Seminar Teil Gygax

Planung und statistische Auswertung biologischer Experimente

Lorenz Gygax
LGygax@proximate-biology.ch

Version 1.3.0
letzte Änderungen: Sommer 2001

"Lehre jemanden den t–Test und er wird für einen Tag glücklich sein; lehre jemanden die Regression und er wird für eine Woche lang glücklich sein; lehre jemanden Statistik und er wird sein ganzes Leben lang Probleme haben." (unbekannter Statistiker)

In diesem Manuskript gibt es zwei verschiedene Einschübe: Überlegungen und Übungen. Die Überlegungen sollen dazu dienen, die Grundlagen gedanklich zu vertiefen und werden im Seminar diskutiert (14–15 Uhr). Die Übungen sollen Gelegenheit bieten, das Gelesene praktisch anzuwenden und können in den Übungen gelöst werden (15–16 Uhr).

Ihr seid die dritten Generation von "Opfern" dieses Skriptes: weitere Hinweise auf Druckfehler, generelle Kritik und Anregungen werden gerne entgegen genommen.

1 Nicht–parametrische Methoden, was ist, was macht ein Test?

1.1 Ein einfaches Beispiel

Anhand eines einfachen Beispiels soll in diesem ersten Kapitel erarbeitet werden, was ein statistischer Test macht und welche Schritte gemacht werden müssen, um ein statistisches Resultat zu erhalten. Ein einfacher Fall für einen statistischen Vergleich ist der Zweistichprobenfall, bei dem man zwei Gruppen miteinander vergleichen möchte.

Bereits hier gibt es zwei verschiedene Möglichkeiten: Nehmen wir an, wir wollen herausfinden, ob Vogelmütter einer bestimmten Art in ihrem zweiten Brutjahr kompetenter werden und deshalb ihre Jungen im zweiten Jahr mit einer höheren Rate füttern. Vorausgesetzt wir erkennen das Alter am Habitus (z. B. an der Gefiederfärbung), können wir im gleichen Jahr ein- und zweijährige Weibchen vergleichen. Da ein Individuum nur in einer Gruppe vorkommt spricht man von sogennant ungepaarten (oder auch "unabhängigen") Gruppen. Wir können auch mehrere einjährige Tiere beobachten und diese Beoabachtungen im naechsten Jahr wiederholen. Wir vergleichen dann dieselben Tiere im Alter von ein und zwei Jahren, was einen sogenannt gepaarten (oder auch "abhängigen") Test notwendig macht. Dieses zweite fiktive Beispiel wollen wir uns nun ein wenig genauer ansehen.

Überlegung 1.1 *Welche Resultate sind hier wohl verlässlicher, die eines gepaarten oder eines ungepaarten Testes? Welche Störvariablen gibt es? Wie würden idealerweise Daten aufgenommen, um diese Frage zu beantworten?*

Überlegung 1.2 *Wieso stehen wohl "abhängig" und "unabhängig" in Anführungsstrichen?*

1.1.1 Computereingabe

Nehmen wir an, dass wir 10 Weibchen, die wir im ersten Jahr beobachtet haben, auch im zweiten Jahr noch finden und wir beobachten, mit welcher Rate (Beutestücke pro Minute) sie Futter eintragen. Die Messungen unserer Zielvariablen, der Eintrag–Rate für das erste Jahr sind 0.30, 0.56, 0.80, 0.95, 1.35, 1.98, 0.75, 0.63, 0.77, 0.82 und für das zweite 0.33, 0.62, 1.22, 1.02, 1.45, 1.94, 0.88, 0.58, 0.98, 1.13. Insbesondere bei komplizierteren Modellen werden in den meisten Programmen alle Werte einer Zielvariable untereinander in eine Kolonne geschrieben. In den folgenden Kolonnen trägt man dann die Werte für die erklärenden Variablen ein. In unserem Beispiel wäre das eine Kolonne mit zwei Einträgen für die beiden Beobachtungsjahre, rsp. Altersstufen und eine Kolonne mit Bezeichnungen für die Individuen. Bei einfachen gepaarten Test werden häufig auch die zusammengehörenden Werte in die gleiche Zeile zweier Kolonnen geschrieben, wie in Tabelle 1, in der auch noch weitere Grössen aufgelistet sind, die wir später brauchen.

1.1.2 Visualisierung

Der erste und sehr wichtige Schritt jeder statistischen Auswertung sollte eine genaue (graphische) Betrachtung der Daten sein. Dies hat verschiedene Zwecke: Man lernt die Struktur der Daten kennen (z. B. was die unabhängigen Replikate sind), was einem hilft, die nötige Statistik auszuwählen und plausibel zu interpretieren. Auch können Resultate, bei denen etwas schief gelaufen ist, eventuell erkannt werden. Fehlerhafte Werte, Ausreisser und schiefe Verteilungen können ebenfalls früh erkannt werden (dies ist insbesondere bei den parametrischen Statistiken

Tabelle 1: Beispiel: Rohdaten für einen einfachen gepaarten Test und einige davon abgeleitete Grössen

1. Jahr	2. Jahr	Differenz	Ränge	Vorzeichen
0.30	0.33	0.03	1	+
0.56	0.62	0.06	4	+
0.80	1.22	0.42	10	+
0.95	1.02	0.07	5	+
1.35	1.45	0.10	6	+
1.98	1.94	0.04	2	−
0.75	0.88	0.13	7	+
0.63	0.58	0.05	3	−
0.77	0.98	0.21	8	+
0.82	1.13	0.31	9	+

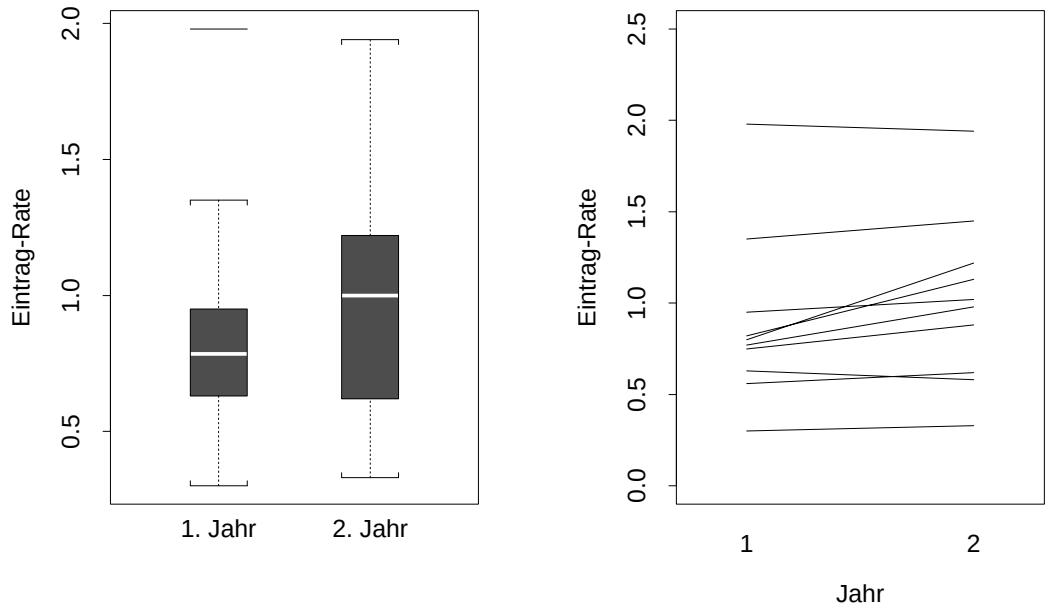


Abbildung 1: Eintrag–Rate versus Alter der Vogelweibchen dargestellt als Boxplot (links) und als eine Liniengraphik (rechts).

wichtig). Ausserdem wird in einem Bericht oder einem Artikel ein statistisches Resultat viel überzeugender sein, wenn man eine intuitive Graphik dazu zeigen kann.

Eine sehr kompakte Darstellung von kategoriellen Daten erlauben die Boxplots, die den Median, das untere und obere Quartil und die Extremwerte darstellen und damit einen guten Eindruck der gesamten Verteilung ergeben (Abb. 1, links). Im Gegensatz zu einer Darstellung von Mittelwert und Standardabweichung impliziert der Boxplot auch keine vorgegebene (Normal–)verteilung. In unserem Beispiel ist der Boxplot nicht sehr eindrücklich, da der Unterschied zwischen den beiden Gruppen sehr klein aussieht. Wenn wir aber die Struktur unserer Datenaufnahme auch in der Graphik benützen, sehen wir, dass beinahe alle Weibchen ihre Fütterungsrate im zweiten Jahr gesteigert haben (Abb. 1, rechts).

1.2 Statistisches Testen

Ein statistischer Test ist immer ein Widerspruchstest, d. h. einer Nullhypothese wird eine Alternativhypothese gegenübergestellt und man will zeigen, dass die Nullhypothese bei gegebenen Daten sehr unwahrscheinlich ist. Dies bedeutet, dass wir nur eine Nullhypothese verwerfen können, nicht jedoch eine Null– oder Alternativhypothese beweisen. Die Nullhypothese in unserem Beispiel lautet, dass sich die Eintragrate im zweiten Jahr nicht von derjenigen im ersten unterscheidet.

Wir möchten also aus unseren Daten eine Grösse ableiten, eine sogenannte "Teststatistik". Für diese Teststatistik müssen wir die Verteilung unter der Nullhypothese herausfinden, d. h. welche Werte der Teststatisik häufig zu erwarten sind und welche selten. Auf Grund dieser Erwartung berechnen wir eine Wahrscheinlichkeit für unser gefundenes Resultat. Ist diese Wahrscheinlichkeit klein (üblicherweise kleiner als 0.05), sagt man, dass die Teststatik im Verwerfungsbereich der Nullhypothese liegt, und man spricht von einem statistisch signifikanten Ergebnis. Wir können den Verwerfungsbereich auch als Fläche unter der Kurve darstellen, die durch die Wahrscheinlichkeitsverteilung der Teststatitik gegeben ist (Abb. 2, oben). Diese Fläche macht bei einem Test auf dem Niveau 5% auch genau 5% der Fläche unter der Kurve aus. Wie das konkret vor sich geht, werden wir gleich an einigen Beispielen erarbeiten.

Zweiseitige Tests schauen, ob die Teststatisik unter den 5% der extremsten Werte der Teststatisik liegt unabhängig davon, auf welcher Seite der Verteilung diese Extreme liegen. Einseitige Tests berücksichtigen 5% der Werte auf nur einer Seite der Verteilung und verdoppeln damit den Verwerfungsbereich auf dieser Seite. Trotzdem gibt es die Konvention, dass bei einseitiger Hypothese ein einseitiger Test gemacht werden darf. Es gibt dazu aber keine mathematisch–statistische Begründung (Abb. 2, oben).

1.2.1 Fehlerarten, Macht

Vergleichen wir zwei Gruppen, kann ein realer Unterschied bestehen oder nicht. Unsere Teststatistik bestätigt diesen Unterschied oder nicht:

	Gibt es einen realen Unterschied?	
	ja	nein
die Statistik ist signifikant	ok	α
die Statistik ist nicht-signifikant	β	ok

In zwei Fällen sind wir zufrieden: Wenn es keinen Unterschied gibt und wir auch keinen gefunden haben und wenn es einen Unterschied gibt und wir ihn gefunden haben. In den anderen beiden Fällen machen wir einen Fehler. Wir machen einen Fehler 1. Art mit Wahrscheinlichkeit α, d. h. mit dieser Wahrscheinlichkeit verwerfen wir eine Nullhypothese, obwohl sie richtig war.

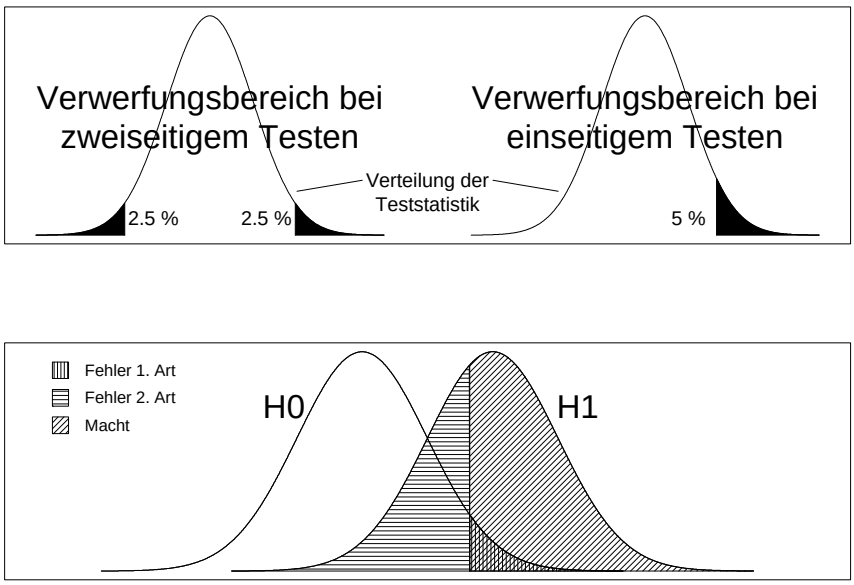


Abbildung 2: oben: Ablehnungsbereiche bei ein– und zweiseitigem Test, unten: Fehler 1. und 2. Art und die Macht unter der Verteilung der Teststatistik gemäss unserer Nullhypothese (H0) und unserer Alternativhypothese (H1) und gegebenem $\alpha = 0.05$.

Diese Grösse kennen wir als Irrtumswahrscheinlichkeit, die im Allgemeinen möglichst klein sein soll. Wir können aber mit Wahrscheinlichkeit β auch ein nicht–signifikantes Resultat erhalten, obwohl eigentlich ein Unterschied vorhanden ist. Mit Macht bezeichnen wir die Grösse $1 - \beta$, die Wahrscheinlichkeit einen Unterschied zu finden, wenn er vorhanden ist. β lässt sich auch als die Wahrscheinlichkeit der Teststatistik unter der Alternativhypothese ausdrücken (Veranschaulichung in Abb. 2, unten).

Überlegung 1.3 *Welche Information(en) braucht man, um die Macht ausrechnen zu können. Wieso wird das wohl so selten gemacht?*

1.3 t–Test

Für die Berechnung des t–Testes vgl. nächstes Kapitel. Beachtet, dass es verschiedene t–Tests gibt für gepaarte und ungepaarte Situationen und für Gruppen mit gleicher oder unterschiedlicher Varianz.

Der t–Test ist ein einfacher Test, der gut geeignet ist, die Berechnung einer Testgrösse vorzuführen. Er ist in der Praxis aber nie zu empfehlen, da er annimmt, dass die Daten innerhalb der Gruppen normalverteilt sind. Alternativ können wir nicht–parametrische Tests benutzen (siehe unten), die nur eine ungefähr symmetrische Verteilung verlangen. Diese Tests sind etwas weniger effizient als der t–Test, wenn normalverteilte Daten vorliegen, d. h. sie finden einen Unterschied nicht ganz so gut. Sie sind aber viel effizienter bei nicht-normalverteilten Daten. Zum Problem der Abschätzung, ob Daten normalverteilt sind oder nicht, vgl. Kapitel über die Residuenanalyse (Regression).

1.4 Vorzeichentest

Beim Vorzeichen– oder Binomialtest betrachten wir nur das Vorzeichen der Differenz unserer gepaarten Werte. Im Beispiel sind dies acht plus und zwei minus (vgl. Tabelle 1). Die Nullhypothese besagt, dass bei jeder Differenz plus und minus zufällig mit gleicher Wahrscheinlichkeit zu finden sind ($p = q = 0.5$). Wir müssen berechnen, wie gross die Wahrscheinlichkeit des gefundenen Resultates unter Annahme der Nullhypothese ist. Dazu benützen wir die Binomialverteilung:

$$\begin{aligned} P[\textbf{beobachtete} + \text{extremere Anzahlen}] &= P[0,\ 1,\ \textbf{2},\ \textbf{8},\ 9 \text{ oder } 10 \text{ mal ein } +] \\ &= 2 \cdot P[0,\ 1 \text{ oder } \textbf{2} \text{ mal ein } +] \\ &= 2 \cdot \left[\begin{pmatrix} 10 \\ 0 \end{pmatrix} 0.5^{10} + \begin{pmatrix} 10 \\ 1 \end{pmatrix} 0.5^{10} + \begin{pmatrix} 10 \\ 2 \end{pmatrix} 0.5^{10} \right] \\ &= 2 \cdot [1 \cdot 0.5^{10} + 10 \cdot 0.5^{10} + 45 \cdot 0.5^{10}] \\ &= 0.11 \end{aligned}$$

Da sowohl plus und minus mit der gleichen Wahrscheinlichkeit von 0.5 auftreten, konnten wir uns in der obigen Rechnung auf eine Seite konzentrieren und die resultierende Wahrscheinlichkeit mit zwei multiplizieren (2. Zeile). Das Resultat bedeutet, dass dem Vorzeichentest zufolge auf einem Niveau von 5% kein signifikanter Unterschied besteht (da $0.11 > 0.05$). Dies ist das Resultat des zweiseitigen Testes. Bei $p = q = 0.5$ können wir das einseitige Resultat erhalten, indem wir unser Ergebis wieder durch zwei teilen. Der Unterschied zwischen den beiden Jahren ist noch immer nicht signifikant, aber sehr knapp ($p = 0.055$, zu einseitigem Teste siehe auch weiter oben).

1.5 Randomisierungstest

Irgendwie ist es ja schade, dass wir unser Wissen über die Grösse der Differenzen nicht berücksichtigen konnten. Das wollen wir mit einem Randomisierungstest tun, aber trotzdem nicht eine gegebene (Normal–)Verteilung für die Differenzen annehmen, wie das der t–Test macht. Wir wollen uns sozusagen die Verteilung einer Teststatistik aus den Daten geben lassen. Dies geht natürlich nur – und das ist eine grosse Einschränkung der Randomisierungsteste – wenn die Daten tatsächlich eine gute Stichprobe aus der wahren Verteilung sind, d. h. diese gut beschreiben.

Wie beim Binomialtest sagen wir uns für die Nullhypothese, dass es Zufall ist, ob wir ein positives oder negatives Vorzeichen der Differenz erhalten. Die Nullhypothese besagt also, dass wir bei jeder Differenz sowohl ein positives als auch ein negatives Vorzeichen finden könnten. Dies ergibt $2^N = 2^{10} = 1024$ Möglichkeiten. Für die Teststatistik berechnen wir jeweils die Summe der Differenzen mit positivem und die Summe der Differenzen mit negativem Vorzeichen; im Beispiel 0.09 für die negativen und 1.33 für die positiven Differenzen (Tabelle 1). Als Teststatistik benutzen wir den kleineren der beiden Werte. Die Verteilung dieser Werte sehen wir in Abb. 3 (links) mit dem beobachteten Wert. Wir finden insgesamt 18 Fälle, deren Teststatistik kleiner oder gleich derjenigen aus unserem Beispiel sind (≤ 0.09). D. h., dass unsere Irrtumswahrscheinlichkeit $18/1024 = 0.017$ beträgt. Hier finden wir also einen signifikanten Unterschied zwischen den Jahren. Dies ist leicht einzusehen, denn die positiven Unterschiede sind viel grösser als die negativen (vgl. Tabelle 1).

Solche Randomisierungstests können auch auf die Situation von ungepaarten Stichproben mit unterschiedlichen Stichprobenumfängen in den beiden Gruppen und auf noch kompiziertere

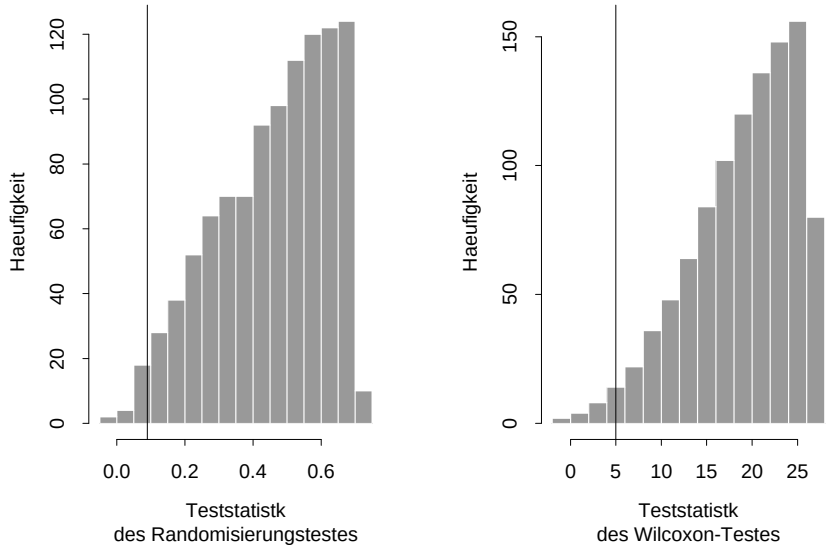


Abbildung 3: Verteilung der Teststatistik für den Randomisierungs (links) und den Wilcoxon Test (rechts). Die vertikale Linie zeigt den beobachteten Wert an.

Fälle wie ANOVA und Regression übertragen werden. Wichtig dabei ist, möglichst viel von der Struktur des Problems im Test beizubehalten. Die komplizierteren solchen Methoden sind häufig unter den Namen "bootstrap" und "jacknife" zu finden.

Überlegung 1.4 *Siehst Du was solche Randomisierungstest sehr beliebt macht? Kannst Du Dir vorstellen, was an ihnen problematisch ist?*

1.6 Rangtest

Beim Rangtest wird genau das Gleiche gemacht wie beim Randomisierungstest, ausser, dass die Differenzen zuerst rangiert werden (vgl. Tabelle 1), bevor die Rangsummen für das positive und das negative Vorzeichen berechnet werden. Auch hier gibt es 1024 Möglichkeiten und wir finden 20 Fälle, die gleich oder extremer sind als unsere Beobachtung. Unsere Irrtumswahrscheinlichkeit ist dann $20/1024 = 0.019$, also immer noch signifikant, aber nicht mehr ganz so stark, wie beim Randomisierungstest. Dies ist so, weil der Rangtest die extrem grossen Differenzen nicht so stark gewichtet, da er nur mit deren Rängen arbeitet.

Im Vergleich zum Randomisierungstest müssen wir bei den Rangtests nicht mehr annehmen, dass die Beobachtungen die wahre Verteilung der Daten sehr gut wiederspiegelt. Nur die Abfolge der Daten (ihre Ränge) müssen richtig sein. Die Rangtests haben auch noch einen praktischen Vorteil: Die Verteilung der Teststatistik ist für ein gegebenes N nicht mehr von den Daten abhängig. Das ermöglicht es, kritische Testgrössen für ein gegebenes N einfach zu tabellieren.

1.7 Zusammenstellung der Rangtests

Wie beim t–Test erwähnt, ist die Effizienz der Rangtests fast ebenso gross wie bei ihren parametrischen Ebenbildern. Wenn also ein statistisches Problem ansteht, das von der Struktur

Tabelle 2: Die gängigen Rangtests

	RANGTESTS (nicht-parametrisch, resampling)			Param. Modelle (Beispiele)
	2 Gruppen	> 2 Gruppen	"einseitig" (Trend)	
ungepaart	Mann-Whitney-U	Kruskal-Wallis	Jonkheere	t-Test, ANOVA, Regression
gepaart	Wilcoxon	Friedman	Page	
Zus'hang	Spearman, Kendall	partielle Korrelation	—	Pearson, Regression

her von den Rangtests gelöst werden kann, gibt es keinen Grund, sich mit den Annahmen der parametrischen Tests "herumzuschlagen" (vgl. spätere Kapitel dieses Kurses). Die gängigen Rangtests und ihre Anwendung finden sich in Tabelle 2.

Überlegung 1.5 *Versuche Dir einige Probleme vorzustellen, bei denen die Struktur der nicht–parametrischen Rangtests nicht genügend ist.*

Bei Vergleichen zwischen mehr als 2 Gruppen (die Situation des Friedman– oder Kruskal–Wallis–Testes) ist man meist daran interessiert zu wissen, zwischen welchen Einzelgruppen es Unterschiede gibt. Dies ist ein Problem des multiplen Testens (vgl. Übung 1.2): Dazu sollte man immer einen spezialisierten Test benützen oder aber zuerst über alle Gruppen vergleichen. Das zweitere klärt ab, ob es zwischen irgendwelchen Gruppen einen signifikanten Unterschied gibt. Die paarweisen Einzelvergleiche zwischen allen Gruppen dienen nur noch als qualitatives Werkzeug um herauszuarbeiten, wo die Signifikanz über alle Gruppen herrührt.

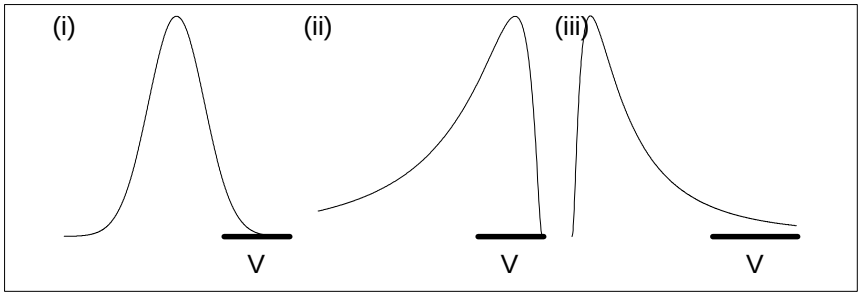
Generell wird auch (zu) oft der χ^2–Test gebraucht. Dieser ist aber sehr anfällig darauf, ob die Zähldaten unabhängig sind voneinander. Man findet meist eine Alternative, wenn man sich überlegt, was denn die eigentlichen Beobachtungseinheiten waren (z. B. Individuen).

1.8 Literaturhinweise

Eine breite, allgemeine und gut verständliche Einführung findet sich in Stahel (1995). Eine Übersicht der gängigsten Methoden mit Berechnungshinweisen findet sich in Siegel (1987), ausführlichere Methoden werden in Bortz *et al.* (1990), Siegel and Castellan (1988) und Zar (1984) beschrieben. Tufte (1999) macht sich ausserordentlich gute Gedanken zu graphischen Darstellungen.

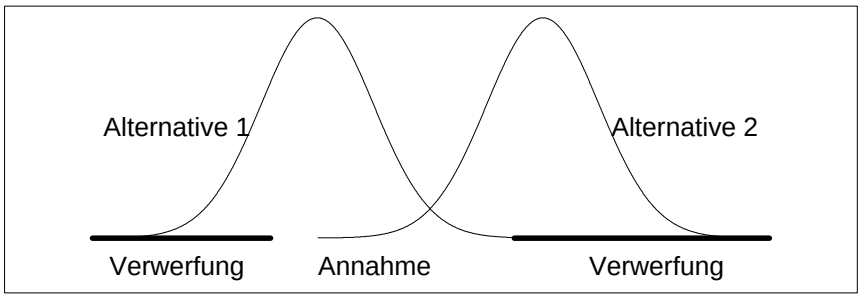
1.9 Übungen

Übung 1.1 *a. In der nachfolgenden Figur sollten die Verteilungen je einer Teststatistik unter der Nullhypothese und der Verwerfungsbereich V zum 5%–Niveau eingezeichnet sein. Welche Figur(en) könnte(n) richtig sein?*



b. In der nächsten Figur sind der Verwerfungsbereich eines Testes und die Verteilung derselben Teststatistik unter 2 Alternativen eingezeichnet.

- *Ist die Macht unter der Alternative 1 etwa 30%, 50% oder 80%?*
- *Ist die Macht unter der Alternative 2 etwa 30%, 50% oder 80%?*



c. Wie ändert sich die Macht unter einer festen Alternative, wenn man das Niveau verkleinert?

Übung 1.2 *Bei der Zusammenfassung der Rangtests wurde das multiple Testen erwähnt. Berechne wie gross die Wahrscheinlichkeit ist, dass man in 20 unabhängigen (!) Tests zufälligerweise ein oder mehrere signifikante Resultate findet (bei einer vorgegebenen Irrtumswahrscheinlichkeit von 0.05).*

Übung 1.3 *Stell Dir vor ein Kollege von Dir untersucht aggressives Verhalten bei weiblichen Schimpansen. Er beobachtet eine Gruppe in Gefangenschaft und zählt für die folgenden drei Situationen, wie oft eines von 10 Fokustieren innerhalb einer halben Minute aggressiv reagiert: wenn mindestens ein anderes adultes Weibchen (und ev. Männchen) näher als 2 m ist, wenn mindestens ein adultes Männchen (aber kein Weibchen) näher als 2 m ist und wenn sich im Umkreis von 2 m nur Jungtiere aufhalten. Die Annahme ist, dass das Aggressionspotential in dieser Reihenfolge der Situationen abnimmt.*

Der Kollege hat nun die Daten so zusammengestellt, dass er über alle Weibchen summiert und nun weiss, wie oft diese Situationen vorkamen und wie oft aggressiv reagiert wurde (eine 2 x 3 Tabelle). Er möchte nun einen χ^2–Test machen, um zu sehen, ob sich die aggressiven Reaktionen nach der Beobachtungshäufigkeit aufteilen, oder ob sich die Situationen bzgl. ihres Aggressionspotentials unterscheiden. Was sagst/rätst Du ihm?

Übung 1.4 *Johnny Cool ist Bodybuilder. Er kauft sich deswegen eine Büchse Eiweisspräparat, das in 80% aller Fälle die Muskelmasse erhöht. Dummerweise kann er (wegen seiner Hyperdark–Filtration–Sonnenbrille) diese Büchse nicht mehr von seiner Ovomaltinenbüchse (erhöht in 30% aller Fälle die Muskelmasse) unterscheiden. Er wählt deshalb willkürlich eine Büchse und testet die Hypothese, dass er das Eiweisspräparat erwischt hat.*

Der Test sieht folgendermassen aus: Er füttert seine 10 Hamster mit dem Inhalt der ausgewählten Büchse. Falls weniger als 6 von ihnen an Gewicht zunehmen, verwirft er seine Hypothese.

Welche Fehlentscheidungen sind möglich? Beschreibe in Worten. Berechne die Wahrscheinlichkeit der verschiedenen Fehlentscheidungen. Wie gross ist die Macht dieses Tests?

2 Spezielle Regression

2.1 Vorbemerkungen

In den nächsten vier Kapiteln lernen wir Erweiterungen der Regression in verschiedenste Richtungen kennen. Viele der Konzepte und Ideen liessen sich auch auf andere Analysemethoden übertragen, sind dort aber (noch) nicht ausgearbeitet und auch die hier vorgestellten Ansätze sind teilweise nur in spezialisierten Programmen implementiert. Zumindest drei dieser Verallgemeinerungen können aber auch in der Biologie sehr wichtig sein!

2.1.1 Verallgemeinerung I: Ausreisser und/oder langschwänzige Verteilungen

Oft gibt es die Situation, dass wir Daten erheben, die nicht exakt normalverteilt sind. Entweder treten eine Reihe von Ausreissern auf oder wir haben Verteilungen, die mehr Daten in den Enden der Schwänze aufweisen als eine Normalverteilung. In einer normalen Regression haben solche extremen Werte einen grossen Einfluss auf die Schätzung z. B. einer Steigung. Wir möchten nun eine Methode finden, die es uns erlaubt, trotz einer solch "verunreinigten" Normalverteilung eine gute Schätzung für (Regressions–)parameter zu finden. Gleichzeitig wird es uns diese Methode, die sogenannte "robuste Regression", auch erlauben, Ausreisser einfacher zu identifizieren.

2.1.2 Verallgemeinerung II: Nicht–parametrische Kurvenschätzung

Bei den meisten statistischen Methoden sind wir darauf beschränkt, dass wir Kurven als Linearkombination unserer Variablen darstellen (also als eine mit Parametern gewichtete Summe). Insbesondere dann, wenn die Daten nicht genau einer Geraden folgen (vielleicht sogar irgendeiner wilden Schlangenlinie), möchte man nicht von vornherein die Einschränkung eines parametrischen Modells auf sich nehmen. Verschiedene Methoden der "nicht–parametrischen Regression" ermöglichen es, einen Kurvenverlauf sozusagen von den Daten zeichnen zu lassen. Die Form der Kurve ist dann nicht eingeschränkt und wir können Kurven finden, die besser zu den Daten passen. Dies ermöglicht es auch, bessere parametrische Beschreibungen einer Kurve zu finden. D. h. dass die Methoden vor allem dann wertvoll sind, wenn man (noch) keine gute Vorahnung hat, wie eine Kurve aussehen wird und ein einfaches lineares Modell nicht passt.

2.1.3 Verallgemeinerung III: Nicht–normalverteilte Daten

Bei der Regression konnten wir mit stetigen oder kategoriellen erklärenden Variablen eine stetige Zielvariable beschreiben, die normalverteilt ist (sprich: deren Abweichungen von einer idealen Geraden normalverteilt sind). Wir waren also recht flexibel, was die erklärenden Variablen angeht, aber ziemlich eingeschränkt, was die Zielvariable angeht.

Es gibt nun eine Familie von Erweiterungen, "Verallgemeinerte lineare Modelle" (Generalized linear Models), die andere Verteilungen der Zielvariable zulassen und die die bisherige Regression als Spezialfall umfassen. Die wichtigsten Gruppen von Modellen, von denen wir eines genauer betrachten werden sind:

Poisson–Regression. Hier sind die Zielvariablen Zähldaten, können also alle ganzzahligen Werte grösser gleich Null annehmen. Wir haben bisher schon solche Daten statistisch behandelt und üblicherweise zuerst wurzeltransformiert. Die Poisson–Regression ist aber das

korrektere Modell, da es davon ausgeht, dass die Daten einer für Zähldaten "natürlichen" Poisson–Verteilung folgen.

Logistische Regression. Diese könnte man als eine Verallgemeinerung der Diskriminanzanalyse betrachten. Die Zielvariable ist auch binär (0 oder 1), wir können aber im Gegensatz zur Diskriminanzanalyse auch Interaktionen von erklärenden Variablen berücksichtigen. Diese Methode werden wir noch eingehender betrachten.

Kumulative Logits. Eine Erweiterung der logistischen Regression: Die Zielvariable ist nicht mehr binär, sondern kategoriell, mit mehr als zwei Stufen, die in eine logische Folge gebracht werden können. Wir können also untersuchen, ob gewisse erklärende Variablen dazu führen, dass wir mit grösserer Wahrscheinlichkeit eine höhere Kategorie finden. Vorstellbar wäre z. B. die Rangklasse von Primaten (tief–, mittel–, hochrangig) mit erklärenden Variablen, wie Familienzugehörigkeit, Erfolge in aggressiven Auseinandersetzungen, etc. zu modellieren.

Kategoriellle Daten. Bei den kategoriellen Daten gibt es eigentlich keine eindeutige Zielvariable mehr. Es geht hier darum zu schauen, ob die Anzahl von Merkmalskombinationen von den Merkmalsstufen abhängig ist. Meist analysiert man hochdimensionale Kontingenztafeln (d. h. der Test ist sozusagen ein χ^2–Test bei dem man mehr als zwei Variablen gleichzeitig berücksichtigen kann).

2.1.4 Verallgemeinerung IV: Nicht–lineare Modelle

Die letze Verallgemeinerung ist eine weitere Möglichkeit, den Einschränkungen der linearen Modelle auszuweichen. Hier werden beliebige Funktionen der erklärenden Variablen als Beschreibung der Zielvariablen zugelassen.

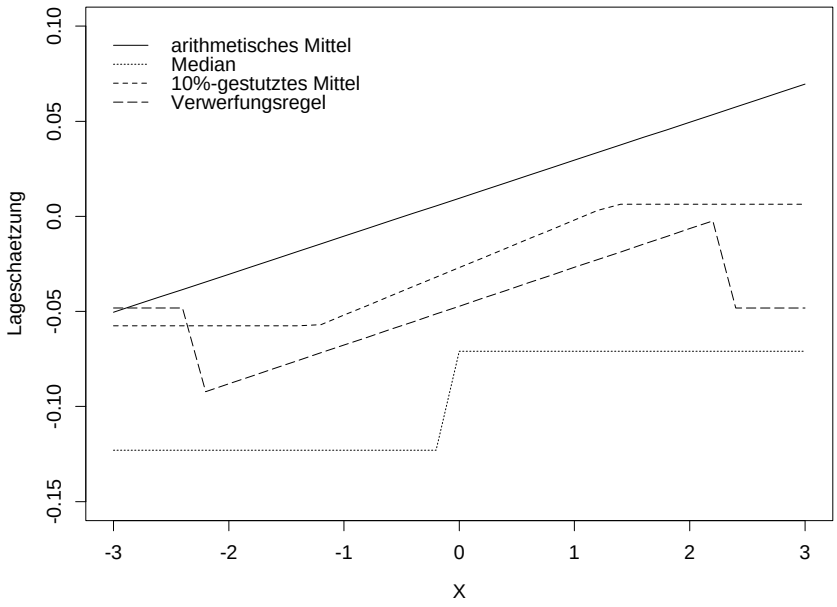


Abbildung 4: Einfluss eines veränderlichen Wertes X auf die Lageschätzung einer Stichprobe.

2.2 Robuste Regression

Die robuste Regression soll uns helfen, mit Daten zu arbeiten, deren Verteilung nicht genau einer Normalverteilung entspricht, in der also entweder lange Schwänze oder einige Ausreisser vorkommen. Trotzdem möchten wir Regressions–Parameter schätzen, die nicht allzusehr von diesen (wenigen) extremen Werten beeinflusst werden.

Die klassische Regression erzielt die optimale Schätzung der Parameter, wenn die Daten exakt normalverteilt sind. Dies ist nicht unbedingt genähert optimal für beobachtete Verteilungen in der Nähe der Normalverteilung. Robuste Schätzer sind unter exakter Normalverteilung nur genähert optimal, aber sie verhalten sich in der Umgebung der vorgesehenen Verteilung immer noch gut.

Leider sind robuste (Regressions–)Methoden in den meisten Programmpaketen noch schlecht oder gar nicht implementiert. Dies obwohl kontaminierte Beobachtungen meist zu erwarten sind! Das heisst, dass es eigentlich vorzuziehen wäre, standardmässig mit robusten Methoden zu arbeiten, gleich wie es empfehlenswert ist, mit Rangtests zu arbeiten, wenn das Problem mit ihnen gelöst werden kann. Hier sollen zwei wichtige Begriffe im Zusammenhang mit Robustheit eingeführt und einige Methoden der Regression vorgestellt werden.

2.2.1 Was heisst Robustheit? Begriffe

Einflussfunktion (influence function). Die Einflussfunktion sagt uns, inwiefern die Veränderung eines (oder mehrerer) Werte in unserer Stichprobe zu einer Veränderung unserer Schätzung eines Parameters führt.

Wir wollen dies an einem ganz einfachen Beispiel betrachten: Wir möchten die Lage einer Stichprobe schätzen (das heisst die typische Grösse eines Messwertes in der Stichprobe). Als Daten nehmen wir eine Stichprobe von 50 normalverteilten Zahlen (mit Mittelwert 0 und Varianz 1). Wir betrachten nun, wie sich verschiedene Lagemasse verändern, wenn wir eine dieser Zahlen, X, beliebig verändern (Abb. 4). Als Lagemasse betrachten wir das gewöhnliche arith-

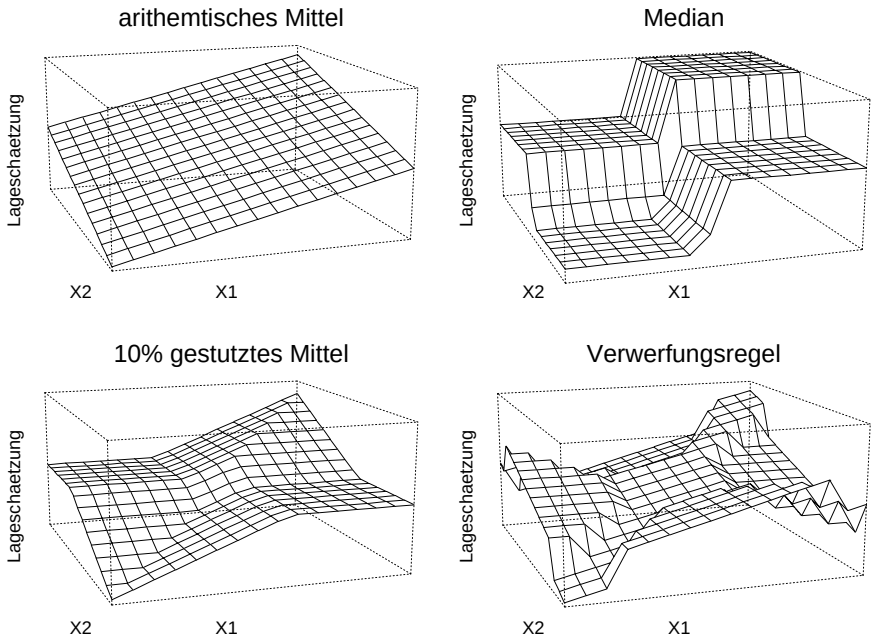


Abbildung 5: Einfluss zweier veränderlicher Wertes X1 und X2 auf die Lageschätzung einer Stichprobe.

metische Mittel, den Median, das 10%–gestutzte Mittel und eine Verwerfungsregel für Ausreisser (z. B.: $|x_i - \bar{X}|/S > 2.18$, wobei $\bar{X}$ das arithmetische Mittel und S die Standardabweichung ist).

Überlegung 2.1 *Versuche den Verlauf der vier Geraden in Abb. 4 (mindestens im Groben) zu verstehen.*

Wir sehen, dass alle Lageschätzungen ausser dem arithmetischen Mittel "beschränkt" sind, d. h. wenn wir X verändern wächst oder sinkt die Einflussfunktion nicht ins Unendliche. Daraus lässt sich eine "empirische Einflussfunktion" definieren: Die Abhängigkeit eines Schätzers von einem (fraglichen) Wert in einer Form, die unabhängig von der Stichprobengrösse ist. Als "Einflussfunktion" bezeichnet man dann diese Abhängigkeit mit einer "unendlich grossen" Stichprobe.

Bei Schätzern mit beschränkter Einflussfunktion können (nicht zu viele) Ausreisser den Schätzwert nur begrenzt beeinflussen, wie wir im Beispiel gesehen haben. Robustheitseigenschaften eines Schätzers lassen sich durch die Sensitivität (das Maximum der Einflussfunktion) beschreiben. Meist ist jedoch nur von Interesse, ob die Funktion begrenzt ist.

Bruchpunkt (breakdown point). Ein weiteres Mass für die Robustheit ist der Anteil einer Stichprobe, der einen Extremwert annehmen muss bis die Robustheit zusammenfällt. Bei einer Stichprobe von 10 Werten genügen nur 2 Ausreisser, um die relativ robusten Methoden des 10%–gestutzten Mittels und der Verwerfungsregel für Ausreisser zusammenbrechen zu lassen (Abb. 5).

Überlegung 2.2 *Versuche den Verlauf der vier Ebenen in Abb. 5 (mindestens im Groben) zu verstehen.*

Wir können uns hier natürlich sofort fragen, ob es denn nicht eine robustere Schätzung der Streuung einer Stichprobe und damit eine bessere Verwerfungsregel gibt. Tatsächlich kann die Streuung mit der Median Absolute Deviation (MAD) geschätzt werden:

$$\text{MAD}(X) = med(X_i - med(X))/0.6745$$

Das heisst wir berechnen von jedem Wert der Stichprobe die Abweichung zum Median der Stichprobe und nehmen dann den Median dieser Abweichungen. Schlussendlich teilen wir durch 0.6745, damit wir den gleichen Wert erhalten wie bei der normalen Streuung falls die Daten normalverteilt sind. Eine bessere Verwerfungsregel ist:

$$(X_i - med(X))/\text{MAD}(X) > 2.18$$

Der Bruchpunkt wird generell als maximaler Anteil der Beobachtungen definiert, die verändert werden können, während die Schätzung beschränkt bleibt ("die Schätzung nicht zusammenbricht"). Dieser Anteil ist < 0.5.

Überlegung 2.3 *Wir haben den Median in unseren Beispielen als eine sehr robuste Lageschätzung erkennen können. Der Bruchpunkt des Medians liegt denn auch beim maximal möglichen Wert von 0.5; versuche dies nachzuvollziehen.*

Die meisten Rangtests sind im Prinzip Tests zum Vergleich von Medianen und damit ausserordentlich robust: ein weiterer Grund sie dort, wo möglich, zu benützen.

2.2.2 Robuste Regression

Alle Lokationsschätzungen im vorherigen Abschnitt können als gewichtetes Mittel verstanden werden. Z. B. sind die Gewichte beim 10%–gestutzen Mittel für 10% der extremsten Werte auf beiden Seiten der Verteilung gleich null und alle anderen gleich eins. Kommen diese Gewichte von einer beschränkten Funktion, also einer Funktion mit irgendeiner Form, ähnlich wie die drei beschränkten Einflussfunktionen in Abb 4, dann haben wir eine robuste Methode vor uns.

Robuste Regressions–Methoden erlauben es, Ausreisser einfach aufzuspüren oder diese bei Interesse an den Parameterschätzungen zu vernachlässigen, da sie einen geringen Einfluss auf die Schätzung haben. *Aber* meist lohnt es sich, Ausreisser genauer zu betrachten, da diese interessante Abweichungen sein können, die uns ein Problem viel besser verstehen lassen.

In zweidimensionalen Problemen, d. h. in Situationen, in denen wir eine erklärende Variable haben, die eine Zielvariable beschreiben soll, sieht man sehr leicht, ob es Ausreisser in der Stichprobe hat und welchen Einfluss diese haben werden (Abb. 6). Ausreisser sind aber insbesondere in hochdimensionalen Problemen, d. h. in solchen, in denen viele erklärende Variablen eine Zielvariable beschreiben sollen, schwierig zu erkennen.

Es gibt vier Haupttypen von robusten Regressionen, deren Eigenschaften in Tabelle 3 aufgelistet sind.

Tabelle 3: Vier Verfahren der robusten Regression

	M–Schätzer	BIF–Schätzer	S–Schätzer	MM–Schätzer
Bruchpunkt	klein	klein	hoch	hoch
Hebelpunkte	problematisch	ok	ok	ok
Effizienz	hoch	hoch	gering	hoch
Berechnung	schnell	schnell	aufwendig	aufwendig

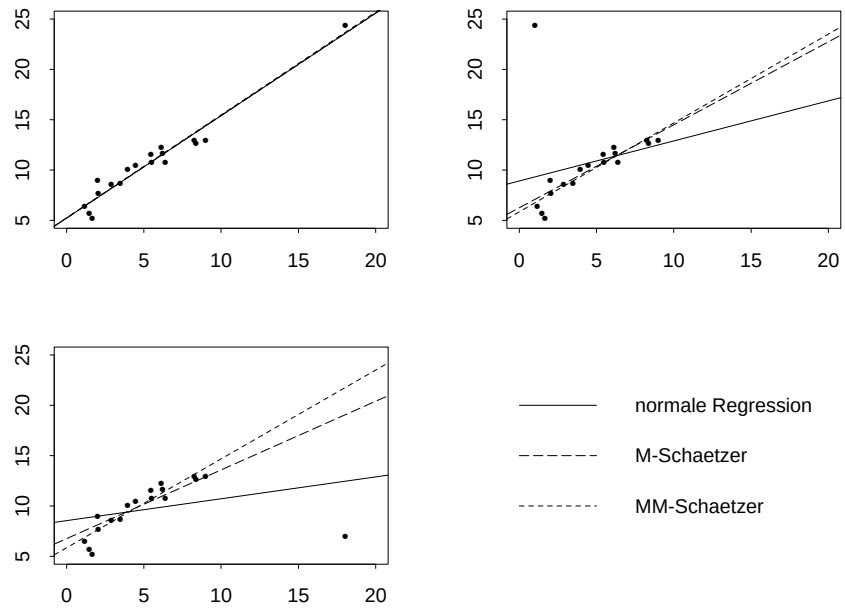


Abbildung 6: Leistung der robusten Regression bei normalverteilten Daten (oben links), bei einem Ausreisser (oben rechts) und bei einem Ausreisser mit grossem Hebelarm (unten links).

M-Schätzer (BIR). Die Regressions–M–Schätzung (auch Huber–Schätzer oder BIR: bounded influence of residuals) kann als gewichtete Kleinste–Quadrate–Schätzung verstanden werden, wobei die Gewichte von einer beschränkten Funktion kommen. Es werden nicht die Quadrate der Residuen, sondern die Summe einer anderen Funktion der Residuen minimiert. Diese robuste Regression ist sehr effizient und relativ einfach, hat aber Probleme, eine Gerade richtig zu schätzen, wenn wir es mit Hebelpunkten zu tun haben. Hebelpunkte sind solche, die weit vom Zentrum der restlichen Punkte weg liegen und daher einen grossen Einfluss auf die Schätzung der Geradenparameter haben können (Abb. 6, linke Seite). Um dem Abhilfe zu schaffen wurden andere Methoden entwickelt:

BIF M-Schätzer. Bei den sogennannten BIF Schätzern (bounding the (total) influence function) haben wir keine Probleme mehr mit den Hebelpunkten, da sie heruntergewichtet werden, wenn sie ein grosses Residuum aufweisen. Der grosse Nachteil bei dieser Methode liegt darin, dass der Bruchpunkt noch sehr klein ist. Er beträgt 1/(Anzahl zu schätzender Parameter), liegt also bei z. B. 7 Parametern schon unter 15%.

S–Schätzer, LMS. In den LMS–Schätzungen (least median of squares estimator) wird nicht mehr die Summe einer Funktion der Residuen, sondern z. B. der Median der Quadrate der Residuen minimiert. Das führt zu einer hohen Robustheit, aber leider auch zu einer schlechten Effizienz, d. h. die Modelle finden Unterschiede bedeutend schlechter als die klassischen Methoden im Falle, dass die Daten tatsächlich normalverteilt sein.

MM–Schätzer. Diese Probleme werden alle mit der MM–Schätzung gelöst (auf Kosten der Computerzeit): wir haben den hohen Bruchpunkt eines S–Schätzers, weil wir von dort unsere Startwerte nehmen, und eine hohe Effizienz, da wir mit ähnlichen Einflussfunktionen arbeiten wie beim M–Schätzer.

Übersicht. Was leisten nun diese Verfahren? Bei normalverteilten Daten gibt es (bis auf die Effizienz) keine Unterschiede in der Schätzung (Abb. 6, oben links). Haben wir jedoch einen Ausreisser (Abb. 6, oben rechts), so sehen wir, dass sowohl die M– als auch die MM–Schätzung gute Resultate liefern, während die Kleinste–Quadrate–Schätzung stark vom Ausreisser beeinflusst wird. Dies sehen wir auch bei einem Ausreisser mit grossem Hebelarm (Abb. 6, unten links): Hier liefert sogar nur die MM–Schätzung eine wirklich vernünftige Gerade.

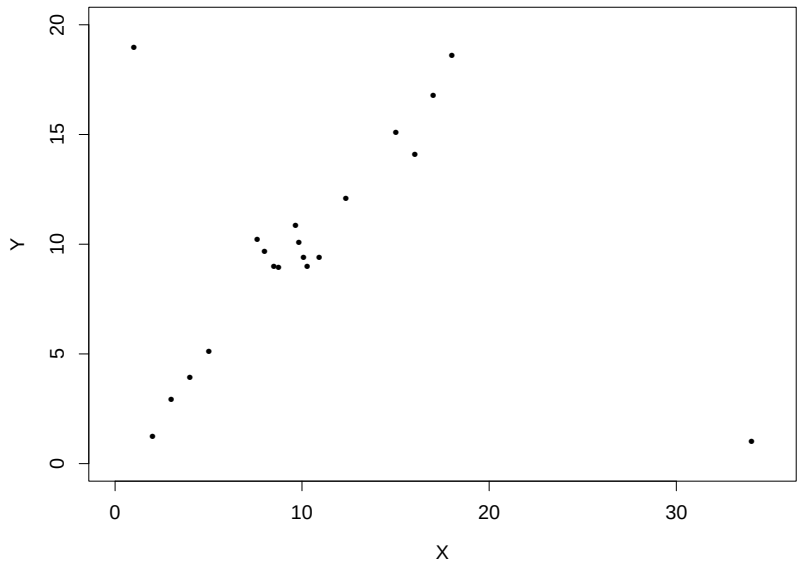
Die Residuenanalyse kann (und soll) man wie bei der gewöhnlichen Regression machen.

2.2.3 Literaturhinweise

Einführungen finden sich in Hoaglin *et al.* (1983) und dem Einleitungskapitel in Hampel *et al.* (1986). Die Klassiker, die auch den ganzen mathematischen Hintergrund mitliefern, sind die restlichen Kapitel in Hampel *et al.* (1986) und Huber (1981).

2.2.4 Übungen

Übung 2.1 *Zeichne in folgende Skizze ein, wie Du erwartest, dass die Kleinste–Quadrate, die M– und die MM–Schätzungen zu liegen kommen:*



2.3 Nicht–parametrische Regression

Das Thema der nicht–parametrischen Regression soll hier nur kurz angedeutet werden, damit Ihr ungefähr wisst, was darunter zu verstehen ist und wann eine Anwendung sinnvoll ist.

2.3.1 Wann ist eine nicht–parametrische Regression sinnvoll?

Wie bereits erwähnt, soll uns die nicht–parametrische Regression dazu verhelfen, die Form des Zusammenhangs von 2 (oder mehreren Grössen) zu sehen, ohne dass wir diese im voraus durch ein parametrisches Modell einschränken. Bevor ich etwas zu den eigentlichen Regressionsverfahren zeige, nehmen wir uns wieder einen einfacheren (eindimensionalen) Fall vor, um einige Konzepte zu erarbeiten.

Wir werden sehen, dass es sich dabei um einen Optimierungsprozess handelt: Wir möchten möglichst glatte Kurven schätzen. Je glatter aber die Kurven sind, desto grösser werden lokale Abweichungen von den Daten (ein sogenannter Bias oder systematischer Fehler). Dies bedeutet, dass wir eigentlich immer einen Kompromiss machen müssen zwischen möglichst hoher Glattheit und möglichst kleinem Bias. Wenn uns diese Methoden nur als Darstellungshilfe dienen, dann können wir diese Optimierung ohne weiteres von Auge machen. Es gibt aber auch Daumenregeln und automatische Optimierungsverfahren.

2.3.2 Glättung

Stellen wir uns wieder einen Datensatz vor: Wir zählen, an wievielen Tagen wir Individuen einer migrierenden Art an einer Futterstelle auf dem Migrationsweg sehen. Uns interessiert, wieviele Individuen jeweils gleich lange Rast machen.

Histogramm. Das wohl geläufigste Hilfsmittel zur Darstellung solcher Häufigkeiten ist das Histogramm. Aber bereits hier gehen zwei Vorgaben implizit in die Darstellung ein: der Punkt, an dem wir mit unseren Klassengrenzen beginnen und die Breite der Klassen (Abb. 7).

Die Höhe der Säulen berechnet sich aus den Gewichten der Beobachtungen: jede Beobachtung innerhalb der aktuellen Klasse hat Gewicht 1 und alle anderen Beobachtungen Gewicht 0. Diese Idee kann man nun verfeinern, indem man jedem Datenpunkt ein Gewicht gibt, das kleiner wird, je weiter man sich vom Datenpunkt entfernt. Dies führt zu sogenannten Kernschätzern:

Kernschätzer. Die Idee ist es, auf jeden Datenpunkt einen Kern zu setzen (Abb. 8), diese zu summieren und dann die Summe der Kerne als Kurve zu zeichnen (Abb. 9).

Wir wollen nun sehen, wie die Daten, die wir oben im Histogramm dargestellt haben aussehen, wenn wir sie mit verschiedenen Kernen und verschiedenen Bandbreiten (eine Eigenschaft der Kerne, die bestimmt, wie schnell die Gewichte neben den Beobachtungen abfallen) darstellen (Abb. 10).

2.3.3 Nicht–parametrische Regressionsverfahren

Diese Ansätze lassen sich auch auf zwei (und mehr) Dimensionen übertragen. Analog zur einfachen Regression sind folgende Ansätze am Gebräuchlichsten (alle basieren eigentlich darauf, dass in X–Richtung die Daten unterteilt werden, und dass man mit diesen Untergruppen etwas tut):

BIN–Glätter. Es werden in festen Fenstern Mittelwerte oder Mediane berechnet (stufenförmig wie Histogramm).

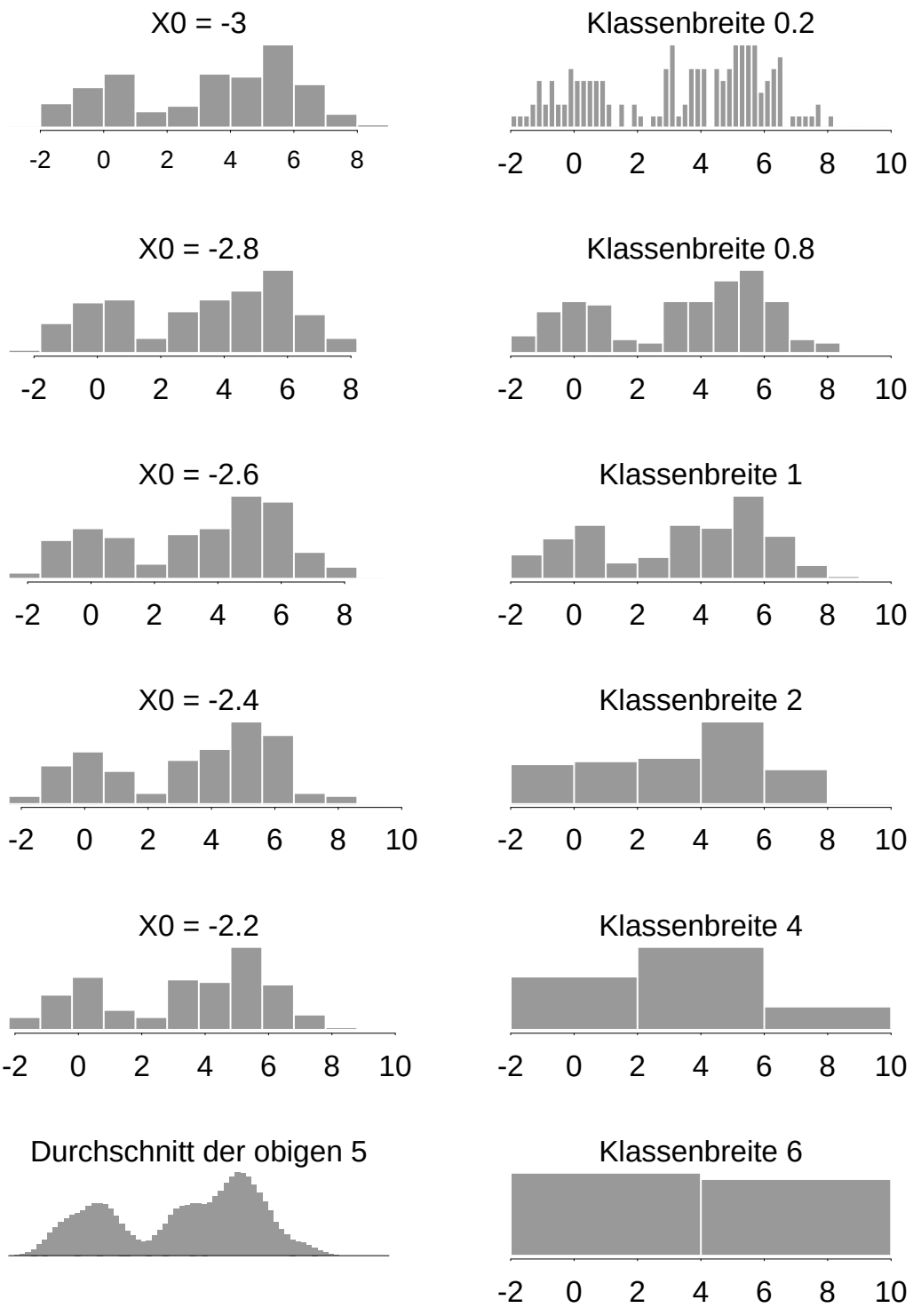


Abbildung 7: Veränderung eines Histogramms in Abhängigkeit des Ortes der Klassengrenzen (links) und der Klassenbreite (rechts).

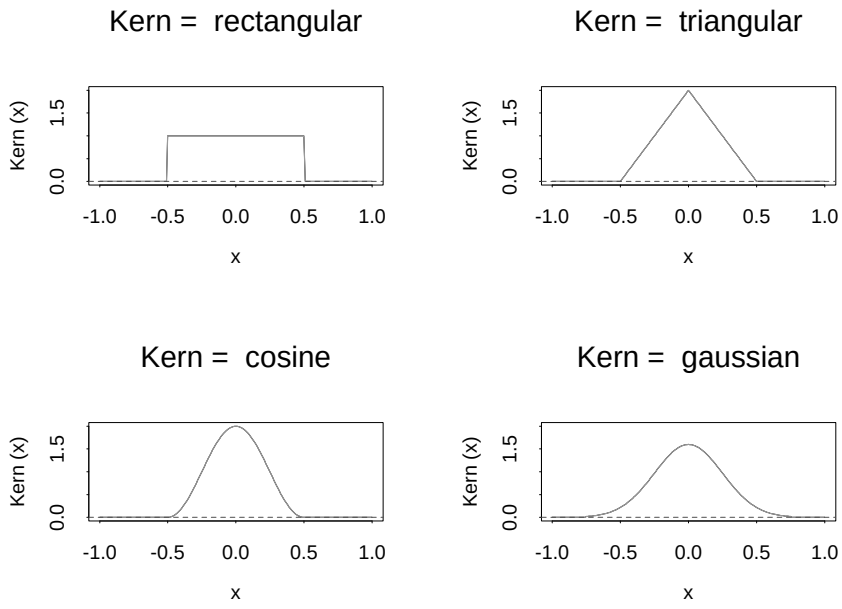


Abbildung 8: Beispiele von vier Kernformen.

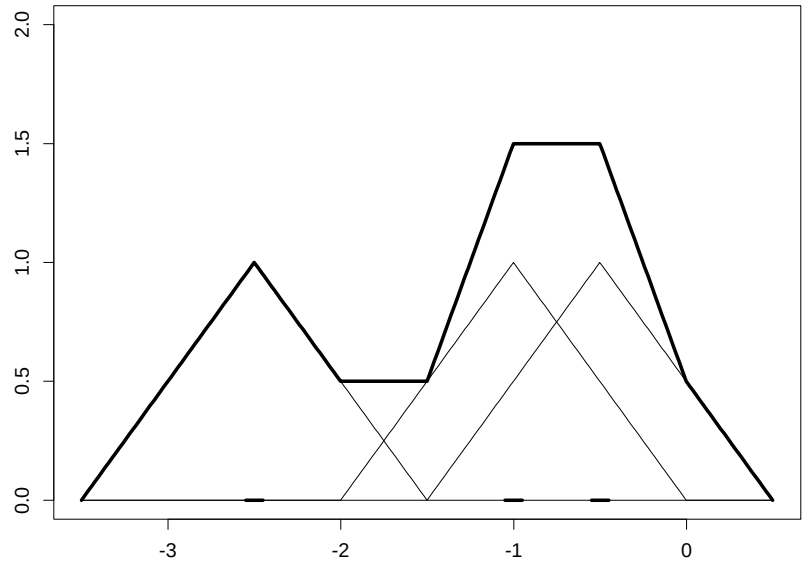


Abbildung 9: Summieren von Kernen dargestellt mit drei Beobachtungen (schwarzer Punkt jeweils im Zentrum des Dreieck–Kernes).

Running Mean. Es werden lokale Mittelwerte errechnet, dabei kann die Lokalität verschieden definiert sein:

- wähle alle Punkte, die in X–Richtung innerhalb einer bestimmten Distanz liegen,
- wähle gleichviele Punkte links und rechts,
- wähle die nächsten k Punkte unabhängig davon, ob sie links oder rechts liegen,
- gewichte die Nachbarn (Spezialfall: running median of 3).

Running Line. Es werden lokale Geraden angepasst.

LOESS. (locally weighted regression) Das gleiche, aber gewichtet.

Kerne. Ansatz wie bei Glättung.

splines. Es werden Kurven, die nicht beliebig biegbar sind (von der mathematischen Formulierung her) an die Daten angebogen (wie eine Schiffsplanke auf den Rahmen für den Rumpf).

Aus diesen Grössen lassen sich im Prinzip auch Vertrauensintervalle und Grössen ableiten, die sagen, wie gut das Modell zu den Daten passt.

2.3.4 Literaturhinweise

Weiterführende Literatur findet sich in Silverman (1986), Scott (1992) und Hastie and Tibshirani (1990).

2.3.5 Übungen

Übung 2.2 *Keine*

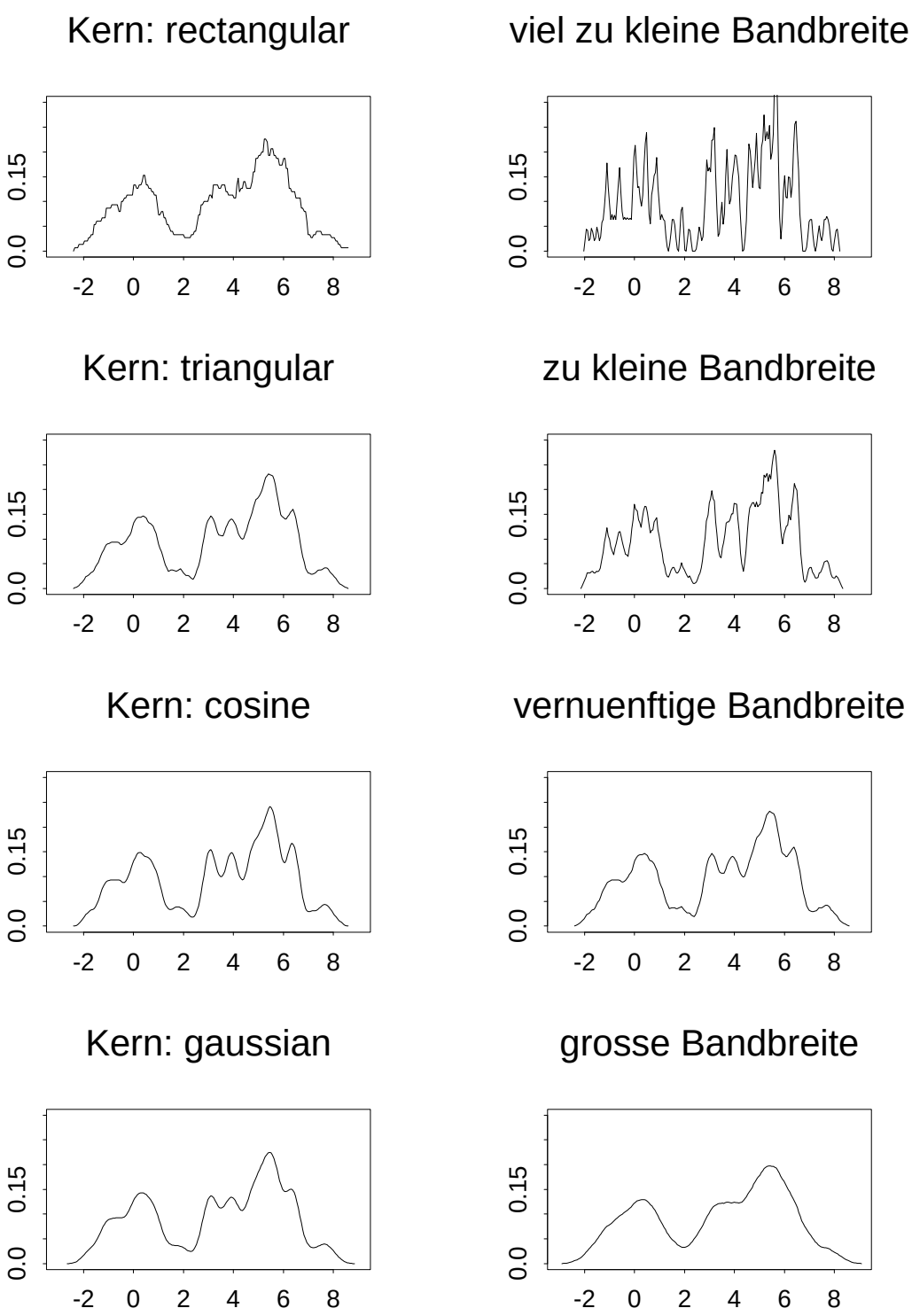


Abbildung 10: Daten von den Histogramms dargestellt mit den verschiedenen Kernen aus Abb. 8 (links) und mit verschiedenen Bandbreiten eines Dreiecks–Kerns (rechts).

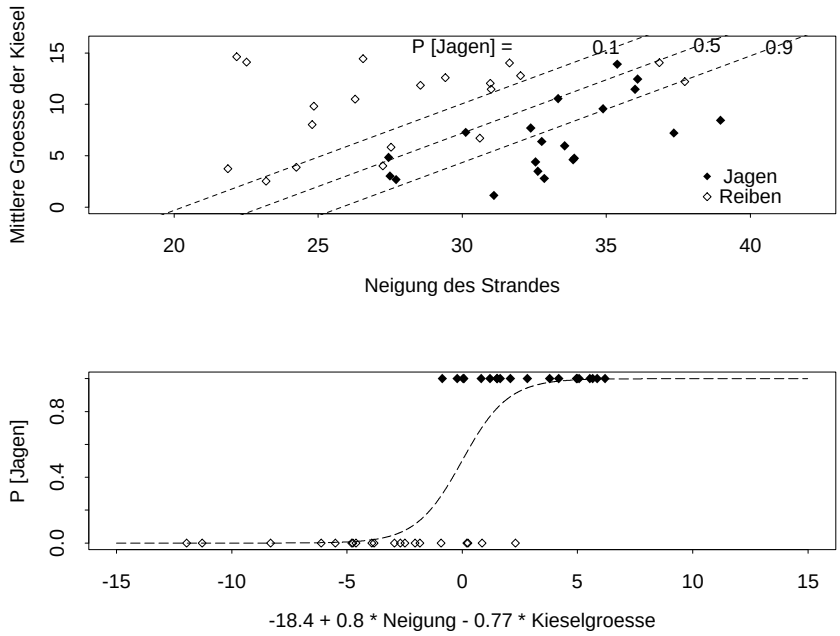


Abbildung 11: Beobachtung von Jagen und Reiben in Abhängigkeit der Strandneigung und der Grösse der Kieselsteine (oben). Die Wahrscheinlichkeiten sind das Resultat der logistischen Regression. Unten: Die Daten und die angepasste Kurve entlang der Achse, welche durch die Parameter der logistischen Regression gegeben wird.

2.4 Logistische Regression

Bei der logistischen Regression geht es um binäre Zielvariablen, d. h. solche die die Werte 0 oder 1 annehmen. Im Prinzip gehören alle Daten mit zwei Gruppen potentiell in diese Sparte von Problemen. Es kann um die Ausprägung eines zweistufigen Merkmals gehen, um ausgefallene und noch funktionierende Maschinen, kranke und gesunde Menschen oder Tiere, um Lebende und Tote, um das Auftreten von Fehlern oder das Vorhandensein eines Merkmals.

Überlegung 2.4 *Versuche Dir einige konkrete Beispiele vorzustellen, wo ein Einsatz der logistischen Regression sinnvoll wäre.*

Wieso können wir nun nicht einfach unsere klassische Regression benützen? Eine Regression mit einer Zielvariablen, die nur die Werte 0 oder 1 annehmen kann, kann als Schätzwerte "unmögliche" Werte kleiner 0 oder grösser 1 erhalten. Keine der bisherigen Transformationen kann hier Abhilfe verschaffen und ein letzter Grund ist, dass ein Modell, das die richtigen Annahmen über die Daten trifft, befriedigender ist.

Überlegung 2.5 *Was waren die Voraussetzungen für eine Diskriminanzanalyse, welche Einschränkungen hatte man dort? Was könnten wir mit der logistischen Regression gewinnen?*

2.4.1 Ein (simuliertes) Beispiel

Nehmen wir an, dass wir das Auftreten von zwei besonderen Verhalten von räuberischen Schwertwalen untersuchen wollen. Es handelt sich bei beiden Verhalten um solche, die sehr nahe am Ufer gezeigt werden: das Jagen von Robben am Strand und ein Sich–Reiben am Untergrund. Nehmen wir weiter an, wir hätten eine Reihe von verschiedenen Stränden, die unterschiedlich

stark geneigt sind und aus verschieden grossen Steinen bestehen (eine ziemlich unrealistische Annahme). Des weiteren seien unsere Beobachtungen an einem Ort durchgeführt worden, wo die Orcas nur vorbeiziehen und wir darum immer wieder andere Gruppen beobachten können (eine noch unrealistischere Annahme, normalerweise müsste man die Gruppenzugehörigkeit berücksichtigen, vgl. gepaarte und ungepaarte Tests). Wir beobachten nun unsere Strände und notieren, ob eines der beiden uns interessierenden Verhalten auftritt. Falls dies passiert, notieren wir die Neigung des Strandes und die mittlere Grösse der Steine. Wir möchten nun wissen, ob sich die Strände, an denen die beiden Verhalten gezeigt werden, unterscheiden und wenn ja, worin. Das Jagen haben wir mit 1, das Reiben mit 0 codiert (Abb. 11).

Bereits in dieser Graphik sehen wir, dass mehr gejagt wird bei steileren Stränden und mehr gerieben bei grösseren Steinen.

2.4.2 Schätzungen und Tests

In unserem Modell möchten wir als Zielvariable die Wahrscheinlichkeit dafür, dass wir eine 1 beobachten, in Abhängigkeit einer Funktion h der erklärenden Variablen setzen. Wir benutzen folgenden Ansatz:

$$\begin{aligned} P[Y=1] &= h(x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}) \\ &= \tilde{h}(\alpha + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_m x^{(m)}), \text{ wobei} \end{aligned}$$

$$\tilde{h}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} \begin{cases} \overset{\eta\to\infty}{\longrightarrow} & 1 \\ \overset{\eta\to-\infty}{\longrightarrow} & 0 \end{cases}$$

Umgekehrt lässt sich schreiben:

$$g(P[Y=1]) = \log\left(\frac{P[Y=1]}{1-P[Y=1]}\right) = \alpha + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_m x^{(m)}$$

Die Funktion g wird "logit"–Funktion genannt und die Verwandtschaft mit der bisherigen Regression lässt sich an der rechten Seite der Gleichung leicht erkennen.

Um unsere Parameter zu schätzen benützen wir die Maximum–Log–Likelihood Methode; dazu bauen wir zuerst eine Funktion, die die Wahrscheinlichkeit unserer Daten beschreiben soll:

$$\text{l–l} = log\left(\prod_{y_j=1} P[Y_j=1] \prod_{y_j=0} (1-P[Y_j=1])\right) = \sum_{y_j=1} \log(\pi_j) + \sum_{y_j=0} \log(1-\pi_j),\ \pi_j = \tilde{h}(\cdots)$$

Diese Funktion wird nun maximiert und wir finden diejenigen Parameter unter denen unser Modell am besten zu den Daten passt. Nun möchten wir natürlich noch berechnen, ob das Modell signifikant etwas zur Beschreibung der Daten beigetragen hat und ob die einzelnen Parameter signifikant verschieden von Null sind.

Die Tests, die man hier anwenden kann sind sogenannte Likelihood–Ratio–Tests, d. h. man bildet Quotienten aus den Likelihoods oder Differenzen der Log–Likelihoods von einem kleinen (mit wenigen erklärenden Variablen) und einem grossen Modell (mit vielen erklärenden Variablen), welches das kleine Modell umfasst. Man kann zeigen, dass die doppelte Differenz χ^2 verteilt ist mit der Anzahl Freiheitsgraden, die der Differenz der Freiheitsgrade der Modelle entspricht. Da man die doppelten Log–Likelihoods braucht, werden von den Statistik–Programmen meist Devianzen ausgedruckt, die genau solche doppelten Log–Likelihoods sind.

Tabelle 4: Relevanter Computer–output von S–Plus zur logistischen Regression der Orca–Verhaltensdaten (slope: Neigung des Strandes, gravel: Durchmesser der Kiesel)

volles Modell: mit Interaktion

```
                   Value Std. Error    t value
 (Intercept) -30.47725573 16.5371528 -1.8429567
       slope   1.20699345  0.5915404  2.0404243
      gravel   0.70430774  1.5859311  0.4440973
slope:gravel  -0.04742437  0.0521893 -0.9086989

    Null Deviance: 55.45177 on 39 degrees of freedom
Residual Deviance: 20.87178 on 36 degrees of freedom
```

volles Modell: ohne Interaktion

```
                  Value Std. Error   t value
(Intercept) -18.3956000  6.7485490 -2.725860
      slope   0.7977799  0.2800214  2.848996
     gravel  -0.7688496  0.2906027 -2.645707

    Null Deviance: 55.45177 on 39 degrees of freedom
Residual Deviance: 21.83712 on 37 degrees of freedom
```

Modell nur mit einer Variablen: Kieselgrösse

```
                 Value Std. Error   t value
(Intercept)  1.8196151 0.81830766  2.223632
     gravel -0.2194826 0.08961215 -2.449251

    Null Deviance: 55.45177 on 39 degrees of freedom
Residual Deviance: 48.50085 on 38 degrees of freedom
```

Modell nur mit einer Variablen: Neigung

```
                 Value Std. Error   t value
(Intercept) -9.8019397   3.291214 -2.978214
      slope  0.3200286   0.106097  3.016377

    Null Deviance: 55.45177 on 39 degrees of freedom
Residual Deviance: 41.6674 on 38 degrees of freedom
```

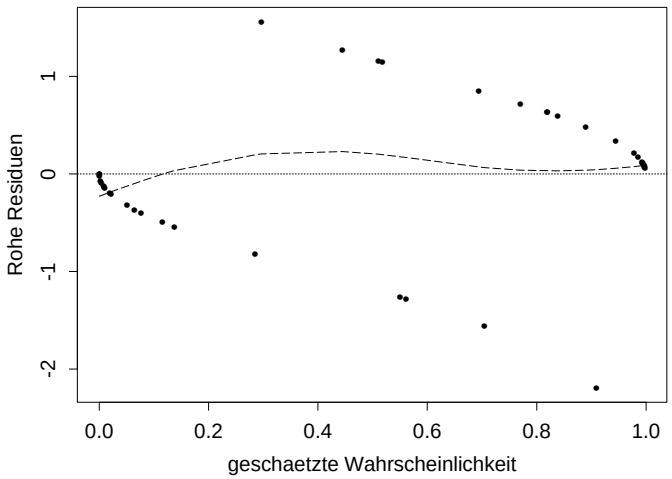


Abbildung 12: Rohe Residuen der logistischen Regression (gestrichelt: ein loess–Glätter).

Was tun wir nun konkret: In Tabelle 4 sehen wir die relevante Information einer logistischen Regression mit S–Plus. Wir sehen neben der Schätzung für die Grösse der Parameter (in der ersten Spalte) eine Schätzung deren Varianz und einen zugehörigen t–Wert. Des weiteren sehen wir die Nulldevianz, die dem kleinstmöglichen Modell entspricht, und die zum spezifischen Modell gehörende (Residual–) Devianz. Das kleinstmögliche Modell ist jenes, für das wir annehmen, dass alle unsere Variablen keinen Einfluss haben, und es somit eine konstante Wahrscheinlichkeit für die Beobachtung einer 1 gibt. Wir können nun zuerst testen, ob die Modelle überhaupt etwas erklären, indem wir jeweils die Residuen–Devianz von der Null–Devianz abziehen und mit der entsprechenden χ^2_{fg} Verteilung mit fg Freiheitsgraden vergleichen:

volles Modell mit Interaktionen	$P[\chi^2_{fg=39-36} \geq 55.45 - 20.87] < 0.0001$
volles Modell ohne Interaktionen	$P[\chi^2_{fg=39-37} \geq 55.45 - 21.84] < 0.0001$
Modell mit nur einer Variablen: Kieselgrösse	$P[\chi^2_{fg=39-38} \geq 55.45 - 48.50] = 0.0084$
Modell mit nur einer Variablen: Neigung	$P[\chi^2_{fg=39-38} \geq 55.45 - 41.67] = 0.0002$

Wir sehen an den Signifikanzen, dass alle Modelle mehr bringen, als wenn wir eine konstante Wahrscheinlichkeit annehmen würden. Um zu sehen, welches Modell nun genügt, müssen wir zwischen den Modellen vergleichen:

ohne Interaktion versus mit Interaktion	$P[\chi^2_{fg=37-36} \geq 21.84 - 20.87] = 0.326$
nur Kieselgrösse versus mit beiden Variablen	$P[\chi^2_{fg=38-37} \geq 48.50 - 21.84] < 0.0001$
nur Neigung versus mit beiden Variablen	$P[\chi^2_{fg=38-37} \geq 41.67 - 21.84] < 0.0001$

Die Nicht–Signifikanz des ersten Testes sagt uns, dass es nichts bringt, wenn wir noch die Interaktion ins Modell nehmen. Die beiden andern Signifikanzen zeigen jedoch, dass es beide der Variablen braucht. Diese Information kann man (asymptotisch) auch an den t–Werten in Tabelle 4 sehen. Am Vorzeichen der Parameter erkennen wir, dass die Wahrscheinlichkeit, Jagen zu beobachten, mit zunehmender Neigung steigt und mit zunehmender Kieselgrösse sinkt. Eine mögliche Interpretation wäre, dass die Orcas beim Jagen besser ins Wasser zurückgelangen können, wenn der Strand steiler ist, aber grössere Steine zum Reiben bevorzugen.

2.4.3 Einige weitere Bemerkungen

Auch in diesen Modellen sollten wir die Residuen anschauen, aber es ist nicht mehr so klar, was wir erwarten sollten, auch gibt es keine durch Diskussion ausgereiften und implementierten Methoden. Ein wichtiger Plot ist noch immer der Tukey–Anscombe–Plot, bei dem wir die Residuen versus den geschätzten Wert auftragen. Um zu sehen, ob der Erwartungswert ungefähr null beträgt, sollte man unbedingt einen Glätter einzeichnen, da Artefakte in diesen Plots entstehen: die Residuen liegen auf zwei Geraden, (Abb. 12). Es gibt auch noch andere Varianten für die Residuen, aber vorläufig bleibt es bei ad–hoc Methoden in der Auswertung von logistischen Regressionen.

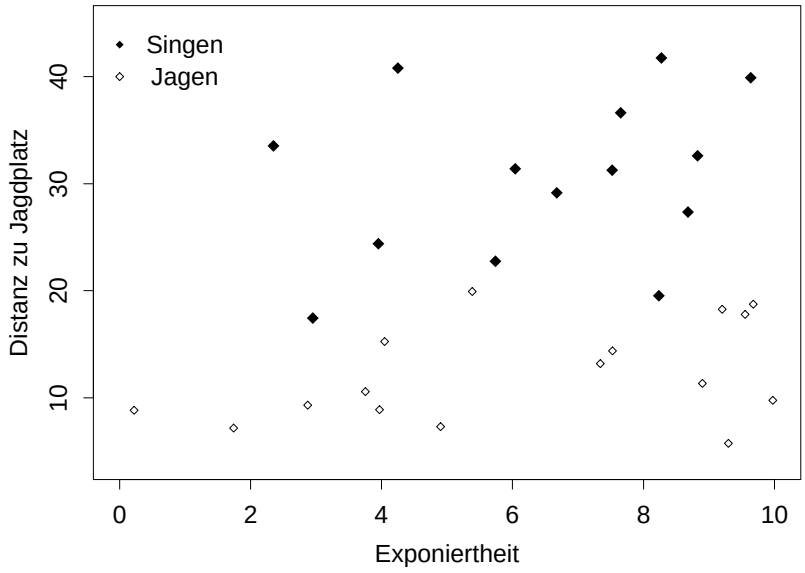
Oft werden Ansätze wie Diskriminanzanalyse und logistische Regression dazu verwendet, Orte, an denen man Tiere gefunden hat mit zufällig ausgewählten Orten zu vergleichen. Es gibt kaum bessere Ansätze, aber man sollte sich immer bewusst sein, dass die logistische Regression darauf ausgelegt ist, Gruppen zu unterscheiden, die sich möglichst wenig überlappen, während Orte, an denen Tiere gefunden werden, immer ein Teilbereich des Angebotes sind (sich die Fundorte also völlig mit gewissen Zufallsorten überdecken).

2.4.4 Literaturhinweise

Weitere Informationen über Generalisierte Lineare Modelle finden sich in McCullagh and Nelder (1989) und speziell im Bezug auf S–plus in Venables and Ripley (1997).

2.4.5 Übungen

Übung 2.3 *Wir beobachten Männchen einer Vogelart und möchten herausfinden, ob sich Orte, an denen der Vogel singt, von solchen unterscheiden, von denen aus er jagt. Wir glauben, dass sich die Orte allenfalls durch ihre Exponiertheit (expo) und die Distanz (dist) zu den Jagdplätzen unterscheiden. Die Daten sind in der folgenden Graphik dargestellt. Zeichne die Linie ein, wo Du etwa erwartest, dass die Wahrscheinlichkeiten einen singenden oder jagenden Vogel zu sehen gleich sind. Was hast Du nun für Erwartungen an die logistische Regression?*



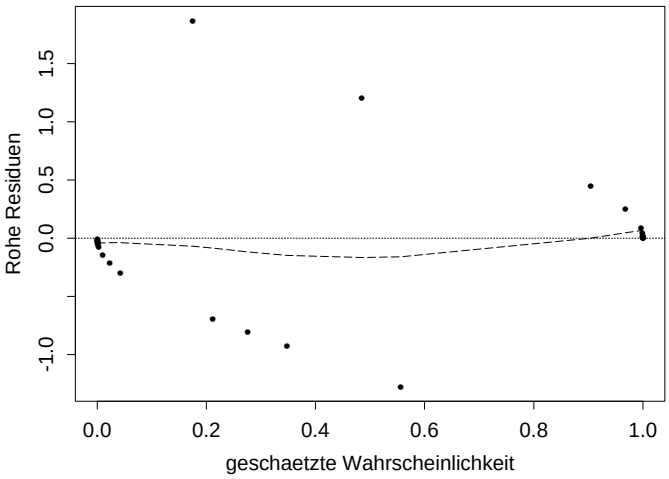
Hier folgt nun eine Auflistung des relevanten Computer–outputs. Welche Modelle und welche Variablen sind signifikant? Wie lässt sich das Interpretieren? Wurden Deine Erwartungen von oben erfüllt?

```
mit Interaktion
                            Value Std. Error    t value
(Intercept) -7.09179695  9.3122856 -0.7615528
       expo -2.12331232  3.1251751 -0.6794219
       dist  0.50733209  0.5764607  0.8800809
  expo:dist  0.08855432  0.1687055  0.5249049

    Null Deviance: 41.4554 on 29 degrees of freedom
Residual Deviance: 6.762643 on 26 degrees of freedom

ohne Interaktion
                             Value Std. Error   t value
(Intercept) -12.3859016  8.6651672 -1.429390
       expo  -0.5137501  0.4054912 -1.266982
       dist   0.8137422  0.4976112  1.635297

    Null Deviance: 41.4554 on 29 degrees of freedom
Residual Deviance: 7.019384 on 27 degrees of freedom

nur mit expo
                         Value Std. Error   t value
(Intercept) -0.42159164  0.9302550 -0.4532001
       expo  0.04560032  0.1350489  0.3376578

    Null Deviance: 41.4554 on 29 degrees of freedom
Residual Deviance: 41.34084 on 28 degrees of freedom

nur mit dist
                         Value Std. Error   t value
(Intercept) -14.0025294  8.0586952 -1.737568
       dist   0.7135532  0.4233168  1.685625

    Null Deviance: 41.4554 on 29 degrees of freedom
Residual Deviance: 8.979398 on 28 degrees of freedom
```

Einige kritische Werte der χ^2_{fg}–Verteilung:

fg	$p = 0.05$	$p = 0.01$	$p = 0.001$
1	*3.84*	*6.63*	*10.83*
2	*5.99*	*9.21*	*13.82*
3	*7.82*	*11.35*	*16.27*

Was sagst Du zu den Residuen?



2.5 Nicht–lineare Regression

Auch das Thema der nicht–linearen Regression soll hier nur kurz angedeutet werden. Wieder geht es darum, dass Ihr ungefähr wisst, was darunter zu verstehen ist und wann eine Anwendung sinnvoll ist.

2.5.1 Wann wird nicht–lineare Regression gebraucht?

Bisher war die Zielgrösse immer eine lineare Funktion der erklärenden Variablen der Form

$$Y_i = \alpha + \beta^{(1)} X_i^{(1)} + \beta^{(2)} X_i^{(2)} + \beta^{(3)} (X_i^{(2)})^2 + \beta^{(4)} \sin(X_i^{(1)}) + \epsilon_i,$$

d. h. es können zwar Funktionen der erklärenden Variablen, wie Potenzen oder der Sinus, in der Formel vorkommen, die generelle Form aber ist eine gewichtete Summe. In den meisten Fällen ist dies auch genügend, insbesondere da, wo man einfach daran interessiert ist, ob eine gewisse erklärende Variable einen Einfluss auf eine Zielvariable hat oder nicht, ohne dass man im Detail daran interessiert ist, welche Form dieser Einfluss hat.

Es kann nun aber vorkommen, dass man aus der Theorie vermutet, dass die Daten einer bestimmten Formel folgen, die man nicht als Linearkombination darstellen kann. Da hilft die nicht–lineare Regression, bei der Variablen auf beliebige Weise miteinander verknüpft werden können. Einfache Beispiele wären das exponentielle (unlimitierte) oder das logistische (limitierte) Wachstum einer Population:

$$\frac{dN}{dt} = gN - sN \rightarrow N(t) = N_0 e^{(g-s)t} \text{ exponentielles Wachstum und}$$

$$\frac{dN}{dt} = rN(1 - \frac{N}{K}) \rightarrow N(t) = \frac{N_0 K e^{rt}}{K + N_0(e^{rt} - 1)} \stackrel{t \rightarrow \infty}{\longrightarrow} K \text{ logistisches Wachstum,}$$

wobei N die Populationsgrösse, N_0 die Populationsgrösse zur Zeit $t = 0$, g die Geburtsrate, s die Sterberate, K die Umweltkapazität und r die Rate, mit der sich die Population K nähert, bezeichnen. Wir können uns also einen Datensatz vorstellen, in dem wir Daten über die Populationsgrösse einer Art $N(t)$ zu verschiedenen Zeitpunkten t haben. Wir fragen uns nun, welches Modell zu dieser Wachstumskurve passt und wollen die restlichen Grössen schätzen (N_0, g, s, r, K).

Weil die Variablen in der nicht–linearen Regression beliebig miteinander verknüpft werden können, gibt es unendlich viele Möglichkeiten solcher Modelle, die an Daten angepasst werden könnten. Daher ist es für diese Art von Modellen unumgänglich, dass die Fomulierung des Modelles aus der Theorie abgeleitet wird.

Überlegung 2.6 *Aus welcher Theorie wurden wohl die Modelle des exponentiellen und des logistischen Wachstums abgeleitet?*

2.5.2 Schätzung der Parameter (Algorithmus), Überprüfung der Voraussetzungen

In den einfachsten Fällen, wie z. B. beim exponentiellen Wachstum, lässt sich ein nicht–lineares System durch Logarithmieren linearisieren:

$$
\begin{array}{ccc}
N(t) = N_0 e^{(b-d)t} & \xrightarrow{\log} & \ln(N(t)) = \ln(N_0) + (b-d)t \\
\downarrow & \text{sind von der Form} & \downarrow \\
Y_i = \theta_1 e^{\theta_2 t_i} + \epsilon_i & & \tilde{Y}_i = \theta_1 + \theta_2 t_i + \tilde{\epsilon}_i \\
& \text{Rücktransformierung} & \downarrow \\
& & Y_i = \tilde{\theta}_1 e^{\theta_2 t_i} e^{\tilde{\epsilon}_i} = \tilde{\theta}_1 e^{\theta_2 t_i} \epsilon_i
\end{array}
$$

Wir können sehen, dass die Fehler beim ursprünglichen Modell (in der linken Spalte) additiv sind und es wird angenommen, dass sie normalverteilt sind. Dies ist auch der Fall im logarithmierten System (zweite Zeile in der rechten Spalte). Wenn wir aber eine Rücktransformation machen, um wieder die ursprünglichen Daten, statt deren Logarithmus zu erhalten, sehen wir, dass der Fehler multiplikativ und $\epsilon = e^{\tilde{\epsilon}}$ sogenannt log–normalverteilt ist, wenn ϵ normalverteilt ist. Je nach System kann also nur das eine oder andere richtig sein. Was richtig ist, und ob ein Modell passt, kann man wie anhin durch das Analysieren der Residuen herausfinden.

Falls wir das System linearisieren können, verwendet man zur Schätzung der Parameter die bisher bekannten Methoden der Regression. Ist eine Linearisierung nicht möglich, kommen die speziellen Methoden der nicht–linearen Regression zum Tragen. Dabei werden die Parameter in einem iterativen Kleinste–Quadrate–Verfahren geschätzt.

Da es sich um ein iteratives Verfahren handelt, brauchen wir Startwerte für die Parameter. Diese müssen entweder bekannt sein oder aus den Daten (graphisch) geschätzt werden. Zusätzlich können gewisse Parameter meist nur Werte in gewissen Bereichen annehmen (Einschränkungen). Da es sich um ein numerisches Verfahren handelt, ist es auch möglich, dass die Iterationen nicht auf bestimmte Parameterwerte konvergieren. In diesem Fall kann man versuchen die Gleichung so umzuformen, dass andere Parameter entstehen (Reparametrisierung).

Das Verfahren nähert "die p–dimensionale (Anzahl Parameter) Hyperfläche lokal durch eine Hyperebene an", d. h. durch die Gleichung wird eine gekrümmte Fläche dargestellt, die lokal durch eine Ebene angenähert wird. Wie gut diese Annäherung ist, lässt sich wieder graphisch überprüfen, mit sogenannten t–Plots und Profilspuren. Die Qualität dieser Approximation ist besonders wichtig für die Konfidenzintervalle der Parameter, d. h. die Tests, ob die Parameter null sein könnten.

2.5.3 Literaturhinweise

Eine Zusammenstellung von Anwendungen, Beispielen und Theorie findet sich in Bates and Watts (1988).

2.5.4 Übungen

Übung 2.4 *Findest Du noch andere Beispiele aus dem Bereich der Biologie, wo Du nicht–lineare Systeme erwarten würdest?*

3 Literatur

References

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis & Its Applications*. John Wiley & Sons, New York.

Bortz, J., Lienert, G. A., and Boehmke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik*. Springer–Verlag, Berlin.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. John Wiley & Sons, New York.

Siegel, S. (1987). *Nichtparametrische Statistische Methoden*. Fachbuchhandlung für Psychologie, Eschborn bei Frankfurt am Main.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw–Hill, New York, 2nd edition.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Stahel, W. (1995). *Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler*. Vieweg, Braunschweig/Wiesbaden.

Tufte, E. R. (1999). *The Visual Display of Quantitative Information*. Graphic Press, Cheshire, Connecticut, 17th printing edition.

Venables, W. N. and Ripley, B. D. (1997). *Modern Appplied Statistics with S–PLUS*. Springer, New York, second edition.

Zar, J. L. (1984). *Biostatistical Analysis*. Prentice Hall, NJ.