Macht Modellieren im Streudiagramm Sinn?

JOACHIM ENGEL, LUDWIGSBURG

Zusammenfassung: *Welche Interpretation hat die aus einem Streudiagramm erhaltene Funktion zur Modellierung des funktionalen Zusammenhangs zweier Variabler? Ausgehend von Beispielen wird argumentiert, dass Modellbildung im Streudiagramm nur dann sinnvoll ist, wenn die probabilistischen Voraussetzungen exakt spezifiziert sind. Weit verbreitete Verfahren wie lineare Regression als Instrument der beschreibenden Statistik (und nicht der Inferenzstatistik) zu sehen, kann zu verhängnisvollen Missverständnissen und Fehlschlüssen führen.*

1 Einleitung

Die Modellierung von Streudiagrammdaten gehört zu den zentralen Themen der angewandten Statistik, auf die kaum ein einführendes Lehrbuch verzichtet. Ausgangspunkt sind n Messpaare (x_1, y_1), ..., (x_n, y_n), die im kartesischen Koordinatensystem als Punkte dargestellt sind. Die Zielsetzung ist eine einfache funktionale Beschreibung des Zusammenhangs zwischen zwei Variablen. Die in der Punktwolke der Ausgangsdaten erhaltene Struktur soll zu einer Funktion $y = f(x)$ komprimiert werden. Dabei zieht man in Betracht, dass die Beobachtungen aufgrund von Messfehlern und Stichprobenfehlern verrauschte Werte einer idealtypischen Variablen sind, der das eigentliche Erkenntnisinteresse gilt. Die Daten werden somit als n Realisierungen eines Paares von Zufallsvariablen (X, Y) aufgefasst. In den meisten Anwendungen ist eine Schätzung der *mittleren Regressionsfunktion* das Ziel, wenn der Zusammenhang zweier Variabler modelliert werden soll. Darunter versteht man den bedingten Erwartungswert der Zufallsvariablen Y gegeben $X = x$: $f(x) = E(Y \mid X = x)$, d.h. die Funktion f gibt den zu erwarteten Wert von Y an, vorausgesetzt die Zufallsvariable X hat den Wert x angenommen. Aber auch andere Ziele wie z.B. die Schätzung einer bedingten Mediankurve sind möglich und sinnvoll (Engel 1998).

Mathematisch sind eine Reihe von elementaren Techniken zum Modellieren im Streudiagramm möglich:

1 Geradenanpassung per Augenmaß

In einer allerersten Annäherung können Schüler zunächst probieren, per Augenmaß eine Gerade an die Punktwolke anzupassen.

2 Objektiver als eine Geradenanpassung per Augenmaß ist die lineare Regression. Diese Methode geht schon auf Gauß zurück und basiert auf der Minimierung eines quadratischen Abstandskriteriums. Die Berechnung von Steigung und y-Achsenabschnitt der Regressionsgerade kann auch ohne Differentialrechnung erfolgen (Hui 1988, Vohmann 1988).

3 Ein robusteres und rechentechnisch leichter durchzuführendes Konzept ist die 3-Schnitt Mediangerade (siehe z.B. Polasek 1995, Engel & Theiss 2001).

4 Wenn der Datensatz keinem linearen Trend folgt, ist eine Geradenanpassung keine angemessene Form der Modellierung. Dann kann versucht werden, eine Funktion aus einer anderen Funktionenklasse (Polynome vorgegebenen Grades, Exponentialfunktion, logistische Funktion etc.) anzupassen. Da die Funktionen der betrachteten Klasse durch einen (auch mehrdimensionalen) Parameter gekennzeichnet sind, spricht man hier von nicht-linearer parametrischer Regression. Auch wenn diese Idee eine direkte Weiterführung der Geradenanpassung ist, sind - je nach Funktionenklasse - die hier verlangten Berechnungen zur Bestimmung des optimalen Parameters sehr kompliziert. Spätestens hier ist der Einsatz von Software gefragt, mit deren Hilfe der beste Parameter bestimmt werden kann.

5 In manchen Situationen gelingt es, die nichtlinearen Daten (x_i, y_i) durch eine einfache monotone Transformation in eine Punktwolke mit linearer Struktur abzubilden: $x_i^* = T(x_i)$, $y_i^* = S(y_i)$. Dann kann man an die transformierten Daten eine Gerade anpassen und das Ergebnis zurücktransformieren, um den Zusammenhang der Ausgangsdaten zu modellieren. Erhält man für die transformierten Daten den linearen Zusammenhang $y^* = a \cdot x^* + b$, so lassen sich die Ausgangsdaten modellieren mittels

$$y = f(x) = S^{-1}\left(a \cdot T^{-1}(x) + b\right).$$

Ob es nun angemessen ist, eine Gerade oder eine Funktion einer anderen vorgegebenen Klasse von Funktionen an eine Datenwolke anzupassen, lässt sich gewöhnlich durch eine Analyse der Residuenplots entscheiden. Da-

runter versteht man ein Streudiagramm, in dem die Abszissen x_i gegen die Residuen der eingepassten Werte $y_i - \hat{y}_i$ dargestellt werden. Lassen sich dabei keine auffallenden Muster erkennen, d.h. streuen die Residuenwerte zufällig und unsystematisch um die horizontale Achse, so kann man an dem Modell festhalten. Andernfalls ist eine Modellierung mit einer anderen Funkionenklasse angesagt. Eine einfache Prüfung der Annahmen eines linearen Modells wird von Cotts (1988) beschrieben. Es ist aber auch möglich, funktionale Zusammenhänge zwischen zwei Variablen ohne Vorgabe einer speziellen Funktionenklasse zu modellieren:

6 Gleitende Mittelwertkurven erlauben das Modellieren im Streudiagramm ohne von vornherein einen bestimmten Funktionstyp als Modellfunktion vorzugeben (Engel 1998, 1999). Sie sind somit flexibler als parametrische Regressionsmethoden, da die Daten selbst die strukturelle Form der Modellfunktion bestimmen. Sie sind explorativ in dem Sinne, dass sie mit minimalen Vorgaben auskommen, um funktionale Abhängigkeiten zu modellieren. Zur Berechnung der gleitenden Mittelwertkurve an einer Stelle t betrachtet man einen Streifen $\{(x, y) \mid t - h < x < t + h\}$ und berechnet das arithmetische Mittel aller Abszissen y_i in diesem Streifen. Wenn jetzt t über eine Menge von dichten Gitterpunkten läuft, so erhält man die gleitende Mittelwertkurve, am Gitter abgegriffen. Die Berechnungen verlangen den Einsatz eines Computers zur Bewältigung des hohen numerischen Aufwandes. Die Streifen- oder Fensterbreite h ist eine feste Zahl, die bei einer Darstellung der Mittelwertkurve am PC experimentell und interaktiv gewählt werden kann. Das Resultat ist ein Funktionsgraph, der einen funktionalen Zusammenhang zwischen den beiden Variablen x und y repräsentiert. Eine kompakte Darstellung der Mittelwertkurve durch einen Funktionsterm ist allerdings nicht möglich. Verschiedene Verfeinerungen in Form von *gewichteten* Mittelwerten, *lokal-linearen Anpassungen* und einer (an einem Optimalitätskriterium orientierten) automatischen Wahl der Fensterbreite h sind möglich, führen aber schnell von der im Grunde einfachen Idee der gleitenden Mittelwerte auf recht komplizierte mathematische Darstellungen (siehe z.B. Engel 1998 und die dort zitierte Literatur). Gleitende Mittelwertkurven sind unter sehr allgemeinen Bedingungen (keine extrem großen oder extrem kleinen Fenster h; der bedingte Erwartungswert ist eine glatte Funktion in x) konsistente Schätzer des bedingten Erwartungswerts $f(x) = \mathrm{E}(Y \mid X = x)$. Für parametrische Kurvenanpassungen gilt diese Feststellung nur, falls der bedingte Erwartungswert $f(x)$ zur betrachteten Funktionenklasse gehört, d.h. eine Geradenanpassung mittels linearer Regression ist nur dann eine konsistente Schätzung der bedingten Erwartungswertkurve, falls $f(x)$ selbst eine lineare Funktion ist.

Unser Augenmerk liegt hier nicht auf der technischen Umsetzung und mathematischen Begründung von Methoden zur Modellierung im Streudiagramm, sondern es soll vielmehr diskutiert werden, ob ihre Anwendung überhaupt Sinn macht. Dadurch soll einer blinden Anwendung statistischer Techniken vorgebeugt werden, die sich ihrer impliziten Voraussetzungen nicht bewusst ist. Auch bei relativ einfachen Anwendungen wie z.B. einer Geradenanpassung nach Augenmaß werden zur sachgerechten Interpretation der Ergebnisse Annahmen in die Analyse hineingesteckt, über die sich Anwender bewusst sein sollen. Die folgenden Beispiele sollen aufzeigen, dass die Konzentration auf die Anwendung von Algorithmen und ihre mathematischen Eigenschaften nicht den Blick für Fragen der Interpretierbarkeit und Sinnhaftigkeit der erhaltenen Resultate versperren darf. Jede Modellierung geht von gewissen vormathematischen Annahmen aus, über deren Gültigkeit sich die Anwender *vor* der Verwendung mathematischer Algorithmen bewusst sein müssen.

2 Das Gewicht von Alligatoren

Viele wild lebende Tierarten können durch Luftaufnahmen beobachtet werden. Informationen über die Anzahl und den Aufenthaltsort sind wichtig, um spezielle Tierarten zu schützen und auch um die Sicherheit der in der Nähe wohnenden Menschen zu garantieren. Einige Charakteristika der Tiere lassen sich per Luftaufnahme vom Flugzeug leicht erfassen, andere Eigenschaften der Tiere sind schwerer zu bestimmen. Die Länge eines Alligators lässt sich im Gegensatz zu seinem Gewicht relativ genau bestimmen. Tabelle 1 gibt die Länge (in cm) und das Gewicht (in kg) von 25 Alligatoren aus Florida wieder.

Aus den Daten soll ein allgemeines Modell

Länge	Gewicht	Länge	Gewicht	Länge	Gewicht
239	58,9	188	23,1	373	289,9
147	12,7	218	36,2	238	49,8
160	15,0	218	40,8	175	16,3
183	38,5	325	165,8	216	38,0
208	36,2	218	37,6	224	31,7
183	27,6	188	24,5	155	19,9
173	17,7	226	38,0	229	48,0
193	19,0	290	89,2	229	46,2
198	25,8				

Tabelle 1: Länge und Gewicht von 25 Alligatoren

entwickelt werden, mit dessen Hilfe das Gewicht eines Alligators aufgrund seiner Länge geschätzt werden kann.

Ein Streudiagramm der Variablen Länge versus Gewicht macht deutlich, dass der gesuchte Zusammenhang nicht linear ist. Abbildung 1 zeigt das Streudiagramm mitsamt (kleinster-Quadrate-) Regressionsgerade. Der dazugehörige Residuenplot ist durch systematische Abweichungen gekennzeichnet. Ein brauchbares Modell für diese Daten wird daher die Nicht-Linearität zu berücksichtigen haben. Motiviert durch sachimmanente Überlegungen (Gewicht ist proportional zum Volumen; Volumen ist dreidimensional, d.h. je länger ein Alligator, desto breiter und dicker ist er auch) bieten sich folgende Modelle an:

1 Anpassung eines Polynoms dritten Grades, d.h
$$\text{Gewicht} = a \cdot \text{Länge}^3 + b \cdot \text{Länge}^2 + c \cdot \text{Länge} + d$$

2 Anpassung eines linearen Modells an die logarithmierten Daten:
$$\log(\text{Gewicht}) = a \cdot \log(\text{Länge}) + b$$
d.h. nach Rücktransformation
$$\text{Gewicht} = b \cdot \text{Länge}^a .$$

Auch die Anpassung des ersten Modells (Polynoms 3.Grades) erfolgt mit linearen Regressionsmethoden, da die zu schätzenden Koeffizienten linear in das Modell eingehen. Beide Modelle können somit mithilfe von Standardsoftware angepasst werden. Abbildung 2 zeigt das Resultat. Denkbar wäre auch eine vereinfachte Version von 1 der Form „$\text{Gewicht} = a \cdot \text{Länge}^3$“, die von einer strikten Proportionalitätsüberlegung (wenn doppelt so lang, dann auch doppelte Dicke und doppelte Breite) ausgeht.

Zumindest für Alligatoren mit einer Länge zwischen 1,5m und 2,5m sind beide Modelle in etwa gleichermaßen gut brauchbar. Lässt sich damit das Gewicht eines Alligators vorhersagen, dessen Länge 2,10m misst? Diese Frage ist zu bejahen, falls die 25 Alligatoren eine repräsentative Stichprobe darstellen für eine Grundgesamtheit, aus der auch der 2,10m lange Alligator entstammt. Im Idealfall trifft dies zu, wenn alle 26 Alligatoren eine Zufallsstichprobe aus derselben Grundgesamtheit sind.

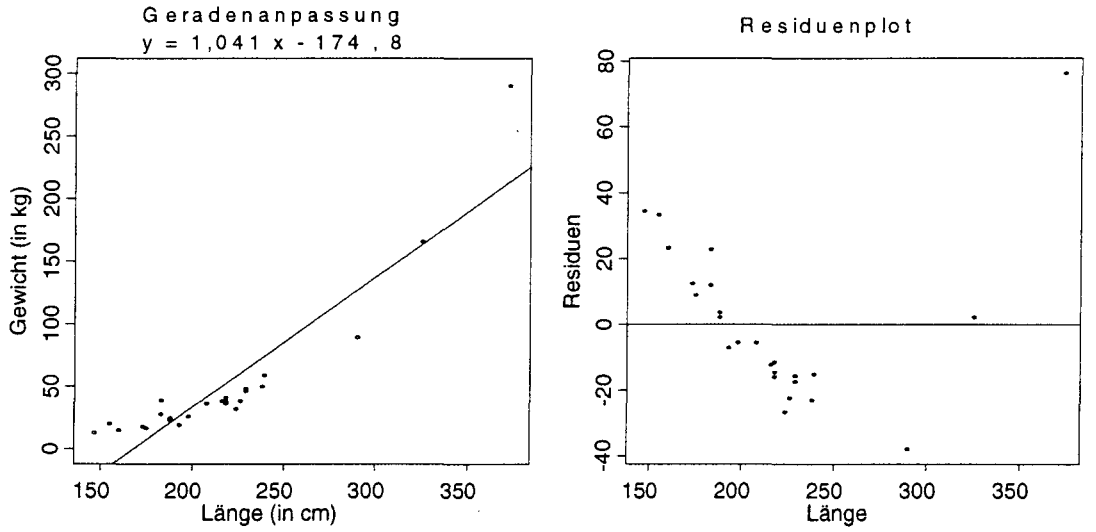

Abbildung 1: Streudiagramm von Länge versus Gewicht mit Regressionsgerade und dazugehörigem Residuenplot von 15 Alligatoren

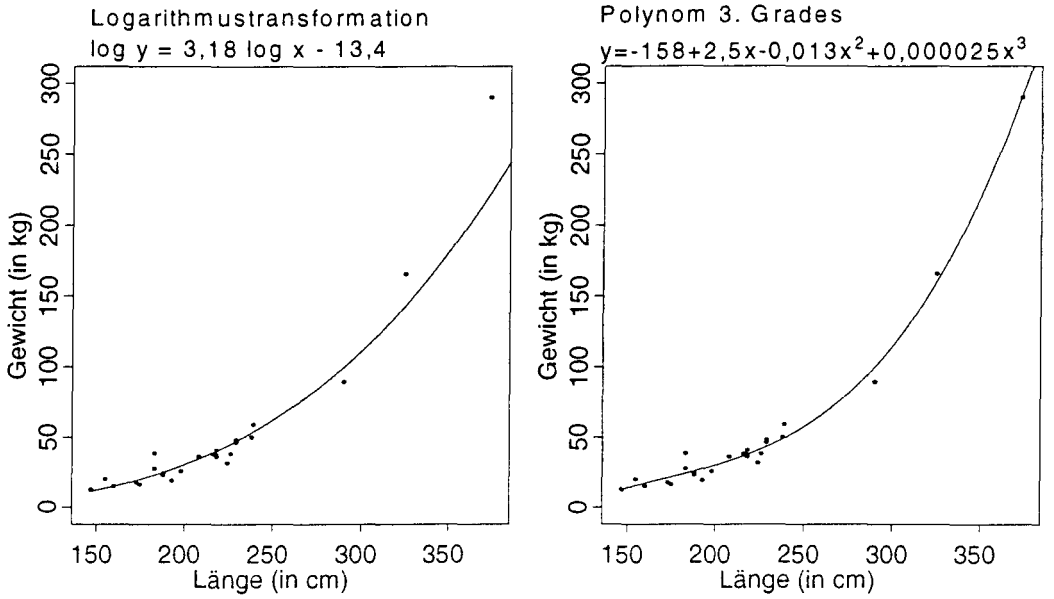

Abbildung 2: Kurvenanpassung mittels Logarithmustransformation und Anpassung eines Polynomes 3. Grades.

3 Radius versus Umfang von Kreisen

Misst man Umfang und Radius einer beliebigen Menge von kreisförmigen Gegenständen (Münze, Teller, Kochtopf, Reifen etc.), so ist der lineare Zusammenhang aufgrund der bekannten Kreisformel $U = 2\pi r$ offensichtlich. Verursacht durch Messfehler werden die Beobachtungen nicht ganz exakt auf der Geraden $y=2\pi x$ liegen, bei halbwegs tauglichen Messinstrumenten streuen die erhaltenen Werte aber eng um diese Gerade. Die Verallgemeinerung von den konkret untersuchten Gegenständen auf die Gesamtpopulation *aller* denkbaren kreisförmigen Gebilde ist hier unproblematisch, da es in diesem Beispiel keinen Stichprobenfehler (englisch: *sampling error*) gibt. Schließlich gilt die Formel für den Kreisumfang für alle Kreise exakt! Allerdings ist der Messfehler (englisch: *non-sampling error*) nicht von konstanter Varianz. Bei extrem großen Kreisen (z.B. Äquator von Gestirnen) wird der Messfehler gewiss größer sein als bei kreisförmigen Gegenständen kleiner oder moderater Größe. Diese Feststellung gilt wohl auch noch für den relativen Messfehler, d.h. den Fehler bezogen auf den Gesamtumfang, da mit den uns verfügbaren Messinstrumenten der Umfang z.B. von Kochtöpfen leichter zu bestimmen ist als die Äquatoriallänge von entfernten Sternen. Sinnvollerweise wird man die Geradenanpassung hier noch unter der Zusatzbedingung einer durch den Ursprung gehenden Geraden durchführen, d.h. wenn $r = 0$ dann soll auch $U = 0$ sein. Abbildung 3 zeigt ein Streudiagramm mitsamt Regressionsgerade für 20 typische kreisförmige Gegenstände. Die Gleichung für die dazugehörige kleinste-Quadrate-Ausgleichsgerade hat hier die Form

$$y = 6{,}3104\,x\ ,$$

was eine gute Annäherung an $y = 2\,\pi\,x$ ist. Eine Verallgemeinerung dieser Formel als Näherungsformel für *alle* Kreise ist wohl akzeptabel, ohne dass näher erklärt werden muss, nach welchem Mechanismus die vorliegende Stichprobe aus der Gesamtpopulation gezogen wurde.

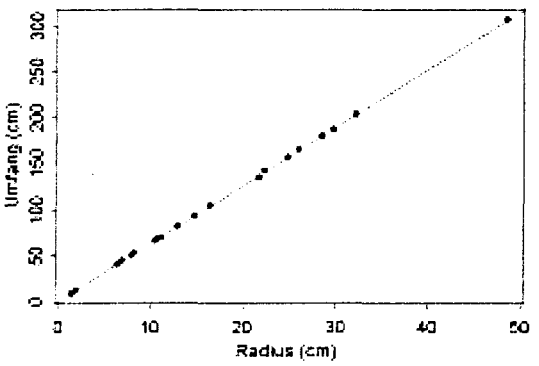

Abbildung 3: Radius und Umfang (beides in cm) für 20 typische kreisförmige Gegenstände mitsamt kleinster-Quadrate Regressionsgeraden.

4 Umfang und Inhalt von Rechtecken

Auch beim Streudiagramm in Abbildung 4 erscheint die Linearitätsannahme für die 25 Beobachtungspunkte problemlos. Die Daten streuen zwar etwas stärker um die kleinste-Quadrate-Ausgleichsgerade, aber systematische Muster sind im Residuen-plot nicht erkennbar (siehe dazu auch Freedman, Pisani und Purves 1980, S. 195). Für die Regressionsgerade errechnet sich rein technisch die Gleichung:

(1) $y = 0{,}59 \cdot x + 6{,}95.$

Lässt sich in ähnlicher Weise wie beim vorangegangenen Beispiel von den Beobachtungen auf eine größere Grundgesamtheit schließen? Genauso wie im vorangegangenen Beispiel sind die Daten Messungen geometrischer Objekte. Die Beobachtungen bestehen aus Umfang und Flächeninhalt von 25 typischen Rechtecken, dargestellt in Abbildung 5.

Können wir somit schließen, dass mit Formel (1) annäherungsweise der Inhalt eines Rechteckes berechnet werden kann, wenn der Umfang U gegeben ist? Kann also Formel (1) die Inhaltsformel $A = a \cdot b$ als Annäherung ersetzen? Diese Überlegung wirft folgende Frage auf: Für welche Rechtecke soll diese Näherung Gültigkeit beanspruchen? Für welche Klasse von Rechtecken sind die Figuren in Abbildung 5 typisch? Können *alle denkbaren* Rechtecke gemeint sein?

Jede Verallgemeinerung, die im Sinne der schließenden Statistik die Qualität ihrer Aussagen präzisieren will, verlangt eine exakte Formulierung ihrer probabilistischen Modellvoraussetzungen. Dies ist hier in der laxen Formulierung von „typischen" Rechtecken (die Formulierung „zufällige Vierecke" wurde bewusst vermieden).

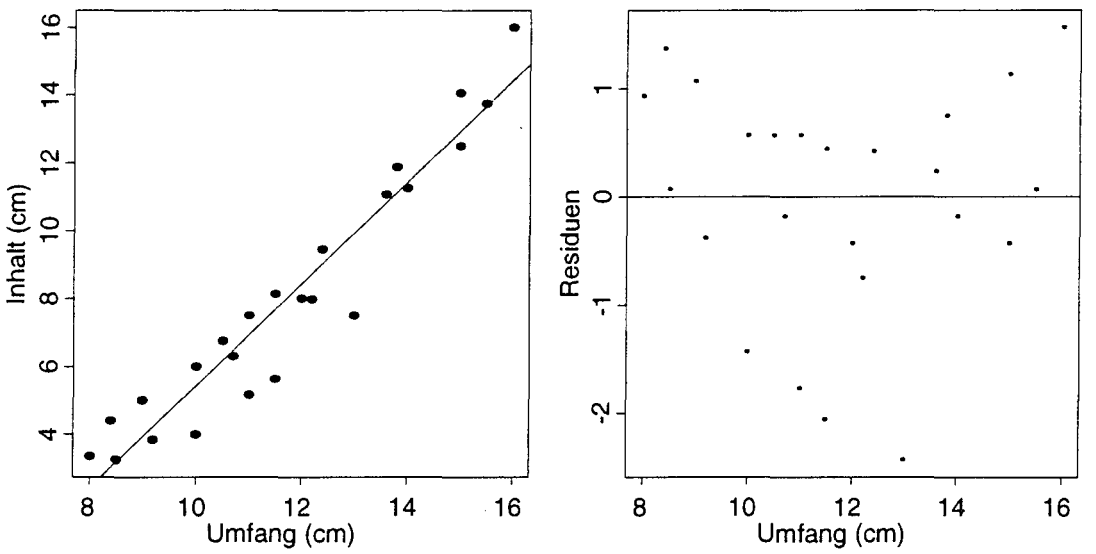

Abbildung 4: Streudiagramnm und Residuenplot für Umfang versus Flächeninhalt von 25 „typischen" Rechtecken mitsamt Regressionsgeraden

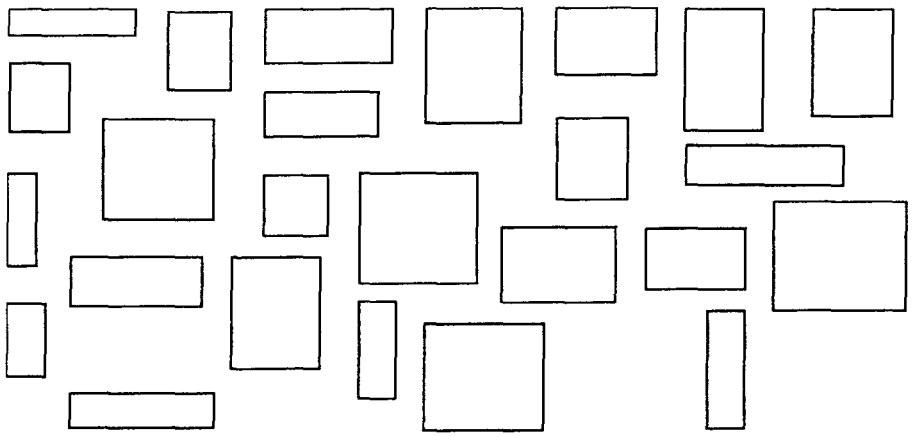

Abbildung 5: 25 typische Rechtecke, deren Umfang und Inhalt im Streudiagramm in Abbildung 4 dargestellt sind.

nicht erfolgt. Von welcher Grundgesamtheit von Figuren sind die Rechtecke in Abbildung 5 eine zufällig gezogene Stichprobe? Man beachte, dass die Vernachlässigung der Modellpräzisierung im vorangegangenen Beispiel vom Kreisradius und Kreisumfang keine Rolle spielt, da ja dort der Stichprobenfehler null ist: Schließlich gilt die Formel $U = 2\pi r$ exakt und ausnahmslos für alle Kreise und Fehler treten nur durch unpräzise Messinstrumente auf. Im Gegensatz dazu lässt sich im Fall der Rechtecke der Stichprobenfehler überhaupt nicht quantifizieren, solange die Gesamtpopulation, auf die sich die Regressionsformel beziehen will, nicht präzisiert ist. Solange dies nicht geschehen ist, kann man Formel (1) bestenfalls als Näherungsformel für den Inhalt genau der Vierecke in Abbildung 5 und keine weiteren Rechtecke ansehen.

Zur Illustration präzisieren wir die Gesamtpopulation und das Zufallsmodell, von dem mit Hilfe eines Zufallsgenerators eine Stichprobe von Vierecken gezogen wird. Wir betrachten Vierecke mit Seitenlängen a und b, wobei a, b Realisierungen normalverteilter Zufallsvariabler A und B mit $A \sim N(2, \frac{1}{2})$ $B \sim N(4, 1)$ sind. Da beide Zufallsvariable negative Werte mit einer Wahrscheinlichkeit von ungefähr $3 \cdot 10^{-5}$ annehmen, kommen in der Simulation negative Seitenlängen praktisch nicht vor. Abbildung 6 zeigt das Streudiagramm für eine vom Zufallsgenerator erzeugte Stichprobe von 200 Rechtecken dieser Grundgesamtheit, mitsamt Regressionsgeraden und gleitender Mittelwertkurve. Vom Streudiagramm her könnte man eine lineare Struktur der Daten für möglich halten[2]. Damit können die Resultate beider Regressionsansätze interpretiert werden als Kurven, die innerhalb der Menge *aller* Vierecke, deren Seitenlängen unabhängige $N(2, \frac{1}{2})$ bzw. $N(4, 1)$ verteilte Zufallsvariablen sind, einen funktionalen Zusammenhang zwischen Umfang und Inhalt herstellen. Die in Abbildung 6 eingezeichnete gleitende Mittelwertkurve lässt sich daher zur Vorhersage des Inhalts „neuer Rechtecke" aus der spezifizierten Grundgesamtheit nutzen.

5 Einkommen und Ausgaben für Nahrungsmitel

Bezogen sich die letzten beiden Beispiele auf künstliche Datensätze, so liegen dem folgenden Beispiel ein umfassender empirisch erhobener Datensatz zugrunde. Abbildung 7 gibt das verfügbare Einkommen und die Ausgaben für Nahrungsmittel von 7081 britischen Haushalten bezogen auf einen Zeitraum von zwei Wochen im Jahr 1979 wieder. Die Haushalte sind einer Zufallsstichprobe entnommen, der jährlich stattfindenden britischen Verbraucherstichprobe (Family Expenditure Survey, ESCR Data Archive, Department of Employment, Statistics Division, Her Majesty's Stationary Office, London). Details dieser Erhebung sind in Kemsley, Redpath und Holmes (1980) beschrieben. Ähnliche Erhebungen werden in allen OECD-Ländern von den offiziellen statistischen Ämtern durchgeführt. Die Ausgaben für ein bestimmtes Gut (hier: Nahrung) werden als Maß für die Nachfrage angesehen. Die funktionale Abhängigkeit der Nachfrage vom Einkommen ist für die

[2]Allerdings ist in diesem Fall auch eine (komplexe) mathematische Analyse möglich, die zeigt, dass $E(A \cdot B \mid 2A + 2B = x)$ keine in x lineare Funktion ist.

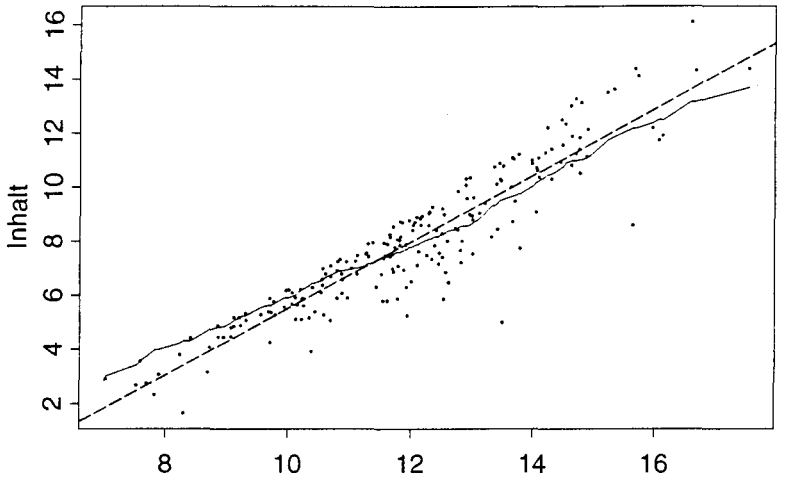

Abbildung 6: Umfang und Inhalt von 200 simulierten Vierecken, deren Länge und Breite Realisierungen unabhängiger $N(2, ½)$ bzw. $N(4, 1)$ verteilter Zufallsvariabler sind. Die durchgezogene Linie stellt eine gleitende Mittelwertkurve, die gestrichelte Linie die Regressionsgerade dar.

Wirtschaftswissenschaften von praktischer und theoretischer Bedeutung (Hildenbrand 1994).

Die umfangreiche Punktwolke in Abbildung 7 ist gekennzeichnet durch eine hohe Streuung, die nach rechts hin (größere Einkommen) noch zunimmt. Dies spiegelt die Tatsache wieder, dass Haushalte mit höherem Einkommen mehr finanziellen Spielraum beim Einkauf als arme Haushalte haben, die ihre finanziellen Ressourcen im wesentlichen für Grundlebensmittel aufwenden müssen. Funktionale Strukturen in Form von Trends sind wegen der hohen Variabilität der Daten mit dem Auge nur sehr schwer wahrnehmbar.

Zur Modellierung von Nachfragekurven haben Wirtschaftswissenschaftler schon vor über 50 Jahren Modelle der Form $y = a\,x + b\,x \log(x)$ vorgeschlagen (Working 1943), d.h. die relativen (auf das Einkommen bezogenen) Anteile („Budgetanteile") y / x sind linear in $\log(x)$. Für einen ausführlicheren Überblick über neuere Ansätze und Methoden zum Schätzen von Nachfragekurven siehe z.B. Engel und Kneip (1996). Abbildung 8 zeigt ein Streudiagramm der Budgetanteile versus der logarithmierten Einkommen.

Die Daten im Streudiagramm von Abbildung 8 streuen noch immer stark. Die von Working (1943) postulierte lineare Struktur ist für die vorliegenden Daten vertretbar, wenn sie auch nicht gerade ins Auge springt. Abbildung 7 zeigt die auf der Geradenanpassung in Abbildung 8 zurücktransformierte Nachfragefunktion

$$f(x) = 0{,}247\,x - 0{,}023\,x \log x\,.$$

zusammen mit einer gleitenden Mittelwertkurve. Im Bild sind die beiden Kurven kaum zu unterscheiden. Beide Ansätze geben hier den funktionalen Zusammenhang in akzeptabler Weise wieder. Trotz der schwer überschaubaren Struktur und der den Daten innewohnenden Variabilität ist das Schätzen der entsprechenden Kurve hier sinnvoll, solange die probabilistischen Voraussetzungen (Zufallsstichprobe aus der Gesamtheit aller britischen Haushalte) klar präzisiert sind.

6 Schlussfolgerungen

Zur Modellbildung im Streudiagramm stehen eine Vielzahl von Techniken zur Verfügung. In empirischen Studien enthaltene Beobachtungen sind immer fehlerhafte Größen, die durch Messfehler, Stichprobenfehler oder den Einfluss weiterer, unberücksichtigter Variabler gestört sind. Die Aufgabe des Modellierens besteht dann darin, aus den *verrauschten Daten* ein Signal $y = f(x)$ zu rekonstruieren. Rein technisch können fast alle verfügbaren Verfahren in jedem Streudiagramm angewandt werden.

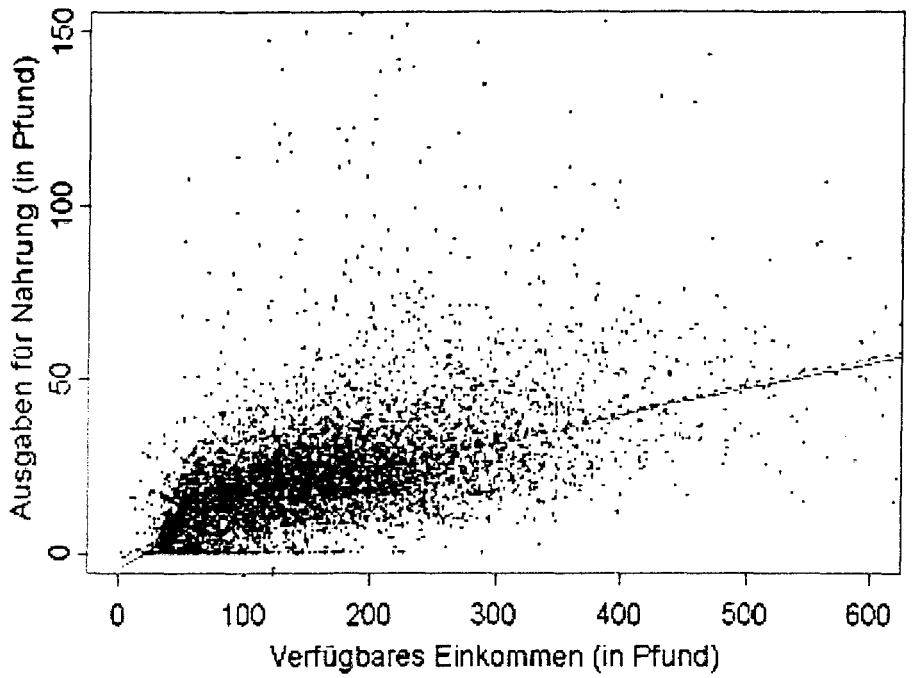

Abbildung 7: Streudiagramm von verfügbarem Einkommen (pro 14 Tagen in britischen Pfund) versus Ausgaben für Nahrung von 7081 britischen Haushalten mitsamt gleitender Mittelwertkurve (durchgezogen) und angepasster Modellfunktion (gestrichelt) der Form $y = a\ x + b\ x\ \log(x)$.

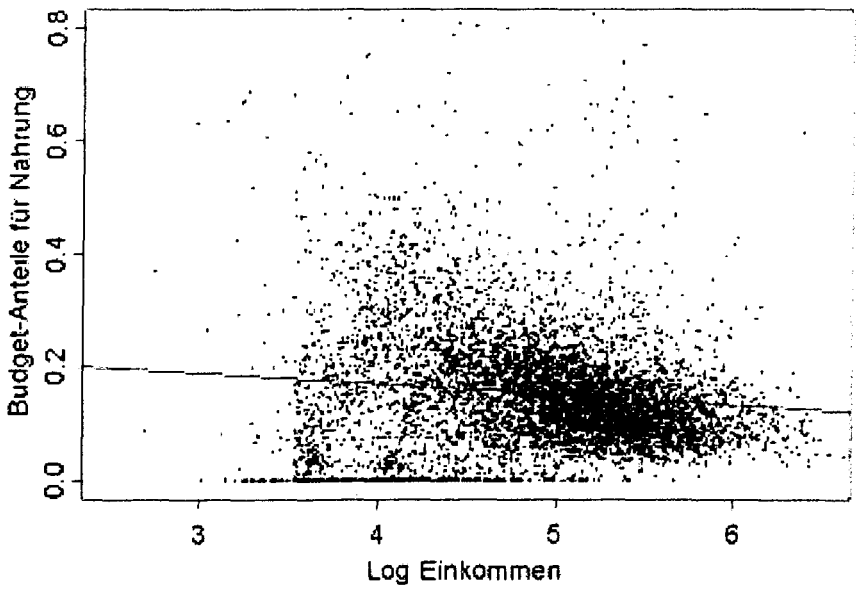

Abbildung 8: Budget-Anteile, d.h. Ausgaben für Nahrung dividiert durch Einkommen versus logarithmierten Einkommen

Zur Interpretierbarkeit der erhaltenen Ergebnisse ist von entscheidender Bedeutung, ob die erlangten Aussagen im Rahmen der beschreibenden Statistik ausschließlich als eine Art Zusammenfassung für die vorliegenden (und keine anderen!) Daten gelten sollen oder ob sie sich im Sinne der schließenden Statistik auf eine größere Grundgesamtheit beziehen. Als eine Komprimierung ist eine Bezugnahme ausschließlich auf die vorliegenden Daten im Rahmen der deskriptiven Statistik immer legitim. Innerhalb der vorliegenden Datenmenge gibt dann die erhaltene Kurve an, in welcher Beziehung der Durchschnittswert der Responzvariablen y zum Wert der Prädiktorvariablen x steht.

Allerdings werden in empirischen Studien Messungen gewöhnlich durchgeführt, um Erkenntnisse über die gerade vorliegenden Beobachtungen hinaus zu gewinnen, um z.B. Vorhersagen für neue Beobachtungen treffen zu können. Modelle werden immer unter dem Aspekt einer gewissen Allgemeingültigkeit gebildet. Ohne weitere wahrscheinlichkeitstheoretische Annahmen kann dann nicht verallgemeinert werden. Das Modell hat sonst keine Gültigkeit über die vorliegenden Daten hinaus. Ein Erkenntnisgewinn für eine umfassendere Grundgesamtheit ist jedoch - wie bei jeder schließenden Statistik - nur möglich, wenn die probabilistischen Voraussetzungen der Stichprobengewinnung genau präzisiert sind. Noch bevor man sich für eine spezielle Modellierungstechnik entscheidet, muss dann das wahrscheinlichkeitstheoretische Modell für die Stichprobe spezifiziert sein. Modellbildung als Abstraktion und Weg zu allgemeingültigen Aussagen basiert auf dem Schließen von der Stichprobe zu einer größeren Grundgesamtheit. Außer wenn Stichprobe und Grundgesamtheit im Hinblick auf das zu untersuchende Merkmal wie im Beispiel „Radius und Umfang von kreisförmigen Gegenständen" deckungsgleich sind, d.h. wenn wie im Beispiel vom Kreisumfang und Radius keinerlei Stichprobenfehler auftreten, müssen die Grundgesamtheit und der Zufallsmechanismus der Stichprobenauswahl präzisiert sein. Sonst können bei der Verallgemeinerung auf eine umfassendere Grundgesamtheit leicht unsinnige und absurde Schlußfolgerungen gezogen werden.

Hingegen ist die Sinnhaftigkeit eines Regressionsmodells völlig unabhängig von der Höhe des Rauschens in den Daten. Bei stark verrauschten Daten ist es zweifellos schwerer (d.h. es werden mehr Daten benötigt, um eine vorgegebene Präzision zu erreichen), eine gute Schätzung und somit ein taugliches Modell zu erhalten. Die erhaltene Regressionskurve lässt sich aber immer als Schätzung des bedingten Erwartungswertes $y = f(x) = E(Y \mid X = x)$ interpretieren, solange das probabilistische Modell präzisiert und die Modellkurve zur Klasse der zulässigen Funktionen gehört (parametrische Regression). Letzteres kann über eine Diagnose der Residuen entschieden werden.

Literatur

Cotts, J.W. (1988): Prüfung der Modellvoraussetzungen bei linearer Regression. Statistik in der Schule, S. 36 - 50.

Engel, J. (1998): Zur stochastischen Modellierung funktionaler Abhängigkeiten: Konzepte, Postulate, Fundamentale Ideen. Mathematische Semesterberichte (2), 95 - 112.

Engel, J. (1999): Von der Datenwolke zur Funktion. Mathematiklehren 97, 60 – 64.

Engel, J.& A. Kneip (1996): Recent Approaches to Estimating Engel Curves. Journal of Economics (63), 187 – 212.

Engel, J.& E. Theiss (2001): Robuste elementare Instrumente zur Analyse von Streudiagrammdaten. Der Mathematisch-naturwissenschaftliche Unterricht.

Freedman, D; Pisani, R& Paves, R (1980): Statistics. Norton: New York.

Hildenbrand, W. (1994): Market Demand. Princeton University Press: Princeton, N.J.

Hui, E. (1988): Lineare Regression ohne Differentialrechnung. Didaktik der Mathematik (2), 94 -98.

Kemsley; Redpath & Holmes (1980) : Family Expenditure Survey Handbook. London: Her Majesty's Stationary Office.

Polasek, W. (1994): Explorative Datenanalyse. Einführung in die deskriptive Statistik. Springer: Berlin.

Vohmann, H.D. (1988): Lineare Regression und Korrelation in einem Einführungskurs über empirische Methoden. Stochastik in der Schule, 3 -16.

Working, H. (1943): Statistical Laws of Family Expenditure. Journal of the American Statistical Association (38), 43 – 56.