PRÜFUNG DER MODELLVORAUSSETZUNGEN BEI LINEARER REGRESSION

von James W. Cotts, Furman

Orginaltitel in "Teaching Statistics" Vol. 9 (1987) Nr. 3:

Checking Model Validity in Linear Regression

Übersetzung: Manfred Borovcnik, Klagenfurt

Kurzfassung: Im Unterricht von Regressions- und Korrelationsrechnung sowie in der Anwendung geht man oft rasch von Ergebnissen im Stil der Beschreibenden Statistik zu Tests oder Vertrauensintervallen über. "Die lineare Beziehung zwischen X und Y ist statistisch gesichert" ist eine typische Aussage. Um überhaupt solche Aussagen in den Zielbereich der Analyse eines Regressionsproblems zu bekommen, müssen eine Reihe von mathematischen Voraussetzungen erfüllt sein. In diesem Artikel wird auf die Art dieser Voraussetzungen eingegangen. Hauptziel ist es, eine elementare, graphische Methode zur Prüfung dieser Modellvoraussetzungen darzustellen und den Leser von der Tauglichkeit dieser Prüfung zu überzeugen.

1. Einleitung

Eines der vordringlichen Ziele von Statistik ist die Suche nach und die Beschreibung von Beziehungen zwischen physikalischen Größen in der Welt, in der wir leben. Diese Beziehungen werden durch Gleichungen ausgedrückt. Dies ist auch das Thema der Regressionsrechnung. In einem einführenden Kurs werden im allgemeinen nur Beziehungen zwischen *zwei* Variablen X und Y behandelt. Anfänglich beschränkt man sich auf lineare Beziehungen.

Das Modell für die Population wird beschrieben durch

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

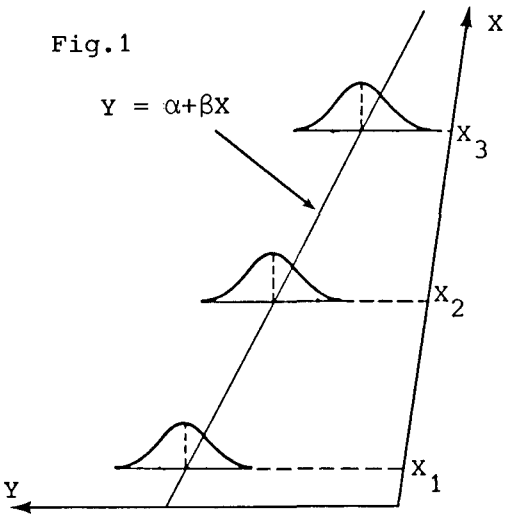
wobei α und β die Regressionskoeffizienten für die Population sind und ε_i die senkrechte Abweichung zwischen dem Punkt

(X_i, Y_i) und der "wahren" Regressiongeraden (für die Population) $Y = \alpha + \beta X$ darstellt. Die Zufallsvariable $\varepsilon_i = Y_i - (\alpha+\beta X_i)$ wird das Residuum im Punkt (X_i, Y_i) genannt. Gewisse Ausnahmen im Modell erlauben, Hypothesen in Bezug auf die Regressionsparameter α und β oder z.B auf $\mu_{Y|X_o}$ (den Erwartungswert von Y zu einem festgelegten Wert X_o) zu testen sowie Vertrauensintervalle dafür anzugeben. (Dabei verwendet man die t-Verteilung). Die üblichen Annahmen dazu sind, daß die ε_i stochastisch unabhängig und identisch verteilt sind, alle ε_i sollen dabei eine Normalverteilung mit Erwartungswert Null und Varianz σ^2 haben, d.h. $\varepsilon_i \sim N(0,\sigma^2)$.

Diese Annahmen werden in einführenden Lehrbüchern gewöhnlich graphisch dargestellt, siehe Fig. 1. Dazu gibt es begleitende Feststellungen:

1) Zu jedem X_i werden die Y-Werte als normalverteilte Zufallsvariablen angenommen.
2) Die Erwartungswerte dieser Normalverteilungen liegen auf der wahren Regressionsgeraden: $Y=\alpha+\beta X$.
3) Die Varianz von Y zu jedem X_i ist dieselbe, d.h. alle Normalverteilungen haben dieselbe Varianz σ^2.

Fig.1



Bemerkung des Übersetzers: Die Voraussetzungen (1) - (3) werden unter der Bezeichnung "lineares Modell" zusammengefaßt. Hat (X, Y) eine zweidimensionale Normalverteilung (die gemeinsame Verteilung sieht wie ein Schlapphut aus), so erfüllen die Verteilungen von Y zu bestimmten Werten X_i genau die Voraussetzungen (1) - (3). Umgekehrt folgt aus den Voraussetzungen (1) - (3) aber nichts über die gemeinsame Verteilung von (X,Y). Insbesondere folgt nicht, daß sie eine Normalverteilung hätten. Ja es ist in (1) - (3) nicht einmal notwendig, X_i als Zufallsvariable festzulegen, X_i könnte einfach ein frei wählbarer Wert sein.

Gewöhnlich beurteilt man mit Hilfe des Stichprobenkorrelationskoeffizienten, ob zwischen X und Y eine lineare Beziehung besteht oder nicht. Anscombe (1973) jedoch gibt nette Beispiele dafür, daß man sich bei Prüfung auf lineare Beziehungen nicht ausschließlich auf den Korrelationskoeffizienten verlassen sollte. In nur wenigen Lehrbüchern wird überhaupt auf die Prüfung der Modellvoraussetzungen (1)-(3) eingegangen.
Hypothesen werden getestet und Vertrauensintervalle werden berechnet (z.B. ob der Korrelationskoeffizient in der Stichprobe signifikant von Null verschieden ist), als ob die Voraussetzungen nie in Frage stünden. Dabei ist eine informelle Überprüfung von (1)-(3) weder schwierig noch jenseits des Aufgabengebiets eines elementaren Lehrgangs. Mein Ziel in diesem Artikel ist es nun, eine graphische Methode zur Überprüfung von (1)-(3) vorzustellen und zu zeigen, daß es wirklich leicht ist, Lernende von der Tauglichkeit dieser Prüfung zu überzeugen.

2. Die Methode

Die wahre Regressionsgerade $Y=\alpha+\beta X$ wird durch die Stichprobenregressionsgerade $\hat{Y}=a+bX$ geschätzt. Diese wird aus den Stichprobendaten $(X_1,Y_1),(X_2,Y_2), \ldots, (X_n,Y_n)$ nach folgenden Formeln berechnet:

$$b = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\Sigma X_i Y_i - \frac{1}{n}\Sigma X_i Y_i}{\Sigma X_i^2 - \frac{1}{n}(\Sigma X_i)^2}, \qquad a = \bar{Y} - b\bar{X}$$

Die Residuen

$$\varepsilon_i = Y_i - (\alpha + \beta X_i)$$

schätzt man durch die Residuen in der Stichprobe (das sind die Abweichungen der Daten von der Regressionsgeraden):

$$e_i = Y_i - (a + bX_i).$$

Setzt man $\hat{Y}_i = a + bX_i$ ($\hat{Y}_i$ ist der "Ausgleichswert" für Y_i), so erhält man:

$$e_i = Y_i - \hat{Y}_i.$$

Wenn die Residuen ε normalverteilt sind, d.h. $\varepsilon \sim N(0,\sigma^2)$, dann kann man die Residuen der Stichprobe, e_i, als Zufallsstichprobe aus $N(0,\sigma^2)$ auffassen. σ^2 schätzt man dabei in der üblichen Weise durch:

$$s^2 = \frac{\Sigma e_i^2}{n-2} = \frac{\Sigma Y_i^2 - a\Sigma Y_i - b\Sigma X_i Y_i}{n-2}$$

(Dieser Varianzschätzer wird üblicherweise mit s_e^2 bzw. $s_{y.x}^2$ bezeichnet. Weil es hier der einzige Varianzschätzer ist, greife ich auf die einfachere Bezeichnung s^2 zurück). Des weiterern sollten e_i/s eine Zufallsstichprobe aus $\varepsilon/\sigma \sim N(0,1)$ darstellen. Die e_i/s werden als "standardisierte Residuen" bezeichnet.

Ziel ist es nun, die Lernenden davon zu überzeugen, daß die Modellvoraussetzungen (1)-(3), wie sie in Fig. 1 dargestellt sind, erfüllt sind, falls die Residuen e_i folgenden Voraussetzungen genügen:

(1*) e_i sind annähernd normalverteilt
(2*) e_i haben Erwartungswert 0,
(3*) e_i haben konstante Varianz σ^2 für alle X_i.

Ferner gilt es, die Lernenden davon zu überzeugen, daß die Voraussetzungen (1*)-(3*) erfüllt sind, falls das Punktdiagramm der e_i gegen X_i (bzw. der e_i/s gegen X_i) ein Zufallsmuster in den Punkten zeigt d.h. man kann kein wie immer geartetes systematisches Muster erkennen. Diese Punktdiagramme werden im folgenden als Residuendiagramme (residual plots) angesprochen.

Die folgende Serie von Abbildungen (Fig. 2 - 5) erweist sich als hilfreich zur Demonstration, wie das Resuduendiagramm aussehen sollte, wenn die Modellvoraussetzungen erfüllt bzw. verletzt sind.

Wenn die Y's normalverteilt mit derselben Varianz für jedes X_i sind und Erwartungswert $\alpha + \beta X_i$ haben, dann sind sie dicht verteilt nahe dem Punkt $\alpha + \beta X_i$ und werden umso spärlicher verteilt, je größer die Distanz von $\alpha + \beta X_i$ wird (Fig. 2a). Für die ε_i, die vertikale Distanz von Y_i zur Regressionsgeraden $Y = \alpha + \beta X$ heißt das: ε_i sind dicht verteilt nahe ihrem Erwartungswert Null und umso spärlicher, je weiter weg man von Null kommt (Fig. 2b). Eine Zufallsstichprobe der ε_i, die e_i, sollte daher im allgemeinen auch diesem Muster folgen. Fig. 2c ist eine Wiederholung von Fig. 2b, jedoch mit mehr Werten von X, eine Zufallsstichprobe der ε_i, die e_i's, ist dabei durch die Zeichen x markiert. In Fig. 2d werden gerade diese Werte der e_i nocheinmal gezeigt, das ist das Residuendiagramm der e_i's gegen X, u.zw. für den Fall, daß alle Annahmen (1)-(3) erfüllt sind. Der "Residuenplot" für die standardisierten Residuen e_i/s gegen X ist identisch mit dem vorhergehenden mit Ausnahme der vertikalen Maßeinheit. In Fig. 2d ist die Punkteverteilung der Residuen bzw. standardisierten Residuen typisch "zufällig", ein solches Bild im "Residuenplot" zeigt also an, daß die Modellvoraussetzungen für die lineare Regression erfüllt sind.

Der Fall, daß Annahme(1) verletzt ist, daß also die Y's nicht zu jedem X_i normalverteilt sind, wird in Fig. 3a aufgegriffen. Fig. 3b stellt das Residuendiagramm der ε_i gegen X_i dar, wiederum ist eine Stichprobe e_i der Residuen durch die Zeichen x markiert. Das Resi-

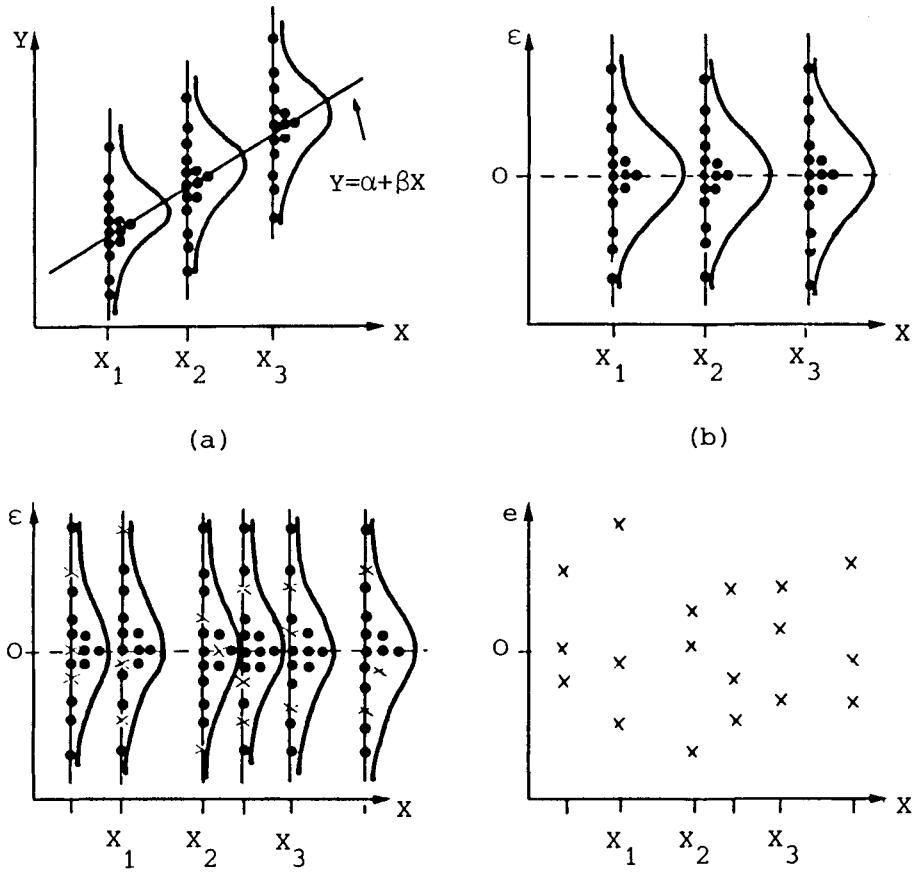


Fig.2

duendiagramm der e_i gegen X_i zeigt in diesem Fall ein typisch nicht-zufälliges Bild, insbesondere was die Häufung der Punkte in einzelnen Gebieten anbelangt (Fig. 3c).

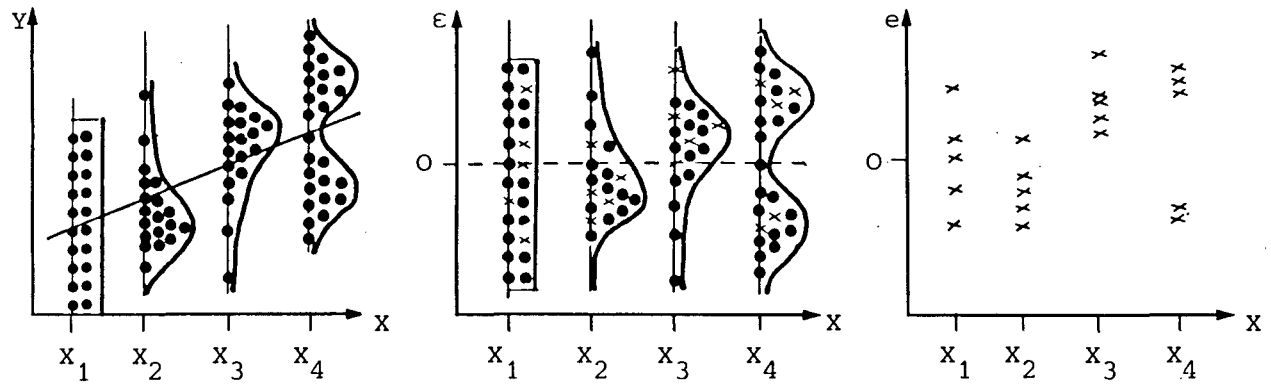


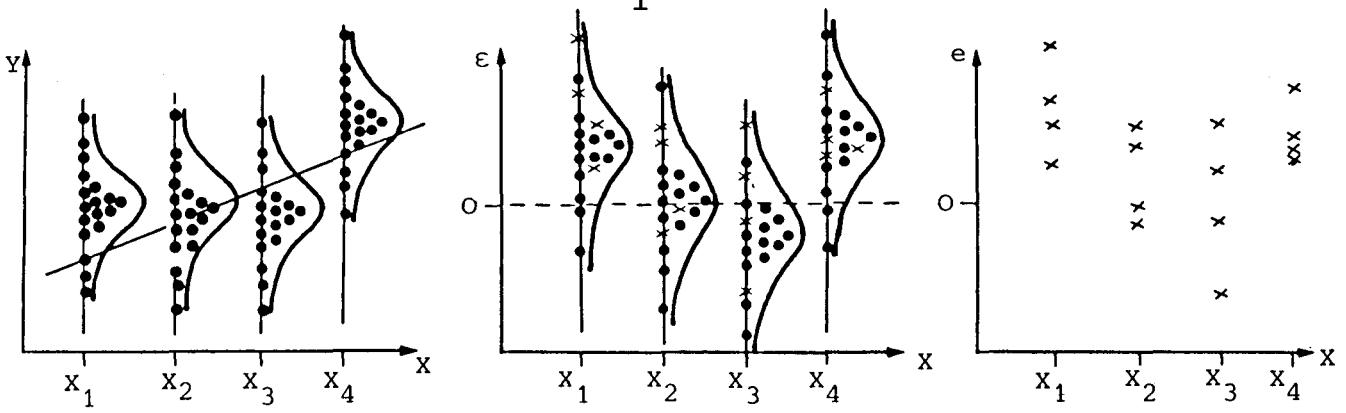
Fig.3

Eine informelle Prüfung der ε bzw. ε/σ auf Normalverteilung kann auch über die σ-Regeln erfolgen: Die Wahrscheinlichkeit für die Intervalle $(\mu-\sigma, \mu+\sigma)$, $(\mu-2\sigma, \mu+2\sigma)$ bzw. $(\mu-3\sigma, \mu+3\sigma)$ beträgt für Normalverteilungen ca. 68 %, 95 % bzw. 99,7 %. Betrachtet man nun die e/s als eine Stichprobe von $N(0,1)$, d.h. falls die Normalverteilungsannahme zutrifft, dann sollten wir 68 %, 95 % bzw. beinahe alle Daten e/s im Intervall (-1,1), (-2,2) bzw. (-3,3) vor-

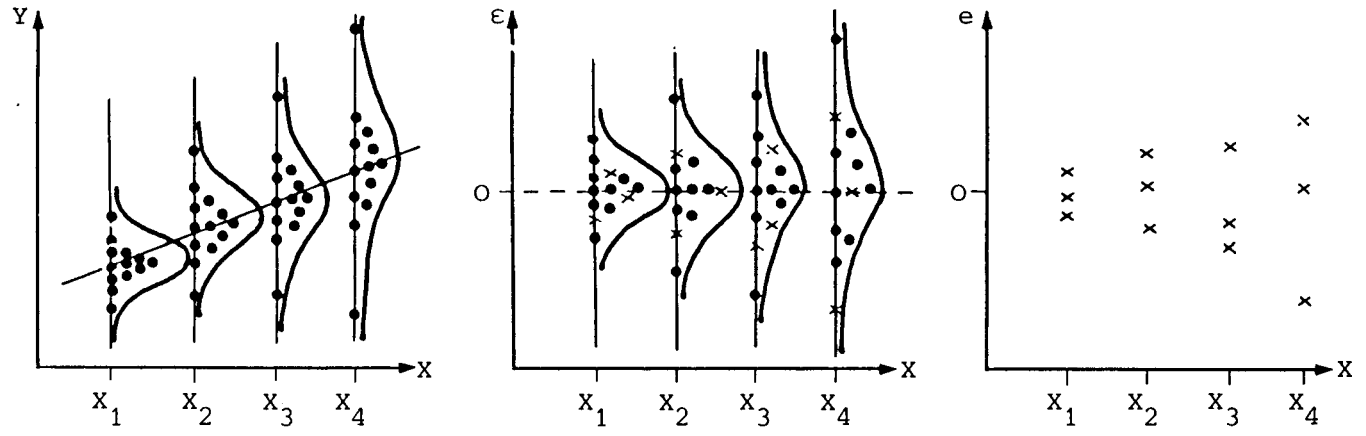
finden. Leider ist diese Methode der Prüfung im Zusammenhang mit Lehrbuchbeispielen nicht sehr zuverlässig, weil gerade die einleitenden Regressionsbeispiele auf kleine Datenmengen zurückgreifen und weil die angegebenen Prozentsätze mit jedem einzelnen Residuum, das dazu zählt oder nicht, beträchtlich schwanken.

Nun zur Annahme (2): Ein typisches Bild für den Fall, daß die Erwartungswerte der Y-Verteilungen zu den jeweiligen X_i nicht alle auf die "wahre" Regressionsgerade $Y = \alpha + \beta X$ fallen, vermittelt Fig. 4a. Die Residuen ε_i würden dann wie in Fig. 4b aussehen, wiederum ist die Stichprobe der ε_i, die e_i's, durch x markiert. Das Residuendiagramm von e gegen X ist in Fig. 4c. Die fehlende Zufälligkeit der Verteilung der Kreuzchen ist offensichtlich. Ferner gilt für den Fall, daß (2) zutrifft, daß der Erwartungswert von ε gleich Null ist (siehe Fig. 2b). Es ist eine leichte algebraische Übung, zu zeigen, daß $\Sigma e_i = 0$ (d.h. $\bar{e} = 0$) für alle Datensätze gilt. Lernende jedoch sind selten mit solchen Σ-Beweisen zufrieden. Man kann sie jedoch davon überzeugen, indem man sie $\bar{e}$ tatsächlich für spezielle Aufgaben berechnen läßt, ihr errechneter Wert für die Summe der Stichprobenresiduen wird nämlich (bis auf Rundungsfehler) Null sein.

Die Annahme konstanter Varianzen, (3), ist in einer Situation wie in Fig. 5a verletzt (Beachten Sie, daß in diesem Bild nur (3) verletzt ist). Figur 5b zeigt den Graphen ε gegen X, wobei wiederum eine Zufallsstichprobe der ε_i (die e_i) mit x markiert ist; Der Residuenplot e gegen X ist in Fig. 5c wiedergegeben. Die fehlende Zufälligkeit ist offensichtlich, es gibt eine systematische Änderung in der vertikalen Streuung der e_i.



(a) (b) (c) Fig.4



(a) (b) (c) Fig.5

Die Verletzung der dritten Annahme kann schwerwiegende Folgen für die Gültigkeit von Hypothesentests und Konfidenzintervallen haben. Weil s die Abweichung von der Regressionsgeraden insgesamt mißt, kann der Fall eintreten, daß es die Standardabweichung von Y zu etwa X_4 beträchtlich unterschätzt. Dies hätte zur Folge, daß man für den (bedingten) Erwartungswert $\mu_{Y|X_4}$ ein erheblich engeres Vertrauensintervall erhielte als es eigentlich ist, die Überdekkungswahrscheinlichkeit (die Sicherheitswahrscheinlichkeit) würde auf diese Weise (zu Unrecht) aufgebläht. Für X_2 würde gerade das Gegenteil des vorher Gesagten zutreffen.

3. Einige Beispiele

Die folgenden Beispiele sollen den Wert der Stichprobenresiduenplots als Werkzeug zur Überprüfung der Modellvoraussetzungen für die Regressionsrechnung demonstrieren.

Beispiel 1: Die Daten weiter unten beziehen sich auf das Geburtsgewicht (in Gramm) von 15 dreiwöchigen Laborversuchstieren sowie auf deren Gewichtszuwachs zufolge einer achtwöchigen speziellen Diät. Von Interesse ist dabei, ob der *Gewichtszuwachs* Y in einem bestimmten Ausmaß vom *Geburtsgewicht* X abhängt oder nicht. Die Regressionsgerade für diesen Datensatz ist $\hat{Y} = 32{,}53 + 1{,}25X$, wobei die Standardabweichung der Residuen $s = 8{,}48$ beträgt.

In der Tabelle sind die $\hat{Y}$-Werte (die Ausgleichswerte auf der Regressionsgeraden), die Residuen und die standardisierten Residuen angeführt. Das Streudiagramm der Daten (Fig. 6a) deutet an, daß

Tabelle 1

Anfangsgewicht X	51	72	54	64	74	60	69	68
Gewichtszunahme Y	110	112	90	119	133	122	115	126
$\hat{Y} = 32{\cdot}53 + 1{\cdot}25X$	96·1	122·2	99·8	112·3	124·7	107·3	118·5	117·2
$e = Y - \hat{Y}$	13·9	−10·2	−9·8	6·7	8·3	4·7	−3·5	8·8
e/s	1·64	−1·20	−1·16	0·79	0·98	0·55	0·41	1·04

Anfangsgewicht X	73	48	57	59	50	52	65
Gewichtszunahme Y	120	90	107	106	82	103	104
$\hat{Y} = 32{\cdot}53 + 1{\cdot}25X$	123·5	92·3	103·5	106·0	94·8	97·3	113·5
$e = Y - \hat{Y}$	−3·5	−2·3	3·5	0·0	−12·8	5·7	−9·5
e/s	−0·41	−0·27	0·41	0·0	−1·51	0·67	−1·12

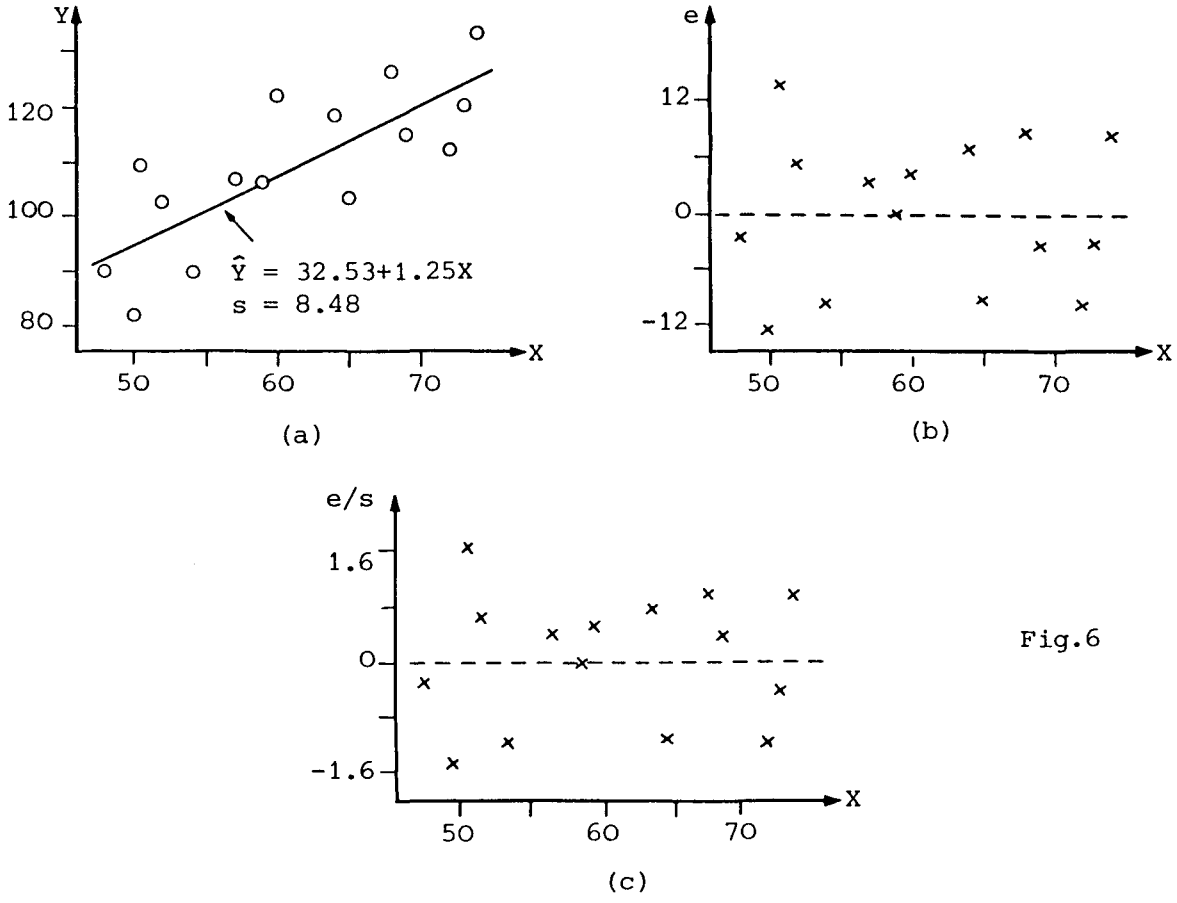
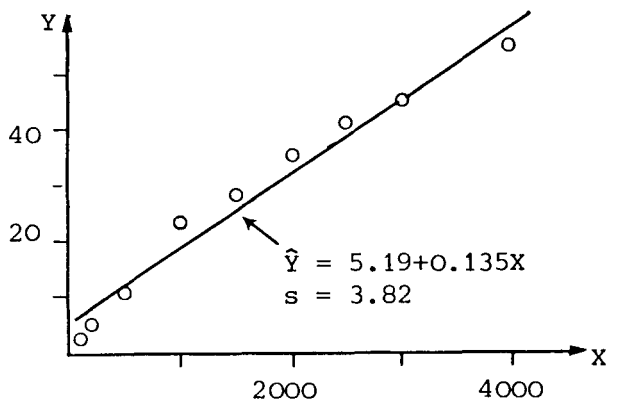


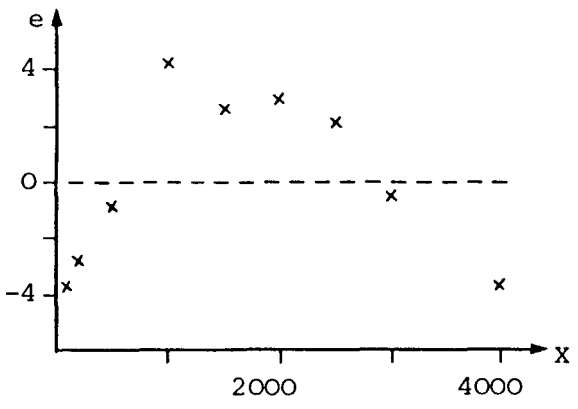
Fig.6

Tabelle 2

Zeit unter Belastung (Stunden) X	100	200	500	1000	1500	2000	2500	3000	4000
Anzahl der Ausfälle (von 100) Y	3	5	11	23	28	35	41	45	55
$\hat{Y} = 5{\cdot}19 + 0{\cdot}135X$	6·6	7·9	11·9	18·6	25·3	32·0	38·7	45·4	58·7
$e = Y - \hat{Y}$	−3·6	−2·9	−0·9	4·4	2·7	3·0	2·3	−0·4	−3·7
e/s	−1·08	−0·87	−0·27	1·33	0·81	0·90	0·69	−0·12	−1·11



(a)



(b)

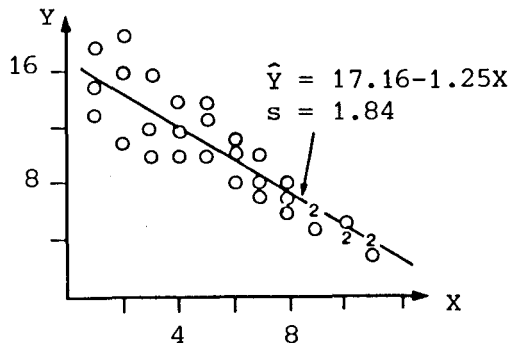
Fig.7

Tabelle 3

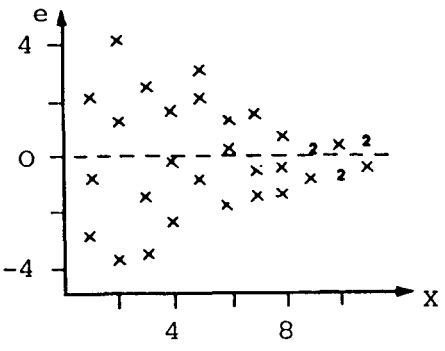
X	1	2	3	4	5	6
Y	15, 13, 18	19, 11, 16	12, 10, 16	14, 10, 12	10, 13, 14	10, 8, 11
$\hat{Y}$	15·9	14·7	13·4	12·1	10·9	9·7
e	−0·9, −2·9, 2·1	4·3, −3·7, 1·3	−1·4, −3·4, 2·6	1·9, −2·1, −0·1	−0·9, 2·1, 3·1	0·3, −1·7, 1·3

X	7	8	9	10	11
Y	8, 7, 10	8, 7, 6	6, 5, 6	5, 4, 4	3, 4, 4
$\hat{Y}$	8·4	7·2	5·9	4·7	3·4
e	−0·4, −1·4, 1·6	0·8, −0·2, −1·2	0·1, −0·9, 0·1	0·3, −0·7, −0·7	−0·4, 0·6, 0·6

$\hat{Y} = 17{\cdot}16 - 1{\cdot}25X$; $s = 1{\cdot}84$.



(a)



(b)

Fig.8

eine lineare Beziehung zwischen *Anfangsgewicht* X und *Gewichtszunahme* Y plausibel erscheint. Die Residuendiagramme e_i gegen X (Fig. 6b) sowie e_i/s gegen X (Fig. 6c) weisen eine unumstrittene Zufälligkeit auf. Es erscheint sinnvoll, daraus zu schließen, daß die Annahmen (1)-(3) für das Regressionsmodell erfüllt sind. Daher ist es zulässig, t-Statistiken zu verwenden, um Hypothesen bezüglich der Parameter α, β oder $\mu_{Y|X_i}$ zu testen oder dafür Vertrauensintervalle zu berechnen.

Beispiel 2: Es ist bekannt, daß ein bestimmtes elektronisches Bauteil manchmal nach Benützung unter bestimmten Belastungsbedingungen ausfällt. Gruppen zu 100 Bauteilen wurden für eine bestimmte, festgesetzte Zeitdauer X dieser Belastung unterzogen. Die Anzahl der Ausfälle (pro Hundert),Y, wurde für jede Belastungsdauer angegeben: Daten in Tabelle 2, Streudiagramm in Fig. 7a.

Der Residuenplot von e gegen X (Fig. 7b) zeigt einen systematischen Trend - dieser ist klar nichtzufällig. Die lineare Regressionsgleichung mag zwar für grobe Schätzungen dienlich sein, Tests und Konfidenzintervalle sollten mit großer Vorsicht betrachtet werden.

Beispiel 3: Versuchsmäuse werden trainiert, daß sie den Weg durch ein Labyrinth finden. Die Wahl einer falschen Abzweigung wird durch einen leichten Schock auf die Nase bestraft. Die Zahl der falschen Abzweigungen im Testlauf, Y, wird der Zahl der Probeläufe, X, gegenübergestellt. Je drei Mäuse hatten 1, 2, 3, . . . , 11 Probeläufe vor dem Testlauf. Die Resultate sind in Tabelle 3 enthalten, wieder mit $\hat{Y}_i$ und e_i.

Schon eine sorgfältige Prüfung des Streudiagramms der Daten (Fig. 8a) deutet eine breitere Streuung der Fehlerzahl bei wenigen Probeläufen im Gegensatz zu vielen an. Der Residuenplot (Fig. 8b) jedoch macht das Problem überdeutlich. Es ist klar, daß eine einzige, globale Varianzschätzung s^2 zu irreführenden Vertrauensintervallen und Hypothesentests führt.

Literatur:

Anscombe, F.J.: Graphs in Statistical Analysis,
In: The American Statistician 27 (1973)