THEORIE UND PRAXIS VON LAD-SCHÄTZERN

Christian Donninger
Camillo Signor
Zbigniew Wasilewski

Forschungsbericht/
Research Memorandum No. 201
Juni 1984

Die in diesem Forschungsbericht getroffenen Aussagen liegen im Verantwortungsbereich der Autoren und sollen daher nicht als Aussagen des Instituts für Höhere Studien wiedergegeben werden.

I N H A L T S V E R Z E I C H N I S

E I N L E I T U N G

Die vorliegende Arbeit geht auf ein Ökonometrie-Seminar am Institut für Höhere Studien zurück. Bei der Behandlung des Themas Least Absolute Deviations Schätzungen stellte sich heraus, daß keine Theoretiker und Praktiker gleichermaßen ansprechende Literatur zu diesem Thema existiert.
Wie im Titel bereits formuliert, sollen in dieser Arbeit beide Bereiche behandelt werden. Dementsprechend sind auch die notwendigen Vorkenntnisse zum Verständnis der einzelnen Beiträge äußerst unterschiedlich.

Beitrag 1 von Ch. Donninger ist eine für Ökonometrie-Praktiker konzipierte Motiviation zur Verwendung von LAD-Schätzern. Ohne auf mathematische Details einzugehen werden eine Reihe von qualitativen Aussagen ("Faustregeln") über LAD-Schätzer präsentiert. Kern dieser Regeln: OLS und LAD verhalten sich wie Mittelwert und Median. Überall dort, wo man im 1-Dimensionalen als Lagemaß den Median verwendet, sollte man bei Regressionen mit LAD - Schätzern arbeiten. Zum Verständnis dieses Beitrages ist im Prinzip nur guter Wille notwendig.

Beitrag 2 von C. Signor beschäftigt sich mit den statistischen Eigenschaften von LAD-Schätzern. Es gelingt ihm zu zeigen, daß die Schätzwerte asymptotisch normalverteilt sind. Damit gibt es gute Gründe, die bei OLS üblichen Tests für die Parameterwerte auch bei LAD anzuwenden. Der Beweis stützt sich auf eine Arbeit von Basset und Koenker, die allerdings einen - nun behobenen - logischen Fehler enthält. Beitrag 2 ist somit der erste vollständige Beweis der asymptotischen Normalverteilung der LAD-Parameter.

Beitrag 3 von Z. Wasilewski beschreibt im Detail die numerischen Verfahren zur Berechnung von LAD-Schätzern. Die beiden ab Herbst 1984 im IAS-System des Instituts für Höhere Studien verfügbaren Algorithmen werden auch an Hand von Beispielen vorgeführt. Mit Hilfe von Simulationen wurde die

Effizienz der Algorithmen abgeschätzt, sowie auf Probleme bei der Berechnung von "entarteten" Ausgangsdaten eingegangen.

Die einzelnen Beiträge sind unabhängig voneinander und können daher einzeln gelesen werden. Den mit LAD-Schätzern weniger vertrauten Lesern sei allerdings Beitrag 1 als Einstieg empfohlen, während für versiertere Leser die Beiträge 2 und 3 von größerem Interesse sein werden.

Wir hoffen mit dieser Arbeit eine Lücke in der Ökonometrie Literatur geschlossen zu haben und den im Titel verkündeten Anspruch zumindest näherungsweise erfüllt zu haben.

Wien, Mai 1984

Christian DONNINGER
Camillo SIGNOR
Zbigniew WASILEWSKI

WAS MAN ÜBER LEAST ABSOLUTE DEVIATION WISSEN SOLLTE

Christian Donninger

Was man über Least Absolute Deviation wissen sollte:

Das folgende Kapitel ist eine Ansammlung von (wichtigen) Faustregeln über LAD-Schätzer. Nicht mehr (und nicht weniger). Dieses Kapitel ist also für den Anwender gedacht, dem technische Feinheiten relativ wenig interessieren. Wohl aber die qualitativen Unterschiede von OLS- und LAD-Schätzern.

1) Warum LAD???

OLS-Schätzung (plus Erweiterungen) gehört heute zum Standardwissen jedes Ökonomen. Bereits bessere Taschenrechner besitzen OLS-Berechnungsroutinen. Welchen Sinn hat dann die zusätzliche Einführung der wesentlich komplizierteren LAD-Schätzer?
Dazu ein Beispiel: Wir nehmen an, wir wollen den beruflichen Erfolg(gemessen als Einkommen) von HAK und Gymnasiumsabsolventen eines bestimmten Jahrgang untersuchen.(Der Einfachheit halber jeweils 5 zufällig gewählte Personen).

Schulabgänger 1976, HAK/Gymnasium Vöcklabruck O.Ö

Tab.1:	Einkommen am 1.1.83 (in 1000S.)	
HAK:	8,11,12,14,25	Mittelwert=14; Median=12
GYMN:	6,14,14,17,20	Mittelwert=14,2; Median=14
Tab.2	Einkommen am 1.1.84 (in 1000S.)	
HAK:	8,11,12,14,85	Mittelwert=26; Median=12
GYMN:	6,14,14,17,20	Mittelwert=14,2; Median=14

Im Jahr 1983 hat sich nur ein Einkommen verändert. Der bisher bereits meistverdienende HAK-Absolvent hat den Aufstieg vom mittleren zum oberen Managment geschafft. Waren zu Beginn des Jahres die Mittelwerte von HAK- und Gymn. Einkommen fast gleich, so bewirkte der Einkommenssprung dieses "Ausreißers" beinahe eine Verdoppelung des HAK-Mittelwertes. Der entsprechende Median veränderte sich nicht. Hier blieben die Relationen zwischen HAK und Gymn. gleich.

REGEL 1: Der Mittelwert reagiert stark auf Ausreißer.
Der Median reagiert nicht auf Ausreißer.

Ob man hier(und in anderen Fällen) den Median oder den Mittelwert als angemesseneres Maß betrachtet, läßt sich auf rein logischer Ebene nicht entscheiden. In unserem Zusammenhang ist nur das stark unterschiedliche Verhalten gegenüber Ausreißern von Bedeutung. Es gilt nämlich der folgende Satz:

SATZ 1: Der Mittelwert ist ein Ordinary Least Square(OLS) Schätzer.
Der Median ist ein Least Absolute Deviation(LAD) Schätzer.

REGEL 2: Die Schätzparameter für die Regression von OLS bzw. LAD verhalten sich bezüglich Ausreißern genauso wie Mittelwert und Median. (OLS reagiert stark, LAD nicht).

Die Frage OLS oder LAD bedeutet also:"Soll der Einfluß von Ausreißern berücksichtigt werden oder nicht?".

In der modernen ökonometrischen Literatur wird häufig von robusten Schätzverfahren gesprochen.

DEF.[1] 1: Ein Schätzverfahren heißt robust, wenn es auf Ausreißer weniger reagiert als OLS.

REGEL 3: LAD ist ein robustes Schätzverfahren.

Anmerkung: Ein anderes auch häufig angewandtes robustes Verfahren wäre: Schneide extrem hohes bzw. niedriges Einkommen bei einer bestimmten Grenze ab und berechne den Mittelwert mit diesen "gestutzten" Werten.

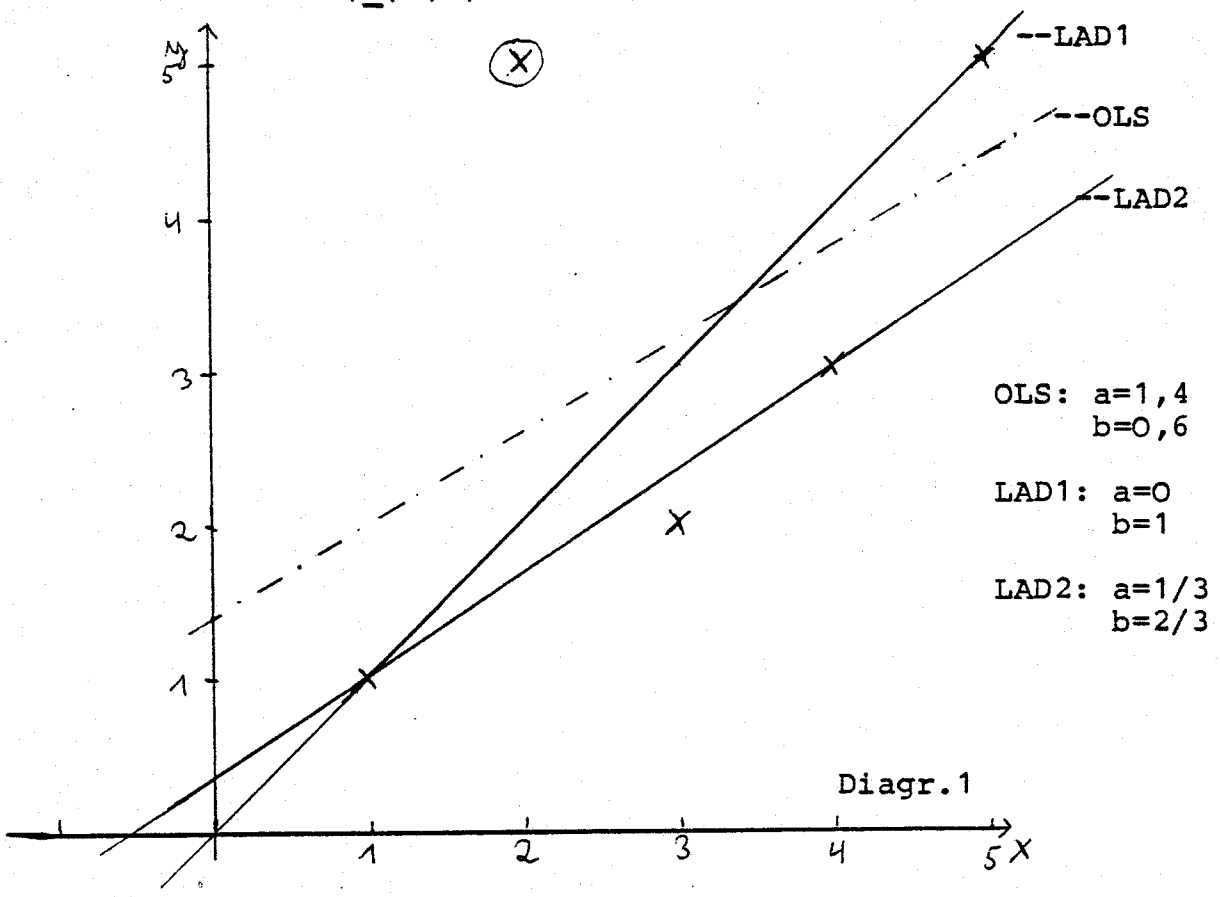
1) Def. 1 ist natürlich keine exakte, mathematische Definition von robust. Sie soll nur anschaulich zusammenfassen, was in den zumeist sehr komplizierten Definitionen "steckt".

2) 2-Dimensionale Regression mit LAD:

Gegeben sei die exogene Variable X und die endogene Variable Y.

X	1,2,3,4,5
Y	1,5,2,3,5

Bsp. 1



Diagr.1

Annahme 1: LAD-Regressionsgerade geht durch Median von X und Y. D.h. Punkt (3,3).

Allgemeine Problemdefinierung:

Gegeben Punkte $\{X_i, Y_i\}$ $i=1,\ldots,n$

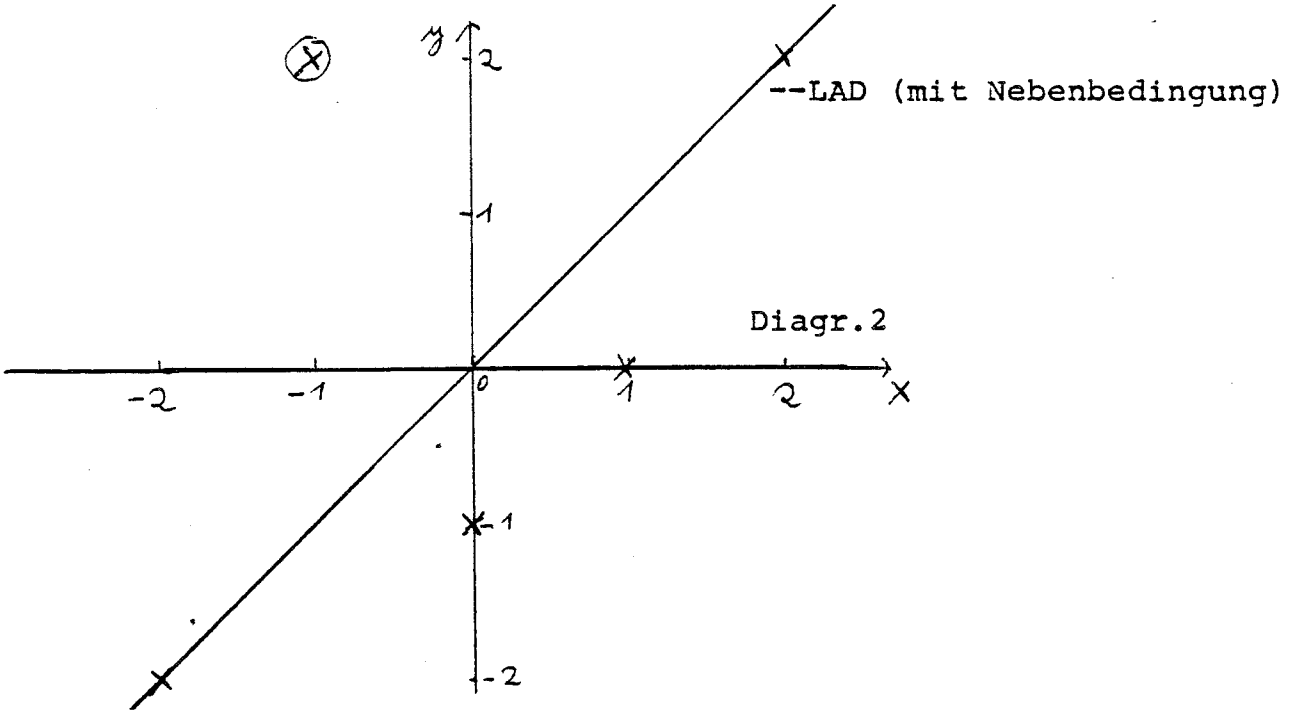
Finde Geradengleichung $\hat{Y}=b.X+a$ so daß

$$S=\sum_{i=1}^{n} |Y_i-\hat{Y}_i| \quad \text{ein Minimum wird. (LAD-Bedingung)}$$

Nach Annahme 1 soll die von uns gesuchte Gerade durch den Median M=(3,3) gehen. (Annahme 1 ist willkürlich und hat nichts mit der LAD-Bedingung zu tun. Sie dient nur zur Vereinfachung des Beispiels).
Zieht man M von jedem Punkt ab, ist der neue Median $\hat{M}=(0,0)$.

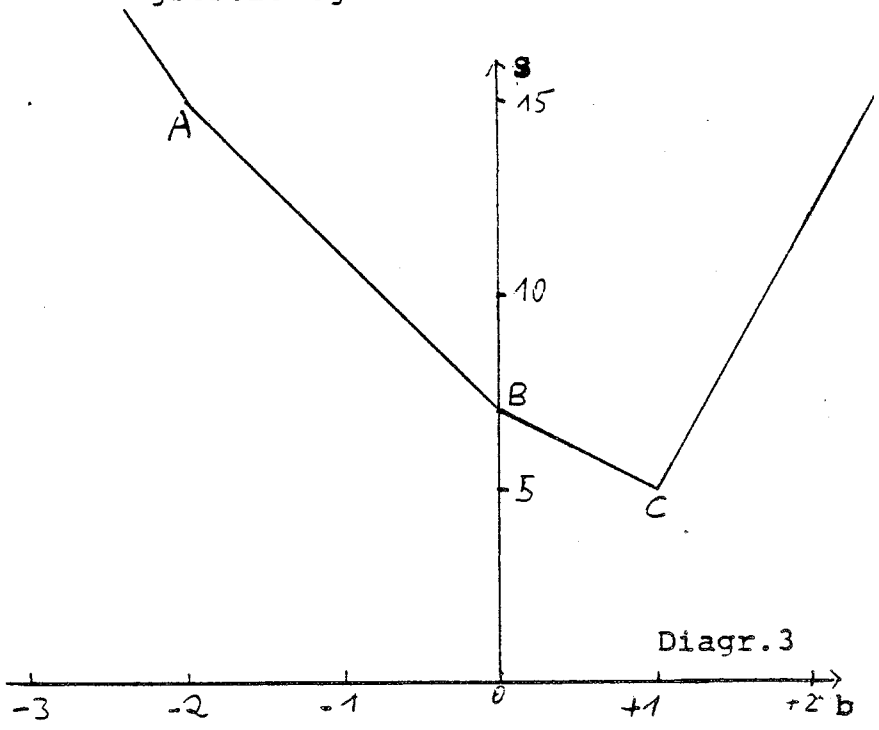
Die neuen Werte $\tilde{X}$ und $\tilde{Y}$ lauten dann:

$\tilde{X}$	-2, <u>-1</u>, 0, 1, 2	Laut Annahme 1 geht die Regressionsgerade (LAD) durch den Ursprung.
$\tilde{Y}$	-2, <u>2</u>, -1, 0, 2	Der konstante Faktor a der Regressionsgeraden $\hat{Y}=b.X + a$ wird Null. Gesucht ist also ein b, sodaß der Ausdruck $S=\sum_{i=1}^{5} \lvert \tilde{Y}_i-\hat{\tilde{Y}}_i \rvert$ ein Minimum wird.



Diagr.2

Die Größe von S hängt nur von b, von der Steigung der Regressionsgeraden ab.



Diagr.3

S, die Summe der absoluten Abweichungen, ist eine konvexe, stückweise lineare Funktion. D.h. Sie bildet die Form eines (verbogenen) U(konvex) und besteht aus einem Zug unterschiedlich geneigter Geraden (stückweise linear).

Die Kurve S hat genau an den Stellen, in denen die Regressionsgerade durch einen Punkt geht, einen Knick.(Das gilt auch für den mehrdimensionalen allgemeinen Fall). Diese "Knicks" verursachen die Schwierigkeiten bei der Berechnung des(der) Regressionsparameters. Normalerweise bestimmt man das Minimum einer Funktion (d.h. das Minimum von S), indem man die Ableitung Null setzt. Auf diese Weise gelangt man auf die relativ einfache Formel für die OLS Parameter. Eine Kurve hat aber in einem "Knick" keine Ableitung. Wie man aber im Diagramm 3 sieht, hat S genau im "Knickpunkt" C ein Minimum. Das ist kein Zufall, sondern gilt immer. Das Minimum liegt immer auf einer Ecke von S.
In unserem einfachen Beispiel braucht man also nur die Ecken, Knickpunkte, A,B,C absuchen und schauen, bei welchen Punkt S den kleinsten Wert hat. Man kann also bei LAD nicht durch Ableitung die Parameter berechnen, sondern muß die Ecken durchsuchen. Das ist in unserem Beispiel sehr einfach, führt aber im Allgemeinen zu sehr rechenaufwendigen Verfahren, die bis vor kurzem die praktische Anwendung von LAD unmöglich machten.

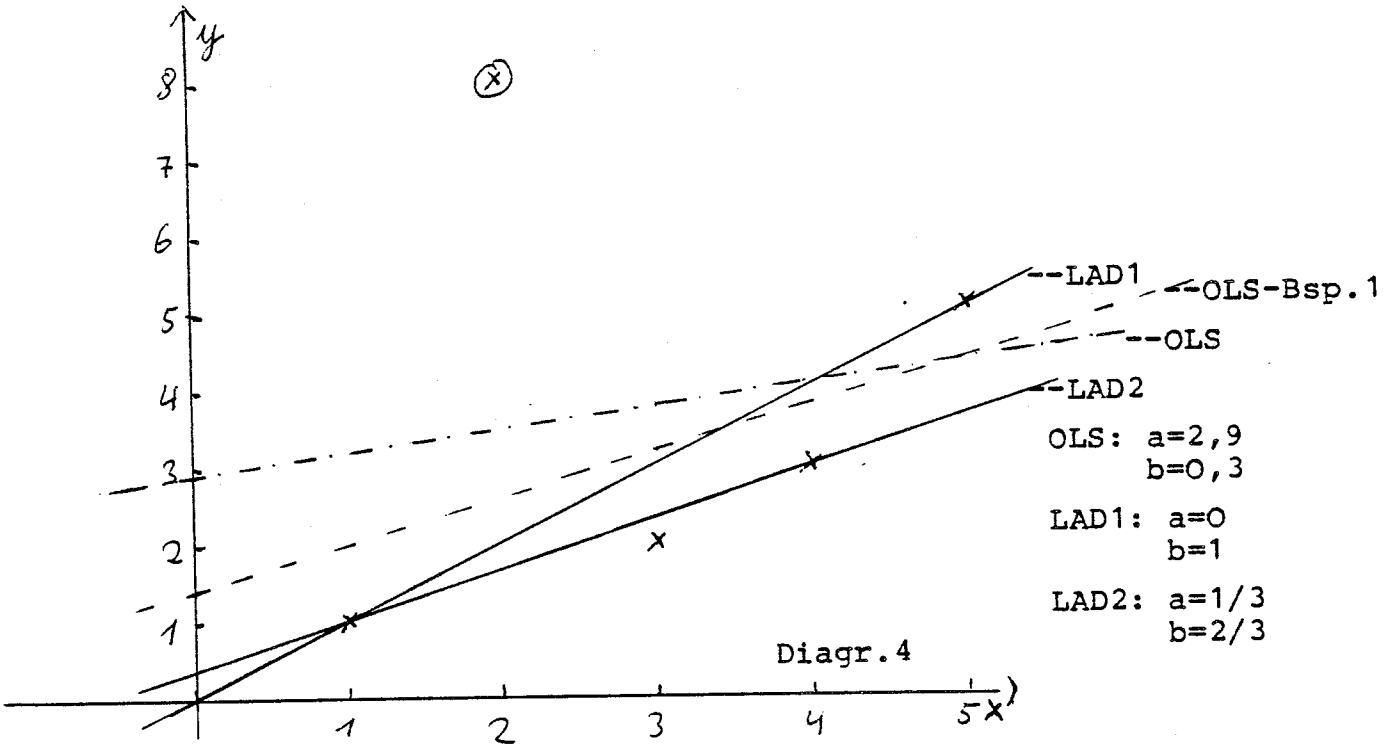
Bsp.2: Analog wie beim Einkommen von HAK und Gymn. Absolventen, soll jetzt gegenüber dem Bsp.1 dieses Absatzes nur ein einziger Wert verändert werden.

X	1,2,3,4,5
Y	1,8,2,3,5

Im Beispiel 1 hat Y_2 den Wert 5, in diedem Beispiel den Wert 8. Alle anderen Größen wurden nicht verändert.(Diagr. 4)

Wie man in Diagramm 4(nächste Seite) sieht, ist die LAD-Schätzung nicht eindeutig!! Jede Gerade die durch den Punkt (1,1) geht und zwischen den beiden Regressionsgeraden liegt, ist ebenfalls eine LAD-Regressionsgerade.Vergleicht man die Parameter von Bsp.1 und Bsp. 2, so erhält man folgende Tabelle.

Tab.3	OLS		LAD1		LAD2	
	a	b	a	b	a	b
Bsp.1	1,4	0,6	0	1	1/3	2/3
Bsp.2	2,9	0,3	0	1	1/3	2/3



Diagr.4

Analog dem Verhältnis von Mittelwert(OLS) und Median(LAD) im 1-ten Abschnitt gilt auch hier:

Die OLS-Parameter wurden durch die Lage des Ausreißers stark verändert.

Die LAD-Parameter wurden durch die Lage des Ausreißers nicht verändert.

Die OLS-Parameter sind (wenn X vollen Rang hat) eindeutig bestimmt.

Die LAD-Parameter müssen nicht eindeutig sein.

3) Der n-dimensionale Fall:

Gegeben sei folgendes Modell:

$Y=X.\beta + u$ wobei $X=n\times k$ - Matrix $n\geq k$
$Y'=X.\hat{\beta}$ $Y=n$-dimensionaler Vektor
$\beta=k$-dimensionaler Vektor der Parameter
$u=n$-dimensionaler Zufallsvektor

Die LAD-Bedingung lautet nun: Gesucht $\hat{\beta}$ so daß :

$$S=\sum_{i=1}^{n} | Y_i-Y'_i | = \sum_{i=1}^{n} |Y_i-X_{i1}\cdot\hat{\beta} -\ldots X_{ik}\cdot\hat{\beta}_k| \quad \text{ist ein Minimum.}$$

Der einzige Unterschied zu OLS besteht darin, daß nicht die

Summe der Abweichungsquadrate, sondern die Summe der Absolutabstände minimiert wird.

E I G E N S C H A F T E N von L A D:

E1: Unabhängig vom Rang von X existiert immer mindestens eine Lösung. (Es können aber auch mehrere sein)

E2: Es existiert mindestens eine Lösung die durch k der n Punkte geht. (Im 2-Dimensionalen gibt es daher immer eine Regressionsgerade, die durch 2 Datenpunkte geht).

E3: Die Lösung ist im Allgemeinen nicht eindeutig. (Für Mathematikfreaks: Mehrere Lösungen haben das Maß 0, was aber nicht heißt - siehe Bsp.1 u. 2 - daß Fälle mit mehrdeutigen Lösungen nicht auftreten können).

E4: Jede konvexe Kombination von Lösungen ist wieder eine Lösung. (In Bsp.1 ist jede Gerade, die zwischen(=konvexe Kombination) den beiden Regressionsgeraden liegt, wiederum eine Regressionsgerade).

E5: Ist N_1 die Anzahl der Punkte über der Regressionsebene
Ist N_2 die Anzahl der Punkte unter der Regressionsebene

Dann gilt: $|N_1-N_2| \leq k$

Im 2-Dimensionalen heißt das z.B.: Die Anzahl der Punkte über und unter der Regressionsgeraden kann sich um nicht mehr als 2 unterscheiden.

Genauso wie der Median die Werte in eine obere und untere Hälfte teilt, trennt die Regressionsebene die Punkte in einen ober- und unterhalb liegenden Teil. Daraus folgt auch, daß die Lage von weit weg liegenden Punkten(=Ausreißern) keinen Einfluß auf die Parameter der Ebene hat.

LAD reagiert nicht auf die Lage von Ausreißern!!

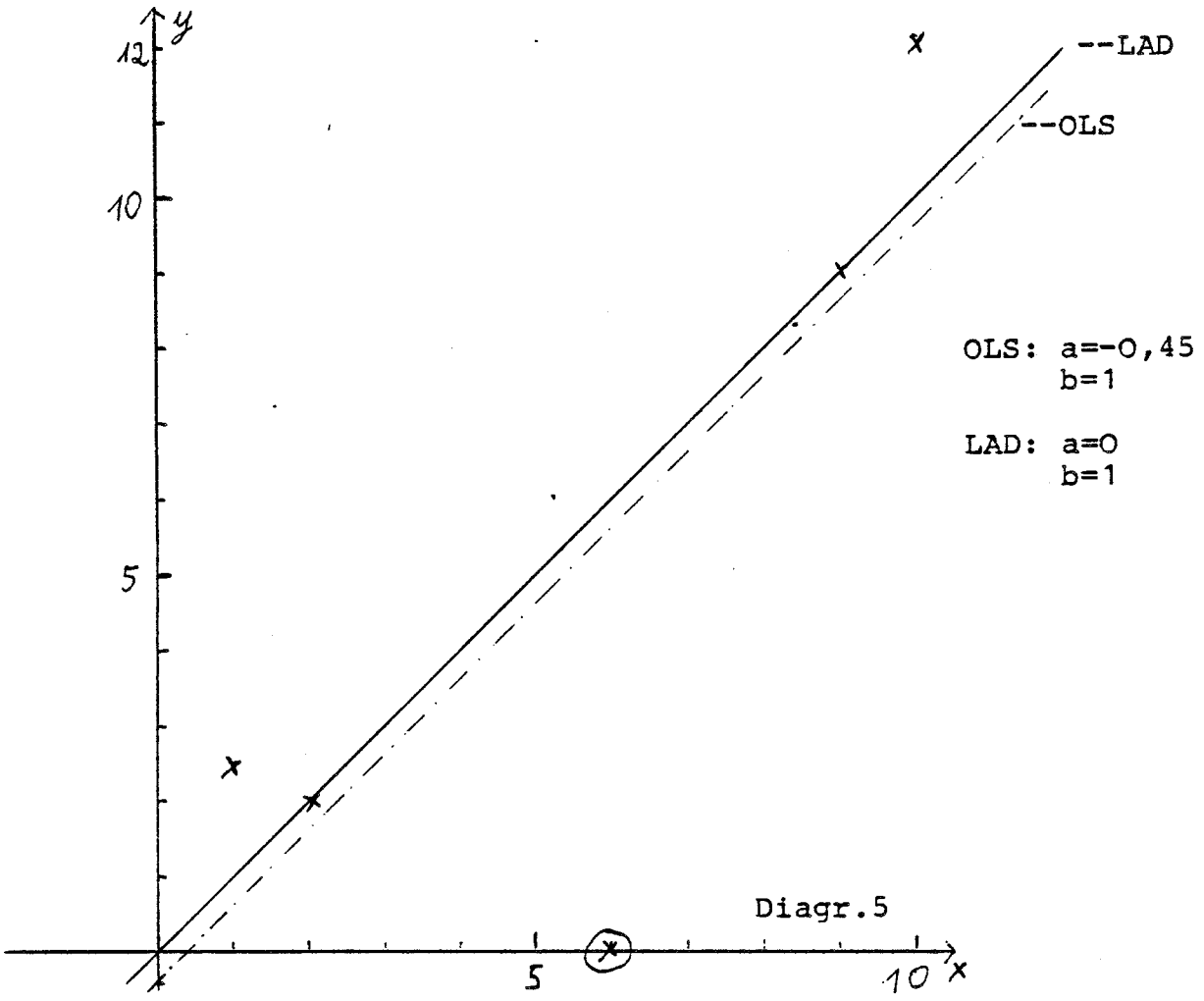
Unterscheiden sich OLS und LAD stark, so liegen Ausreißer vor!!

!!ACHTUNG!! Aus ähnlichen Werten von LAD und OLS kann umgekehrt nicht!!! geschlossen werden, daß

keine Ausreißer vorliegen. Die Berechnung von OLS und LAD, sowie der Vergleich der Parameterwerte, ist nicht als Test für das Fehlen von Ausreißern geeignet.

Bsp. 3: Es liegt Ausreißer vor, aber OLS und LAD unterscheiden sich wenig:

X	1;2;6;9;10
Y	2,5;2;0;9;12



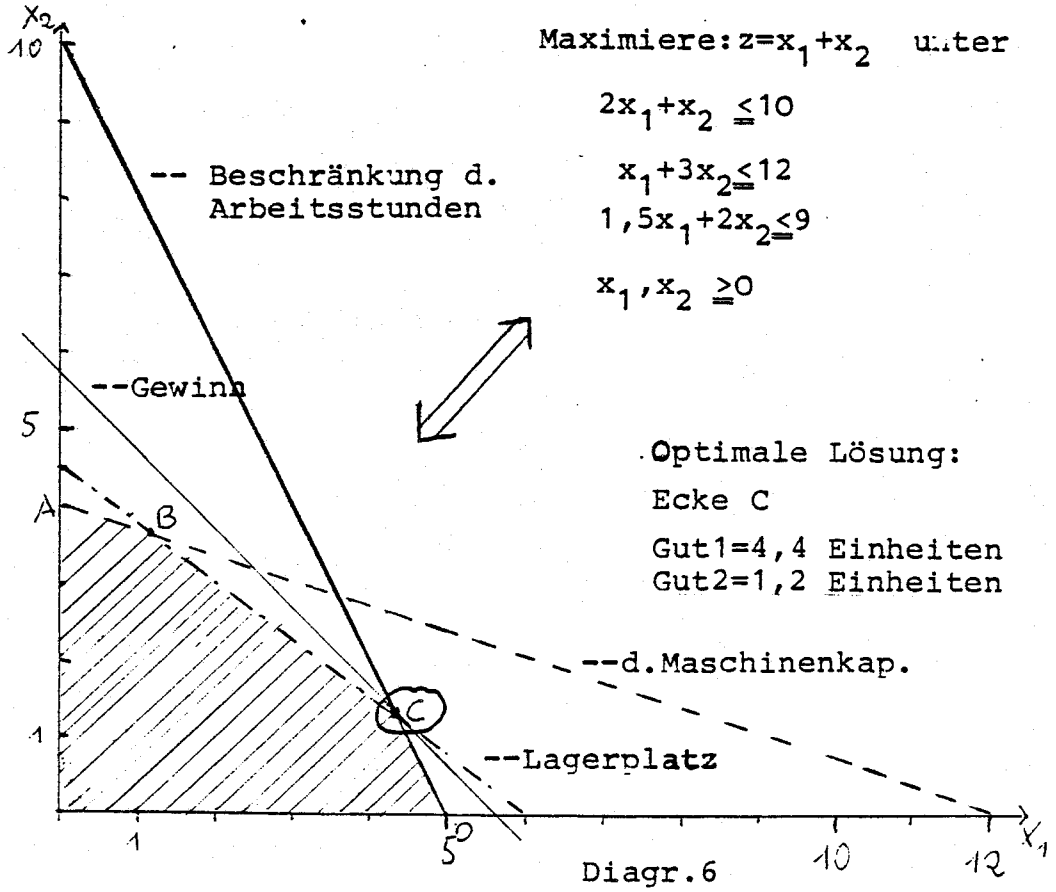
Diagr.5

Die OLS und LAD Regressionsgerade haben exakt dieselbe Steigung und unterscheiden sich auch in der Konstanten wenig. Trotzdem ist der Punkt (6,0) ein offensichtlicher Ausreißer, der weit vom allgemeinen Trend liegt.

4) Lineare Optimierung:

Der Leser wird sich (hoffentlich) fragen,was Lineare Optimierung mit der Schätzung von Regressionsparametern zu tun hat. Die Lineare Optimierung ist sowohl der Schlüssel für Berechnung der Parameter von LAD, als auch für die Ableitung der im vorigen Abschnitt aufgezählten qualitativen Eigenschaften. Dieser Abschnitt soll damit die Idee vermitteln, wie man auf die bereits angeführten "Faustregeln" kommt und wie die Parameter vom Prinzip her berechnet werden.

Eine typische (vereinfachte) Problemstellung der Linearen Optimierung ist folgende Situation:
Eine Fabrik erzeugt Gut1 und Gut2. Insgesamt sind 10 Arbeitseinheiten, 12 Maschineneinheiten und 9 Lagerplatzeinheiten vorhanden. 1 Einheit von Gut1 benötigt: 2Ae,1Me,1,5Le
1 Einheit von Gut2 benötigt: 1Ae,3Me,2Le
Gut1 und Gut2 erbringen per Einheit denselben Gewinn.
Wieviel muß man von Gut1 und Gut2 - ohne die Kapazitäten zu überschreiten - erzeugen, um den Gewinn zu maximieren?



Diagr.6

Allgemeine Problemdefinition:

Primales Lineares Programm:

(1) Maximiere $z=c_1.x_1+\ldots+c_n.x_n$ unter den Restriktionen

(2)
$$\begin{aligned} &a_{11}.x_1+\ldots+a_{1n}.x_n \leq b_1 \\ &\vdots \\ &a_{m1}.x_1+\ldots+a_{mn}.x_n \leq b_m \\ &x_1,\ldots,x_n \leq 0 \end{aligned}$$

Kompakter geschrieben: $A.\underline{x} \leq \underline{b} \quad \underline{x \geq 0}$

Duales Lineares Programm:

(3) Minimiere $w=b_1.u_1+\ldots+b_m.u_m$ unter den Restriktionen

(4) $\underline{u}.A^t \geq \underline{c} \quad u_1,\ldots,u_m \geq 0$

Die duale Version des vorhergehenden Beispiels lautet:

$$\begin{aligned} &u_1.2 + u_2 + 1{,}5.u_3 \geq 1 \\ &u_1.1 + u_2.3 + u_3.2 \geq 1 \quad u_1,u_2,u_3 \geq 0 \\ \text{Minimiere } &w=10.u_1 + 12.u_2 + 9.u_3 \end{aligned}$$

Mit Ungleichungen zu rechnen ist in der Regel nicht sehr angenehm. Man führt daher eine Schlupfvariable $\underline{y \geq 0}$ ein. Dadurch werden die Ungleichungen in (2) und (4) Gleichungen.

(2´)
$$\begin{aligned} &a_{11}.x_1+\ldots.+a_{1n}.x_n+y_1 &&= b_1 \\ &a_{21}.x_1+\ldots.+a_{2n}.x_n+ \ y_2 &&= b_2 \\ &\vdots \\ &a_{m1}.x_1+\ldots.+a_{mn}.x_n+ &&y_m = b_m \end{aligned}$$

Analog mit $-y_i$ für das duale Problem.

Def.2: Eine Lösung $\underline{x}$ ($\underline{u}$) heißt zulässig, wenn $A.\underline{x<b}$ und $\underline{x>0}$ ($\underline{u}.A^t \geq \underline{c}$ und $\underline{u>0}$)

Def.3: Eine Lösung heißt zulässige Basislösung, wenn $\underline{x}$ zulässig und eine Ecke ist.(Im Bsp. Punkte A, B,C,D)

Def.4: $\underline{x}$ ($\underline{u}$) heißt optimale Lösung, wenn $\underline{x}$ ($\underline{u}$) zulässig und $z=c_1.x_1+\ldots+c_n.x_n$ maximal (w ist minimal)

Die folgenden Sätze bilden das Fundament der Linearen Optimierung und der Berechnung von LAD Parametern. (Eine genaue Darstellung findet man in [6])

Satz 2: Existiert eine zulässige Lösung und ist die Zielfunktion z beschränkt, dann existiert mindestens eine optimale Lösung.

Satz 3: (Existenztheorem) Besitzen das primale und duale Problem eine zulässige Lösung, dann existiert mindestens eine optimale Lösung

Man braucht also nur einen Vektor $\underline{x}$ und $\underline{u}$ finden mit $A.\underline{x} \geqq b, \underline{x} \geqq 0$ und $\underline{u}.A^t \leqq \underline{c}, \underline{u} \geqq \underline{0}$ um zu überprüfen, ob es eine optimale Lösung gibt.

Satz 4: (Dualitätstheorem) Ist $\underline{x}$ eine optimale Lösung, dann existiert ein zulässiges $\underline{u}$, so daß $z=w$. Genauso umgekehrt: Ist $\underline{u}$ optimal, dann existiert ein zulässiges $\underline{x}$ mit $w=z$.

Aus diesem Satz ergibt sich sofort, daß die Zielfunktionen z und w des primalen und dualen Problems im Optimum die gleichen Werte annehmen. Man kann damit jeweils das leichter zu lösende Problem berechnen.

Satz 5: Wenn eine optimale Lösung existiert, so liegt mindestens eine optimale Lösung auf einer Ecke.

Dieser Satz ist für die Berechnung der optimalen Lösung von besonderer Bedeutung. Man braucht "nur" die Ecken durchsuchen. Das Problem besteht im faktoriellen Wachstum der Eckenanzahl. Ist A eine $n \times m$-Matrix, so gibt es $\binom{n}{m}$ Ecken. Z.B. gibt es in der Produktion 50 Beschränkungen und 10 verschiedene Produkte, so ist die Anzahl der Ecken 10-Milliarden.

Der am häufigsten verwendete Rechengang ist der Simplexalgorithmus. Alle anderen Algorithmen sind Abwandlungen dieses Grundschemas.

Der Simplexalgorithmus beginnt bei zulässiger Basislösung (=Ecke). Sucht benachbarte Ecke, bei der z am meisten zunimmt. Sucht von dort nächste derartige Ecke ... solange bis man zu einer Ecke kommt, von der aus man sich nicht mehr verbessern kann. Diese Ecke ist eine optimale Lösung.

Obwohl bei diesem Verfahren nicht alle Ecken(nur ein Bruchteil) abgesucht werden müssen, ist der Rechenzeitaufwand noch immer erheblich.Das Durchsuchen einer Ecke erfordert eine Reihe von Rechenschritten.

5) Lineare Optimierung und LAD:

Die LAD Beingung läßt sich auch in folgender Weise schreiben:

Finde $\tilde{\beta}_1,\ldots,\tilde{\beta}_k$ mit:

$$(1)\qquad \begin{matrix} X_{11}\cdot\tilde{\beta}_1+\ldots+X_{1k}\cdot\tilde{\beta}_k+e_1=y_1 \\ \vdots \\ X_{n1}\cdot\tilde{\beta}_1+\ldots+X_{nk}\cdot\tilde{\beta}_k+e_n=y_n \end{matrix}$$

so daß $S=\sum_{i=1}^{n}|e_i|$ ist minimal.

Die Residuen haben hier die Rolle der Schlupfvariablen, die Datenmatrix X jene der Matrix A, der LAD-Parametervektor $\underline{\tilde{\beta}}$ jene des Vektors $\underline{x}$ und $\underline{y}$ jene von $\underline{b}$ übernommen.
Allerdings können die $\tilde{\beta}_1,\ldots,\tilde{\beta}_k,e_1,\ldots,e_n$ negativ sein.
Die Funktion S, die minimiert werden soll, ist nur stückweise linear. Das Problem läßt sich - auf Kosten des Rechenaufwandes - lösen. Jede negative Zahl kann als Summe von zwei nichtnegativen Zahlen geschrieben werden (z.B.: $-5=0-(+5)$)

$\tilde{\beta}_i=\tilde{\beta}_i'-\tilde{\beta}_i`$ wobei $\tilde{\beta}_i'=0$ oder $\tilde{\beta}_i`=0$

$e_i=e_i'-e_i`$ wobei $e_i'=0$ oder $e_i`=0$

Durch diese Verdoppelung der Variablen läßt sich die LAD-Parameterschätzung auf ein Lineares Programm von Abs. 4 zurückführen.

$$(*)\qquad \begin{matrix} X_{11}\cdot\tilde{\beta}_1'-X_{11}\cdot\tilde{\beta}_1`+\ldots+X_{1k}\cdot\tilde{\beta}_k'-X_{1k}\cdot\tilde{\beta}_k`+e_1'-e_1`=y_1 \\ \vdots \\ X_{n1}\cdot\tilde{\beta}_1'-X_{n1}\cdot\tilde{\beta}_1`+\ldots+X_{nk}\cdot\tilde{\beta}_n'-X_{nk}\cdot\tilde{\beta}_n`+e_n'-e_n`=y_n \end{matrix}$$

$$z=\sum_{i=1}^{n}e_i'+\sum_{i=1}^{n}e_i` \quad \text{ist Minimum} =S$$

Die Bedingung z ist ein Minimum läßt sich, indem man mit -1 multipliziert, elementar in die allgemein übliche Form, z sei ein Maximum, bringen.

Damit ist die LAD-Schätzung von n Daten und k Variablen auf Lineare Optimierung mit n Gleichungen und 2.(n+k) Variablen zurückgeführt.Da die Anzahl der Variablen "künstlich" (um keine negativen Zahlen zu bekommen) verdoppelt wurde, versuchen die verwendeten Algorithmen durch eine Reihe von "Tricks" diesen Faktor bei der Berechnung wieder zu eleminieren.(Siehe dazu Kap.3).
Die im Abschnitt 3 behandelten Eigenschaften von LAD-Schätzern folgen (unmittelbar) aus den Sätzen 2-5 der Linearen Optimierung. Z.B. lautet Eigenschaft 1: Unabhängig vom Rang von X existiert immer mindestens eine Lösung. Satz 2 lautet: Existiert eine zulässige Lösung und ist die Zielfunktion beschränkt, dann existiert mindestens eine optimale Lösung. Setzt man die LAD-Parameter $\beta_1,\ldots,_k$ gleich Null, so erhält man in (*) eine zulässige Lösung mit $e_i^{+}-e_i^{-}=y_i$. Da der Absolutabstand nicht kleiner Null sein kann ist die Zielfunktion z (S) nach unten beschränkt, woraus aus Satz 2 sofort folgt, daß es eine optimale Lösung(=LAD-Parameter) gibt.
Analog kann man auch die anderen Eigenschaften von LAD auf die Sätze 2-5 zurückführen.

Mit dem Simplexalgorithmus und seinen Varianten existieren auch allgemein implementierte Berechnungsalgorithmen.
Wenn LAD-Schätzungen in der Ökonometrie bisher eher die Ausnahme als die Regel waren, so liegt dies wohl zum überwiegenden Teil am notwendigen, bis vor kurzem noch nicht handhabbaren, Rechenaufwand. Mit der Lösung dieses Problems und der in diesem paper ebenfalls geleisteten Herleitung der statistischen Eigenschaften von LAD-Parametern, gibt es eigentlich nur ein (wichtiges) Argument gegen die gleichberechtigte Anwendung von OLS und LAD: "Des hamma no nie so gmacht".(auf "Deutsch": Das haben wir noch nie so gemacht) Es ist aber sicherlich zu erwarten, daß dieses Argument mit der Zeit seine Schärfe verlieren wird und LAD-Schätzungen zum anerkannten und üblichen Instrumentarium der Ökonometrie werden.

A N H A N G:

LITERATURVERZEICHNIS:

I) LAD:

(1) ††† L.D.TAYLOR: Minimizing the Sum of Absolute Errors in P.Zarembka(Hg.): Frontiers in Economics p.171-190

(2) † O.J.KARST: Linear Curve Fitting Using Least Deviations JASA, March 1958, p.119-132

(3) H.M.WAGNER: Linear Programming Techniques for Regression Analysis, JASA, March 1959, p.206-212

(4) V.A.SPOSITO,W.C.SMITH: On a Sufficient and a Necessary Condition for L1 Estimation, Journal of Applied Statist. 1976,no 2,p.154-157

II) LINEARE OPTIMIERUNG:

(5) † G.B.DANTZIG: Linear Optimatzion and Extensions,1963

(6) ††† H.W.KUHN,A.W.TUCKER(Hg.): Linear Inequalities and Related Systems, Annals of Mathemetics Studies 38, Princeton 1956

Anmerkung: † bedeutet: Buch ist lesenswert
††† bedeutet: Pflichtlektüre für Interessierte

STATISTISCHE EIGENSCHAFTEN VON LAD

Camillo Signor

Allgemeine Formulierung des Problems:

Hat man für $1 \leq t \leq T$ Beobachtungen (y_t, x_t) wobei $y_t \in R$, $x_t = (x_{t1}, .., x_{tk}) \in R^k$ so ist eine funktionale Beziehung

$y_t = g(x_t, \beta) + u_t \qquad t \in N$ gesucht wobei

$g: R^k \times R^k \to R$ und u_t Zufallsvariable welche ein Wahrscheinlichkeitsmaß auf den Borelmengen in R induzieren.

Ist $S_T(b) = \sum_{t=1}^{T} \rho(u_t) = \sum_{t=1}^{T} \rho(y_t - g(x_t, b)) \qquad b \in R^k$

so heißt $\tilde{\beta}_T$ Schätzer für β $\Leftrightarrow$ $S_T(\tilde{\beta}_T) = \min\{S_T(b) : b \in R^k\}$

Im Standard-Regressionsmodell ist $g(x_t, b) = \sum_{l=1}^{k} x_{tl} b_l$

Für $\rho(u) = u^2$ erhält man die OLS-Schätzung; hier wird nun der Fall $\rho(u) = |u|$ behandelt, das ist der LAD-Schätzer (Least Absolute Deviation).

Gesucht ist also eine Schätzfunktion $\tilde{\beta}_T$ für β mit der Eigenschaft $\sum_{t=1}^{T} |y_t - \sum_{l=1}^{k} x_{tl}\beta_l| = \min\{\sum_{t=1}^{T} |y_t - \sum_{l=1}^{k} x_{tl} b_l| : b \in R^k\} \quad (T \in N)$

Wie üblich sei vorausgesetzt, daß für $T \geq k$ $rg(X_T) = k$ gilt,

wobei $X_T = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ x_T \end{pmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \cdot & & \\ \cdot & & \\ x_{T1} & \cdots & x_{Tk} \end{bmatrix}$

Dann ist $X_T' X_T$ positiv definit und daher invertierbar.

Für $T \geq k$ sei $\kappa_T = \{(t_1,..,t_k): 1 \leq t_i \leq T \text{ f.i=1,..,k \& } t_i \neq t_j \text{ f.i}\neq j\}$

und ferner $H_T = \{ h \in \kappa_T : rg(X_T(h)) = k\}$. Hierbei ist $X_T(h)$ jene kxk-Matrix welche aus X_T durch Streichung der Zeilen x_t mit $t \notin h$ entsteht; analog ist $y(h)$ jener kx1-Vektor der aus

$y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_T \end{bmatrix}$ durch Streichung der y_t mit $t \notin h$ hervorgeht.

Man beachte,daß nach der Voraussetzung über den Rang H_T nichtleer ist. Für den Beweis der folgenden Aussage benötigen wir diese Voraussetzung jedoch gar nicht.Es gilt nämlich

Prop.1: Der LAD-Schätzer $\tilde{\beta}_T$ existiert für alle $T \geq 1$.

Bew.: (a) Sei zunächst $rg(X_T) = k$.Dann existiert $h \in H_T$ und $\{x_t: t \in h\}$ ist Basis des R^k .Klarerweise gilt

$$0 \leq \inf\{\sum_{t=1}^{T} |y_t - x_t b| : b \in R^k\} \leq \sum_{t=1}^{T} |y_t| =: M ;$$

zu jedem $t \in h$ läßt sich eine Orthogonalbasis $B_t = \{x_t, a_1^t, .., a_{k-1}^t\}$ des R^k finden.Also besitzt $b \in R^k$ eine Darstellung

$b = \alpha_t x_t + \sum_{j=1}^{k-1} \alpha_j^t a_j^t$ sodaß $bx_t = \alpha_t |x_t|^2$ und daher

$|y_t - x_t b| = |y_t - \alpha_t |x_t|^2|$ .Wähle $\alpha > 0$ derart,daß $|y_t - \lambda |x_t|^2| > M$

für alle $|\lambda| > \alpha$ und alle $t \in h$ und sei weiter

$W = \{ \sum_{t \in h} \lambda_t x_t : |\lambda_t| \leq \alpha\}$; dann ist W kompakt und für $b \notin W$

$b = \sum_{s \in h} \lambda_s x_s$ gibt es $t \in h$ mit $|\lambda_t| > \alpha$ .Für $s \in h, s \neq t$ ist

$x_s = \sum_{i=1}^{k-1} r_{is} a_i^t$ und daher $b = \lambda_t x_t + \sum_{s \neq t} \lambda_s \cdot \sum_i r_{is} a_i^t$ und weiter

$x_t b = \lambda_t |x_t|^2$ mit $|\lambda_t| > \alpha$ sodaß

$\sum_{s=1}^{T} |y_s - x_s b| \geq |y_t - \lambda_t |x_t|^2| > M$ folgt. Dies gilt für alle $b \notin W$;

Da W kompakt und S_T stetig sind besitzt S_T ein Minimum $\tilde{\beta_T} \in W$ und da

$S_T > M$ auf W^c (dem Komplement von W) ist dieses Minimum global.

(b) Ist nun $rg(X_T) = r \leq k$ dann existiert eine

k×k-Matrix B derart daß $X_T B = (\hat{X_T}, \Theta_{(T \times k-r)})$ mit $rg(\hat{X_T}) = r$; dabei ist

$\hat{X_T}$ eine T×r-Matrix und $\Theta_{(T \times k-r)}$ eine Nullmatrix. Nach Teil (a) des Beweises existiert $\hat{\beta} \in R^r$ sodaß $\sum_{t=1}^{T} |y_t - \sum_{j=1}^{r} x_{tj} \hat{\beta_j}|$ minimal ist.

Ist nun $\beta = \begin{bmatrix} \hat{\beta} \\ \beta' \end{bmatrix}$ wobei $\beta' \in R^{k-r}$ beliebig so nimmt $\sum_{t=1}^{T} |y_t - x_t \beta|$

ebendiesen minimalen Wert an, d.h. jedes solche $\beta \in R^k$ liefert ein Minimum auf R^k; dann ist aber $B\beta$ Stelle eines Minimums von S_T.

□

Wollen nun annehmen, daß in unserem Regressionsmodell

$$y_t = x_t \beta + u_t = \sum_{l=1}^{k} x_{tl} \beta_l + u_t \qquad (t \in N)$$

die Störgrößen u_t unabhängig und gemäß

$p(y_t \leq y) = p(u_t \leq y - x_t \beta) = F(y - x_t \beta)$ verteilt sind.

Def.: Sei u eine Zufallsvariable;

med u := $\{ m \in R: p(u \leq m) \geq \frac{1}{2}$ & $p(u \geq m) \geq \frac{1}{2} \}$; jedes $m \in$ med u heißt ein Median von u.

Ist F,die Verteilungsfunktion (df) von u stetig in $m \in \text{med } u$ so gilt $p(u = m) = 0$ und daher $F(m) = p(u \leq m) = \frac{1}{2}$.

Ist ferner $m \in \text{med } u \Longrightarrow 0 \in \text{med}(u-m)$ sodaß man ohne Beschränkung der Allgemeinheit(oBdA) $0 \in \text{med } u$ annehmen kann, da dies durch Translation zu erreichen ist.
In unserem Modell kann man diese Translation erhalten indem man annimmt daß die erste Spalte von X_T $1_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ ist,

d.h. daß $x_{t1} = 1$ für alle $t \in N$ gilt.

Betrachten im folgenden eine Folge von eindeutig bestimmten LAD-Schätzern $(\tilde{\beta}_T)_{T \in N}$. Diese erfüllen also

$$\min \{ \sum_{t=1}^{T} |y_t - x_t b| : b \in R^k \} = \sum_{t=1}^{T} |y_t - x_t \tilde{\beta}_T| \quad (T \in N) \qquad (1)$$

Zusätzlich seien folgende Annahmen erfüllt:

(A1) F ist stetig,hat Dichtefunktion f und f ist im Median 0 stetig und positiv.

(A2) $\lim_{T \to \infty} \frac{1}{T} X_T' X_T = Q$ existiert und ist positiv definit.

Lemma 1: Sei $B_T^{\sim}$ die Lösungsmenge des Minimalproblems (1); dann gilt: $rg(X_T) = k \Longrightarrow$ es existiert $h \in H_T$ mit $X_T(h)^{-1} y(h) = b(h) \in B_T^{\sim}$ (2);ferner ist $B_T^{\sim}$ die konvexe Hülle aller Lösungen von (1) der Form (2).

Lemma 1 folgt aus der Formulierung von (1) als Lineares Programmierungsproblem.

Führen noch folgende Notation ein:

$$\|v\| = \max\{|v_j| : 1 \leq j \leq k\} \qquad (v \in R^k)$$

$$C[\delta,\varepsilon] = \{c \in R^k : \|c - \delta\| \leq \varepsilon\} \; ; \quad C(\delta,\varepsilon) = \{c \in R^k : \|c - \delta\| < \varepsilon\}$$

$$\operatorname{sgn}^{\sim}(u;v) = \begin{cases} \operatorname{sgn}(u) & \text{falls } u \neq 0 \\ \operatorname{sgn}(v) & \text{sonst} \end{cases}$$

$$(3)\quad \zeta(h,v) := \sum_{t \in h^c} \zeta_t(h,v) = \sum_{t \in h^c} \operatorname{sgn}^{\sim}(y_t - x_t b(h); -x_t X_T(h)^{-1} v)\, x_t X_T(h)^{-1}$$

also $\zeta : H_T \times R^k \to R^k$. Zunächst noch einige

<u>Bemerkungen:</u> Vergleicht man $\beta_T^{\sim}$ mit dem OLS-Schätzer $(X_T'X_T)^{-1}X_T'y$ so sieht man

(a) $\beta_T^{\sim}$ ist nicht Linearkombination aller y_t sondern nur von k von ihnen; daher appliziert auch der Satz von Gauß-Markov nicht auf LAD.

(b) $\beta_T^{\sim}$ verläuft durch genau k der T Beobachtungen. In gewissem Sinne ignoriert $\beta_T^{\sim}$ also Information, jedoch dient die gesamte Stichprobe zur Auswahl jener k Punkte.

<u>Lemma 2:</u> (a) Ist $S: R^k \to R$ konvex so gilt:

$\frac{dS}{dw}(b_o) \geq 0$ für alle $w \neq 0 \iff$ S hat Minimum in b_o.

Ist die Ungleichung strikt so ist das Minimum eindeutig.

(b) Für $S_T(b) = \sum_{t=1}^{T} |y_t - x_t b|$ gilt auch die Umkehrung:

S_T hat eindeutig bestimmtes Min. in $b_o \iff \frac{dS_T}{dw}(b_o) > 0$ für alle $w \neq 0$.

(S heißt <u>konvex</u> $\rightleftharpoons$ $S(rb+(1-r)b') \leq rS(b)+(1-r)S(b')$ für alle $r \in [0,1]$, für alle b,b').

Bew.: (a) ($\Longrightarrow$): Angenommen es gibt $b_1 \in R^k$ mit $S(b_1) \underset{(\leq)}{<} S(b_o)$ $(b_1 \neq b_o)$, dann sei $w := b_1 - b_o$; dann folgt für die Richtungsableitung

$$\frac{dS}{dw}(b_o) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon \|w\|}(S(b_o + w) - S(b_o)) =$$

$$\lim_{\varepsilon \to 0} (S(b_o + \varepsilon b_1 - \varepsilon b_o) - S(b_o))/\varepsilon\|w\| = \lim_{\varepsilon \to 0} (S(\varepsilon b_1 + (1-\varepsilon) b_o) - S(b_o))/\varepsilon\|w\| \leq$$

$$\lim_{\varepsilon \to 0} (\varepsilon S(b_1) + (1-\varepsilon) S(b_o) - S(b_o))/\varepsilon\|w\| = \lim_{\varepsilon \to 0} \varepsilon.(S(b_1) - S(b_o))/\varepsilon\|w\| =$$

$$(S(b_1) - S(b_o))/\|w\| \underset{(\leq)}{<} 0 \quad , \text{ ein Widerspruch zur Voraussetzung.}$$

($\Longleftarrow$): Die schwache Umkehrung gilt sogar ohne Annahme der Konvexität. Hat nämlich S ein Minimum in b_o und ist $w \neq 0$ so gilt für alle $\varepsilon > 0$ $(S(b_o + \varepsilon w) - S(b_o))/\varepsilon\|w\| \geq 0$ und für $\varepsilon \to 0$ folgt $\frac{dS}{dw}(b_o) \geq 0$.

(b) Wegen (a) nur noch die Richtung ($\Longrightarrow$) zu zeigen.

Habe also S genau ein Minimum in b_o; dann gilt wegen (a) $\frac{dS}{dw}(b_o) \geq 0$ für alle $w \neq 0$. Angenommen $\frac{dS}{dw}(b_o) = 0$ für ein $w \neq 0$;

Sei $T_1 = \{t \in N : 1 \leq t \leq T,\ y_t = x_t b_o\}$ (dann ist $T_1 \supsetneq h \in H_T$)

und sei $T_2 = \{1, .., T\} - T_1$. Dann ist

$$\frac{dS}{dw}(b_o) = \sum_{t \in T_1} |x_t w| - \sum_{t \in T_2} \operatorname{sgn}(y_t - x_t b_o) x_t w = 0 \quad \text{nach Annahme}$$

$$\text{sowie } S(b_o + \varepsilon w) = \sum_{t \in T_1} \underbrace{|y_t - x_t b_o - \varepsilon x_t w|}_{= |\varepsilon x_t w|} + \sum_{t \in T_2} \operatorname{sgn}(y_t - x_t b_o - \varepsilon x_t w)(y_t - x_t b_o - \varepsilon x_t w$$

(dabei $\varepsilon > 0$)

Da sgn nur bei 0 unstetig ist gilt für genügend kleines $\varepsilon > 0$

$\operatorname{sgn}(y_t - x_t b_o - \varepsilon x_t w) = \operatorname{sgn}(y_t - x_t b_o)$ $(\neq 0)$ für alle $t \in T_2$ und daher

$$S(b_o + \varepsilon w) = \varepsilon \sum_{t \in T_1} |x_t w| + \sum_{t \in T_2} \operatorname{sgn}(y_t - x_t b_o)(y_t - x_t b_o) -$$

$$- \varepsilon \sum_{t \in T_2} \operatorname{sgn}(y_t - x_t b_o) x_t w = \sum_{t \in T_2} |y_t - x_t b_o| + \varepsilon \frac{dS}{dw}(b_o) =$$

$$\sum_{t \in T_2} |y_t - x_t b_o| = S(b_o) \text{ im Widerspruch zur Eindeutigkeit.}$$

□

Man beachte,daß eine Aussage wie Lemma 2(b) für in b_o differenzierbares S nicht gelten kann;in diesem Fall ist vielmehr die Gültigkeit von $\frac{\partial S}{\partial b_i}(b_o) = \frac{dS}{de_i}(b_o) = 0$ notwendig für das Vorliegen eines Minimums($e_i = (\delta_{ij})_{j=1}^{k}$ = i-ter Einheitsvektor) und zwar für alle i = 1,..,k .

Ferner sei angemerkt,daß $\frac{dS}{dw}(b(h)) > 0 \iff \zeta(h,v)v < \sum_{j=1}^{k} |v_j|$

($w = X_T(h)^{-1}v$ gesetzt)

gilt.

Lemma 3: (a) Seien $X_1,..,X_n,Y$ unabhängige Zufallsvariable,Y habe stetige stetige df und sei $g:R^n \to R$ Borelmeßbar;dann gilt $p(Y = g(X_1,..,X_n)) = 0$.

(b) $\zeta(h,v) \in C\{0,1\}$ (bzw. $C(0,1)$) für alle $v \neq 0 \implies$ $b(h)=X(h)^{-1}y(h) \in \tilde{B}$ (bzw.auch eindeutig).

(c) Seien $\tilde{\delta} = \tilde{\beta} - \beta$,$d(h)=b(h) - \beta = X(h)^{-1}u(h)$,($\tilde{\beta}$ eindeutige Lösung) $E_1(h,\delta,\varepsilon) = \{u \in R^T: d(h) \in C(\delta,\varepsilon)\} = \{u(h): d(h) \in C(\delta,\varepsilon)\}$, $E_2(h) = \{u: \zeta(h,v) \in C(0,1)$ für alle $v \neq 0\}$ dann gilt: $p(\tilde{\delta} \in C(\delta,\varepsilon)) = p\{\bigcup_{h \in H} (E_1(h,\delta,\varepsilon) \cap E_2(h))\}$.

(Das Subskript T wird hier und im folgenden Beweis aus Gründen der Einfachheit der Notation weggelassen,da T hier als fix angenommen wird)

Bemerkung: Die unter (c) geforderte Eindeutigkeit,ebenso wie die auf Seite 4 verlangte Eindeutigkeit der $(\tilde{\beta}_T)_{T \in N}$ ist nicht unbedingt notwendig. Die folgenden Beweise verlaufen analog wenn man diese Bedingung fortläßt und an den entsprechenden Stellen die offenen C(.,.) durch abgeschlossene C{.,.} bzw. strikte Ungleichungen durch schwache ersetzt.

Bew.: (a) Wir zeigen: Sind X und Y unabhängige Zufallsvariable,
Y mit stetiger df $F_2 \Rightarrow p(X=Y)=0$.
Dann folgt mit $X:= g(X_1,..,X_n)$ das Gewünschte; man beachte, daß $g(X_1,..,X_n)$ und Y dann unabhängig sind (siehe etwa [1] Satz 45B.)

Sei also F_1 die df von X, F_2 die von Y, $Z:= (X,Y)$ und
$\Delta = \{(x,y) \in R^2 : x = y\} \Rightarrow p(X=Y) = p(Z \in \Delta) = pZ^{-1}(\Delta) =$
$\int_\Delta d(pX^{-1} \times pY^{-1}) = \int\int_{R^2} 1_\Delta(x,y)\, dF_2(y)\, dF_1(x)$; die letzten beiden Gleichungen gelten wegen der Unabhängigkeit bzw. wegen des Satzes von Fubini.
Zeigen noch: $\int_{-\infty}^{+\infty} 1_\Delta(x,y)\, dF_2(y) = 0$ für alle $x \in R$.

Nun gilt $\int_{-\infty}^{+\infty} 1_\Delta(x,y)\, dF_2(y) \leq \int_{]x-\varepsilon, x]} dF_2(y) = F_2(x) - F_2(x-\varepsilon) \to 0$
$(\varepsilon \to 0)$ wegen Stetigkeit von F_2.

(b) Zeigen zunächst:
$\zeta(h,v) \in C[0,1]$ (bzw. $C(0,1)$) $\Rightarrow$ $\zeta(h,v)v \underset{(<)}{\leq} \sum_{j=1}^{k} |v_j|$
für alle $v \neq 0$ (4)

Sei $X(h)^{-1} = (x^{lj})_{l,j=1}^{k}$ dann ist
$\zeta(h,v) = \left(\sum_{l=1}^{k} \sum_{t \in h^c} \operatorname{sgn}^{\sim}(y_t - x_t b(h); -x_t X(h)^{-1} v) x_{tl} x^{lj} \right)_{j=1}^{k}$

Also $\zeta(h,v) \in C[0,1]$ (bzw. $C(0,1)$) $\Rightarrow \left| \sum_{l=1}^{k} \sum_{t \in h^c} \operatorname{sgn}^{\sim}(\ldots) x_{tl} x^{lj} \right| \underset{(<)}{\leq} 1$
für alle $j=1,..,k$

und daher $\zeta(h,v)v \leq |\zeta(h,v)v| \leq \sum_{j=1}^{k} \left| \sum_{l} \sum_{t} \operatorname{sgn}^{\sim}(\ldots) x_{tl} x^{lj} \right| |v_j| \underset{(<)}{\leq} \sum_{j=1}^{k} |v_j|$

Wegen Konvexität von S_T genügt es nach Lemma 2(a) zu zeigen, daß
$\frac{dS_T}{dw}(b(h)) \underset{(>)}{\geq} 0$ für alle $w \neq 0$; nun ist $\frac{dS_T}{dw}(b(h)) =$
$- \sum_{t=1}^{T} \operatorname{sgn}^{\sim}(y_t - x_t b(h); -x_t w) x_t w = \sum_{t \in h} \operatorname{sgn}(x_t w) x_t w - \sum_{t \in h^c} \operatorname{sgn}^{\sim}(y_t - x_t b(h); -x_t w) x_t w$
$\underset{(>)}{\geq} 0$

nach (4), wobei $w = X(h)^{-1} v$ gesetzt wurde.

(c) Sei $\zeta_0(h) = \sum_{t\in h^c} \operatorname{sgn}(y_t - x_t b(h)) x_t X(h)^{-1}$

Zeigen zunächst: $p(\zeta(h,v) = \zeta_0(h) \text{ für alle } v\neq 0) = 1$ (5)

Für $t\in h^c$ ist nach Teil (a) $p(\{u: y_t = x_t b(h)\}) = p(\{u: u_t = x_t d(h)\}) = 0$

da $x_t d(h)$ Linearkombination nur von u_s' mit $s\in h$ ist; also folgt

$p(y_t \neq x_t b(h)) = 1$ für alle $t\in h^c$ und daher

$p(\zeta(h,v) \neq \zeta_0) = p(\bigcup_{t\in h^c} \{y_t = x_t b(h)\}) \leq \sum_{t\in h^c} p(y_t = x_t b(h)) = 0;$

dies liefert (5).

Ferner gilt: $\zeta_0(h)v \underset{(<)}{\leq} \sum_{j=1}^{k} |v_j|$ für alle $v\neq 0 \Leftrightarrow \zeta_0(h) \in C\{0,1\}$ (bzw. $C(0,1)$) (6)

($\Rightarrow$): Ist i der Index mit maximalem $|\zeta_0(h)_i|$ so wähle man

$v_i = \operatorname{sgn}\zeta_0(h)_i$ und $v_j = 0$ für $j\neq i$.

Die Umkehrung beweist man wie in Teil (b).

Damit erhält man $p(\zeta(h,v) \in C(0,1) \text{ für alle } v\neq 0) = p(\zeta(h,v) \in C(0,1) \,\&\, \zeta_0 = \zeta \text{ f.a.} v\neq 0) =$

$= p(\zeta_0 = \zeta \text{ f.a.} v\neq 0 \,\&\, \{b(h)\} = B^{\sim}).$ (7)

Die erste Gleichung folgt aus (5); für die zweite beachte man, daß aus Lemma 2(b) und (6) folgt: $\zeta_0 = \zeta$ f.a. $v\neq 0$ & $\zeta \in C(0,1) \Leftrightarrow \zeta_0 = \zeta$ f.a. $v\neq 0$ & $\{b(h)\} = \{\beta^{\sim}\} = B^{\sim}$ (7')

Nun ist $\|\delta^{\sim} - \delta\| = \|\beta^{\sim} - \beta - \delta\| < \varepsilon \Leftrightarrow \beta^{\sim} = b(h_o)$ & $\|d(h_o) - \delta\| < \varepsilon \Leftrightarrow$

$u \in E_1(h_o, \delta, \varepsilon)$ wobei $\{\beta^{\sim}\} = \{b(h_o)\} = B^{\sim}$ für genau ein $h_o \in H$

(Eindeutigkeit wurde vorausgesetzt) Man hat demnach

$p(\delta^{\sim} \in C(\delta,\varepsilon)) = \bar{p}(E_1(h_o,\delta,\varepsilon) \,\&\, \{b(h_o)\} = B^{\sim}) =$ (wegen (5))

$\bar{p}(E_1(h_o,\delta,\varepsilon) \,\&\, \{b(h_o)\} = B^{\sim} \,\&\, \zeta(h_o,v) \equiv \zeta_0(h_o)) =$ (wegen (7'))

$p(E_1(h_o,\delta,\varepsilon) \,\&\, \zeta \equiv \zeta_0(h_o) \,\&\, \zeta \in C(0,1) \text{ f.a.} v\neq 0) =$ (wegen (5))

$p(E_1(h_o,\delta,\varepsilon) \,\&\, \zeta \in C(0,1) \text{ f.a.} v\neq 0) = p(E_1(h_o,\delta,\varepsilon) \cap E_2(h_o)) \leq$

$p(\bigcup_{h\in H} (E_1(h,\delta,\varepsilon) \cap E_2(h)))$.

Die umgekehrte Ungleichung gilt nicht nur stochastisch:
Ist nämlich $\zeta(h,v) \in C(0,1)$ f.a.$v\neq 0$ & $\|d(h)-\delta\| < \varepsilon \Rightarrow$
$\{b(h)\} = \{\beta^{\sim}\} = B^{\sim}$ & $\|b(h)-\beta-\delta\| = \|\delta^{\sim}-\delta\| < \varepsilon$ sodaß also
$p(\bigcup_{h\in H}(E_1(h,\delta,\varepsilon) \cap E_2(h))) \leq p(\delta^{\sim} \in C(\delta,\varepsilon))$ gilt.

□

Bemerkung: Es ist instruktiv den Sonderfall k=1 im Zusammenhang mit diesen Ergebnissen zu betrachten.Man hat dann

$$X_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \qquad H_T = \{\{i\} : 1 \leq i \leq T\} \qquad \beta_T^{\sim} = y(h) = y_j \qquad (h = \{j\})$$

Nach Lemma 3(b) ist dann $\{y_j\} = \{\beta_T^{\sim}\} = B_T^{\sim}$ falls

$|\sum_{t\neq j} \operatorname{sgn}^{\sim}(y_t - y_j; w)| < 1$ für alle $w\neq 0$ d.h. $\sum_{t\neq j} \operatorname{sgn}^{\sim}(y_t - y_j; w) = 0$

für alle $w\neq 0$; das bedeutet,daß y_j eindeutiger Stichproben-Median ist falls $T \equiv 1 \mod 2$ gilt. Ist umgekehrt T gerade,so ersieht man aus (5) und Lemma 2(a) daß es mit Wahrscheinlichkeit 1 ein ganzes Intervall von Stichprobenmedianen gibt.

Auskunft über lineare Eigenschaften des LAD-Schätzers gibt

Lemma 4: $\beta_T^{\sim}(y,X_T) = \beta^{\sim} \in B_T^{\sim}(y,X_T)$ hat folgende Eigenschaften:

(a) $\beta^{\sim}(\lambda y, X) = \lambda\beta^{\sim}$ $\quad (\lambda \in R)$

(b) $\beta^{\sim}(y+Xz, X) = \beta^{\sim} + z$ $\quad (z \in R^k)$

(c) $\beta^{\sim}(y, XA) = A^{-1}\beta^{\sim}$ $\quad A$: k×k-Matrix,nichtsingulär

(d) $\beta^{\sim}(X\beta^{\sim}+Du^{\sim}, X) = \beta^{\sim}$ $\quad u^{\sim}=y-X\beta^{\sim}$, D: T×T-Diagonalmatrix mit $D_{tt} \geq 0$.

Bew.: Sei $\psi(b,y,X_T) = \sum_{t=1}^{T} |y_t - x_t b|$ dann gelten folgende Beziehungen:

(a) $|\lambda|\psi(b,y,X_T) = \psi(\lambda b, \lambda y, X_T)$

(b) $\psi(b,y,X_T) = \psi(b+z, y+X_T z, X_T)$

(c) $\psi(b,y,X_T) = \psi(A^{-1}b, y, X_T A)$

Für (d) beachte man $\frac{d\psi}{dw}(\tilde{\beta_T}) = -\sum_{t=1}^{T} \text{sgn}^{\sim}(y_t - x_t\tilde{\beta_T}; -x_t w)x_t^{w} \geq 0$

für alle $w \in R^k$

Nun gilt $\text{sgn}^{\sim}(x_t\tilde{\beta_T} + D_{tt}(y_t - x_t\tilde{\beta_T}) - x_t\tilde{\beta_T}; -x_t w)x_t w \leq$

$\text{sgn}^{\sim}(y_t - x_t\tilde{\beta_T}; -x_t w)x_t w$ da $D_{tt} \geq 0$ und daher

$$0 \leq -\sum_{t=1}^{T} \text{sgn}^{\sim}(y_t - x_t\tilde{\beta_T}; -x_t w)x_t w \leq -\sum_{t=1}^{T} \text{sgn}^{\sim}(D_{tt}(y_t - x_t\tilde{\beta_T}); -x_t w)x_t w$$

sodaß die Behauptung aus Lemma 2(a) folgt. □

Bevor das Hauptresultat formuliert wird noch einige

Bemerkungen:

(1) Ist $\|X_T\| := \max\{|x_{tj}| : 1 \leq t \leq T, 1 \leq j \leq k\}$ so gilt

$$\|X_T\| = o(\sqrt{T}) \qquad (8)$$

(siehe [4] S 226 ff.)

(2) Sei $Z = (Z_1, .., Z_k)$ ein Zufallsvektor mit charakteristischer Funktion $\phi(t_1, .., t_k) = E\{\exp(i \sum_{j=1}^{k} t_j Z_j)\}$ $(i = \sqrt{-1})$

Man sagt Z hat <u>nonlattice-Verteilung</u> $\rightleftharpoons$

$|\phi(t_1, .., t_k)| < 1$ für alle $(t_1, .., t_k) \neq 0$

In [5] und [6] wird gezeigt: Ist $\mu_T(x,\varepsilon)$ das Ws-Maß der Menge $\{y \in R^k : x_j \leq y_j \leq x_j + \varepsilon, 1 \leq j \leq k\}$ bezüglich einer df F_T (dabei $x \in R^k, \varepsilon > 0, T \in N$) und ist $(b_T)_{T \in N}$ Folge reeller Zahlen sodaß $\lim_{T \to \infty} F_T(b_T x) = F(x)$ so gilt:

Ist F_T nonlattice für alle $T \geq T_o$, g Dichtefunktion von F $\Rightarrow$

$\mu_T(x,1) = b_T^{-k} g(0) + o(b_T^{-k})$. Speziell für $b_T = \sqrt{T}$ folgt

$\mu_T(x,1) T^{-k/2} \to g(0) \qquad (T \to \infty)$

(Per Definition ist g Dichtefunktion auf $R^k \rightleftharpoons \int_{R^k} g(x)dx = 1$ &

$g(x) \geq 0 \quad f.a.x \in R^k$)

(3) In [7] wird gezeigt:

Ist $(g_T)_{T \in N}$ eine Folge von Dichtefunktionen auf R^k und gilt

$g_T \to g$ a.e. $(T \to \infty)$ (d.h.Konvergenz bis auf Borelmengen in R^k mit Borelmaß 0; "almost everywhere")

so gilt: Ist g Dichtefunktion $\Rightarrow \int_B g_T(x)dx \to \int_B g(x)dx \qquad (T \to \infty)$

gleichmäßig auf allen Borelmengen B.

Def.: Sei $(Z_T)_{T \in N}$ eine Folge von Zufallsvektoren auf R^k und sei $(F_T)_{T \in N}$ die zugehörige Folge der df's; F df auf R^k

$Z_T \xrightarrow{L} F \quad (T \to \infty)$ (Convergence in Law) $\rightleftharpoons$

$F_T \to F \quad (T \to \infty)$ auf allen Stetigkeitsstellen von F und auch bei $\pm\infty$.

Nach diesen Vorbereitungen kommen wir zu folgendem

Satz: Sei $(\tilde{\beta_T})_{T\geq N}$ eine Folge von eindeutig bestimmten LAD-Schätzern und gelte (A1) & (A2); dann gilt:
$\sqrt{T}(\tilde{\beta_T} - \beta) \xrightarrow{L} N(0,\Sigma) \quad (T \to \infty)$.
Dabei ist $\Sigma = (1/\omega^2)Q^{-1}$ die Kovarianzmatrix der Grenzverteilung, mit $\omega = 2f(0)$.

Bew.:

1.Schritt:

Wir zeigen: Ist ϕ_T die Dichtefunktion von $\sqrt{T}(\tilde{\beta_T} - \beta)$ so gilt

$$\phi_T(\delta) = T^{-k/2} \sum_{h\in H_T} |\det(X_T(h))| \prod_{t\in h} f(T^{-1/2}x_t\delta).p(Z_T(\delta,h)\in C(0,1)) \tag{9}$$

Dabei ist $Z_T(\delta,h) = \sum_{t\in h^c} z_t(\delta,h) = \sum_{t\in h^c} \operatorname{sgn}(u_t - T^{-1/2}x_t\delta)x_t X_T(h)^{-1}$ (10)

Seien wie in Lemma 3 $\delta := \tilde{\beta_T} - \beta$, $d(h) = b(h) - \beta = X_T(h)^{-1}u(h)$, $E_1(h,\delta,\varepsilon)$ und $E_2(h)$ wie dort definiert; ferner sei angemerkt, daß $y_t - x_t b(h) = u_t - x_t d(h)$ (und das $= 0$ für $t\in h$) gilt.

Sei $M = k\,\|X_T\|$ und $E_3(h,\delta,\varepsilon) := \{u\in R^T: |u_t - x_t| > \varepsilon M \text{ f.a.} t\in h^c\}$

Klarerweise $p(E_1 \cap E_2) = p(E_1 \cap E_2 \cap E_3) + p(E_1 \cap E_2 \cap E_3^c)$ und wegen $|u_t - x_t| < \varepsilon M$ f.a.$t\in h$ für $u\in E_1$

($|u_t - x_t\delta| = |u_t - x_t d(h) + x_t(d(h) - \delta)| = |x_t(d(h) - \delta)| < \varepsilon M$) folgt

$$p\Big(\bigcup_{h\in H}(E_1 \cap E_2 \cap E_3)\Big) = \sum_{h\in H} p(E_1 \cap E_2 \cap E_3) =$$

$$\sum_{h\in H} p(E_1)p(E_2 | E_1\, E_3)p(E_3 | E_1) \tag{11}$$

Sei weiter $E_2'(\delta,h) := \{u\in R^T: Z_T(\sqrt{T}.\delta,h) \in C(0,1)\}$ so gilt jedenfalls für genügend kleines $\varepsilon > 0$: $p(E_2') = p(E_2 | E_1 \cap E_3)$ (12)

Aus Lemma 3(c) erhält man

$p(\delta^{\sim} \in C(\delta,\varepsilon)) = p(\bigcup_{h\in H}(E_1 \cap E_2 \cap E_3) \cup (E_1 \cap E_2 \cap E_3^c))$ und bezeichnet

m_k das Lebesgue-Maß auf R^k so erhält man die Dichte g

des von $\delta^{\sim}$ erzeugten Ws-Maßes als Radon-Nikodym-Ableitung

$\lim_{\varepsilon\to 0} p(\delta^{\sim}\in C(\delta,\varepsilon))/m_k(C(\delta,\varepsilon)) = g(\delta)$ und unter Beachtung

von $p(E_3(h,\delta,\varepsilon)) \to 1 \quad (\varepsilon\to 0)$ folgt aus (11) und (12)

$$g(\delta) = \lim_{\varepsilon\to 0} p(\delta^{\sim}\in C(\delta,\varepsilon))/m_k(C(\delta,\varepsilon)) = \lim_{\varepsilon\to 0} \sum_{h\in H} p(E_2') \frac{p(E_1(h,\delta,\varepsilon))}{m_k(C(\delta,\varepsilon))} =$$

$$= \sum_{h\in H} |\det(X(h))| \cdot \prod_{t\in h} f(x_t\delta) \cdot p(E_2'(h,\delta))$$

Für die letzte Gleichung beachte man daß

$p(E_1) = p(\{u: u(h) \in X(h)C(\delta,\varepsilon)\})$; normiert man mit $T^{-1/2}$

erhält man (9).

2.Schritt:

Nun wird gezeigt, daß ϕ_T gegen die Dichtefunktion einer

k-dimensionalen Normalverteilung konvergiert.

Sei $Z_T'(\delta,h) := \frac{1}{2\sqrt{T}} Z_T(\delta,h)$; wir beachten, daß

$p(z_t(\delta,h) = x_t X(h)^{-1}) = 1 - F(T^{-\frac{1}{2}} x_t\delta)$

$p(z_t(\delta,h) = -x_t X(h)^{-1}) = F(T^{-\frac{1}{2}} x_t\delta)$ gilt und

wegen $\sum_{t\in h^c} x_t' x_t = O(T)$ (A2) und (8) folgt aus dem

Zentralen Grenzwertsatz $T^{-\frac{1}{2}} Z_T(\delta,h) = T^{-\frac{1}{2}} \sum_{t\in h^c} z_t(\delta,h) \xrightarrow{L}$

(13)

$$N(-2f(0)\delta' QX(h)^{-1}, X'(h)^{-1} Q X(h)^{-1})$$

Seien nun für $T\in N$ G_T die von Z_T' $(=Z_T'(\delta,h))$ erzeugten df's und p_T die entsprechenden Ws-Maße auf R^k so gilt gemäß (13)

$$G_T \overset{L}{\to} G \quad (T\to\infty) \quad \text{wobei} \quad G \sim N(-f(0)\delta'QX(h)^{-1}, \tfrac{1}{4}X'(h)^{-1}QX(h)^{-1})$$

Ferner seien $C_T = C(0,\frac{1}{2\sqrt{T}}) = \{c\in R^k: \|c\| < \frac{1}{2\sqrt{T}}\}$,dann haben die C_T Lebesgue-Maß $m_k(C_T) = T^{-k/2}$.

Sind nun die G_T nonlattice für alle $T\geq T_o$ so erhält man gemäß Bemerkung (2) für die Radon-Nikodym-Ableitung

$$\lim_{T\to\infty} p_T(C_T)/m_k(C_T) = \lim_{T\to\infty} T^{k/2}.p(Z_T'\in C_T) = g(0) \qquad (14)$$

wobei g die Dichte der Grenzverteilung G ist.

Man hat also $T^{k/2}.p(Z_T'\in C_T) = (2\pi)^{-k/2}\det(\frac{1}{4}X'(h)^{-1}QX(h)^{-1})^{-1/2}.$

$$.\exp\{-\tfrac{1}{2}f^2(0)\delta'QX(h)^{-1}(\tfrac{1}{4}X'(h)^{-1}QX(h)^{-1})^{-1}X'(h)^{-1}Q'\delta\}$$

$$+ o(1) =$$

$$(2\pi)^{-k/2}.2^k|\det(X(h))|\det(Q)^{-1/2}.$$

$$.\exp\{-\tfrac{1}{2}f^2(0)4\delta'QX(h)^{-1}(X(h)Q^{-1}X'(h))X'(h)^{-1}Q\delta\} + o(1) =$$

(beachte $Q = Q'$)

$$(2\pi)^{-k/2}.2^k|\det(X(h))|\det(Q)^{-1/2}.\exp\{-\tfrac{1}{2}(2f(0))^2\delta'Q\delta\} + o(1)$$

Da f in O stetig nach (A1) folgt $\prod_{t\in h} f(T^{-1/2}x_t\delta) = f^k(0)+o(1)$ (15)

und unter Beachtung von $Z_T'\in C_T \Longleftrightarrow Z_T \in C(0,1)$ erhält man durch Substitution von (15) in (9) daraus

$$\phi_T(\delta) = T^{-k/2} \sum_{h\in H} |\det(X(h))| (f^k(0)+o(1)) . T^{-k/2} \{(2\pi)^{-k/2} 2^k .$$

$$. |\det(X(h))| \det(Q)^{-1/2} . \exp(-\tfrac{1}{2}\omega^2 \delta' Q\delta) + o(1)\} =$$

$$= T^{-k} \sum_{h\in H} \det(X(h))^2 \omega^k \det(Q)^{-1/2} (2\pi)^{-k/2} \exp(-\tfrac{1}{2}\omega^2 \delta' Q\delta) + o(1)$$

wobei $\omega = 2f(0)$ gesetzt wurde. Beachtet man noch

$\sum_{h\in H} \det(X(h))^2 = \det(X'X)$ (siehe [3] S.474) folgt

$$\phi_T(\delta) = \underbrace{T^{-k} \det(X'X)}_{=\det(\frac{X'X}{T}) \to \det(Q) \quad (T\to\infty) \text{ nach (A2)}} \det(Q)^{-1/2} \omega^k (2\pi)^{-k/2} \exp(-\tfrac{1}{2}\omega^2 \delta' Q\delta) + o(1)$$

sodaß schließlich

$$\phi_T(\delta) \to (2\pi)^{-k/2} \det(\omega^2 Q)^{1/2} . \exp(-\tfrac{1}{2}\omega^2 \delta' Q\delta) \qquad (16)$$

$(T\to\infty)$

folgt und die rechte Seite von (16) ist eben die Dichtefunktion der Normalverteilung $N(0, Q^{-1}/\omega^2)$.

Nun folgt das gewünschte Resultat aus Bemerkung (3).

Sind die G_T Lattice-Verteilungen so gilt (14) nur bis auf einen Proportionalitätsfaktor, der sich aus dem Maß der Gitterpunkte in C_T ergibt und durch Summation über H erhält man eine gleichmäßig beschränkte Dichtefunktion proportional zu $\exp(-\frac{1}{2}\omega^2 \delta' Q\delta)$. Nun folgt die Konvergenz der Integrale $\int_{R^k} \phi_T(\delta)\, dm_k(\delta)$ aus dem Satz von Lebesgue über die beschränkte Konvergenz und der Satz ist bewiesen.

□

Bevor wir aus dem eben bewiesenen Resultat noch eine Aussage über Konsistenz und Erwartungstreue herleiten eine

Def.: Sei $(Z_T)_{T\in N}$ eine Folge von Zufallsvektoren auf (R^k,p); $(Z_T)_{T\in N}$ heißt stochastisch beschränkt (Notation $Z_T = O_p(1)$) $\rightleftharpoons$ für alle $\varepsilon>0$ existiert $K>0$ & $T_\varepsilon \in N$ sodaß $p(\|Z_T\| \not< K) \leq \varepsilon$ für alle $T \geq T_\varepsilon$ (Dabei ist $\|Z\| := \max\{|Z_i| : 1\leq i\leq k\}$).

Wir zeigen nun folgendes

Korollar: (a) $E(u) = 0 \Rightarrow E(\tilde{\beta_T}) = \beta$ (Erwartungstreue)

(b) Unter den Voraussetzungen des Satzes gilt:

$\operatorname{plim}_{T\to\infty} \tilde{\beta_T} = \beta$ (Konsistenz).

Bew.: (a) $E(\tilde{\beta_T}) = E(X_T(h)^{-1}y(h)) = E(X_T(h)^{-1}(X_T(h)\beta+u(h))) =$

$\beta + E(X_T(h)^{-1}u(h)) = \beta + X_T(h)^{-1}E(u(h)) = \beta$.

(b) Zeigen: Ist F eine df mit $Z_T \xrightarrow{L} F$ $(T\to\infty)$ so gilt

$Z_T = O_p(1)$.

Sei dazu $\varepsilon>0$ beliebig und seien $(x_1,..,x_k)$ und $-(y_1,..,y_k) \in R^k$ derart daß $x_i, y_i \geq 0$ für $1\leq i\leq k$ und $F(x_1,..,x_k)-F(-y_1,..,-y_k)>1-\varepsilon$ Außerdem sei F an diesen beiden Punkten stetig (Man beachte daß eine df bis auf höchstens abzählbar viele Hyperebenen der Form $x_i = c_{ij}$ $(j\in N, i=1,..,k)$ Stetigkeitsstellen besitzt). Sei $K:=\{\max\ x_i, y_i : 1\leq i\leq k\}$; ist F_T die von Z_T erzeugte df so gilt für genügend großes $T (T\geq T_\varepsilon)$:

$$1-\varepsilon < F_T(x_1,..,x_k)-F_T(-y_1,..,y_k)$$
$$\downarrow \qquad\qquad \downarrow \qquad (T\to\infty)$$
$$F(x_1,..,x_k)- F(-y_1,..,y_k) \qquad \text{also}$$

$$1-\varepsilon < p(-y_i < Z_{T,i} \leq x_i, 1\leq i\leq k) \leq p(|Z_{T,i}| \leq K, 1\leq i\leq k) \leq p(\|Z_T\| \leq K)$$

und daher $p(\|Z_T\| > K) \leq \varepsilon$ für alle $T \geq T_\varepsilon$.

Ist nun $Z_T = \sqrt{T}.Y_T$ und $Z_T = O_p(1) \Rightarrow \operatorname{plim}_{T\to\infty} Y_T = O$:

Sind nämlich $\varepsilon, \delta > 0 \Rightarrow p(\|Y_T\| > \frac{K}{\sqrt{T}}) \leq \varepsilon$ f.a. $T \geq T_o$

und $\frac{K}{\sqrt{T}} < \delta$ für alle $T \geq T_1$, daher $p(\|Y_T\| > \delta) \leq p(\|Y_T\| > \frac{K}{\sqrt{T}}) \leq \varepsilon$

für alle $T \geq \max\{T_o, T_1\}$ also $\operatorname{plim}_{T\to\infty} Y_T = O$.

Wählen wir $Y_T = \beta_T^{\sim} - \beta$ so folgt das Gewünschte.

□

L I T E R A T U R :

[1] Halmos,P.R.,Measure Theory,
D.Van Nostrand,New York,1950.

[2] Bassett,G.Jr. & Koenker,R.,
Asymptotic Theory of Least Absolute Error Regression
Journal of the Am.Statistical Ass.,Vol.73(1978)pp.618-622.

[3] Knuth,Art of Computer Programming,Vol.1

[4] Malinvaud,E.,Statistical Methods of Econometrics,
North-Holland Publishing Co.,New York,1970.

[5] Shepp,L.A.,A Local Limit Theorem,
Annals of Mathematical Statistics,35,pp.419-423,1964.

[6] Stone,Charles,A Local Limit Theorem for Nonlattice
Multi-dimensional Distribution Functions,
Annals of Math.Statistics,36,pp.546-551,1965.

[7] Scheffé,Henry,A Useful Convergence Theorem for
Probability Distributions,
Annals of Math.Statistics,18,pp.434-438,1947.

ALGORITHMS FOR COMPUTATION OF THE LAD-ESTIMATES

Zbigniew Wasilewski

ALGORITHMS FOR COMPUTATION OF THE LAD-ESTIMATES

Computation of the Least Absolute Deviation estimates is in general a linear programing problem that was originaly formulated by Charnes,Cooper and Ferguson(1955),Wagner(1958) and Fischer(1961).Recently more efficient algorithms have been provided by Barrodale and Roberts (1973),Abdelmalek(1974), Narula and Wellington(1976) and Snyder(1978),and an iterative reweighted least squares method has been suggested by Schlosmacher (1973) and Fair(1974).

On the basis of the survey of the literature and numerical experiments,we decided to implement in the IAS system two of these algorithms,namely this proposed by Narula and Wellington and this given by Snyder.

The first one determines the values of the unknown regression coefficients that minimize the sum of absolute errors after shifting the origin to the point of means.The estimates one obtains in such a way are not in general the same as these obtained from original data,unless the point of means coincides with one of the observations in the solution bassis.Nevertheless such restricted LAD estimates are still unbiased when the distribution of the error term in estimated regression model is symetric about zero(see Taylor(1974)).

The linear programing formulation of the problem when the origin passes through the mean is

$$\text{(1)} \qquad \text{minimize} \quad TAE = \sum_1^n (e_i^+ + e_i^-)$$

$$\text{(2)} \qquad \text{subject to} \quad \sum_j^k (b_j^+ - b_j^-)(x_{ij} - \bar{x}_j) + e_i^+ - e_i^- = y_i - \bar{y} \qquad i=1,2,\ldots,n$$

$$\text{(3)} \qquad e_i^+ , e_i^- \geq 0 , b_j^+, \bar{b}_j \geq 0 \qquad j=1,2,\ldots,k$$

Where $e_i = e_i^+ - e_i^-$ $\quad b_j = b_j^+ - b_j^-$

$\bar{x}_j = \Sigma_1^n x_{ij}/n \quad \bar{y} = \Sigma_1^n y_i/n \quad b_o = \bar{y} - \Sigma_{j=1}^k b_j \bar{x}_j$

This problem is in the case of algorithm due to Narula and Wellington solved by means of dual simplex method. This type of simplex procedure in comparison with ordinary one starts with the initial basic solution,with optimal value of the objective function but the coresponding constraints are not fulfiled.in successive steps the procedure evaluates this solution to achieve the fulfilness of the given constraints by as small as possible change of the value of the objective function.In order to have possibly good view into the structure of the procedure let us rewrite the formulatoon of the problem in following way.

Select such a set $\{ b_1, \ldots, b_k \}$ of coefficients of independemt variables in the linear approximation and set $\{ e_1, \ldots, e_n \}$ of the corresponding approximation errors relative to the observed values of the dependent variable,that minimizes the objective function TAE

(1) $\quad TAE = \sum_{i=1}^k c_i' (b_i^+ + b_i^-) + \sum_{i=1}^n c_i'' (e_i^+ + e_i^-)$

subject to

$$b_1^+\tilde{x}_{11} - b_1^-\tilde{x}_{11} + b_2^+\tilde{x}_{12} - b_2^-\tilde{x}_{12} + \ldots + e_1^+ - e_1^- = \tilde{y}_1$$

$$b_1^+\tilde{x}_{21} - b_1^-\tilde{x}_{21} + b_2^+\tilde{x}_{22} - b_2^-\tilde{x}_{22} + \ldots + e_2^+ - e_2^- = \tilde{y}_2$$

(2) $\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$

$\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$

$$b_1^+\tilde{x}_{n1} - b_1^-x_{n1} + b_2^+\tilde{x}_{n2} - b_2^-\tilde{x}_{n2} + \ldots + e_n^+ - e_n^- = \tilde{y}_n$$

(3) $\quad e_i^+ , e_i^- > 0 \qquad b_j^+ , b_j^- > 0 \qquad j=1,\ldots,k$

where $e_i = e_i^+ - e_i^-$ $\quad b_j = {}^+ - b_j^-$

and $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$ $\quad \tilde{y}_i = y_i - \bar{y}$

The constraints (2') consist of n equations in 2(n+k) unknowns and therefore normally possess an infinite number of solutions.These equations cam be solved for any n variables in terms of the remaining n-2k variables.These n variables are called basic variables and remaining n+2k variables are called non basic variables.Note that setting non basic variables equall to zero the equations (2') can be used to calculate the solution.This type of the solution is called a basic solution.The simplex iterations can be performed within an array of dimensions (n+1),(2(k+n)+1), where n-number of observations and k number of regressor variables which initially has the following form:

$c_i^{'} \; c_i^{''}$	⟶	0	0		0	0	1	1		1	1	
$c_i^{'b} \; c_i^{''b}$	BASIS ↓	b_1^+	b_1^-	$\ldots b_k^+$	b_k^-	e_1^+	e_1^-	$\ldots e_n^+$		e_n^-		$\tilde{y}$
1	e_1	$\tilde{x}_{11}$	$-\tilde{x}_{11}$	$..\tilde{x}_{1k}$	$-\tilde{x}_{1k}$	1	-1	0		0		$\tilde{y}_1$
.	.	.	.	.	.	.	.	.		.		.
.	.	.	.	.	.	.	.	.		.		.
1	e_n	$\tilde{x}_{n1}$	$-\tilde{x}_{n1}$	$..\tilde{x}_{nk}$	$-\tilde{x}_{nk}$	0	0	1		-1		$\tilde{y}_n$
$c_j - c_i^b x_{ij}$		0	0	...0	0	0	2	0		2		

The initial basis solution has the form:
$b_j^+ = b_j^- = 0$ $j=1,..,k$, $e_i^- = 0$, $e_i^+ = y_i$ $\quad i=1,...,n$ and cores-ponding value of the objective function $TAE = \Sigma_{i=1}^{n} e_i^+ = \Sigma_{i=1}^{n} y_i = 0$.

This initial solution does not however fulfiled the constraints (3'),because some of $\tilde{y}_i$ are always negative due to the fact that algorithm operates on the deviations from the mean values of original variables.The next basic solution is find by means of simplex substitution of one of the basic variable.This is completed in following steps.

S1. Determine the variable to leave the basis choosing this one,which coresponds to the smallest negative value of $\tilde{y}_i$.

S2. Determine the variable to enter the basis choosing this one,which affects the increase of the objective function as small as possible.

In order to do this,let us note that the repleacement of the i-th basic variable by j-th non-basic variable causes such an evaluation of the simplex table,that the column coresponding to the variable which enter the basis will after transformation posses 1 in the i-th row and 0 in the others, Such a transformation is called a simplex pivot. The effect of this transformation on the value of the objective function is given by

$$\Delta(TAE) = \frac{\tilde{y}_{ipivot}}{\tilde{x}_{ipivot,jpivot}} c_{jpivot} +$$

$$+ \sum_{i \in b-\{ipivot\}} (\tilde{y}_i - \frac{\tilde{y}_{ipivot}\tilde{x}_{i,jpivot}}{\tilde{x}_{ipivot,jpivot}} c_i)$$

$$- c_{ipivot}\, \tilde{y}_{ipivot} - \sum_{i \in b-\{ipivot\}} c_i \tilde{y}_i =$$

$$= \frac{\tilde{y}_{ipivot}}{\tilde{x}_{ipivot,jpivot}} (c_{jpivot} - \sum_{i \in b} \tilde{x}_{i,jpivot}\, c_i)$$

where $c_{i \in b}$ -coefficients of basic variables in objective function

ipivot - index of the pivot row in simplex table,

jpivot - index of the pivot column in simplex table.

In order to minimize Δ TAE we must, due to negativity of $\tilde{y}_{ipivot}$, choose such a variable to enter the basis (i.e. such a pivot column j=jpivot) which minimizes the value

$$\frac{c_j - \sum_{i \in b} \tilde{x}_{i,j} c_i}{\tilde{x}_{ipivot,j}}$$

We must however limited our choice only to negative values of $\tilde{x}_{ipivot,j}$, because the value of the new basic variable which will be equall $\tilde{y}_{ipivot} / \tilde{x}_{ipivot,jpivot}$ ought to be positive, to satisfy constraints (3').

After choosing variable to enter the basis we finished iteration in step 3 and return to S1.

S3. Perform an ordinary simplex pivot.

This is achieved by putting for all j,

$$\tilde{x}_{i,j} = \begin{cases} \tilde{x}_{i,j} / \tilde{x}_{ipivot,jpivot} & i=ipivot \\ \tilde{x}_{i,j} - \dfrac{\tilde{x}_{ipivot,j}}{\tilde{x}_{ipivot,jpivot}} \tilde{x}_{i,jpivot} & i=1,..n \ , i \neq ipivot \end{cases}$$

$$\tilde{y}_i = \begin{cases} \tilde{y}_i / \tilde{x}_{ipivot,jpivot} & i=ipivot \\ \tilde{y}_i - \dfrac{\tilde{y}_{ipivot}}{\tilde{x}_{ipivot,jpivot}} \tilde{x}_{i,jpivot} & i=1,..,n \ , \ i \neq ipivot \end{cases}$$

$$c_j - \sum_{i \in b} \tilde{x}_{i,j} c_i = (c_j - \sum_{i \in b} \tilde{x}_{i,j} c_i) - \frac{\tilde{x}_{ipivot,j}}{\tilde{x}_{ipivot,jpivot}} \cdot$$

$$\cdot (c_{jpivot} - \sum_{i \in b} \tilde{x}_{i,jpivot} c_i)$$

The algorithm terminates with optimal solution when in step 1 there are not any negative values of $\tilde{y}_i$. The final values of b_i and e_i are equal 0 if index "i" coresponds to non-basic variables; $\tilde{y}_i^*$ if "i" coresponds to basic variables b_i^+ or e_i^+ and $-\tilde{y}_i$ if "i" coresponds to basic variables b_i^- or e_i^-, where $\tilde{y}_i^*$ denotes $\tilde{y}_i$ value after last iteration. Constant term is computed from $b_o = \bar{y} - \sum_{j=1}^{k} b_j \bar{x}_j$.

For computational purposes original simplex table can be extremally reduced. It is sufficient to store only k+1 columns, obtaining in result condensed table of dimensions (n+1) (k+1). This can be achieved due to the fact that in original table there are (n+k) duplicate columns differing only in sign, and we need not store the basic columns. The result of the simplex transformations on the variable which leave the basis is simply writen in the column coresponding to variable which enter the basis. The corespondence betwen columns and rows of the simplex matrix and respective indices of non-basic and basic variables is keeping by means of two working one dimensional arrays. The i-th element of the first array points the number of variable which coresponds to i-th column of the matrix (i-th non-basic variables). The i-th element of the second array points the number of variable which corespond to i-th row of the simplex matrix (i-th basic variable). The negative sign of the index points on the e_i^- or b_i^- variable.

It is also worthly to note, that in the case, when the leaving and entering the basis variables differ only in sign there is no need to perform simplex pivot. In this case we change only the signs of elements in pivot row and values of $c_j - \sum_{i \in b} \tilde{x}_{ij} c_i$.

The ilustration of the use of the procedure gives the following example.

Exemple 1

Table 1 contains a sample of obserwations on two variables x and y. The problem is to fit a regression line of the form

$$\hat{y} = a + bx$$

which minimizes $\sum_i |y_i - \hat{y}_i|$.

Table 1

x	-3	-2	-2	-1	4	4	$\bar{x} = 0$
y	-8.5	-6.5	-5	-2.5	12	13.5	$\bar{y} = 0.5$

We start with the following simplex table based on the deviations from the mean for x and y.

c ↓	basis ↓	0_{b^+}	0_{b^-}	$1_{e_1^+}$	$1_{e_1^-}$	$1_{e_2^+}$	$1_{e_2^-}$	$1_{e_3^+}$	$1_{e_3^-}$	$1_{e_4^+}$	$1_{e_4^-}$	$1_{e_5^+}$	$1_{e_5^-}$	$1_{e_6^+}$	$1_{e_6^-}$	$y_i - \bar{y}$
1	e_1^+	-3	3	1	-1	0	0	0	0	0	0	0	0	0	0	-9
1	e_2^+	-2	2	0	0	1	-1	0	0	0	0	0	0	0	0	-7
1	e_3^+	-2	2	0	0	0	0	1	-1	0	0	0	0	0	0	-5.5
1	e_4^+	-1	1	0	0	0	0	0	0	1	-1	0	0	0	0	-3
1	e_5^+	4	-4	0	0	0	0	0	0	0	0	1	-1	0	0	11.5
1	e_6^+	4	-4	0	0	0	0	0	0	0	0	0	0	1	-1	13
$c_j - \sum_i c_i x_{ij}$		0	0	0	2	0	2	0	2	0	2	0	2	0	2	

The informations contained in this table can be in practice stored in the following condenced form, which is possible due to the fact ,that all variables are duplicated and in pairs differ only in sign.

Basis	non-basis var. b^+	$y_i-\bar{y}$
e_1^+	-3	-9
e_2^+	-2	-7
e_3^+	-2	-5.5
e_4^+	-1	-3
e_5^+	4	11.5
e_6^+	4	13
$c_j-\sum_i c_i x_{ij} = 0$		TAE= 0

As the variable which leaves the basis we choose e_1^+ ,because it coresponds to the smallest negative value of$(y_i-\bar{y})$equal -9.The basis enters the variable b^+, which characterizes the greatest value of $(c_j-\sum c_i x_{ij})/x_{ipivot,j}$, for all $x_{ipivot,j}$ smaller than 0.Note,that for variables, which have their "pair-variables" in basis (i.e. have -1 in pivot row and zeros elsewhere) the value of this ratio is always equal -2,which simplyfies comparisons.

After simplex pivot we obtain the table:

Basis	non-basis var. e_1^+	$y_i-\bar{y}$
b^+	-0.333	3
e_2^+	-0.666	-1
e_3^+	-0.666	0.5
e_4^+	-0.333	0
e_5^+	1.333	-0.5
e_6^+	1.333	1
$c_j-\sum c_i x_{ij} = 0$		TAE= 0

In the following iterations we have respectively:

III. $\min (y_i-\bar{y})=-1$, thus ipivot = 2

$$\max \frac{c_j-\sum_i c_i x_{ij}}{x_{ipivot,j}} = 0 \text{ for } e_2^+$$

Basis	non-basic var. e_2^+	$y_i-\bar{y}$
b_1^+	-0.5	3.5
e_1^+	-1.5	1.5
e_3^+	-1	1.5
e_4^+	-0.5	0.5
e_5^+	2	-2.5
e_6^+	2	-1
$c_j-\sum_i c_i x_{ij}$	0	0

IV. $\min(y_i-\bar{y})=-2.5$, thus ipivot= 5

$$\max \frac{c_j-\sum_i c_i x_{ij}}{x_{ipivot,j}} = 0 \text{ for } e_2^- \quad (\text{because } x_{ipivot,j} \text{ for } e_2^+ \text{ is positive})$$

Note that all elements in pivot column e_2^+ must be taken with opposite sign during transformations.

Basis	non-basic var. e_5^+	$y_i-\bar{y}$
b_1^+	0.25	2.875
e_1^+	0.75	-0.375
e_3^+	0.5	0.25
e_4^+	0.25	-0.125
e_2^-	-0.5	1.25
e_6^+	-1	1.5
$c_j-\sum_i c_i x_{ij}$	1	3

V. $\min(y_i-\bar{y}) = -0.375$ thus ipivot = 2

$$\max \frac{c_j-\sum c_i x_{ij}}{x_{ipivot,j}} = -1.333 \text{ for } e_5^-$$

Basis	non-basic var. e_1^+	$y_i-\bar{y}$
b_1^+	-0.333	3
e_5^-	-1.333	0.5
e_3^+	-0.666	0.5
e_4^+	-0.333	0
e_2^-	0.666	1
e_6^+	1.333	1
$c_j-\sum_i c_i x_{ij}$	1.333	3

This is the final solution because all of values $y_i-\bar{y}$ are positive. The coresponding value of the objective function is equal 3 and values of residuals and regression parameters are equal respectively :
$e_1=0$, $e_2=-e_2^-=-1$, $e_3=e_3^+=0.5$, $e_4=e_4^+=0$, $e_5=e_5^-=-0.5$, $e_6=e_6^+=1$, $b=b^+=3$ and $a=\bar{y}-\sum_{i=1}^{k} \bar{x}_i b_i = 0.5-0 = 0.5$.
The estimated fited line is of the form

$$\hat{y} = 0.5 + 3x$$

Snyder's LAD-estimation algorithm

Algorithm proposed by Snyder (1978) is more general than the previous one,because it operates on original data without the requirement that the fitted line passes through the mean.It is also based on the simultaneous equations substitution procedure,but uses an effective method for selection a pivot equation for substitution procedure,based on the analysis of the gradient of TAR with respect to the changes of the value of the chosen non-basic variable.

The algorithm does not use artificial non-negativity conditions and thus solves the minimization problem for n-k unknowns and n constraints

$$(4) \qquad \sum_{j=1}^{k} x_{ij} b_j + e_i = y_i \qquad i = 1,2,..,n$$

The algorithms begins by solving for the error variables in term of the regrresion coefficients to give

$$(5) \qquad e_i = y_i - \sum_{j=1}^{k} x_{ij} b_j \qquad i= 1,2,..,n$$

where the variables on the left hand side of the equation represents basic variables and variables on the right hand side of the equation represent non-basic variables. During the algorithm some of the error variables replace the regression coefficients on the right hand side of the equations as non-basic variables, giving in general the system of equations

$$(6) \qquad V_i = g_{io} - \sum_{j=1}^{k} g_{ij} Z_j \quad . \; i = 1,2,..,n$$

where V_i and Z_j may represent error or regression coefficient variables being in given iteration respectively basic and non-basic variables.

The algorithm uses a systematic procedure for generating new basic solutions by varying one of the non-basic variables at a time from its zero position, and analyzing the effect of this shift on the value of total absolute error. All computations are realized by means of the n x (k+2) table

Basis	non-basic var. Z_1 Z_2 ..... Z_{k+1}	g_{io}
V_1	g_{11} g_{12} ..... g_{1k+1}	g_{1o}
.	. . ..	.
.	. . ..	.
V_n	g_{n1} g_{n2} ..... g_{nk+1}	g_{no}

which initialy is of the form

Basis	non-basic var. a b_1 .....b_k	g_{io}
e_1	1 x_{11}.....x_{1k}	y_1
e_2	1 x_{21}.....x_{2k}	y_2
.	. .....	.
.	. .....	.
e_n	1 x_{n1}.....x_{nk}	y_n

Each iteration of the algorithm begins by selecting one of the non-basic variables,say Z_q,for variation and finding such a value of Z_q which minimizes objective function TAE where $TAE = \sum_i |e_i|$,while the others non-basic variables staing to be equal zero.

In order to gain better insight into the way in which it is done,note that the TAE function is piecewise linear and convex,with vertices at these points where one or more errors equal zero. Using the relation (6) for $V_i = e_i$, the vertex occurs at the point $Z_q = r_i$ where $r_i = g_{io}/g_{iq}$, ($i = 1,2,..,n$, $g_{iq} = 0$).The derivative

$$\frac{d\, e_i}{d\, Z_q} = \mathrm{sgn}(Z_q - r_i)\,|g_{iq}| \tag{7}$$

is positive and equal to $|g_{iq}|$ to the right of this point and it is negative and equal $-|g_{iq}|$ to the left of this point. Thus the derivative increases by $2|g_{iq}|$ when moving from left to right through the point in which $e_i = 0$.The derivative of the objective function TAE with respect to Z_q equals simply the sum of derivatives of its components i.e.

$$\frac{d\ TAE}{d\ Z_q} = \sum_i \mathrm{sgn}\,(Z_q - r_i)\,|g_{iq}| \quad ; i = 1,2,..,n \tag{8}$$

Due to convexity, the TAE function achieves its minimum with respect to Z_q, for such a value of Z_q which indicates the change of the sign of the respective derivative d TAE/d Z_q , when moving from the left to the right at a coresponding vertex. To simplify the search of such a value Z_q we examine at first the derivative d TAE/d Z_q at the point $Z_q = 0$, to choose , if any, the proper direction of the further improvement. The negativity of the right derivative $dTAE/dZ^{+}_{q/Z_q=0}$ indicates, that a search to the right may yield better basic solution, i.e. increase of Z_q from its current level causes the decrease of the TAE value, at a rate of $dTAE/dZ_q$ until the next vertex is reached. The positivity of the left derivative indicates on points in the left direction for improvement.

The algorithm then moves from vertex to vertex in the direction of improvement recursively recalculating the derivatives as it passes through each vertex. This is done until a point is found, where the derivative changes the sign or stays to be equal zero moving from the left to right. Such a point represents the best basic solution that can be achieved by varying Z_q at this stage of algorithm. The coresponding error which equals zero is then forced to become a non-basic variable using conventional simultaneous equations substitution procedure. This yields a new system of equations and therefore complets the iteration.

Ocasionally when a new iteration begins the derivative of the TAE function immediately to the left and the right of the point $Z_q = 0$ is non-positive and non-negative respectively. In these circumstances no benefit can be derived by changing Z_q from its current value of zero and an attempt is made to find another non-basic variable which can be varied to give better result.

Thus the optimal solution is achieved in following steps:

S1. Select one non-basic variable, say Z_q, for variation and compute the values of right and left derivative of TAE at $Z_q = 0$

$$(9) \quad \frac{dTAE}{dZ_q^+ |_{Z_q=0^+}} = \sum_{i=1}^{n} sgn^+(-g_{io})\, g_{iq} \quad \text{,where } sgn^+(x) = \begin{cases} 1 & x \geqslant 0 \\ -1 & x < 0 \end{cases}$$

$$(10) \quad \frac{dTAE}{dZ_q^- |_{Z_q=0^-}} = \sum_{i=1}^{n} sgn^-(-g_{io})\, g_{iq} \quad \text{,Where } sgn^-(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leqslant 0 \end{cases}$$

and $g_{io} = 0$ if it-h residual is non-basic residual
$g_{iq} = 0$ if it-h residual is non-basic residual and $i \neq q$.

If $dTAE/dZ_q^+ < 0$ go to S2 ;

If $dTAE/dZ_q^- > 0$ go to S3.

S2. Order the positive values of the points in which vertices occur (i.e. the values of $r_i = g_{io}/g_{iq}$ for basic residual variables and $g_{iq} \neq 0$) in increasing manner and calculate recursively for each vertex the value of the new derivative

$$(11) \quad \frac{dTAE}{dZ_q^+ |_{Z_q=r_{i+1}}} = \frac{dTAE}{dZ_q^+ |_{Z_q=r_i}} + 2|g_{iq}|$$

unless the derivative changes its sign or stays to be equal zero. Then go to S4.

S3. Order the negative values of the points in which vertices occur in the decreasing manner and calculate recursively for each vertex the value of the new derivative

$$(12) \quad \frac{dTAE}{dZ_q^- |_{Z_q=r_{i+1}}} = \frac{dTAE}{dZ_q^- |_{Z_q=r_i}} - 2|g_{iq}|$$

unless the derivative changes its sign or stays to be equal zero.Then go to S4.

S4. Set the i-th row as a pivot row and transform the data using simultaneous equations substitution procedure. In result the variable z_q will be the new basic variable and the residual corresponding to the point r_i in which $dTAE/dZ_q$ changed its sign will be non-basic variable equal zero.Begin the new iteration in S1.

Algorithm terminates when in step S1 none of the conditions $dTAE/dZ_q^+ < 0$, $dTAE/dZ_q^- > 0$ is fulfiled.

The ilustration of this algorithm is given below in the example 2.

Example 2

Using data from the example 1,we will estimate coefficients a and b of the regression line $\hat{y} = a + bx$ by means of the Snyder's LAD-algorithm.

The initial table is of the form

Basis	non-basic var. a	b	g_{io}
e_1	1	-3	-8.5
e_2	1	-2	-6.5
e_3	1	-2	-5
e_4	1	-1	-2.5
e_5	1	4	12
e_6	1	4	13.5

Selecting regression coefficient "a" as a variable entering the basis we have:

$$\frac{dTAE}{da^+|_{a=0^+}} = 2 \quad , \quad \frac{dTAE}{da^-|_{a=0^-}} = 2$$

This result indicates,that by decreasing "a" from its zero level the TAE declines at a rate of 2 until the next vertex is reached,at $a=g_{4o}/g_{41} = -2.5$,where $e_4 = 0$.

At this point all but $d|e_4|/da$ remain unchanged.In fact $d|e_4|/da$ changes from 1 to -1,i.e. decreases by 2.Since dTAE/da also decreases by 2 to zero it follows , that the minimum of TAE with respect to"a"occurs at this point. We let e_4 become non-basic variable instead of"a"transforming the initial table by means of simplex simultaneous equations substitution procedure.In result we obtain the new simplex table of the form:

Basis	non.basic var. e_4	b	g_{io}
e_1	-1	-2	-6
e_2	-1	-1	-4
e_3	-1	-1	-2.5
a	1	-1	-2.5
e_5	-1	5	14.5
e_6	-1	5	16

Selecting "b" as a variable entering the basis we have at b=0,$dTAE/db^+ = -14$, which points, that the improvement of the solution can be achieved in point coresponding to the positive values of $b = r_i = g_{io}/g_{i2}$ for basic residuals' variables,i.e. for b=2.9,b=3,b=3.2 and b=4.Ordering these values in increasing manner and recursively recalculating the derivative $dTAE/db^+$ in each point going from left to the right we obtain:

$$\frac{dTAE}{db^+}\Big|_{b=2.9} = -14 + 10 = -4 \quad \text{and} \quad \frac{dTAE}{db^+}\Big|_{b=3} = -4 + 4 = 0$$

This indicates that TAE achieved its minimum with respect to "b" for b = 3 , where $e_1 = 0$.Thus variable e_1 leaves the basis and after simplex transformations we obtain the new table of the form:

Basis	non-basic var. e_4	e_1	g_{io}
b	-0.5	0.5	3
e_2	-1.5	0.5	-1
e_3	-1.5	0.5	0.5
a	0.5	0.5	0.5
e_5	1.5	-2.5	-0.5
e_6	1.5	-2.5	1

Comparing the values of the right and the left derivatives of TAE with respect to e_4 and e_1 respectively we have

$$\frac{dTAE}{de_4^+|e_4=0^+} = 1 \quad , \quad \frac{dTAE}{de_4^-|e_4=0^-} = -1$$

$$\frac{dTAE}{de_1^+|e_1=0^+} = 1 \quad , \quad \frac{dTAE}{de_1^-|e_1=0^-} = -1$$

which indicates , that the current value of zero for these two non-basic variables is the best.This therefore is the optimal solution,and the estimated regression line is of the form $\hat{y} = 0.5 + 3x$.

Comparison of the algorithms

The results obtained in examples 1 and 2 shows, that both algorithms lead to the same solution,if there exists unique solution ofcourse,but the computational effort can be different for each of algorithms.The simulation experiments seem to point on the greater computational efficency of the Snyders algorithm in comparison with the algorithm of Narula and Wellington and the rewighted least squers method due to Schlosmacher.The last one can be easely implemented on the basis of the standard least squares method,but it is very time-consuming method for computations.

In order to compare the behaviour of the algorithms with respect to various levels of bad condition of the data, used to estimation, some numerical experiments have been carried out. The experimental data were generated by means of algorithm proposed by Kennedy, Gentle and Sposito (1977) which allow the choice of:

1. the column means of the data matrix for all columns after the first, which coresponds to the constant term in regression line;
2. the order of magnitude of the condition number of the data matrix;
3. the L1-solution vector for estimated regression coefficient vector;
4. the deviations about the fitted hyperplane.

The datasets (X:y) with prescribed numerical condition and given distribution of the deviations were forming as follows:

$$y_t = x_t'\beta \qquad t=1,\ldots,k$$

$$y_t = \begin{cases} x_t'\beta - e_t & t=k+1,k+3,\ldots,n-1 \\ x_t'\beta + e_t & t=k+2,k+4,\ldots,n \end{cases}$$

where k denotes the number of estimated parameters,
$n-k$ is even, $e_t > 0$ for $t=k+1,k+2,\ldots,n$ and
$x_t = x_{t+1}$ for $t=k+1,k+3,\ldots,n-1$.

The estimates LADNW obtained by algorithm of Narula and Wellington were compared with these obtained by means of algorithm due to Snyder (LADS) and estimates IRLAD obtained by means of iteratively rewighted least squares method. The main results of the experiments for the regression model $y = X\beta + \varepsilon$ with X matrix of the dimensions $n=25, k=5$, are given in table 2.

Table 2

Mean relative errors of estimation

$$MRE(b)=\frac{1}{5}\sum_{i=1}^{5}\frac{b_i-\beta_i}{\beta_i}$$

with respect to degree of bad condition of X

index of bad condition	estimators IRLAD	LADNW	LADS
24	.00000	.00000	.00000
500	.00038	.00004	.00000
1500	.35560	.00008	.00000
7000	36.62842	.00046	.00000

From the table 2 results , that the greatest resistance on the bad condition of the data matrix X characterizes the estimator obtained by algorithm due to Snyder.Algorithm of Narula and Wellington gives the estimates which are only a bit worse from this point of view.Estimates obtained by rewighted least squares method characterized increase of the estimation error together with increase of degree of bad condition of matrix X beyond certain bounds.The mean relative error of estimation depends in the last case on the number of iterations allowed by user.

References

1. Barrodale,I.,Roberts,F.D.K.(1973):An Improved Algorithm for Discrete L1 Linear Approximations SIAM J.Numer.Anal.,Vol.10 No.5
2. Karst,O.J.(1958):Linear Curve Fitting Using Least Deviations,J.Amer.Statist.Ass.pp 118-132
3. Kennedy,W.J.,Gentle,J.E.,Sposito,V.A.(1977): A Computer Oriented Method For Generating Test Problems For L1 Regression,Commun.Statist.-Simula Computa.,B6(1),pp 21-27
4. Narula,S.C.,Wellington,J.F.(1976):Multiple Linear Regression With Minimum Sum of Absolute Errors, Appl.Statist.Ser.C,pp 106-111
5. Snyder,R.D.(1978):Regression Analysis with the Absolute Error Criterion.Working Paper.Monash Univrsity,Department of Econometrics and Operations Research.
6. Taylor,L.(1974):Estimation by Minimizing the Sum of Absolute Errors.In Frontiers in Econometrics P.Zarembka edit. N.Y.
7. Wagner,H.M.(1958):Linear Programming Techniques for Regression Analysis,J.Amer.Statist.Ass. 56, pp.206-212