Naturwissenschaft

Christian Hennig

Datenanalyse mit Modellen für Cluster linearer Regression

Doktorarbeit / Dissertation

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Dieses Werk sowie alle darin enthaltenen einzelnen Beiträge und Abbildungen sind urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsschutz zugelassen ist, bedarf der vorherigen Zustimmung des Verlages. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen, Auswertungen durch Datenbanken und für die Einspeicherung und Verarbeitung in elektronische Systeme. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe (einschließlich Mikrokopie) sowie der Auswertung durch Datenbanken oder ähnliche Einrichtungen, vorbehalten.

Copyright © 1997 Diplom.de
ISBN: 9783832421571

https://www.diplom.de/document/217988

Christian Hennig

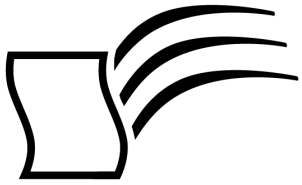
Datenanalyse mit Modellen für Cluster linearer Regression

Diplom.de

Christian Hennig

Datenanalyse mit Modellen für Cluster linearer Regression

Dissertation
an der Universität Hamburg
Fachbereich Mathematik
Institut für Mathematische Stochastik
Mai 1997 Abgabe



Diplomarbeiten Agentur
Dipl. Kfm. Dipl. Hdl. Björn Bedey
Dipl. Wi.-Ing. Martin Haschke
und Guido Meyer GbR

Hermannstal 119 k
22119 Hamburg

agentur@diplom.de
www.diplom.de

Hennig, Christian: Datenanalyse mit Modellen für Cluster linearer Regression /
Christian Hennig - Hamburg: Diplomarbeiten Agentur, 2000
Zugl.: Hamburg, Universität, Dissertation, 1997

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtes.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, daß solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden, und die Diplomarbeiten Agentur, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

Dipl. Kfm. Dipl. Hdl. Björn Bedey, Dipl. Wi.-Ing. Martin Haschke & Guido Meyer GbR
Diplomarbeiten Agentur, http://www.diplom.de, Hamburg 1999
Printed in Germany

Diplomarbeiten Agentur

Wissensquellen gewinnbringend nutzen

Qualität, Praxisrelevanz und Aktualität zeichnen unsere Studien aus. Wir bieten Ihnen im Auftrag unserer Autorinnen und Autoren Wirtschaftsstudien und wissenschaftliche Abschlussarbeiten – Dissertationen, Diplomarbeiten, Magisterarbeiten, Staatsexamensarbeiten und Studienarbeiten zum Kauf. Sie wurden an deutschen Universitäten, Fachhochschulen, Akademien oder vergleichbaren Institutionen der Europäischen Union geschrieben. Der Notendurchschnitt liegt bei 1,5.

Wettbewerbsvorteile verschaffen – Vergleichen Sie den Preis unserer Studien mit den Honoraren externer Berater. Um dieses Wissen selbst zusammenzutragen, müssten Sie viel Zeit und Geld aufbringen.

http://www.diplom.de bietet Ihnen unser vollständiges Lieferprogramm mit mehreren tausend Studien im Internet. Neben dem Online-Katalog und der Online-Suchmaschine für Ihre Recherche steht Ihnen auch eine Online-Bestellfunktion zur Verfügung. Inhaltliche Zusammenfassungen und Inhaltsverzeichnisse zu jeder Studie sind im Internet einsehbar.

Individueller Service – Gerne senden wir Ihnen auch unseren Papierkatalog zu. Bitte fordern Sie Ihr individuelles Exemplar bei uns an. Für Fragen, Anregungen und individuelle Anfragen stehen wir Ihnen gerne zur Verfügung. Wir freuen uns auf eine gute Zusammenarbeit

Ihr Team der *Diplomarbeiten* Agentur

Dipl. Kfm. Dipl. Hdl. Björn Bedey
Dipl. Wi.-Ing. Martin Haschke
und Guido Meyer GbR

Hermannstal 119 k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

agentur@diplom.de
www.diplom.de

Als Dissertation angenommen vom Fachbereich Mathematik der Universität Hamburg

auf Grund der Gutachten von Prof. Dr. Konrad Behnen
und Prof. Dr. Dietmar Pfeifer.

Hamburg, den 16.5.1997

Prof. Dr. Reiner Hass
Sprecher des Fachbereichs Mathematik

Diese Arbeit ist Prof. Gerhard „Heik“ Portele vom Zentrum für Hochschuldidaktik gewidmet. Er hat mir nahegebracht und vorgelebt, daß es beim Lehren in der Hochschule möglich ist, Autonomie zu gewähren, auf Machtausübung zu verzichten und in lebendigem Kontakt mit den Lehrenden zu bleiben. Prof. Portele starb am 10.7.1996 an Krebs.

Danksagung: Mein größtes Dankeschön gilt meinem Betreuer Prof. Konrad Behnen für Aufgeschlossenheit und Anregungen sowie für seine kritischen Anmerkungen über Darstellung und Lesbarkeit. Von diesen Hinweisen werden vermutlich die Leser/innen nicht nur dieser Arbeit sehr profitieren. Prof. Dietmar Pfeifer danke ich, daß er die Arbeit des Zweitgutachters übernommen hat. Ein weiterer besonderer Dank geht an Vanessa Didelez, Gabi Schneider, Dr. Silvelyn Zwanzig und Dr. Lutz Mattner, die Teile der Arbeit durchgesehen und wertvolle Kommentare zu Form und Inhalt gegeben haben. Zuletzt möchte ich allen weiteren Menschen danken, die sich für meine Arbeit interessiert, mir Literatur- und andere sinnvolle Hinweise gegeben oder mich in anderer Weise unterstützt haben.

Abstract[1]

A linear regression can be modeled by a family of distributions $(P_{\beta,\sigma^2} : \beta \in I\!R^{p+1}, \sigma^2 \in I\!R^+)$ for $(x, y) \in I\!R^{p+1} \times I\!R$, where $y = x'\beta + u$, u independent of x and distributed normal or symmetrically about 0 with variance σ^2.

This thesis deals with the analysis of datasets $(x_i, y_i) \in I\!R^{p+1} \times I\!R, i = 1, \ldots, n$. A linear regression distribution P_{β,σ^2} is treated as a distribution for one cluster, i.e. linear regression distributions with different parameters $(\beta_i, \sigma_i^2), i = 1, \ldots, s$ are supposed to be adequat for different parts of the dataset. Furthermore, there can be outliers in the data for which no such model is appropriate.

Various models for such data are introduced, especially mixture models of the form $\sum_{i=1}^{s} \epsilon_i P_{\beta_i,\sigma_i^2}$. Maximum Likelihood estimation of the parameters (β_i, σ_i^2) is discussed. New proposals for estimating the number of clusters s are given.

Sufficient conditions for the identifiability of the parameters are derived. Counter-examples are given in some situations where the conditions do not hold.

As a new method, Fixed Point Cluster Analysis (FPCA) is introduced. It enables the analysis of data with unknown number of clusters s and outliers. FPCA bases on the identification of outliers and can be generalized to other clustering problems. A Fixed Point Cluster (FPC) corresponds to a subset of $I\!R^{p+1} \times I\!R$ and should contain points (x, y) which belong together in some sense. Every FPC corresponds to parameters $(b, s^2) \in I\!R^{p+1} \times I\!R^+$ which can be interpreted as estimation of the regression parameters (β_i, σ_i^2). FPC are defined for datasets and distributions.

Convergence of an algorithm for the computation of FPC for given datasets is proven.

Distributions of the form $(1-\epsilon)P_{\beta_0,\sigma_0^2} + \epsilon H^*$ are considered. P_{β_0,σ_0^2} here is interpreted as a distribution for a linear regression cluster. H^* is a distribution on $I\!R^{p+1} \times I\!R$, e.g. a mixture of other P_{β,σ^2}. The existence of FPC is shown under various assumptions on H^* and ϵ. The parameters of these FPC lie in a bounded neighborhood of (β_0, σ_0^2). For homogenous regression distributions ($\epsilon = 0$) exists one and only one FPC. It has parameters (β_0, σ_0^2).

In a simulation study FPCA and two Maximum Likelihood procedures are compared.

[1]Eine deutsche Zusammenfassung findet sich auf Seite 179.

Inhaltsverzeichnis

1 Einführung

1.1 Das Problem

Der folgende Datensatz findet sich auf Seite 26 von Rousseeuw und Leroy (1988). Er enthält die von Belgien aus geführten internationalen Telefongespräche (in 10 Millionen) in den Jahren 1950-1973.

Nr.	Telefonate (y)	Jahr (x)		Nr.	Telefonate (y)	Jahr (x)
1	0.44	50		13	1.61	62
2	0.47	51		14	2.12	63
3	0.47	52		15	11.9	64
4	0.59	53		16	12.4	65
5	0.66	54		17	14.2	66
6	0.73	55		18	15.9	67
7	0.81	56		19	18.2	68
8	0.88	57		20	21.2	69
9	1.06	58		21	4.3	70
10	1.20	59		22	2.4	71
11	1.35	60		23	2.7	72
12	1.49	61		24	2.9	73

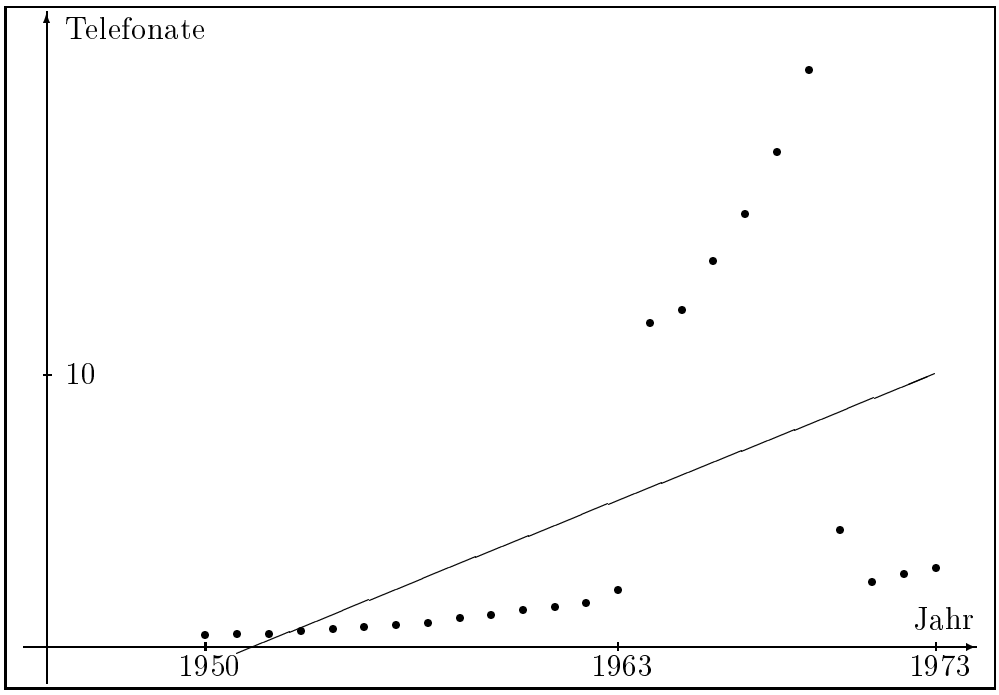


Abbildung 1: Telefondatensatz

In Abbildung 1 fällt sofort auf, daß sich die Telefonate in den Jahren von 1964-1970 grundsätzlich anders verhalten als die Mehrheit der Daten. Der Zusammenhang zwischen Jahr und Telefonatezahl sieht für die Jahre 1950-1962 und 1971-1973 annähernd linear

aus. Auf Nachfrage erfuhren Rousseeuw und Leroy, daß 1964-1969 nicht die Telefonate, sondern die Minuten gezählt wurden, die die Telefonate insgesamt dauerten. 1963 und 1970 wurden beide Verfahren teilweise angewendet.

In der robusten Statistik wurde dieser Datensatz häufig diskutiert als Beispiel für eine lineare Regression mit mehreren Ausreißern. Berechnet man den Kleinste-Quadrate-Schätzer (KQ) zum Modell

$$y = \beta_1 x + \beta_2 + u,\ E(u) = 0,$$

so paßt die resultierende Gerade fast keinen Punkt gut an (steigende Linie in der Abbildung 1).

Bemerkung 1.1 *Gegeben sei ein Datensatz* $(\mathbf{X}, y)$*,*

$$\mathbf{X} = (x_1, \ldots, x_n)',\ y = (y_1, \ldots, y_n)',\ x_i \in I\!R^{p+1},\ y_i \in I\!R\ \forall i = 1, \ldots, n.$$

Dann ist der KQ-Schätzer $\hat{\beta}_{KQ} \in I\!R^{p+1}$ *definiert durch*

$$\sum_{i=1}^{n} (y_i - x_i' \hat{\beta}_{KQ})^2 \stackrel{!}{=} \min .$$

Das heißt, falls $(\mathbf{X}'\mathbf{X})^{-1}$ *existiert,* $\hat{\beta}_{KQ} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$*. Im obigen Fall der Regression mit Achsenabschnitt* β_2 *werden die* $x_i = (x_{i1}, x_{i2})$ *als Elemente aus* $I\!R^2$ *interpretiert, wobei immer* $x_{i2} = 1$*.*

Wählt man aber einen robusten Regressionsschätzer wie zum Beispiel „Least Median of Squares“ (siehe Rousseeuw und Leroy (1988)), so wird eine Gerade geschätzt, die nur die Mehrheit der Jahre gut anpaßt, in denen die Gespräche gezählt wurden. Die Daten werden also unterteilt in „gute“ Daten und „Ausreißer“. Aber was ist mit den Daten von 1964-69? Sie sind ja nicht falsch, sondern nur andersartig. Besteht bei ihnen vielleicht auch ein einfacher linearer Zusammenhang? Da es so wenige sind, ist das vom optischen Eindruck her nicht klar. Inhaltlich wird man zumindest einen approximativ linearen Zusammenhang bei den Gesprächslängen vermuten, falls Linearität für die Anrufsanzahlen vorausgesetzt wird.

Das Thema dieser Arbeit ist die Clusteranalyse von Daten aus linearen Regressionen. Das heißt: Es geht darum, Gruppen von Daten zu finden, wobei Daten zusammen eine Gruppe bilden sollen, wenn sie durch denselben linearen Zusammenhang zwischen der (möglicherweise mehrdimensionalen) Regressorvariable x und der (eindimensionalen) abhängigen Variablen y erzeugt wurden. Zur Modellierung der Daten einer Gruppe soll also ein klassisches lineares Regressionsmodell (siehe (2.1) in Abschnitt 2) adäquat sein. Zu beachten ist dabei, daß hier im Unterschied zur häufigsten Verwendung des Wortes „Cluster“ (Klumpen) die Zusammengehörigkeit von Punkten nicht direkt mit ihrem Abstand voneinander zusammenhängt. Das ist in Abbildung 1 zum Beispiel zu sehen, wenn man den Punkt für 1973 betrachtet, der weiter vom 1950er Punkt entfernt ist als von sämtlichen „Ausreißern“.

Zu diesem Ziel werden zunächst Maximum Likelihood- und andere bekannte Ansätze untersucht. Dann führe ich im Hauptteil der Arbeit die Fixpunktclusteranalyse ein, die speziell zur Clusteranalyse bei Clustern unterschiedlicher Art und Präsenz von Ausreißern dienen soll.

1.2 Modelle für die Clusteranalyse (Teil I)

Der übliche stochastische Zugang zu einem Clusteranalyse-Problem ist die Formulierung eines möglichst einfachen Clustermodelles[2]. Innerhalb dieses Modells kann dann nach Schätzern mit guten Eigenschaften für die Regressions- und Störskalenparameter der einzelnen Cluster gesucht werden.

Es gibt zwei unterschiedliche Methoden, stochastische Modelle für die Clusteranalyse zu formulieren: Mischmodelle, d.h. Modelle, bei denen die Punkte unabhängig identisch verteilt sind. Die Werte werden mit festgelegten, aber unbekannten Wahrscheinlichkeiten aus unterschiedlichen Populationen erzeugt. In Modellen mit fester Zuordnung sind die Punkte unterschiedlicher Cluster dagegen unterschiedlich verteilt und die Zugehörigkeit eines Punktes zu einem Cluster wird als fester, unbekannter Modellparameter behandelt. In Abschnitt 2 werden die unterschiedlichen Modelle vorgestellt.

Ein Spezialfall der zweiten Modellvariante sind Wechselpunktprobleme („changepoint problems"), über die im Regressionsfall am meisten bekannt ist. In einem Wechselpunktmodell ändern sich die Regressionsparameter in Abhängigkeit von der Zeit oder anderen Regressoren. In Abschnitt 3.1 wird ein kurzer Überblick über die diesbezügliche Literatur gegeben. Ein solches Modell könnte auch für den Telefondatensatz benutzt werden. Allerdings wird in der Literatur über Wechselpunktprobleme normalerweise nicht vorgesehen, daß ein System wieder in den alten Zustand zurückspringt (im Datensatz nach 1970).

Weiter wurden Kleinste-Quadrate- und Maximum Likelihood(ML)-Schätzer für den Fall vorgeschlagen, daß die Zugehörigkeit der Punkte zu den Clustern als unabhängig von den Regressoren vorausgesetzt wird. Diese Ansätze werden in den Abschnitten 3.2 und 3.3 diskutiert. Über die theoretischen Eigenschaften dieser Schätzer gibt es bislang im Regressionsfall kaum wesentliche Resultate. Ein großer Teil der Literatur befaßt sich mit der Entwicklung konvergenter Algorithmen zur Berechnung der Schätzer. Für die Schätzung der Clusterzahl wird häufig die Minimierung von informationsbasierten Kriterien vorgeschlagen, für die es aber nur wenig theoretische Rechtfertigung gibt.

Im allgemeinen kann über Abhängigkeiten zwischen Regressoren und Regressionsparametern keine einfache Voraussetzung gemacht werden. Clustermodelle mit fester Zuordnung ohne die restriktiven Voraussetzungen des Wechselpunktproblems wurden bislang nur im Lokationsproblem[3] behandelt. In Abschnitt 3.4 übertrage ich einen ML-Ansatz von Scott und Symons (1971) auf den linearen Regressionsfall. Abschnitt 3.5 stellt kurz alternative Ansätze zur Behandlung des Regressions-Clusterproblemes vor. In den Teilen von Abschnitt 3 wird ein Überblick über die bisher vorhandene Literatur zur Problemstellung gegeben.

Eine wesentliche Voraussetzung für Resultate über konsistente Schätzungen in clustererzeugenden Modellen ist die Identifizierbarkeit der Modellparameter: Die Parameterwerte, die eine bestimmte Verteilung definieren, müssen eindeutig sein. In den Abschnitten 4 bis 6 wird die Identifizierbarkeit der vorgestellten Modelle untersucht.

[2]Unter einem „Modell" verstehe ich eine Familie von Verteilungen $\{P_\theta, \theta \in \Theta\}$ auf einem Raum mit einer $\sigma-$Algebra, üblicherweise $(\mathbb{R}^d, \mathbb{B}^d)$. Mit „Verteilung" meine ich P_θ für ein bestimmtes θ.

[3]Wenn in dieser Arbeit vom Lokationsproblem die Rede ist, dann ist die Analyse von Daten mit Modellen gemeint, in denen unterschiedliche Teilmengen der Datenpunkte (Cluster) durch Verteilungen der Form $F\left[\mathbf{A}(y-b)\right], y \in \mathbb{R}^p$ mit unterschiedlichen Parametern $b \in \mathbb{R}^p$ (Lokations-, Lageparameter) beschrieben werden sollen. Der Modellparameter $\mathbf{A} \in \mathbb{R}^{p^2}$ kann fest, frei, bekannt oder unbekannt sein.

Es stellt sich heraus, daß häufig nicht alle Parameter identifizierbar sind. Zum Beispiel sind im Modell mit fester Zuordnung die Zuordnungsparameter nicht identifizierbar. Für die Identifizierbarkeit von Regressions- und Störskalenparametern werden hinreichende Bedingungen an die Regressoren hergeleitet.

Der Telefondatensatz wirft aber Probleme auf, die mit der skizzierten Herangehensweise schwerlich zu lösen sind:

- Es ist nicht klar, ob ein Modell mit mehreren Clustern dem Datensatz angemessener ist als ein Modell mit einer Mehrheit von Daten aus demselben Regressionsmodell und einer Minderheit nicht näher spezifizierter Ausreißer.
- Es ist nicht klar, ob der Zusammenhang in allen Clustern linear ist.
- Es ist nicht klar, ob es Punkte gibt, die sinnvollerweise zu gar keinem oder mehreren Clustern dazugerechnet werden sollten. Was ist mit den Jahren 1963 und 1970, als die Zählung umgestellt wurde?

Diese Probleme tauchen nicht nur im Falle der Telefondaten auf. Welches Modell für einen gegebenen Datensatz angemessen ist, weiß man von vornherein nie.

1.3 Exkurs: Angemessenheit von Modellen

Für die Motivation der späteren Abschnitte spielt die Funktion von Modellen in der Datenanalyse eine große Rolle. Daher möchte ich kurz die Vorstellungen skizzieren, die für meine Arbeit maßgeblich sind.

Der Satz „Ein bestimmtes Modell ist angemessen für einen Datensatz“ bedeutet sinnvollerweise nicht: „Der Datensatz ist von einer Verteilung dieses Modells generiert worden.“ Eine solche Aussage würde sich auf keine Weise verifizieren lassen, und es ist kaum vorstellbar, daß sie jemals stimmen könnte. Davies (1995) schreibt:

The term „adequate“ reflects the philosophy that a model is not true nor even treated as true. The model is regarded as being adequate for some given purpose. (...) The adequacy region specifies those probability models whose samples typically look like the actual data.

Die „adequacy region“ ist Davies' Ansatz, Angemessenheit formal zu definieren. „Typically look like“ bedeutet hier, daß der Datensatz eine - je nach Interpretationsziel definierte - Eigenschaft hat, die Datensätze aus dem entsprechenden Modell mit hoher Wahrscheinlichkeit haben.

„Angemessenheit“ hat bei Davies also zwei Aspekte:

- Erzeugt man künstlich Daten aus einer geeigneten Verteilung eines angemessenen Modells, so sollen diese Daten dem vorliegenden Datensatz ähnlich sehen.
- Der Begriff „ähnlich“ ist subjektiv. Ob ein vorgegebener Datensatz einem typischen Modelldatensatz „ähnlich“ sieht, hängt von Ähnlichkeitskriterien ab, die man selbst wählen muß.

Ein dritter wichtiger Aspekt ist, daß ein angemessenes Modell dafür geeignet sein sollte, die Fragen zu beantworten, die man an den Datensatz hat.

Zum Beispiel wäre ein homogenes lineares Regressionsmodell mit normalverteiltem Störterm u für den Telefondatensatz nicht angemessen: Die Residuen sind in auffälliger Weise und entgegen den Modellvoraussetzungen abhängig vom Regressor x (was zu formalisieren wäre, um Davies' Ansatz anzuwenden). Ein lineares Regressionsmodell für die Jahre 1950-1963 und 1971-1973 wird dagegen nach Davies' Kriterien kaum für unangemessen gehalten werden können. Es kann jedoch nicht alle Fragen an den Datensatz beantworten, wenn man sich dafür interessiert, wie die restlichen Jahren genau zu interpretieren sind. Eine Mischung aus zwei linearen Regressionsmodellen kann vermutlich die Jahre 1964 und 1970 nicht „angemessen" anpassen. In Abschnitt 10.1 wird der Telefon-Datensatz als Anwendungsbeispiel für die in dieser Arbeit betrachteten Verfahren diskutiert.

Es gibt auch Datensätze, die - bis auf die Verwendung eines Zufallszahlengenerators - tatsächlich aus einer Mischung mehrerer linearer Regressionen stammen, wobei aber diese Mischung mit statistischen Methoden kaum von einem geeigneten homogenen Modell oder einer Mischung mit ganz anderen Parameterwerten zu unterscheiden ist. Zum Beispiel ist x in Abbildung 2 verteilt nach $\mathcal{N}_{(0,1)}$, mit Wahrscheinlichkeit 0.5 ist $y = 0.5x + u$, mit derselben Wahrscheinlichkeit $y = -0.5x + u$, wobei u verteilt nach $\mathcal{N}_{(0,2)}$ erzeugt worden ist. Das erzeugende Modell ist dem Datensatz sicher nach Davies'schen Kriterien angemessen. Dennoch bringt die Analyse des Datensatzes mit den Regressionsparametern eines solchen Mischmodells offenbar keine anschaulich brauchbare, interpretierbare Vorstellung von den Daten.

Diese Diskussion soll verdeutlichen, was gemeint ist, wenn in dieser Arbeit von „Angemessenheit" die Rede ist. Das Wort wird allerdings informell benutzt. Das formale Konzept der „adequacy region" wird nicht weiter verwendet.

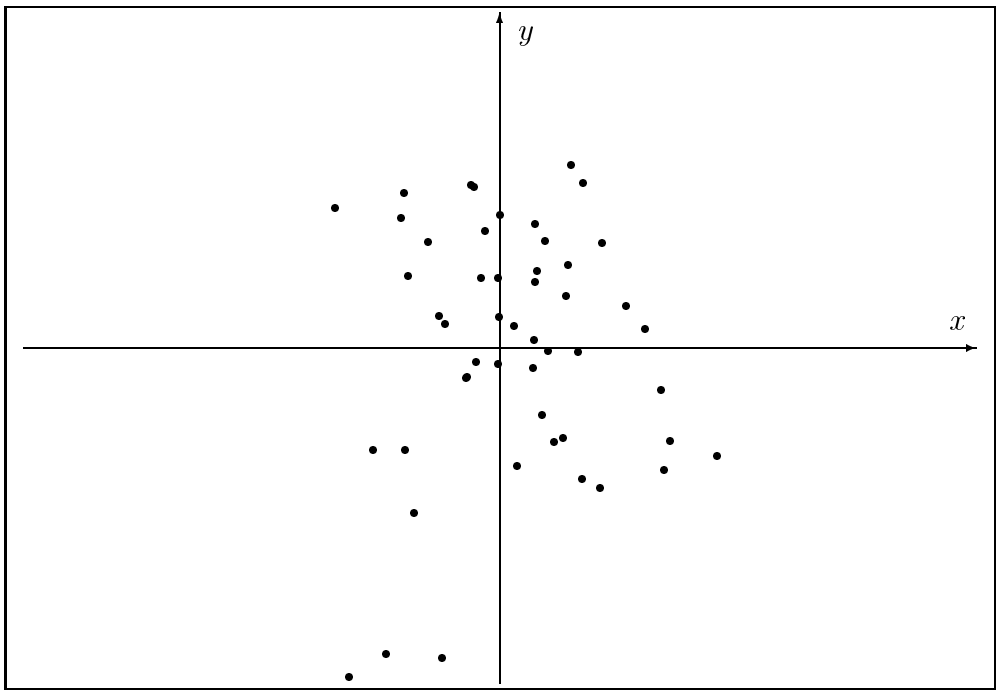


Abbildung 2: Gibt es hier Cluster?