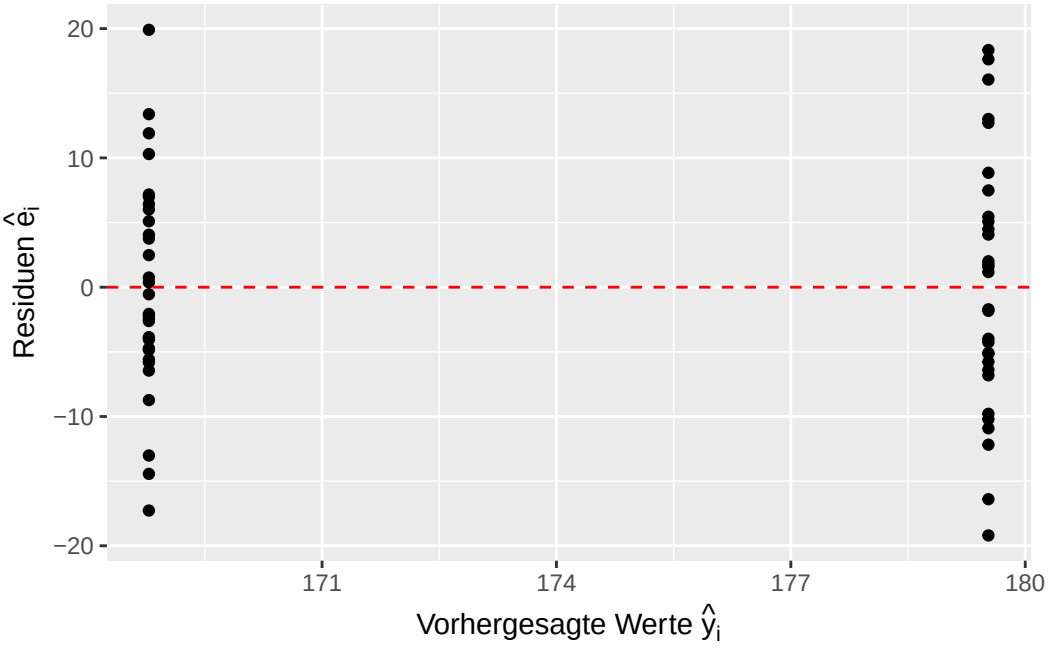


Figure 14.3: Residuen

14.12 Wen's interessiert - t-Wert

Seien beide Gruppen gleich groß (n) mit $N = n_m + n_w = 2 \times n$. Der t-Wert für β_1 berechnet sich aus $t = \frac{b_1}{s_b}$ mit:

$$s_b = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-2} \frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

Dadurch, das die x_i entweder gleich 0 oder 1 sind, ist $\bar{x} = 0.5$ und die Abweichungsquadrate im zweiten Term sind alle gleich $\frac{1}{4}$.

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N}\left(x_i - \frac{1}{2}\right)^2 = \sum_{i=1}^{N}\frac{1}{4} = \frac{N}{4} = \frac{2n}{4} = \frac{n}{2}$$

Der ersten Term kann mit etwas Algebra und der Definition für die Stichprobenvarianz s^2 auf die gewünschte Form gebracht werden.

$$\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N-2} = \frac{\sum_{i=1}^{n}(\overbrace{y_{im} - \bar{y}_m}^{Mnner})^2 + \sum_{i=1}^{n}(\overbrace{y_{iw} - \bar{y}_w}^{Frauen})^2}{2(n-1)} = \frac{(n-1)s_m^2 + (n-1)s_w^2}{2(n-1)} = \frac{s_m^2 + s_w^2}{2}$$

14.13 Wen's interessiert - $\beta_1 = \mu_w - \mu_m$

Mit $s_x^2 = \frac{N\frac{1}{4}}{N-1} = \frac{N}{4(N-1)}$

$$\begin{aligned}
b_1 &= \frac{cov(x,y)}{s_x^2} \\
&= \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{N-1}\frac{4(N-1)}{N} \\
&= 4\frac{\sum_{i=1}^{n}(y_{im} - \bar{y})\frac{-1}{2} + \sum(y_{iw} - \bar{y})\frac{1}{2}}{N} \\
&= \frac{4}{2}\frac{\sum_{i=1}^{n}(y_{iw} - \bar{y}) - \sum_{i=1}^{n}(y_{im} - \bar{y})}{2n} \\
&= \frac{\sum_{i=1}^{n} y_{iw}}{n} - \frac{n\bar{y}}{n} - \frac{\sum_{i=1}^{n} y_{im}}{n} + \frac{n\bar{y}}{n} \\
&= \bar{y}_w - \bar{y}_m = \Delta
\end{aligned}$$

14.14 Wen's interessiert - $\beta_0 = \mu_m$

Mit $b_1 = \Delta = \bar{y}_w - \bar{y}_m$:

$$\begin{aligned}
b_0 &= \bar{y} - \Delta \times \bar{x} \\
&= \frac{\sum_{i=1}^{N} y_i}{N} - \Delta \times \frac{1}{2} \\
&= \frac{\sum_{i=1}^{n} y_{im} + \sum_{i=1}^{n} y_{iw}}{2n} - \frac{1}{2}(\bar{y}_w - \bar{y}_m) \\
&= \frac{1}{2}\frac{\sum_{i=1}^{n} y_{im}}{n} + \frac{1}{2}\frac{\sum_{i=1}^{n} y_{iw}}{n} - \frac{1}{2}\bar{y}_w + \frac{1}{2}\bar{y}_m \\
&= \frac{1}{2}\bar{y}_m + \frac{1}{2}\bar{y}_w - \frac{1}{2}\bar{y}_w + \frac{1}{2}\bar{y}_m \\
&= \bar{y}_m
\end{aligned}$$

14.15 Können auch mehr als zwei Stufen verwendet werden?

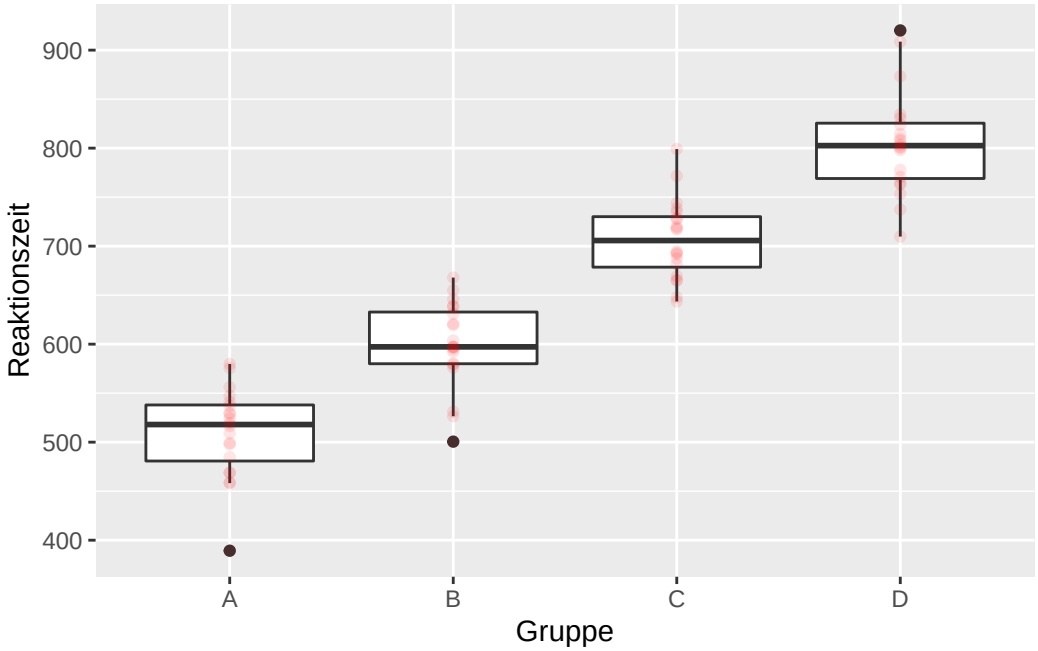


Figure 14.4: Ein Reaktionszeitexperiment mit vier Stufen A, B, C und D

Table 14.5: Gruppenmittelwerte, Standardabweichung und Unterschiede zu Stufe A

Gruppe	$\bar{y}_j$	s_j	Δ_{j-A}
A	509.53	45.66	
B	599.68	43.57	90.15
C	706.94	40.49	197.41
D	805.09	52.51	295.56

	x1	x2	x3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

14.16 Deskriptive Daten

14.17 Reaktionszeitexperiment als lineares Modell

14.17.1 Modell

$$y_i = \mu_A + \Delta_{B-A}x_1 + \Delta_{C-A}x_2 + \Delta_{D-A}x_3 + \epsilon_i$$

14.17.2 Dummyvariablen

14.18 Nochmal allgemeiner

Mit K Faktorstufen werden (K-1) Dummyvariablen $x_1, x_2, \ldots, x_{K-1}$ benötigt. Eine Stufe wird als Referenz definiert. Die x_1 bis x_{K-1} kodieren die Abweichungen der anderen Stufen von dieser Stufe.[4]

[4]Diese Art der Kodierung wird auch als treatment Kodierung bezeichnet.

	x_1	x_2	...	x_{K-1}
Referenz ($j = 1$)	0	0		0
$j = 2$	1	0	...	0
$j = 3$	0	1	...	0
$j = K$	0	0	...	1

Table 14.6: Modellfit

	$\hat{\beta}$	s_e	t	p
(Intercept)	509.526	10.235	49.784	<0.001
groupB	90.150	14.474	6.228	<0.001
groupC	197.414	14.474	13.639	<0.001
groupD	295.561	14.474	20.420	<0.001

Table 14.7: ANOVA-Tabelle

	Df	SSQ	MSQ	F	p
group	3	988935.1	329645.04	157.35	<0.001
Residuals	76	159221.0	2095.01		

14.19 Reaktionszeitexperiment mit `lm()`

```
mod <- lm(rt ~ group, data)
```

14.20 Ausblick

```
anova(mod)
```

14.21 Kombination von kontinuierlichen und nominalen Variablen

14.22 Modellansatz

- Aus gender (K = 2) wird eine **Dummyvariable**
- Frauen werden (zufällig) als Referenz genommen

$$Y_i = \beta_{ta=0,x_1=0} + \Delta_m \times x_1 + \beta_{ta} \times ta + \epsilon_i$$

$$x_1 = \begin{cases} 0 \text{ wenn weiblich} \\ 1 \text{ wenn männlich} \end{cases}$$

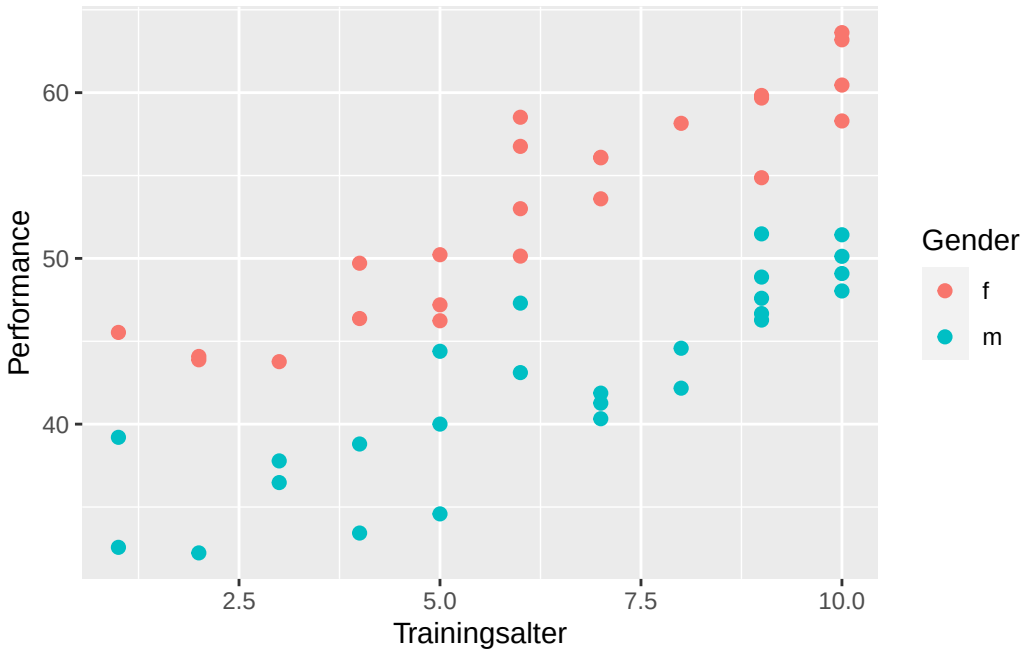


Figure 14.5: Hypothetische Leistungsentwicklung in Abhängigkeit vom Alter und Gender

Table 14.8: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	41.181	1.083
gender_fm	-10.877	0.805
ta	1.927	0.145
$\hat{\sigma}$	2.845	

14.23 Modellieren mit `lm()`

```
mod <- lm(perf ~ gender_f + ta, lew)
```

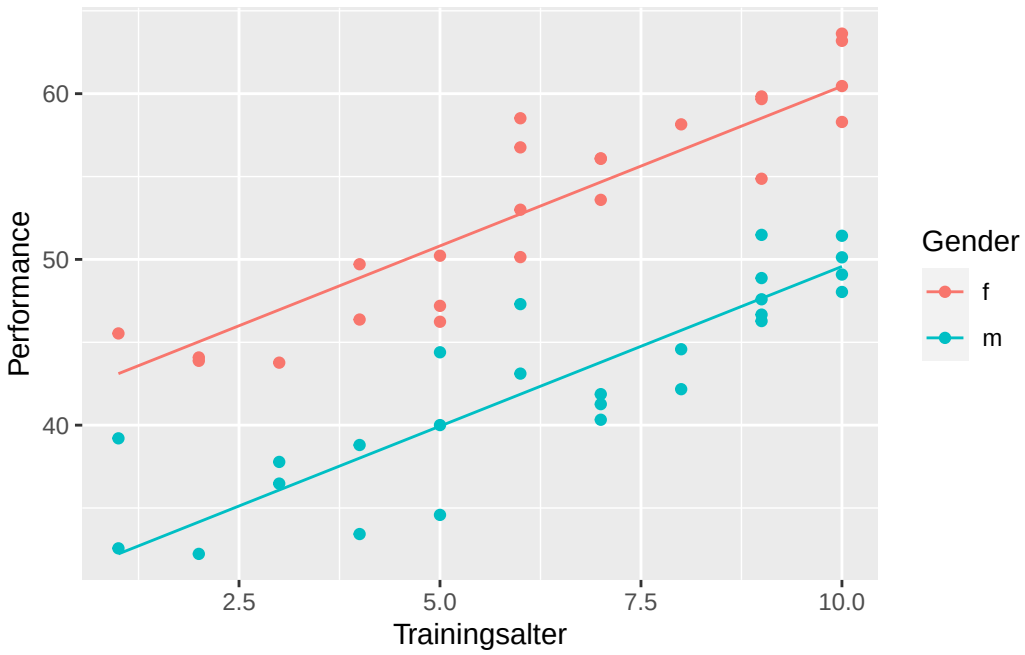


Figure 14.6: Leistungsentwicklung in Abhängigkeit vom Alter und Gender

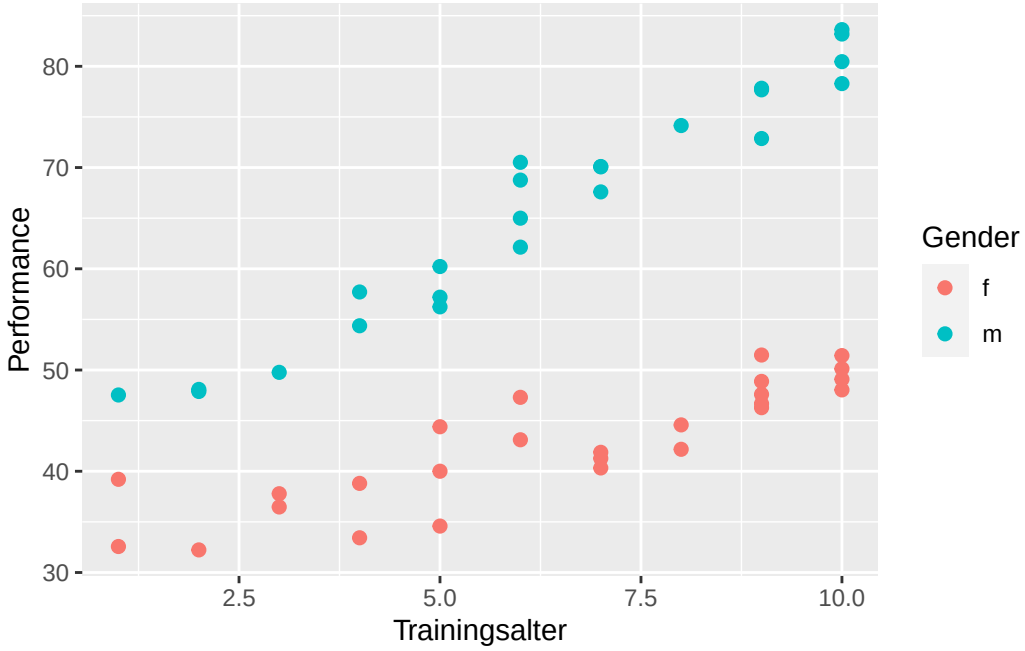


Figure 14.7: Leistungsentwicklung in Abhängigkeit vom Alter und Gender

Table 14.9: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	31.354	1.370
gender_fm	8.575	2.010
ta	1.763	0.195
gender_fm:ta	2.362	0.290
$\hat{\sigma}$	2.828	

14.24 Die resultierenden Graden

14.25 Interaktion zwischen kontinuierlichen und nominalen Variablen

14.26 Ansatz für ein Interaktionsmodell

Das vorhergehendes Modell wird um einen Interaktionsterm erweitert.

$$y_i = \beta_{ta=0,x_1=0} + \Delta_m \times x_1 + \beta_{ta} \times ta + \beta_{ta \times gender} \times x_1 \times ta + \epsilon_i$$

14.27 Interaktionsmodell mit `lm()`

```
mod <- lm(perf ~ gender_f * ta, lew)
```

14.28 Regressionsgeraden

14.29 Zum Nacharbeiten

Kutner et al. (2005, 313–19)

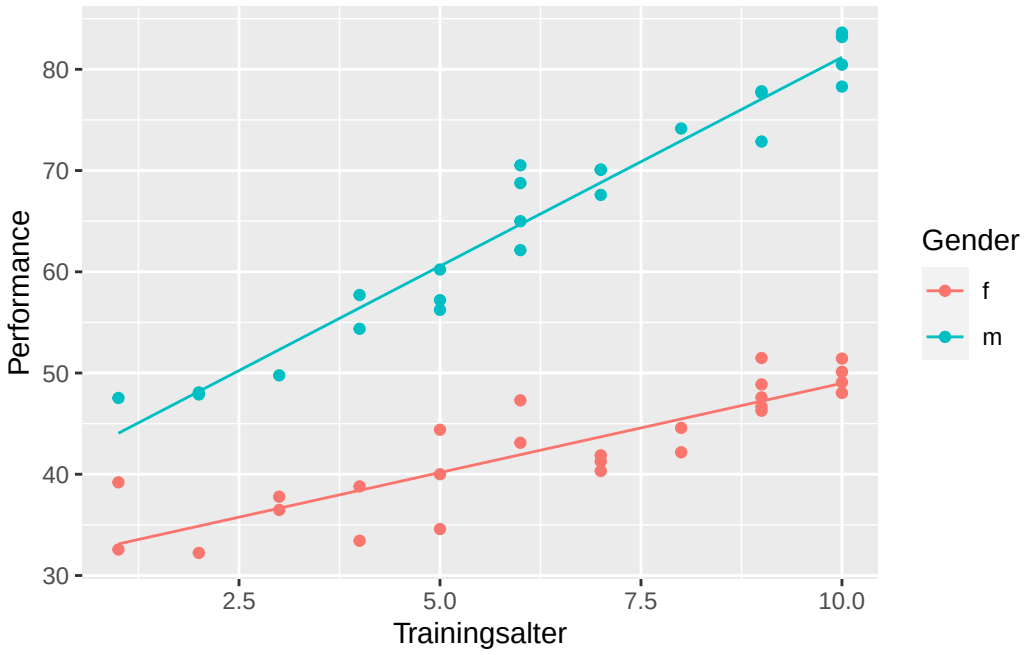


Figure 14.8: Leistungsentwicklung in Abhängigkeit vom Alter und Gender

15 Modellhierarchien

15.1 Einfaches Modell

```
mod0 <- lm(y ~ x, simple)
summary(mod0)
```

```
Call:
lm(formula = y ~ x, data = simple)

Residuals:
      1       2       3       4
-0.5817  0.9898 -0.2345 -0.1736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8414     0.7008   2.628    0.119
x             0.4574     0.3746   1.221    0.346

Residual standard error: 0.8376 on 2 degrees of freedom
Multiple R-squared:  0.4271,    Adjusted R-squared:  0.1406
F-statistic: 1.491 on 1 and 2 DF,  p-value: 0.3465
```

15.2 Einfaches Modell

```
mod0 <- lm(y ~ x, simple)
summary(mod0)
```

```
Call:
lm(formula = y ~ x, data = simple)

Residuals:
      1       2       3       4
-0.5817  0.9898 -0.2345 -0.1736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8414     0.7008   2.628    0.119
x             0.4574     0.3746   1.221    0.346

Residual standard error: 0.8376 on 2 degrees of freedom
Multiple R-squared:  0.4271,    Adjusted R-squared:  0.1406
F-statistic: 1.491 on 1 and 2 DF,  p-value: 0.3465
```

15.3 Abweichungen ... noch mal

15.3.1 Sum of squares of error

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Typischerweise beinhaltet ein Modell zum berechnen der $\hat{y}_i$ verschiedene Parameter. Bei der einfachen Regression zum Beispiel β_0 und β_1 (#Modellparameter $p = 2$) .

15.3.2 Freiheitsgrade (degrees of freedom) von SSE

$$dfE := n - p$$

Die *effektive* Anzahl der Beobachtungen um die Varianz σ^2 abzuschätzen.

15.4 MSE als Schätzer für σ^2

15.4.1 Mean squared error MSE

$$MSE = \frac{SSE}{dfE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p}$$

Als Schätzer $\hat{\sigma}^2$ für σ^2 aus $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

15.4.2 Parallel zur Berechnung der Stichprobenvarianz

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^{2} (y_i - \bar{y})^2$$

wo s^2 ein Schätzer für die Varianz von y ist.

15.5 Genereller Linearer Modell Testansatz[1]

15.5.1 Idee

Wir bauen uns eine Teststatistik die die Verbesserung in der Vorhersage (= Reduktion der Fehlervarianz) als Metrik verwendet. Modelle werden in eine Hierarchie gesetzt mit einfacheren Modellen untergeordnet zu komplexeren Modellen.

15.5.2 Leitfrage:

Bringt mir die Aufnahme zusätzlicher Modellparameter eine Verbesserung in der Vorhersage von Y bzw. bezüglich der Aufklärung der Varianz in Y?

[1] Kutner et al. (2005), p.72

15.6 Genereller Linearer Modell Testansatz - Full model

Beispiel einfache lineare Regression

15.6.1 Volles Modell

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

15.6.2 Residualvarianz SSE(F)

$$\text{SSE(F)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

mit $p = 2, dfE(F) = n - 2$

15.7 Genereller Linearer Modell Testansatz - Reduced model

15.7.1 Reduziertes Modell

$$Y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

15.7.2 Residualvarianz SSE(R)

$$\text{SSE(R)} = \sum_{i=1}^{n} (y_i - \beta_0)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{SSTO}$$

mit $p = 1, dfE(R) = n - 1$

Im Allgemeinen gilt: $SSE(F) \leq SSE(R)$

15.8 Link: Reduziertes Modell und Stichprobenvarianz

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} (y_i - \beta_0)^2 = \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 + \beta_0^2) \\
0 &= \frac{\mathrm{d}}{\mathrm{d}\beta_0} \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 + \beta_0^2) \\
0 &= \sum_{i=1}^{n} (-2y_i + 2\beta_0) = -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \beta_0 \\
n\beta_0 &= \sum_{i=1}^{n} y_i \\
\beta_0 &= \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} \rightarrow \frac{SSE}{n-1} = \hat{\sigma}^2 = s^2
\end{aligned}$$

15.9 Genereller Linearer Modell Testansatz

Annahme: Das reduzierte Modell ist korrekt. Dann sollte

$$\text{SSE(R)} - \text{SSE(F)}$$

eher klein sein (Beide Modelle haben einen gleich guten fit).

Annahme: Das reduzierte Modell ist falsch: Dann sollte

$$\text{SSE(R)} - \text{SSE(F)}$$

eher groß sein (Das reduzierte Modell kann die Daten nicht so gut fitten wie das komplizierte Modell)

15.10 Genereller Linearer Modell Testansatz - Teststatistik

Wenn das reduzierte Modell korrekt ist, dann lässt sich zeigen, dass:

$$MS_{\text{test}} = \frac{\text{SSE(R)} - \text{SSE(F)}}{\text{dfE(R)} - \text{dfE(F)}}$$

ein Schätzer für die Varianz σ^2 ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$) ist.

Wenn das reduzierte Modell korrekt ist, dann ist auch das volle Modell korrekt. Daher ist dann:

$$\text{MSE(F)} = \frac{\text{SSE(F)}}{\text{dfE(F)}}$$

auch ein Schätzer für σ^2

15.11 F-Wert als Teststatistik

$$F = \frac{MS_{\text{test}}}{MSE(F)} = \frac{\frac{\text{SSE(R)} - \text{SSE(F)}}{\text{dfE(R)} - \text{dfE(F)}}}{\frac{\text{SSE(F)}}{\text{dfE(F)}}}$$

15.12 Verteilung der F-Statistik

$$F = \frac{MS_{\text{test}}}{MSE(F)} \sim F(\text{dfE(R)} - \text{dfE(F)}, \text{dfE(F)})$$

15.13 Hypothesentest mit F-Wert

[2]

[2] In R: `df()`, `pf()`, `qf()`, `rf()`

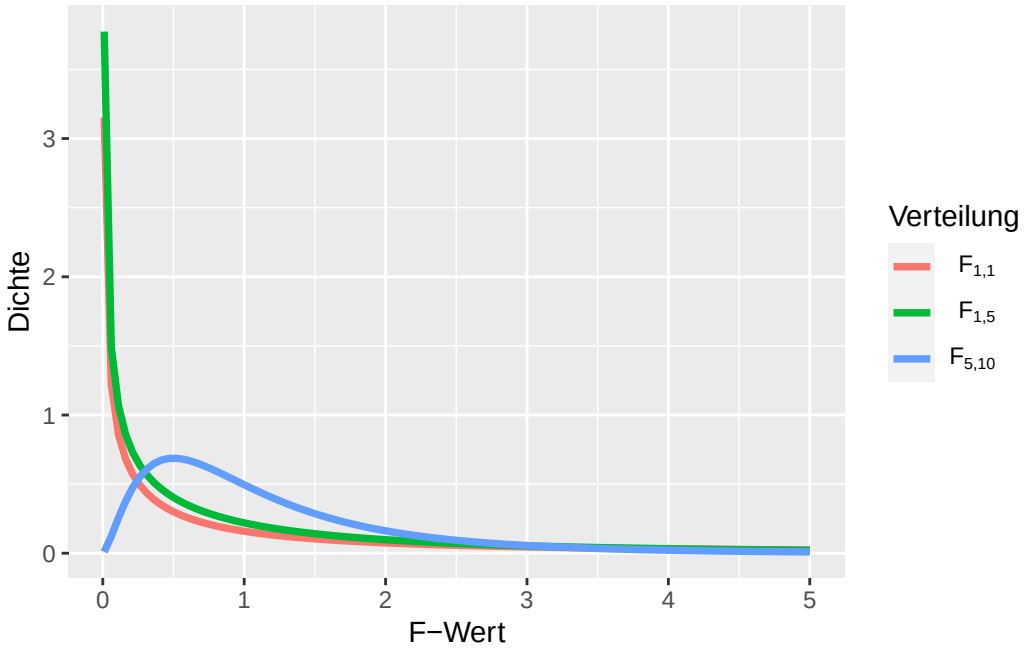


Figure 15.1: Beispiele für die F-Verteilung mit verschiedenen Freiheitsgraden df_1, df_2

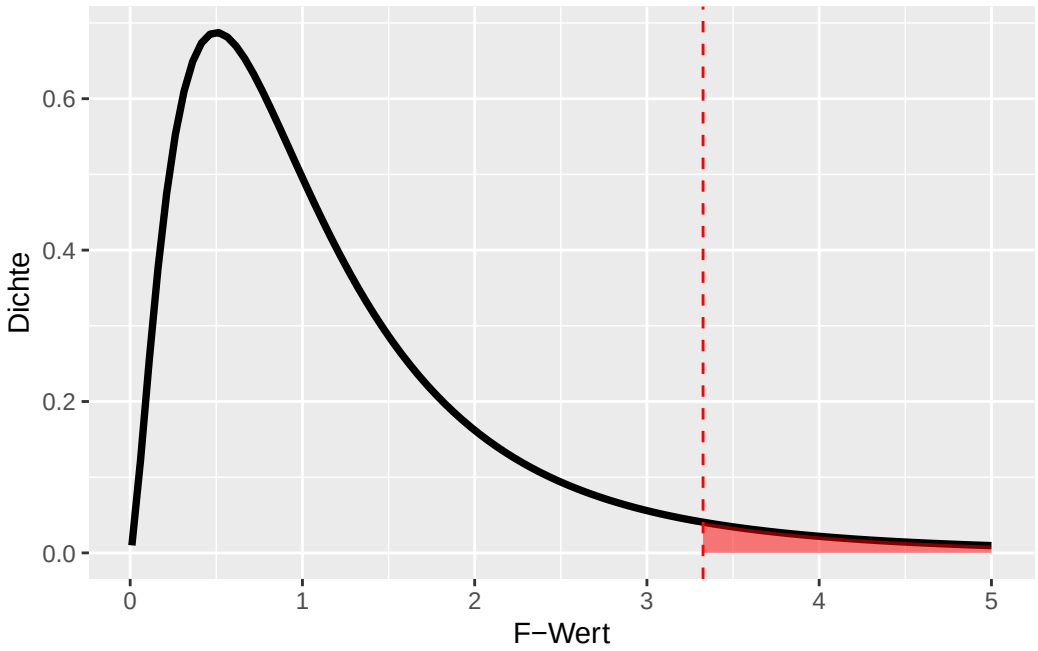


Figure 15.2: F-Verteilung mit $df_1 = 5, df_2 = 10$ und kritischem Wert bei $\alpha = 0.05$

15.14 Teilziel

- Durch den Vergleich von Modellen kann die Verbesserung/Verschlechterung der Modellvorhersage statistisch Überprüft werden
- Alternativ: Brauchen ich zusätzliche Parameter oder reicht mir das einfache Modell?

15.15 Beispiel: Candy-Problem

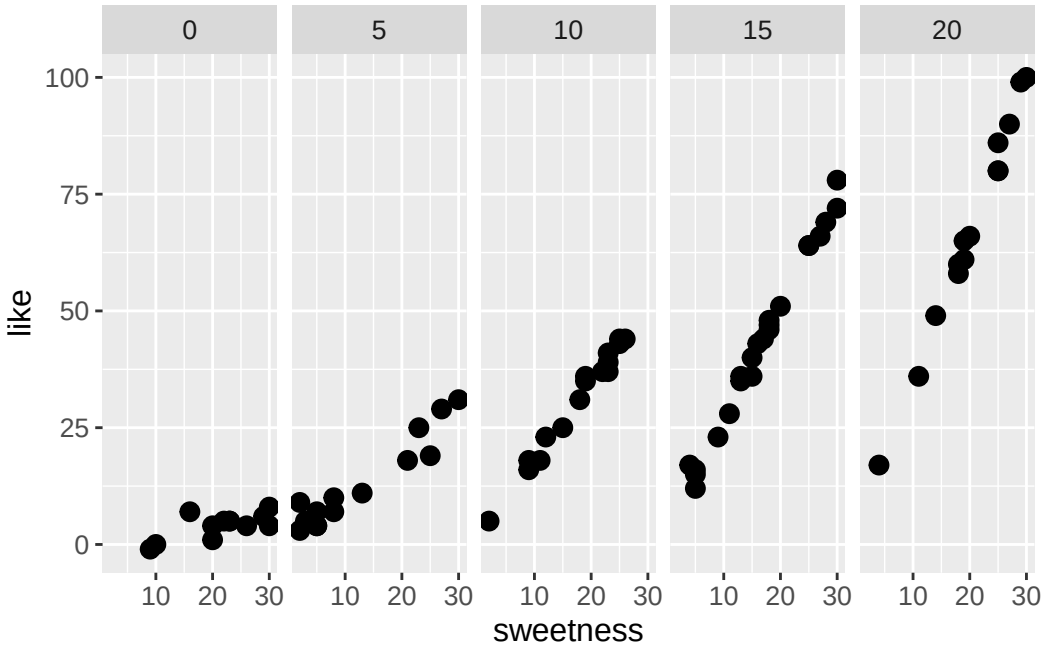


Figure 15.3: Zusammenhang zwischen der Präferenz für ein Bonbon und dem Süßgrad für verschiedene Weichheitsgrade

15.16 Modelle als Hierarchien auffassen

15.16.1 Full model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

15.16.2 Hierarchie

$$
\begin{aligned}
m_0 &: y_i = \beta_0 + \epsilon_i \\
m_1 &: y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \\
m_2 &: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \\
m_3 &: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i
\end{aligned}
$$

Es gilt: $m_0 \subseteq m_1 \subseteq m_2 \subseteq m_3$

15.17 Modelle als Hierarchien auffassen in R

In R:

```
mod_0 <- lm(like ~ 1, candy)
mod_1 <- lm(like ~ sweetness, candy)
mod_2 <- lm(like ~ sweetness + moisture, candy)
mod_3 <- lm(like ~ sweetness * moisture, candy)
```

15.18 Vergleich m_0 gegen m_1

$$
\begin{aligned}
m_0 &: y_i = \beta_0 + \epsilon_i \\
m_1 &: y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i
\end{aligned}
$$

```
anova(mod_0, mod_1)
```

Table 15.1: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ 1	77				
Model 2: like ~ sweetness	76	1	16251.47	32.96	0

15.19 Vergleich m_1 gegen m_2

$$
\begin{aligned}
m_1 &: y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \\
m_2 &: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i
\end{aligned}
$$

```
anova(mod_1, mod_2)
```

Table 15.2: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ sweetness	76				
Model 2: like ~ sweetness + moisture	75	1	32219.64	459.94	0

15.20 Vergleich m_2 gegen full model m_3

$$m_2 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$
$$m_3 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

```
anova(mod_2, mod_3)
```

Table 15.3: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ sweetness + moisture	75				
Model 2: like ~ sweetness * moisture	74	1	4914.29	1070.72	0

15.21 Vergleich full model m_3 gegen minmales Modell m_0

$$m_0 : y_i = \beta_0 + \epsilon_i$$
$$m_3 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

```
anova(mod_0, mod_3)
```

Table 15.4: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ 1	77				
Model 2: like ~ sweetness * moisture	74	3	53385.4	3877.17	0

15.22 In `summary()` m_3 gegen m_0

```
Call:
lm(formula = like ~ sweetness * moisture, data = candy)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0867 -1.6999  0.1977  1.1862  6.0127
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.340570   1.099250   0.310  0.75757
sweetness         0.170331   0.053384   3.191  0.00208 **
moisture          0.162394   0.095668   1.697  0.09381 .
sweetness:moisture 0.149407   0.004566  32.722  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.142 on 74 degrees of freedom
Multiple R-squared:  0.9937,    Adjusted R-squared:  0.9934
F-statistic:  3877 on 3 and 74 DF,  p-value: < 2.2e-16
```

15.23 Eine nominale Variable mit vier Stufen

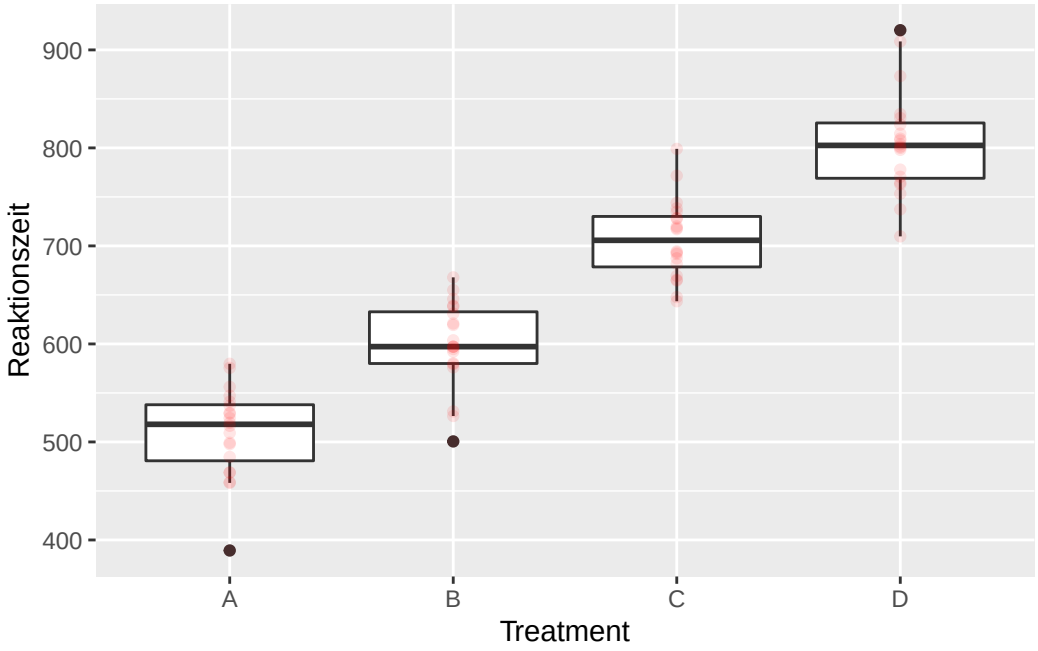


Figure 15.4: Ein Reaktionszeitexperiment mit vier Stufen A, B, C und D

15.24 Früher - Analysis of Variance (ANOVA bzw. AOV)

$$s^2_{zwischen} = \frac{1}{K-1}\sum_{j=1}^{K} N_j(\bar{x}_{j.} - \bar{x})^2$$

$$s^2_{innerhalb} = \frac{1}{N-K}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(x_{ji} - \bar{x}_{j.})^2 = \frac{1}{N-K}\sum_{j=1}^{K}(N_j - 1)s^2_j$$

$$F = \frac{\hat{\sigma}^2_{zwischen}}{\hat{\sigma}^2_{innerhalb}} \sim F(K-1, N-K)$$

15.25 ANOVA in R

```
mod_aov <- aov(rt ~ group, rt_tbl)
summary(mod_aov)
```

Table 15.5: Ausgabe mit aov()

term	df	sumsq	meansq	statistic	p.value
group	3	988935.1	329645	157.3	0
Residuals	76	159221.0	2095		

15.26 Ansatz mittels Modellhierarchien

15.26.1 Full model

$$y_i = \beta_0 + \beta_{\Delta_{B-A}} x_1 + \beta_{\Delta_{C-A}} x_2 + \beta_{\Delta_{D-A}} x_3 + \epsilon_i$$

15.26.2 Reduced model

$$y_i = \beta_0 + \epsilon_i$$

Wenn das reduced model die Daten gleich gut fittet wie das full model ⇒ Information über das Treatment verbessert meine Vorhersage von y_i nicht.

15.27 Model fit - Full model

```
mod <- lm(rt ~ group, rt_tbl)
```

Table 15.6: Modellfit

	$\hat{\beta}$	s_e	t	p
(Intercept)	509.526	10.235	49.784	<0.001
groupB	90.150	14.474	6.228	<0.001
groupC	197.414	14.474	13.639	<0.001
groupD	295.561	14.474	20.420	<0.001

15.28 `anova()` mit nur einem Modell

```
anova(mod)
```

Table 15.7: Äquivalent zum Vergleich full gegen reduced model

term	df	sumsq	meansq	statistic	p.value
group	3	988935.1	329645	157.3	0
Residuals	76	159221.0	2095		

15.29 Zum Nacharbeiten

Christensen (2018, 57–64)

Part IV

Das allgemeine lineare Modell

16 Synthese

Literatur

Altman, Douglas G, and J Martin Bland. 1995. "Statistics Notes: Absence of Evidence Is Not Evidence of Absence." *Bmj* 311 (7003): 485.

Altman, Naomi, and Martin Krzywinski. 2015a. "Points of Significance: Multiple Linear Regression." *Nature Methods* 12 (12): 1103–4.

———. 2015b. "Points of Significance: Simple Linear Regression." *Nature Methods* 12 (11).

———. 2016a. "Points of Significance: Analyzing Outliers: Influential or Nuisance." *Nature Methods* 13 (4): 281–82.

———. 2016b. "Points of Significance: Regression Diagnostics." *Nature Methods* 13 (5): 385–86.

Christensen, Ronald. 2018. *Analysis of Variance, Design, and Regression: Linear Modeling for Unbalanced Data*. CRC Press.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge.

Cumming, Geoff. 2013. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge.

Debanne, Thierry, and Guillaume Laffaye. 2011. "Predicting the Throwing Velocity of the Ball in Handball with Anthropometric Variables and Isotonic Tests." *Journal of Sports Sciences* 29 (7): 705–13.

Fox, John. 2011. *An r Companion to Applied Regression*. 2nd ed. SAGE Publication Inc., Thousand Oaks.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill Irwin New York.

McElreath, Richard. 2016. *Statistical Rethinking, a Bayesian Course with Examples in r and Stan*. 1st ed. Boca Raton: CRC Press.

Spiegelhalter, David. 2019. *The Art of Statistics: Learning from Data*. Penguin UK.

Wasserstein, Ronald L, and Nicole A Lazar. 2016. "The ASA Statement on p-Values: Context, Process, and Purpose." Taylor & Francis.

Wild, Christopher J, and Georg AF Seber. 2000. *Chance Encounters: A First Course in Data Analysis and Inference*. Wiley Press.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134 (2): 557–98.

Index