

# Diagnostik von Regressionsschätzungen bei kleinen Stichproben

Ben Jann

24. März 2005

## Zusammenfassung

Wie alle anderen statistischen Verfahren konzentriert sich auch die Methode der Regression nur auf die Analyse ausgewählter Aspekte vorliegenden Datenmaterials. Entsprechend sind zu gegebenen Regressionsergebnissen ganz unterschiedliche Datenkonstellationen denkbar, wovon aber für die Interpretation der Ergebnisse nicht alle unproblematisch sind. So besteht besonders bei kleinen Stichproben die Gefahr, dass die Regressionsschätzung entscheidend von einzelnen Extremwerten abhängt, was die Verlässlichkeit der daraus abgeleiteten Schlussfolgerungen beeinträchtigt. In diesem Beitrag werden deshalb anhand von Beispielen einige einfache grafische und formale Instrumente zur Diagnose einflussreicher Datenpunkte vorgestellt, die im Prozess der Datenanalyse standardmässig angewendet werden sollten. Weiterhin werden nach Identifikation „atypischer“ Datenpunkte zu verfolgende Analysestrategien diskutiert.

*Schlagwörter:* Ausreisser, einflussreiche Daten, kleine Stichproben, Regressionsdiagnostik, robuste Regression

## Regression Diagnostics in Small Samples

Like all other statistical procedures, regression methods focus on the analysis of only select aspects of the data material at hand. Accordingly, a given set of regression results may result from many different data constellations, not all of which lend themselves to straightforward interpretation. Particularly in small samples, regression estimates may, for example, depend critically on a few extreme data points, a situation in which the reliability of the conclusions derived from the regression analysis is greatly reduced. In this contribution, I will therefore introduce a few simple graphical and formal instruments for the diagnosis of influential data. These tools should be applied to data analysis as a matter of course. The steps to be taken after the identification of “atypical” cases are also discussed.

*Keywords:* Influential Data, Outliers, Regression Diagnostics, Robust Regression, Small Samples

## **Inhaltsverzeichnis**

<b>1</b>	<b>Viele Wege führen nach Rom</b>	<b>3</b>
<b>2</b>	<b>Grafische Beurteilung der Daten</b>	<b>5</b>
2.1	Partielle Regressionsplots . . . . .	6
2.2	Partielle Residuenplots . . . . .	7
<b>3</b>	<b>Formale Identifikation einflussreicher Datenpunkte</b>	<b>11</b>
3.1	Leverage . . . . .	11
3.2	Residuen . . . . .	12
3.3	Einfluss auf die Koeffizienten und Vorhersagewerte . . . . .	14
3.3.1	Gesamteinfluss . . . . .	15
3.3.2	Einfluss auf individuelle Koeffizienten . . . . .	16
3.4	Einfluss auf die Präzision der Schätzung . . . . .	17
3.5	Übersicht und weitere Ansätze . . . . .	18
<b>4</b>	<b>Reputationseffekte in Internet-Auktionen: Ein Anwendungsbeispiel</b>	<b>21</b>
<b>5</b>	<b>Diskussion</b>	<b>30</b>

*We are usually happier about asserting a regression relation if the relation is still apparent after a few observations (any ones) have been deleted – that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two. (Anscombe 1973: 18)*

*An apparently wild (or otherwise anomalous) observation is a signal that says: “Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study.” (Kruskal 1960: 1)*

## 1 Viele Wege führen nach Rom

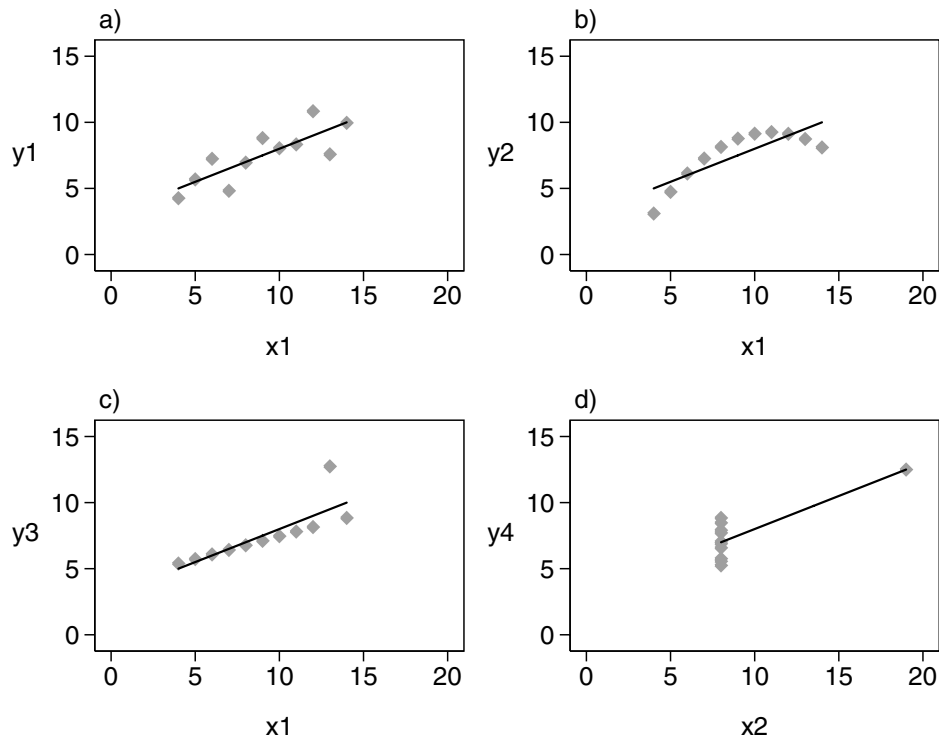
Die Methode der (linearen) Regression ist zweifelsohne eines der am häufigsten verwendeten Verfahren zur statistischen Behandlung von sozialwissenschaftlichen Fragestellungen. Trotz der sehr günstigen allgemeinen Eigenschaften und der unumstrittenen Nützlichkeit des Verfahrens ist es aber keineswegs so, dass eine Regressionsschätzung vorliegendes Datenmaterial umfassend beschreiben würde. Viele Eigenschaften der Daten bleiben verborgen und können an den üblicherweise im Rahmen einer Regressionsanalyse berechneten Zahlen nicht abgelesen werden. Entsprechend sind zu gegebenen Regressionsergebnissen ganz unterschiedliche Datenkonstellationen denkbar.

Ein eindruckliches Beispiel hierfür liefert Anscombe (1973): Die Daten, die den vier Diagrammen in Abbildung 1 zu Grunde liegen, führen jeweils zu den gleichen Schätzergebnissen. Die Regressionsgerade lautet in allen Diagrammen  $Y = 3.0 + 0.5X$  und eine ganze Reihe weiterer Kennwerte ist jeweils praktisch identisch: Die Mittelwerte und Standardabweichungen der Variablen sind gleich ( $\bar{X} = 9.0$ ,  $s_X = 3.32$ ,  $\bar{Y} = 7.5$ ,  $s_Y = 2.03$ ) und in allen vier Fällen werden 67 % der Varianz der abhängigen Variablen erklärt. Zudem ergeben sich für die Regressionsparameter immer die gleichen Standardfehler (ca. 1.125 und 0.118), das heisst, auch die inferenzstatistischen Schlüsse (Hypothesentests über den Einfluss von  $X$  auf  $Y$ ), die man aus den Ergebnissen ziehen würde, sind identisch.

Weder an den normalerweise in Publikationen angegebenen Kennzahlen zu einer Regressionsschätzung noch anhand der Standardausgabe eines Statistikprogramms sind die vier Fälle also unterscheidbar. Dennoch würde man wohl nur von den Daten in Diagramm (a) vorbehaltlos behaupten, dass sie durch die Regressionsgleichung adäquat beschrieben werden. Die Daten in Diagramm (b) beruhen offensichtlich auf einem nichtlinearen Zusammenhang, in Diagramm (c) gibt es einen einzelnen Datenpunkt der von einem sonst sehr deutlichen Zusammenhang abweicht und in Diagramm (d) wird eine Regressionsschätzung überhaupt erst durch den Datenpunkt am rechten Rand der Abbildung ermöglicht. Ohne diesen Punkt wäre die Varianz der  $X$ -Variablen gleich null und somit die Steigung der Regressionsgeraden nicht bestimmbar.

Das Beispiel von Anscombe verdeutlicht die Notwendigkeit, die einer Regressionsschätzung zu Grunde liegenden Daten etwas genauer unter die Lupe zu nehmen und somit „Regressionsdiagnostik“ zu betreiben. Einerseits werden bei der Schätzung von Regressionsmodellen verschiedene Annahmen getroffen, die mehr oder weniger stark verletzt sein können (in Abbildung 1b, zum Beispiel, wird die Annahme eines linearen Zusammenhangs zwischen  $Y$  und  $X$  verletzt), andererseits kann sich das Problem stellen, dass einzelne Messwerte in auffälliger Weise vom Rest der Daten abweichen (wie in Abbildung 1c) und unter Umständen einen sehr

Abbildung 1: Anscombes Quartet



starken Einfluss auf die Ergebnisse der Regressionsschätzung nehmen (Abbildung 1d), was die Gültigkeit der Ergebnisse zumindest in Frage stellt.

Besonders in kleinen bis moderaten Stichproben können die Resultate einer Regression ganz entscheidend von einzelnen Extremwerten abhängen und die Daten sollten standardmässig bezüglich des Vorliegens einer solchen Situation evaluiert werden. Ich werde deshalb in diesem Beitrag einige einfache grafische und formale Instrumente zur Diagnose von atypischen Datenkonstellationen vorstellen, wobei das Schwergewicht auf der Identifikation einflussreicher Datenpunkte liegen wird.<sup>1</sup> Nach einer kurzen Übersicht zur Notation werden im zweiten Abschnitt einige Plots zur grafischen Veranschaulichung der Datenstruktur vorgestellt. Abschnitt 3 befasst sich mit den wichtigsten Masszahlen zur Diagnose von Ausreissern und einflussreichen Fällen und in Abschnitt 4 wird eine exemplarische Anwendung der vorgestellten Instrumente durchgespielt. Abschnitt 5 fasst kurz zusammen und diskutiert sinnvolle Analysestrategien.

*Grundlegendes Modell und Notation:* Im Folgenden wird von einem multiplen linearen Regressionsmodell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{mit} \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I} \quad (1)$$

ausgegangen, wobei  $\mathbf{y}$  ein  $n \times 1$  Vektor der Werte der abhängigen Variablen,  $\mathbf{X}$  eine  $n \times k$  Matrix der Werte der Regressoren (inklusive Konstante),  $\boldsymbol{\beta}$  ein  $k \times 1$  Vektor der unbekannt Koeffizienten und  $\boldsymbol{\epsilon}$  ein  $n \times 1$  Vektor der unbekannt Fehler ist. Es gelten die üblichen Annahmen (vgl. z. B. Cook und Weisberg 1999; Fox 1997; Greene 2003; Wooldridge 2003), insbesondere dass

<sup>1</sup>Die wichtigsten Grundlagenwerke zu diesem Thema sind Belsley et al. (1980), Cook und Weisberg (1982a) und Chatterjee und Hadi (1988); zugängliche Einführungen liefern Fox (1991) und Bollen und Jackman (1990). In deutscher Sprache finden sich entsprechende Abschnitte z. B. in Brüderl (2000) oder Kohler und Kreuter (2001).

der Erwartungswert der Fehler gleich null ist und die Fehler unabhängig sind und konstante Varianz  $\sigma^2$  aufweisen. Der Parametervektor  $\beta$  wird auf Grundlage der Kleinste-Quadrate-Methode (OLS) geschätzt als

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2)$$

und die Residuen (die geschätzten Fehler) ergeben sich als

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} \quad (3)$$

Die Kovarianz-Matrix der OLS-Koeffizienten schliesslich wird berechnet als

$$\text{Cov}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \quad (4)$$

wobei

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n - k} \quad (5)$$

ein Schätzer für die Fehlervarianz ist.

## 2 Grafische Beurteilung der Daten

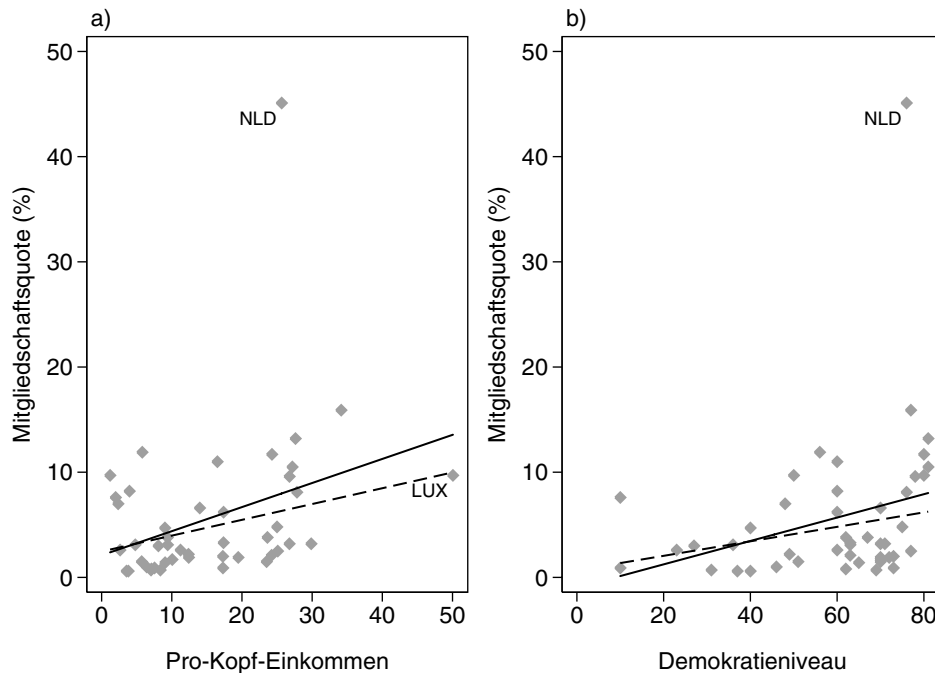
Einen guten ersten Überblick über die Beschaffenheit des Datenmaterials kann man sich meistens mit Hilfe von Grafiken verschaffen. Grafiken liefern zwar keine exakten Kennwerte, können aber sehr viel Information verdichten und auf einen Blick erfassbar machen. Im Falle einer Einfachregression kann zum Beispiel bereits ein einfaches Streudiagramm zwischen  $X$  und  $Y$  viele Eigenschaften der Daten wie etwa das Vorhandensein von Ausreißern oder auch Nichtlinearität oder Heteroskedastizität offen legen.

Die Streudiagramme in Abbildung 2 zeigen ein extremes Beispiel für einen Ausreisser. Die Daten stammen aus einer Studie von Dalton (2004), in der es darum geht, die länderspezifischen Unterschiede in der Mitgliedschaftsquote in Umweltorganisationen zu erklären. Bei den Mitgliedschaftsquoten handelt es sich um aggregierte Werte aus dem World Values Survey (1999–2002, je nach Land). Dalton untersuchte unter anderem, inwieweit der Zulauf von Umweltorganisationen vom Wohlstand eines Landes (gemessen am kaufkraftbereinigten Pro-Kopf-Einkommen) und vom Grad der Demokratisierung (gemessen mit einem Indikator für die Pressefreiheit) abhängt. Klar erkennbar ist der Ausreisser Holland mit einer phänomenalen Mitgliederrate von rund 45 % und es stellt sich die Frage, inwieweit Holland überhaupt mit den übrigen Ländern verglichen werden kann. Es handelt sich offensichtlich um einen Spezialfall (oder um einen „Messfehler“ z. B. aufgrund eines nicht vergleichbaren Messinstruments).<sup>2</sup> Auffällig ist zudem in Diagramm (a) der Datenpunkt für Luxemburg, bei dem es sich um einen – allerdings weniger dramatischen – Ausreisser bezüglich der  $X$ -Achse, also des Pro-Kopf-Einkommens, handelt.

Neben bivariaten Streudiagrammen können viele weitere Grafiken zur Beurteilung von Regressions-schätzungen verwendet werden (eine sehr ausführliche Übersicht gibt Schnell 1994). Beispiele sind Streudiagramme der Residuen gegen die Vorhersagewerte zur Ermittlung von Heteroskedastizität und Histogramme oder Normal-Quantil-Plots der Residuen zur Beurteilung der Normalitätsannahme. Zur Diagnose von Ausreißern und einflussreichen Datenpunkten sind vor allem zwei Plots von Bedeutung und sollen deshalb hier genauer besprochen werden: partielle Regressionsplots und partielle Residuenplots.

<sup>2</sup>Dalton (2004) berichtet denn auch sämtliche Resultate jeweils unter Einschluss wie auch unter Ausschluss von Holland.s

Abbildung 2: Mitgliedschaft in Umweltgruppen in Abhängigkeit von Wohlstand und Demokratie-niveau (die gestrichelte Linie entspricht der Regressionsgerade unter Ausschluss von NLD)



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

## 2.1 Partielle Regressionsplots

Für Modelle mit mehreren unabhängigen Variablen können Streudiagramme zwischen  $Y$  und den unabhängigen Variablen ein irreführendes Bild vermitteln, weil sie nur die bivariaten Zusammenhänge widerspiegeln, nicht jedoch die *partiellen* Einflüsse der Regressoren. Als Ergänzung sollte hier die Verwendung von partiellen Regressionsplots (*partial regression plot*, Bollen und Jackman 1990; auch *added variable plot*, Cook und Weisberg 1982a, oder *partial regression leverage plot*, Belsley et al. 1980) in Betracht gezogen werden.

Ein partieller Regressionsplot für die Variable  $X_j$  in einem multiplen Regressionsmodell (1) wird erstellt, indem die Residuen einer Regression von  $Y$  auf  $X_{\ell \neq j}$ , das heisst auf alle anderen Regressoren *ausser*  $X_j$ , gegen die Residuen einer Regression von  $X_j$  auf  $X_{\ell \neq j}$  abgetragen werden. Es werden also zwei Modelle

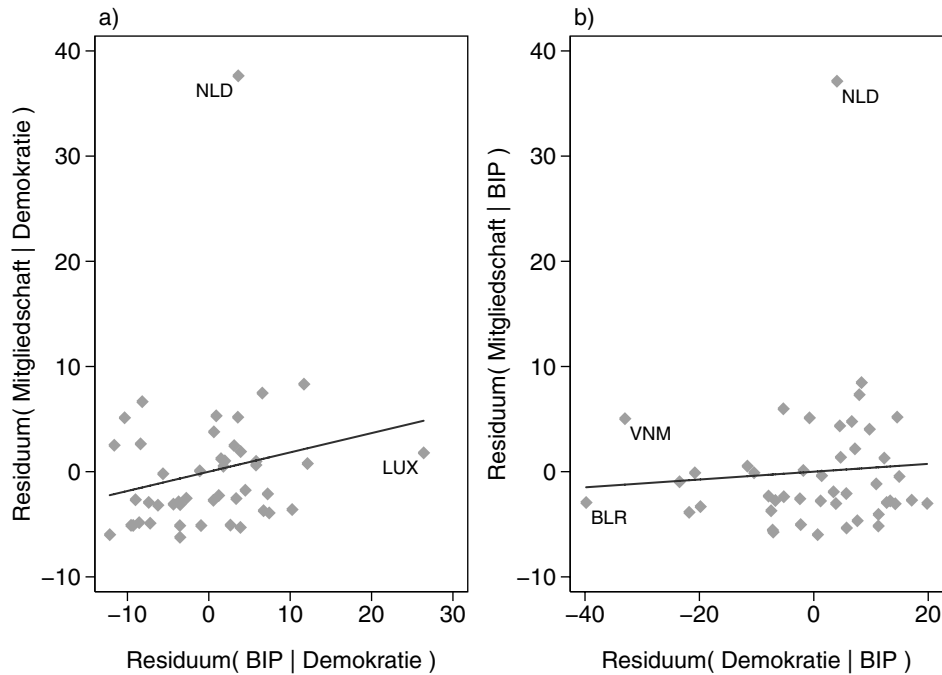
$$\mathbf{y} = \mathbf{X}_{(-j)}\boldsymbol{\alpha} + \boldsymbol{\nu} \quad \text{und} \quad \mathbf{x}_j = \mathbf{X}_{(-j)}\boldsymbol{\gamma} + \boldsymbol{\nu}$$

geschätzt und dann deren Residuen  $\hat{\boldsymbol{\nu}}$  und  $\hat{\boldsymbol{\nu}}_j$  in einem Streudiagramm abgebildet ( $\mathbf{X}_{(-j)}$  bezeichnet die Matrix der Werte der Regressoren ausschliesslich  $X_j$ ). Dies ist sinnvoll, da der Steigungskoeffizient des Modells

$$\hat{\boldsymbol{\nu}} = b_0 + b_1 \hat{\boldsymbol{\nu}}_j + \mathbf{e} \tag{6}$$

gerade dem Koeffizienten für  $X_j$  in der multiplen Regression entspricht, also  $\hat{b}_1 = \hat{\beta}_j$ . Zudem gilt  $\hat{\mathbf{e}} = \hat{\boldsymbol{\nu}}$ , das heisst, die Residuen von (6) sind identisch mit den Residuen aus der multiplen Regression (1). Von der Struktur her ist ein partieller Regressionsplot also immer noch ein einfaches Streudiagramm zwischen  $Y$  und  $X_j$ , ausser dass die beiden Variablen um den Teil, der

Abbildung 3: Partielle Regressionsplots für die Mitgliedschaft in Umweltgruppen in Abhängigkeit von Wohlstand und Demokratieniveau



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

durch die anderen Regressoren  $X_{\ell \neq j}$  erklärt wird, bereinigt wurden: Anstatt der ursprünglichen  $Y$  und  $X_j$  Werte werden die korrigierten Werte

$$\hat{v} = \mathbf{y} - \mathbf{X}_{(-j)}\hat{\boldsymbol{\alpha}} \quad \text{und} \quad \hat{w} = \mathbf{x}_j - \mathbf{X}_{(-j)}\hat{\boldsymbol{\gamma}}$$

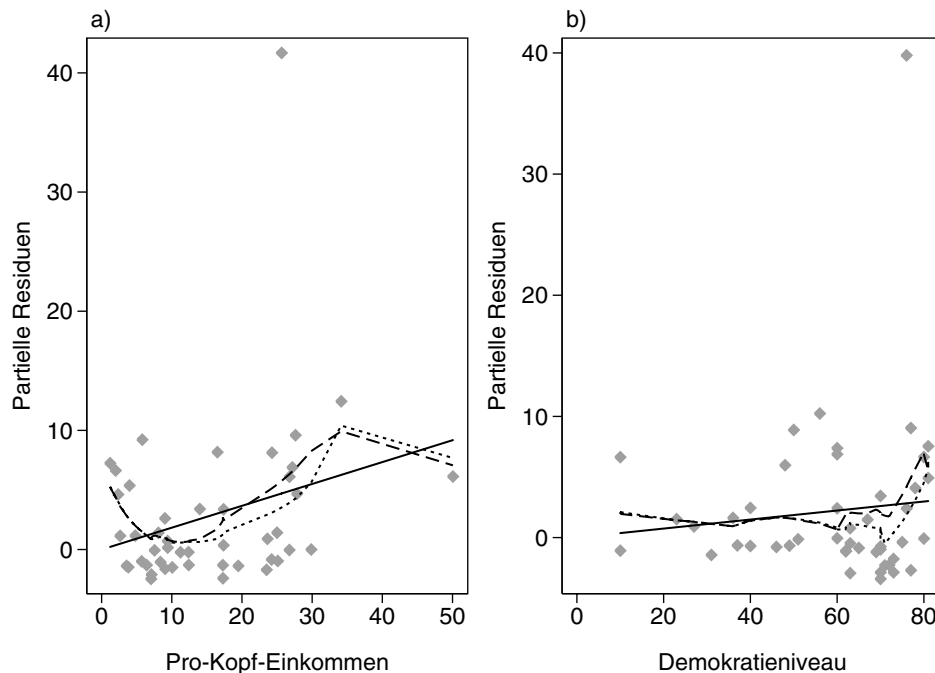
verwendet.

Abbildung 3 zeigt die partiellen Regressionsplots für die Daten von Dalton (2004), wenn die Mitgliedschaftsrate in Umweltgruppen simultan auf das Pro-Kopf-Einkommen und das Demokratieniveau regressiert wird. Auffallend ist erstens, dass die Steigung der Geraden im Plot für das Demokratieniveau (Diagramm b) im Vergleich mit dem bivariaten Streudiagramm deutlich abgenommen hat (diese Aussage ist hier allerdings nur möglich, da die Skalierungen der Achsen im Vergleich zum bivariaten Streudiagramm etwa gleich geblieben sind): Unter Kontrolle des Pro-Kopf-Einkommens (BIP) ist der Einfluss auf die Mitgliedschaftsquote praktisch null. Zweitens treten in Diagramm (b) zwei Ausreißer bezüglich der  $X$ -Achse stärker hervor. Es handelt sich um die zwei Länder in der Stichprobe, die die geringste Pressefreiheit aufweisen: Vietnam (VNM) und Weissrussland (BLR).

## 2.2 Partielle Residuenplots

Als Alternative und Ergänzung zu den partiellen Regressionsplots können auch partielle Residuenplots (*partial residual plot*; auch: *component-plus-residual plot*) berechnet werden, wobei sich diese allerdings etwas weniger gut zur Identifikation von einflussreichen Datenpunkten eignen. Ihre Stärke liegt darin, dass sie nicht-lineare Zusammenhänge relativ gut sichtbar ma-

Abbildung 4: Partielle Residuenplots für die Mitgliedschaft in Umweltgruppen



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

chen. Die Bezeichnung *partial residual* bzw. *component-plus-residual* kommt daher, dass den Residuen der durch die Variable  $X_j$  erklärte Teil der Differenz  $Y_i - \bar{Y}$  wieder beigefügt wird, also

$$q_j = \hat{\epsilon} + \hat{\beta}_j \mathbf{x}_j \quad (7)$$

Diese „partiellen“ Residuen werden dann gegen die Werte von  $X_j$  abgetragen. Auch hier entspricht die Steigung der Regressionsgeraden in der Grafik gerade wieder dem Koeffizienten  $\hat{\beta}_j$  im multiplen Modell. Zur Verdeutlichung der Zusammenhangsform wird häufig eine Lowess-Kurve (*locally weighted scatterplot smoother*, Cleveland 1979; zur Berechnung auch z. B. Schnell 1994: 112f. oder Fox 1991: 85ff.) eingetragen.<sup>3</sup>

Abbildung 4 zeigt die partiellen Residuenplots für die Daten von Dalton (2004). Eingezeichnet ist jeweils eine Lowess-Kurve unter Berücksichtigung aller Datenpunkte (gestrichelt) und eine Lowess-Kurve unter Ausschluss von Holland (gepunktet). Bezüglich des Pro-Kopf-Einkommens (Diagramm a) zeichnet sich ein eher u-förmiger Zusammenhang ab, wobei Lu-

<sup>3</sup>In einer Variation des partiellen Residuenplots, die nichtlineare Zusammenhänge u. U. noch stärker verdeutlicht, wird für die zu untersuchende Variable zusätzlich ein quadratischer Term eingesetzt (*augmented partial residual plot*, Mallows 1986). Es wird also ein Modell

$$Y_i = \sum_{\ell \neq j} \beta_\ell X_{i\ell} + \gamma_1 X_{ij} + \gamma_2 X_{ij}^2 + v_i$$

geschätzt und die (erweiterten) partiellen Residuen berechnen sich als

$$\hat{q}_{ij} = \hat{v}_i + \hat{\gamma}_1 X_{ij} + \hat{\gamma}_2 X_{ij}^2, \quad i = 1, \dots, n$$

Man beachte, dass anders als bei den einfachen partiellen Residuenplots für jede darzustellende Variable ein eigenes Modell geschätzt werden muss. Ferner spiegelt die Regressionsgerade im erweiterten partiellen Residuenplot nicht mehr den Koeffizienten  $\hat{\beta}_j$  wider.



xemburg (ganz rechts) nicht ins Bild passt; im Plot für das Demokratieniveau (Diagramm b) verläuft die Lowess-Kurve bis zu Skalenwert 70 praktisch horizontal und steigt danach stark an. Um diesen Ergebnissen gerecht zu werden, könnte das Modell erweitert werden mit einem quadratischen Term für das Pro-Kopf-Einkommen ( $BIP$ ) und einem Strukturbruch im Effekt des Demokratieniveaus ( $D$ ) bei Skalenwert 70, also

$$Y_i = \beta_0 + \beta_1 BIP_i + \beta_2 BIP_i^2 + \beta_3 D_i + \beta_4 D_i^{(70+)} + \epsilon_i \quad (8)$$

$$\text{mit } D_i^{(70+)} = \begin{cases} D_i - 70 & \text{falls } D_i > 70 \\ 0 & \text{sonst} \end{cases}$$

Tatsächlich verbessert sich die Anpassung des Modells dadurch spürbar. Modell 1a in Tabelle 1 zeigt das Ausgangsmodell mit den zwei linearen Termen. Der Anteil erklärter Varianz verdoppelt sich beinahe, wenn das Modell gemäss (8) erweitert wird (Modell 2a; die Modellverbesserung ist allerdings bei einem Signifikanzniveau von 5 % knapp nicht signifikant). Eine Betrachtung der Koeffizienten zeigt, dass zwar der Strukturbruch im Effekt des Demokratieniveaus gut abgebildet wird, sich der erwartete u-förmige Einfluss des Pro-Kopf-Einkommens jedoch nicht manifestiert. Dies ist offensichtlich auf den Ausreisser Luxemburg mit dem ungewöhnlich hohen Pro-Kopf-Einkommen zurückzuführen. Die Modelle 1b und 2b in Tabelle 1 zeigen zum Vergleich die Regressionsergebnisse, wenn Luxemburg und der  $Y$ -Ausreisser Holland von den Berechnungen ausgeschlossen werden. Während sich das lineare Ausgangsmodell nicht massgeblich verändert, ist die durch die Erweiterung erzeugte Verbesserung der Modellanpassung nun ziemlich dramatisch und hoch signifikant (Erhöhung von  $R^2$  um ca. 250 %). Abbildung 5 illustriert die Zusammenhänge in Modell 2b mit Hilfe von modifizierten partiellen Residuenplots.<sup>4</sup> Eingezeichnet sind auch die partiellen Prognosefehler für die beiden ausgeschlossenen Punkte (hervorgehoben).<sup>5</sup>

Aufgrund der guten Modellanpassung und der signifikanten Koeffizienten könnte man nun versucht sein, nur Modell 2b als Forschungsergebnis zu berichten. Das wäre hier aber natürlich höchst unzulässig. Erstens wurden die modellierten Zusammenhangsformen vollkommen ad hoc und unter Abwesenheit jeglicher inhaltlicher Überlegungen ermittelt. Es müssten zumindest plausible inhaltliche Erklärung mitgeliefert werden (Warum sollte der Einfluss des Volkseinkommens parabolisch sein? Wieso sollte das Demokratieniveau erst ab Wert 70 einen Einfluss haben?). Zweitens wurden die beiden Datenpunkte, die am schlechtesten zum Modell passen, willkürlich eliminiert. Da erstaunt es kaum, wenn sich die Modellanpassung verbessert. Drittens sind die Fallzahlen eher zu gering um solche statistischen Kürübungen zu vollführen – das Eis auf dem man sich bewegt ist dünn. Relativ geringe Änderungen der Daten (kleinere Messfehler, leicht unterschiedliche Operationalisierung der Variablen, etc.) oder zum Beispiel die Aufnahme einer zusätzlichen Variablen können die Ergebnisse unter Umständen radikal verändern. Trotz dieser Einwände ist es aber dennoch eindrücklich, wie sehr das Modell durch

<sup>4</sup>Das Verfahren zur Erstellung der Plots ist hier ähnlich wie bei den *erweiterten* partiellen Residuenplots (vgl. Fussnote 3), da jeweils zwei Parameter an einem Zusammenhang beteiligt sind. Das heisst, die partiellen Residuen werden berechnet als  $Q_{i,BIP} = \hat{\epsilon}_i + \hat{\beta}_1 BIP_i + \hat{\beta}_2 BIP_i^2$  und  $Q_{i,D} = \hat{\epsilon}_i + \hat{\beta}_3 D_i + \hat{\beta}_4 D_i^{(70+)}$  und die eingezeichneten Kurven entsprechen  $f(BIP) = \hat{\beta}_1 BIP + \hat{\beta}_2 BIP^2$  und  $f(D) = \hat{\beta}_3 D + \hat{\beta}_4 D^{(70+)}$ .

<sup>5</sup>Das Auftreten von Luxemburg in Diagramm 5b als Ausreisser (am unteren Rand) ist auf die schlechte Vorhersage durch den parabolischen Effekt des Pro-Kopf-Einkommens in Diagramm 5a zurückzuführen. Dies verdeutlicht, nebenbei bemerkt, die Abhängigkeit der partiellen Residuenplots untereinander. Wird bezüglich eines Regressors  $X_j$  eine Veränderung am Modell vorgenommen, ändern sich die partiellen Residuenplots auch für alle anderen Variablen  $X_{\ell \neq j}$ .

Tabelle 1: Verschiedene Regressionsschätzungen zum länderspezifischen Anteil Mitglieder in Umweltgruppen (Standardfehler in Klammern)

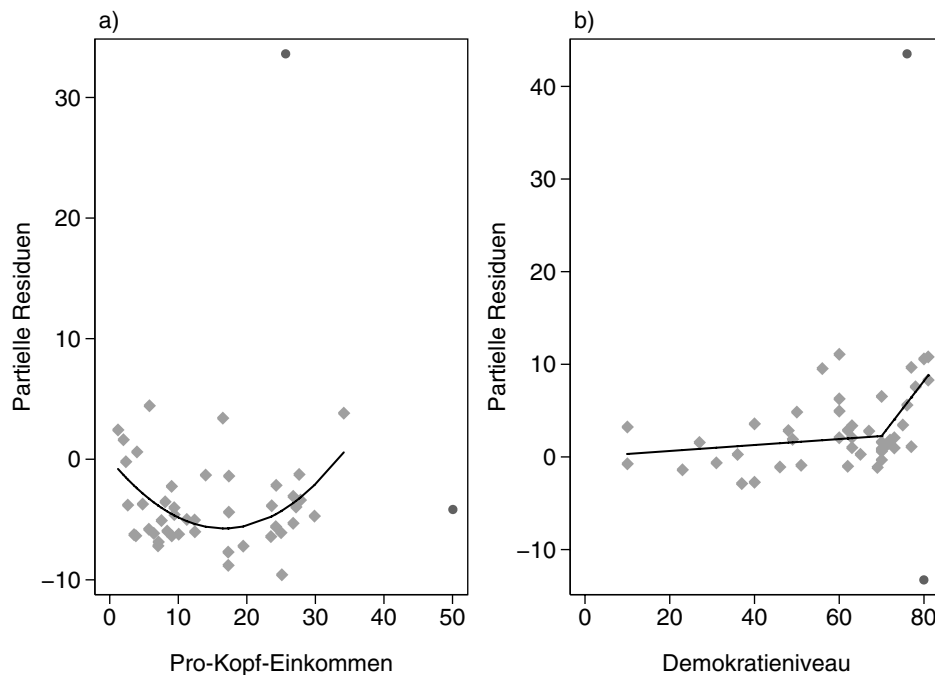
	Alle Fälle		Ohne NLD, LUX	
	Modell 1a	Modell 2a	Modell 1b	Modell 2b
Pro-Kopf-Einkommen	0.183 (0.135)	-0.014 (0.336)	0.128 (0.090)	-0.689** (0.253)
Pro-Kopf-Einkommen <sup>2</sup> /100		-0.011 (0.702)		2.067** (0.740)
Demokratieniveau	0.037 (0.078)	0.011 (0.082)	0.018 (0.045)	0.032 (0.039)
Demokratieniveau 70+		0.981* (0.422)		0.563* (0.210)
Konstante	0.596 (3.627)	3.277 (3.685)	1.765 (2.031)	5.661** (1.858)
$R^2$	0.120	0.226	0.129	0.450
$F$ -Test ( $p$ -Wert) <sup>a</sup>		0.073		0.000
$N$	46	46	44	44

Quelle: Daten aus Dalton (2004), eigene Berechnungen.

<sup>a</sup> Test für die Verbesserung der Erklärungskraft im Vergleich zu Modell 1

Signifikanzen: \*  $p < 0.5$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  (zweiseitig)

Abbildung 5: Partielle Residuenplots zu Modell 2b in Tabelle 1



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

Ausschluss von nur zwei Datenpunkten hier „optimiert“ werden konnte. Ferner wird verdeutlicht, wie stark ein einzelner Punkt (hier Luxemburg) auf die passende Zusammenhangsform Einfluss nehmen kann.

### 3 Formale Identifikation einflussreicher Datenpunkte

Ausreisser werden klassischerweise als Datenpunkte definiert, die stark abweichen vom Grossteil der anderen Punkte (vgl. z. B. Beckman und Cook 1983, Barnett und Lewis 1995). Nicht jeder Ausreisser ist aber unbedingt ein Problem bei der Schätzung eines Regressionsmodells. Es geht also nicht einfach darum zu ermitteln, ob gewisse Fälle beispielsweise unüblich grosse oder kleine  $Y$ -Werte aufweisen. Vielmehr ist das Ziel, diejenigen Punkte zu identifizieren, die einen besonderen Einfluss auf die Regressionsergebnisse ausüben. Belsley et al. (1980: 11) definieren solche Fälle wie folgt:

„An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of the various estimates (coefficients, standard errors,  $t$ -values, etc.) than is the case for most of the other observations.“

Ein offensichtlicher Ansatz zur Bestimmung des Einflusses von Beobachtungen, der diese Definition sehr direkt umsetzt, liegt in der Messung der Veränderung der Schätzergebnisse, wenn ein Punkt (oder eine Gruppe von Punkten) aus der Datenmatrix gelöscht wird. Der Ausschluss eines Datenpunktes kann sich dabei auf unterschiedliche Kennzahlen der Regressionsschätzung auswirken: „The question ‘Influence on what?’ is, therefore, an important one“ (Chatterjee und Hadi 1986: 380). Entsprechend findet sich in der Literatur eine ganze Reihe von Masszahlen, die den Einfluss auf unterschiedliche Zielgrössen operationalisieren oder unterschiedlich standardisiert sind. Die meisten dieser Masszahlen lassen sich jedoch als Funktion zweier zentraler Elemente verstehen: der Abweichung eines Punktes bezüglich des Vorhersagewertes der abhängigen Variablen sowie der relativen Lage des Punktes bezüglich der Werte der unabhängigen Variablen und somit der „Hebelwirkung“ auf die Regressionsfunktion. Für die Bestimmung des Einflusses eines Datenpunktes ist also wohl von Bedeutung, inwieweit es sich um einen Ausreisser bezüglich der  $Y$ -Werte handelt (*fit-outlier*), jedoch auch, ob der Datenpunkt extreme  $X$ -Werte aufweist (*factor-outlier*).

#### 3.1 Leverage

Die Leverage  $h_i$  („Hebelwirkung“) erfasst den potentiellen Einfluss eines Datenpunktes  $(Y_i, \mathbf{x}_i)$  auf die Vorhersagewerte  $\hat{Y}$ . Gegeben sei wiederum ein lineares Regressionsmodell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Durch Einsetzen des OLS-Parameterschätzer  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  erhält man die Vorhersagewerte für  $Y$ :

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Man erkennt, dass die Vorhersagewerte von  $Y$  als Funktion der beobachteten  $Y$ -Werte verstanden werden können, was durch die folgende Formulierung verdeutlicht wird:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad \text{mit} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (9)$$

Die  $n \times n$  Matrix  $\mathbf{H}$  wird als „hat matrix“ bezeichnet weil sie die  $Y$ -Werte in die  $\hat{Y}$ -Werte überführt (Hoaglin und Welsch 1978; Chatterjee und Hadi 1986 hingegen sprechen von einer „Prädiktionsmatrix“). Als Projektionsmatrix ist  $\mathbf{H}$  symmetrisch und idempotent, das heisst  $\mathbf{H}' = \mathbf{H}$  und  $\mathbf{H}^2 = \mathbf{H}$ , und es gilt

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 = \sum_{j=1}^n h_{ij}^2 \quad \text{mit } h_{ii} \in [0, 1]$$

Ein Diagonalelement  $h_i = h_{ii}$  lässt sich somit als ein zusammenfassendes Mass für den potentiellen Einfluss von  $(Y_i, \mathbf{x}_i)$  auf die Vorhersagewerte interpretieren und wird Leverage (Hebelwirkung) oder „Hut-Wert“ (*hat value*) genannt.

Wie man an (9) erkennt, ist  $h_i$  lediglich eine Funktion der beobachteten  $X$ -Werte und hat mit den Werten der abhängigen Variablen nichts zu tun. Tatsächlich kann  $h_i$  als relative (quadratische) Distanz zwischen  $\mathbf{x}_i$  und  $\bar{\mathbf{x}}$ , dem Vektor der Mittelwerte der Regressoren, verstanden werden, was durch die Berechnungsformel für  $h_i$  im Falle eines einfachen Regressionsmodells mit nur einem Regressor verdeutlicht wird:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

Für Modelle mit mehreren Regressoren lässt sich dies allgemein formulieren als

$$h_i = \frac{1}{n} + \frac{MD_i^2}{n-1}$$

mit  $MD_i$  als der Mahalanobis-Distanz der  $X_i$ -Werte vom Vektor der Mittelwerte, also

$$MD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' C(X)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

(vgl. z. B. Rousseeuw und van Zomeren 1990; Weisberg 1985: 112; Velleman und Welsch 1981;  $C(X)$  ist die Kovarianz-Matrix der  $X$ -Variablen). Extremwerte auf den unabhängigen Variablen führen somit zu hoher Leverage.

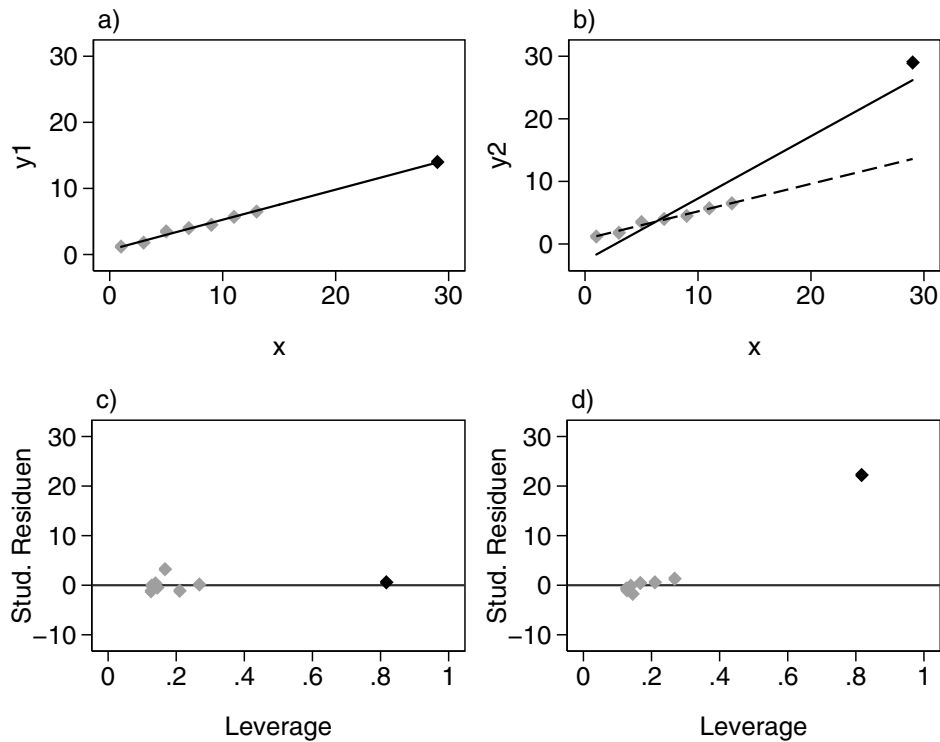
Zur Identifikation von Datenpunkten mit beachtenswerter Leverage wird i. d. R. der Schwellenwert  $h_i > 2k/n$  verwendet (die Schranke entspricht dem zweifachen Durchschnitt der  $h_i$ , da  $\sum_{i=1}^n h_i = k$ , vgl. Hoaglin und Welsch 1978; siehe auch Belsley et al. 1980: 17). Hohe Leverage wirkt sich indes aber nur dann auf die Schätzer für die Regressionsparameter aus, wenn zusätzlich  $Y_i$  selbst „aus der Reihe tanzt“, das heisst, bei der Leverage handelt es sich nur um ein Mass für den *potentiellen* Einfluss. Dies wird in Abbildung 6 veranschaulicht. Für den markierten Punkt in den Diagrammen (a) und (b) besteht jeweils eine sehr hohe Leverage ( $h_i = 0.82$ ), nur in Diagramm (b) hat der Punkt aber auch einen starken Einfluss auf die Lage der Regressionsgeraden. Wird der Punkt bei der Berechnung weggelassen, verläuft die Regressionsgerade viel flacher (gestrichelte Linie).

### 3.2 Residuen

Um Datenpunkte zu identifizieren, die einen starken Einfluss auf die Lage der Regressionsgeraden ausüben, muss also neben der Leverage auch noch berücksichtigt werden, inwieweit es sich bei den korrespondierenden  $Y$ -Werten um „Ausreisser“ handelt, wofür die Residuen

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = [\mathbf{I} - \mathbf{H}]\mathbf{y}$$

Abbildung 6: Leverage und Einfluss



herangezogen werden können. Man muss dabei aber bedenken, dass hohe Leverage zu tendenziell kleineren Residuen führt. Grosse Hebelwirkung hat ja gerade zur Folge, dass die Regressionsgerade näher an den entsprechenden Datenpunkt „herangezogen“ wird. Auch wenn für die Fehler  $\epsilon_i$  Homoskedastizität besteht, also  $Var(\epsilon_i) = \sigma^2$ , lässt sich zeigen, dass die Residuen  $\hat{\epsilon}_i = Y_i - \hat{Y}_i$  unterschiedliche Varianzen aufweisen, nämlich  $Var(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$  (vgl. z. B. Cook und Weisberg 1982a: 10ff.). Hohe Leverage führt also aufgrund der Hebelwirkung zu geringerer Residualvarianz und es liegt nahe, die Residuen mit Hilfe ihrer geschätzten Varianz zu standardisieren, also

$$\hat{\epsilon}'_i = \frac{\hat{\epsilon}_i}{\hat{\sigma} \sqrt{1 - h_i}} \quad (10)$$

Allerdings sind  $\hat{\sigma}$  und  $\hat{\epsilon}_i$  nicht unabhängig voneinander, so dass die so genannten studentisierten Residuen i. d. R. vorgezogen werden.<sup>6</sup> Diese sind definiert als

$$\hat{\epsilon}^*_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_i}} \quad (11)$$

<sup>6</sup>Das Attribut „studentisiert“ (*studentized*) geht zurück auf William Sealey Gosset (1867–1937), der unter dem Pseudonym „Student“ publizierte.

wobei  $\hat{\sigma}_{(-i)}$  den Schätzer für  $\sigma$  unter Ausschluss von Fall  $i$  bezeichnet.<sup>7</sup> Für  $\hat{\sigma}_{(-i)}$  kann die folgende Berechnungsformel hergeleitet werden

$$\hat{\sigma}_{(-i)}^2 = \frac{(n-k)\hat{\sigma}^2}{n-k-1} - \frac{\hat{\epsilon}_i^2}{(n-k-1)(1-h_i)} \quad (12)$$

Die studentisierten Residuen  $\hat{\epsilon}_i^*$  entsprechen somit lediglich einer monotonen Transformation der standardisierten Residuen  $\hat{\epsilon}_i'$  (Atkinson 1981). Sie haben jedoch den Vorteil, dass sie grosse Abweichungen stärker hervorheben und unter den Standardannahmen der linearen Regression einer  $t$ -Verteilung mit  $n-k-1$  Freiheitsgraden folgen. Als Schwelle zur Identifikation von beachtenswerten Residuen wird i. d. R.  $|\hat{\epsilon}_i^*| > 2$  empfohlen. Unter Normalbedingungen befinden sich ca. 5 % der Datenpunkte in diesem Bereich.

### 3.3 Einfluss auf die Koeffizienten und Vorhersagewerte

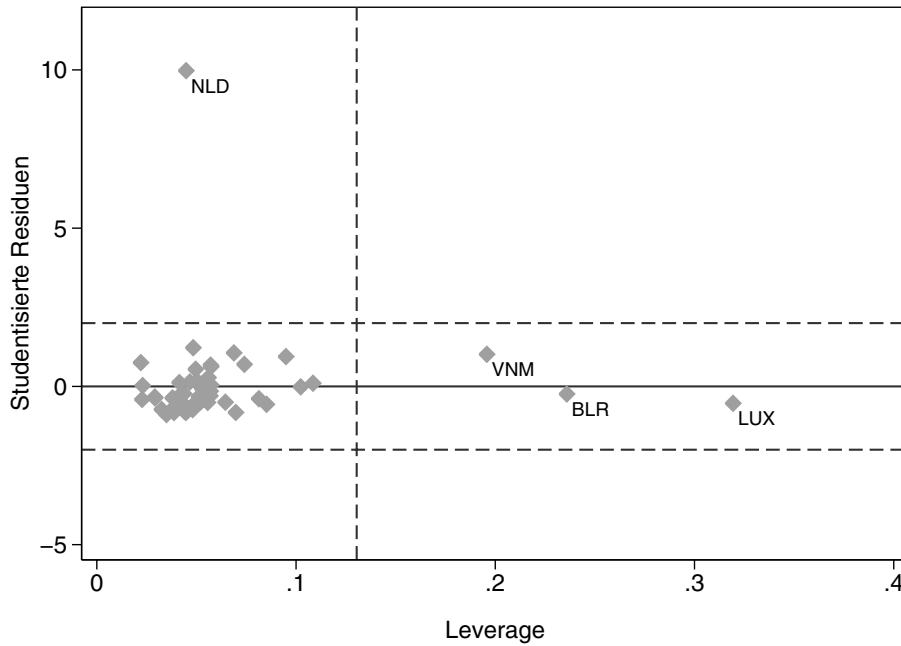
Erst durch die gemeinsame Betrachtung der Leverage  $h_i$  und der studentisierten Residuen  $\hat{\epsilon}_i^*$ , beispielsweise wenn die beiden Grössen in einem Streudiagramm gegeneinander abgetragen werden, können Datenpunkte identifiziert werden, die einen grossen Einfluss auf die Regressions-schätzung ausüben.<sup>8</sup> In Abbildung 6 ist dies exemplarisch veranschaulicht: nur der markierte Punkt in Diagramm (b) hat einen starken Einfluss auf die Lage der Regressionsgeraden, da er im Plot mit den Residuen und der Leverage auf beiden Achsen hohe Werte aufweist (Diagramm d). Am Vergleich der Diagramme (b) und (d) erkennt man zudem sehr schön die Beziehung zwischen den unstandardisierten und den studentisierten oder standardisierten Residuen: Der markierte Punkt liegt aufgrund der hohen Leverage kaum weiter weg von der Regressionsgerade als andere Punkte, das studentisierte Residuum, bei dem um den Einfluss der Leverage korrigiert wird, ist jedoch sehr gross.

Abbildung 7 zeigt ferner ein Streudiagramm der Leverage und studentisierten Residuen für die in Abbildung 2 vorgestellten Daten (es handelt sich um die  $h_i$ - und  $\hat{\epsilon}_i^*$ -Werte, die man bei man bei Regression der Mitgliedschaftsquoten in Umweltgruppen auf das Pro-Kopf-Einkommen und das Demokratieniveau erhält). Zusätzlich sind als gestrichelte Linien die oben angegebenen Schwellenwerte eingetragen ( $\hat{\epsilon}^* = \pm 2$  und  $h = 2k/n = 2 \cdot 3/46 = 0.13$ ). Man kann erkennen, dass der massive  $Y$ -Ausreisser Holland (NLD) für die Lage der Regressionsparameter gar nicht unbedingt ein so gravierendes Problem darstellt, da er eine relativ geringe Leverage hat. Weiterhin ist das im bivariaten Streudiagramm für den Regressor „Pro-Kopf-Einkommen“ identifizierte Luxemburg (LUX) zwar ein Punkt mit hoher Leverage, der tatsächliche Einfluss ist aber ebenfalls gering aufgrund des kleinen Residuums. Weitere, allerdings ebenfalls eher problemlose Extremwerte bezüglich der Leverage sind Vietnam (VNM) und Weissrussland (BLR).

<sup>7</sup>Bezüglich der Benennung von  $\hat{\epsilon}_i'$  und  $\hat{\epsilon}_i^*$  herrscht ein ziemliches Durcheinander. Während einige Autoren im ersten Fall von den „standardisierten“ Residuen sprechen, wenden Cook und andere auch hier den Begriff „Studentized residuals“ oder genauer „internally Studentized residuals“ an (z. B. Cook und Weisberg 1982a; Beckman und Cook 1983; Weisberg 1985). Für  $\hat{\epsilon}_i^*$  findet man neben dem Begriff „studentized residual“ (Belsley et al. 1980; Velleman und Welsch 1981; Fox 1991) auch „externally Studentized residual“ (Cook und Weisberg 1982a), „cross-validatory residual“ (Atkinson 1981) oder „jackknifed residual“ (Rousseeuw und Leroy 1987). In SPSS wird  $\hat{\epsilon}_i'$  als „Studentized residual“ (SRESID) und  $\hat{\epsilon}_i^*$  als „Studentized deleted residual“ (SDRESID) bezeichnet. Der Begriff „Standardized residual“ (ZRESID) bezieht in sich SPSS sich auf den Ausdruck  $\hat{\epsilon}_i/\hat{\sigma}$ .

<sup>8</sup>Eine andere Darstellungsform mit ähnlichem Informationsgehalt ist der L-R-Plot (*leverage residual plot*), bei dem  $h_i$  gegen  $\hat{\epsilon}_i^2/(n-k)\hat{\sigma}^2$  abgetragen wird (Gray 1986, 1989a).

Abbildung 7: Studentisierte Residuen und Leverage für das Modell zur Erklärung länderspezifischer Mitgliederrate in Umweltgruppen



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

### 3.3.1 Gesamteinfluss

Man kann nun versuchen, das Zusammenspiel von Leverage und studentisierten Residuen mit Hilfe einer einzigen Masszahl für den Gesamteinfluss eines Datenpunktes auf die Regressions-schätzung auszudrücken. Die am häufigsten verwendeten derartigen Statistiken sind Cooks  $D$  (Cook 1977, 1979; Cook und Weisberg 1980, 1982a) und die Masszahl  $DFFITs$  (Belsley et al. 1980). Cooks  $D$  wurde entwickelt als Mass für die Veränderung im geschätzten Parametervektor, wenn eine Beobachtung  $i$  aus dem Datensatz eliminiert wird, und ist definiert als

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' \mathbf{X}' \mathbf{X} (\hat{\beta}_{(-i)} - \hat{\beta})}{k \hat{\sigma}^2}, \quad i = 1, \dots, n$$

wobei  $\hat{\beta}_{(-i)}$  den Parameterschätzer unter Ausschluss des Datenpunktes  $i$  symbolisiert. Alternativ lässt sich  $D_i$  auch als Veränderung in den Vorhersagewerten interpretieren, wenn Punkt  $i$  bei der Schätzung weggelassen wird, also

$$D_i = \frac{(\hat{\mathbf{y}}_{(-i)} - \hat{\mathbf{y}})' (\hat{\mathbf{y}}_{(-i)} - \hat{\mathbf{y}})}{k \hat{\sigma}^2}$$

mit  $\hat{\mathbf{y}}_{(-i)}$  als Vektor der Vorhersagewerte bei Verwendung der Schätzers  $\hat{\beta}_{(-i)}$ , also  $\hat{\mathbf{y}}_{(-i)} = \mathbf{X} \hat{\beta}_{(-i)}$  (Cook und Weisberg 1982a; Chatterjee und Hadi 1986). Die Masszahl kann aber auch einfach als Funktion der Leverage  $h_i$  und des standardisierten Residuums  $\hat{\epsilon}_i'$  formuliert werden

$$D_i = \frac{\hat{\epsilon}_i'^2}{k} \cdot \frac{h_i}{1 - h_i} = \frac{1}{k} \cdot \frac{\hat{\epsilon}_i'^2}{\hat{\sigma}^2(1 - h_i)} \cdot \frac{h_i}{1 - h_i} \quad (13)$$

(vgl. Cook und Weisberg 1982a; Hocking 1983; Bollen und Jackman 1990; Fox 1991).

Die Masszahl *DFFITs* hat einen leicht anderen Fokus. Sie misst die (standardisierte) Veränderung des Vorhersagewertes  $Y_i$ , wenn Punkt  $i$  für die Schätzung des Parametervektors weggelassen wird, und ist definiert als

$$DFFITs_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{\hat{\sigma}_{(-i)} \sqrt{h_i}} = \frac{\mathbf{x}_i(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{\hat{\sigma}_{(-i)} \sqrt{h_i}}, \quad i = 1, \dots, n$$

(Belsley et al. 1980). Auch diese Statistik lässt sich mit Hilfe der Leverage und der Residuen ausdrücken, nämlich

$$DFFITs_i = \hat{\epsilon}_i^* \sqrt{\frac{h_i}{1-h_i}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)} \sqrt{1-h_i}} \sqrt{\frac{h_i}{1-h_i}} \quad (14)$$

(Belsley et al. 1980; Hocking 1983; Bollen und Jackman 1990; Fox 1991). Aus (13) und (14) folgt zudem

$$D_i = \frac{1}{k} \left( DFFITs_i \frac{\hat{\sigma}_{(-i)}}{\hat{\sigma}} \right)^2$$

Die beiden Masse sind also trotz der unterschiedlichen Ausgangslage sehr ähnlich. Es sollte folglich nicht eine bedeutende Rolle spielen, welche der Statistiken man verwendet. Der einzige substantielle Unterschied ist, dass bei *DFFITs* das studentisierte Residuum verwendet wird, bei Cooks *D* jedoch das standardisierte, was *DFFITs* etwas sinnvoller erscheinen lässt (vgl. auch Bollen und Jackman 1990: 266). Als kritische Bereiche werden i. d. R.  $|DFFITs| > 2 \sqrt{k/n}$  und folglich  $D > 4/n$  vorgeschlagen.

Tabelle 2 zeigt die Einflussstatistiken für einige Fälle aus dem Beispiel von Dalton (2004). Von den im Plot mit den studentisierten Residuen und der Leverage identifizierten Punkten (Abbildung 7) übersteigt nur Holland die kritischen Werte für Cooks *D* und *DFFITs* aufgrund des enormen Residuums. Die Leverage-Extremwerte Luxemburg, Vietnam und Weissrussland jedoch liegen relativ gut im linearen Trend der übrigen Daten, haben also kleine Residuen, und werden durch die beiden Masse nicht als einflussreiche Datenpunkte identifiziert (wobei allerdings Vietnam den kritischen Wert fast erreicht).

### 3.3.2 Einfluss auf individuelle Koeffizienten

Die Masse Cooks *D* und *DFFITs* drücken den Einfluss eines Datenpunktes auf den Schätzer des *gesamten* Parametervektors aus. Häufig möchte man jedoch wissen, wie stark sich die individuellen Koeffizienten in Abhängigkeit von einzelnen Datenpunkten verändern. Zur Bestimmungen des Einflusses eines Datenpunktes  $i$  auf den Schätzer für einen Koeffizienten  $\beta_j$  ist die Differenz  $\hat{\beta}_j - \hat{\beta}_{j(-i)}$  massgebend. Belsley et al. (1980) schlagen die folgende Statistik vor:

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{SE_{(-i)}(\hat{\beta}_j)} \quad (15)$$

wobei mit  $SE_{(-i)}(\hat{\beta}_j)$  der Standardfehler von  $\hat{\beta}_j$  unter Ausschluss von Punkt  $i$  gemeint ist (zu den genauen Berechnungsformeln siehe Belsley et al. 1980 oder Chatterjee und Hadi 1986). Die Ergebnisse, die mit der *DFBETAS*-Statistik gewonnen werden, sind relativ unübersichtlich, da für jeden Datenpunkt  $k$  Werte zu berechnen sind. Zur Betrachtung der Werte können deshalb Streudiagramme sehr nützlich sein (vgl. Abschnitt 4). Als Kriterium für beachtenswerte Punkte



Tabelle 2: Ausreisser im Modell zur länderspezifischen Mitgliederrate in Umweltgruppen (Modell 1a in Tabelle 1)

Land	Stud. Residuen	Leverage	Cooks $D$	$DFFITs$	$COVR.$	$DFBETAS$	
						$BIP$	$D$
ZWE	0.101	0.109	0.000	0.035	1.203	0.004	-0.025
BLR	-0.239	0.236	0.006	-0.133	1.399	-0.065	0.123
LUX	-0.534	0.319	0.045	-0.366	1.545	-0.336	0.159
VNM	1.016	0.196	0.084	0.501	1.241	0.147	-0.423
NLD	9.977	0.045	0.473	2.161	0.029	0.725	0.466
Schwelle	$\pm 2.000$	0.130	0.087	$\pm 0.511$	$1 \pm 0.196$	$\pm 0.295$	$\pm 0.295$

Quelle: Daten aus Dalton (2004), eigene Berechnungen.

Anmerkungen: In der Tabelle sind alle Punkte enthalten, die für mindestens eine der aufgeführten Statistiken im kritischen Bereich liegen. Die nach den im Text angegebenen Regeln berechneten kritischen Werte sind im Fuss der Tabelle aufgeführt. BIP: Pro-Kopf-Einkommen; D: Demokratieniveau.

geben Belsley et al. (1980) den Bereich  $|DFBETAS| > 2/\sqrt{n}$  an. Bollen und Jackman (1990) schlagen alternativ den Bereich  $|DFBETAS| > 1$  vor: so werden Punkte identifiziert, die den Parameterschätzer um mehr als einen Standardfehler verändern.

Wie oben angesprochen, werden Vietnam, Luxemburg und Weissrussland im Beispiel von Dalton trotz ihrer grossen Leverage-Werte durch die Statistiken für den Gesamteinfluss nicht als beachtenswerte Punkte identifiziert. Die Betrachtung der  $DFBETAS$ -Werte gibt hier ein differenzierteres Bild (Tabelle 2). Zwar erscheint Weissrussland auch bezüglich der einzelnen Koeffizienten als nicht besonders bedeutend, Luxemburg und Vietnam jedoch nehmen je auf einen Koeffizienten spürbar verminderten Einfluss. Wird Luxemburg weggelassen, erhöht sich der Effekt des Pro-Kopf-Einkommens; bei Ausschluss von Vietnam steigt der Einfluss des Demokratieniveaus. Den stärksten Einfluss auf beide Koeffizienten hat jedoch, kaum überraschend, der Ausreisser Holland.

### 3.4 Einfluss auf die Präzision der Schätzung

Die bisher besprochenen Masse beziehen sich auf den Einfluss auf den Koeffizientenvektor der Regressionsschätzung bzw. die Vorhersagewerte. Einzelne Datenpunkte können aber auch einen bedeutenden Einfluss auf die Grösse der Standardfehler der Parameterschätzer, das heisst auf die Präzision der Modellschätzung haben. Eine Masszahl, die die relative Veränderung des (quadrierten) Volumens des gemeinsamen Konfidenzraumes aller Parameter bei Ausschluss der Beobachtung  $i$  approximiert und somit als Mass für den Einfluss auf die Schätzgenauigkeit dienen kann, wurde von Belsley et al. (1980) definiert als

$$COVRATIO_i = \frac{\det \left[ \hat{\sigma}_{(-i)}^2 (\mathbf{X}'_{(-i)} \mathbf{X}_{(-i)})^{-1} \right]}{\det \left[ \hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1} \right]}$$

das heisst, als der Quotient aus der Determinanten der Kovarianz-Matrix des Modells ohne Punkt  $i$  und des vollen Modells. Mit Hilfe von (11) und (12) lässt sich auch hier zeigen, dass die Masszahl als eine Funktion der Leverage und Residuen dargestellt werden kann, nämlich

$$COVRATIO_i = \left( \frac{\hat{\sigma}_{(-i)}^2}{\hat{\sigma}^2} \right)^k \cdot \frac{1}{1 - h_i} = 1 \left/ \left( \frac{n - k - 1 + \hat{\epsilon}_i^{*2}}{n - k} \right)^k (1 - h_i) \right. \quad (16)$$

Wert 1 für  $COVRATIO_i$  bedeutet, dass kein Einfluss auf die Standardfehler besteht. Werte grösser 1 weisen darauf hin, dass sich der Datenpunkt positiv auf die Schätzgenauigkeit auswirkt (Verkleinerung des Konfidenzraumes im Vergleich mit dem Modell ohne den Datenpunkt). Dies ist zum Beispiel für Punkte mit hoher Leverage  $h_i$  und kleinem studentisiertem Residuum  $\hat{\epsilon}_i^*$  der Fall. Werte kleiner 1 bedeuten, dass sich der Konfidenzraum durch Hinzunahme des Datenpunktes vergrössert, das heisst, die Schätzung wird ungenauer (z. B. bei kleinem  $h_i$  und grossem  $\hat{\epsilon}_i^*$ ). Gemäss Belsley et al. (1980) sind Punkte mit

$$|COVRATIO_i - 1| > 3k/n$$

beachtenswert.

Tabelle 2 zeigt die fünf Fälle im Beispiel von Dalton, die diese Schwelle überschreiten. Es handelt sich einerseits um die vier bereits bezüglich der anderen Masszahlen besprochenen Punkte: Das grosse Residuum von Holland führt zu einer massiven Verschlechterung der Präzision der Modellschätzung (durch Ausschluss von Holland fällt der Standardfehler des Modells, also  $\hat{\sigma}$ , um fast 50 % von 6.9 auf 3.8; entsprechend halbieren sich auch die Standardfehler der einzelnen Koeffizienten beinahe). Für Luxemburg, Weissrussland und Vietnam ist die Situation gerade entgegengesetzt. Es handelt sich um Punkte mit eher kleinen Residuen dafür aber hoher Leverage. Sie wirken sich somit positiv auf die Schätzgenauigkeit aus. Der fünfte identifizierte Punkt ist Zimbabwe, ein Punkt, der praktisch auf der Regressionsfläche liegt. Das geringe Residuum in Kombination mit einer relativ hohen Leverage (Zimbabwe hat nach Luxemburg, Weissrussland und Vietnam die vierthöchste Leverage) führt hier ebenfalls zu dem günstigen Einfluss auf die Präzision der Regressionsschätzung.

### 3.5 Übersicht und weitere Ansätze

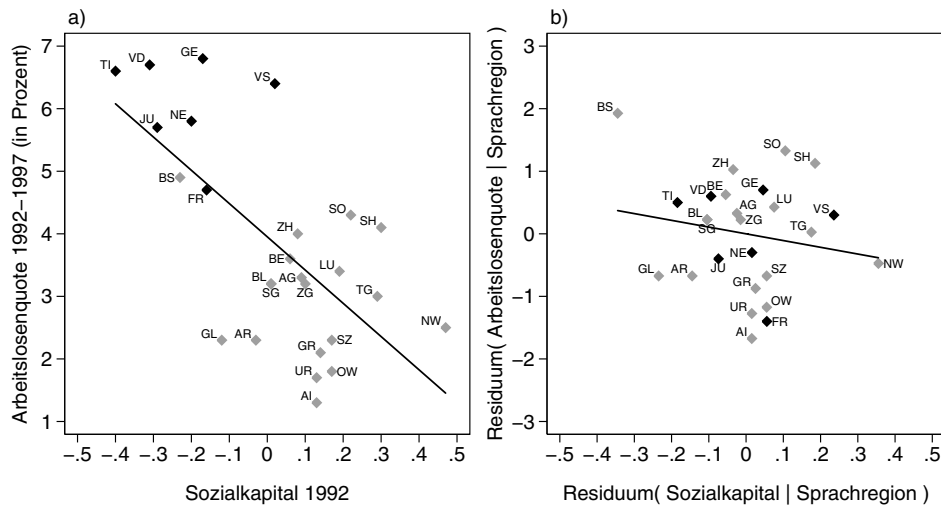
Tabelle 3 fasst die besprochenen Masszahlen nochmals zusammen und gibt zudem Hinweise zur Verfügbarkeit der Masszahlen in den häufig verwendeten Statistikprogrammen SPSS und Stata. In der Literatur werden noch weitere Einflussmasse und Varianten vorgeschlagen (z. B. Andrews und Pregibon 1978, Draper und John 1981 und Welsch 1982; vgl. auch etwa die Übersichten in Hocking 1983 und Chatterjee und Hadi 1986), eine Analyse der hier besprochenen Grössen in Verbindung mit den grafischen Methoden sollte aber normalerweise ausreichen. Wenn es jedoch darum geht, die optimale Zusammenhangsform mit Hilfe von Transformationen der Variablen zu finden, können zusätzliche Statistiken von Nutzen sein, die den Einfluss einzelner Datenpunkte auf die Transformation quantifizieren (vgl. z. B. Atkinson 1982, 1983, 1986, Cook und Wang 1983, Carroll und Ruppert 1987 und Hinkley und Wang 1988). Ferner beschränken sich die hier besprochenen Methoden auf die lineare Regression. Für entsprechende Instrumente im Rahmen der logistischen Regression siehe z. B. die Einführungen in Hosmer und Lemeshow (2000) und Menard (2002) sowie die Grundlagentexte von Pregibon (1981), Landwehr et al. (1984), Jennings (1986) und Fowlkes (1987) sowie Lesaffre und Albert (1989) für multinomiale Modelle. Zur Regressionsdiagnostik in der Ereignisanalyse siehe z. B. Hosmer

Tabelle 3: Übersicht über die Masszahlen zur Identifikation einflussreicher Daten

Mass	Formel	kritischer Bereich	Verfügbarkeit <sup>a</sup>
Potentieller Einfluss (Leverage)	$h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$	$h_i > 2k/n$	SPSS: LEVER (Leverage values) Stata: leverage, hat (Leverage)
Standardisierte Residuen	$\hat{\epsilon}_i' = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}$	$ \hat{\epsilon}_i'  > 2$	SPSS: SRESID (Studentized residuals) Stata: rstandard (Standardized residuals)
Studentisierte Residuen	$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$	$ \hat{\epsilon}_i^*  > 2$	SPSS: SDRESID (Studentized deleted residuals) Stata: rstudent (Studentized residuals)
Gesamteinfluss auf Koeffizienten	$D_i = \frac{\hat{\epsilon}_i^2}{k} \cdot \frac{h_i}{1-h_i}$	$D_i > 4/n$	SPSS: COOK (Cook's Distance) Stata: cooks (Cook's Distance)
	$DFFFITS_i = \hat{\epsilon}_i^* \sqrt{\frac{h_i}{1-h_i}}$	$ DFFFITS_i  > 2\sqrt{k/n}$	SPSS: SDFIT (Standardized DFFit) Stata: dffits (DFITS)
Einfl. auf einzelne Koeffizienten	$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_j^{(-i)}}{SE_{(-i)}(\hat{\beta}_j)}$	$ DFBETAS_{ij}  > 2/\sqrt{n}$	SPSS: SDBETA (Standardized DfBeta(s)) Stata: dfbeta () (DFBETA)
Einfluss auf die Präzision	$COVRATIO_i = \left(\frac{\hat{\sigma}_{(-i)}^2}{\hat{\sigma}^2}\right)^k \cdot \frac{1}{1-h_i}$	$ COVRATIO_i - 1  > 3k/n$	SPSS: COVRATIO (Covariance Ratio) Stata: covratio (COVRATIO)

<sup>a</sup> Angegeben ist die jeweilige Bezeichnung in der Befehlssyntax sowie in Klammern die Beschriftung in den Dialogfeldern. Die Masszahlen können in SPSS mit Hilfe der SAVE-Option im REGRESSION-Befehl gespeichert werden. In Stata werden die Masszahlen mit dem predict-Befehl berechnet.

Abbildung 8: Streudiagramm zwischen Arbeitslosenquote und Sozialkapital (a) und partieller Regressionsplot unter Kontrolle der Sprachregion (b)



Quelle: Daten aus Freitag (2000), eigene Berechnungen.

und Lemeshow (1999) sowie Cain und Lange (1984) und Pettitt und Bin Daud (1989). Allgemeine Ansätze im Rahmen der generalisierten linearen Modelle (GLM; McCullagh und Nelder 1989) liefern Polasek (1984), Wang (1985), Vos (1991) und O'Hara Hines und Carter (1993).

Ein weiterer Aspekt, der bisher ausgeklammert wurde, betrifft das Problem, dass sich einflussreiche Datenpunkte gegenseitig „verdecken“ können. Die besprochenen Masse beziehen sich alle auf den Effekt auf die Regressionsergebnisse, wenn ein *einzelner* Datenpunkt eliminiert wird. Es kann aber auch sein, dass multiple Ausreisser in Gruppen auftreten und durch die klassischen Diagnosemasse aufgrund des so genannten Maskierungseffektes nicht erkannt werden (die Eliminierung eines einzelnen Punktes der Gruppe ändert nicht viel an den Regressionsergebnissen, weil die restlichen Punkte ihren Einfluss weiterhin ausüben). Die besprochenen Masse wie z. B. Cooks  $D$  lassen sich zwar relativ problemlos generalisieren für die Betrachtung des simultanen Einflusses mehrerer Beobachtungen (vgl. z. B. Cook und Weisberg 1982b: 347ff.; Chatterjee und Hadi 1988: 185ff.; Belsley et al. 1980: 31ff.). Dies ist allerdings meistens relativ nutzlos für die Identifikation von einflussreichen Gruppen, weil es sehr viele Möglichkeiten zur Gruppierung der Daten gibt. Verschiedene Ansätze mit dem Problem umzugehen wurden vorgeschlagen (vgl. etwa Gray und Ling 1984, Rousseeuw und van Zomeren 1990, Ritschard und Antille 1992, Crawford et al. 1995, Pena und Yohai 1995, Peña und Prieto 2001, Hennig 2003, Marchette und Solka 2003), können aber hier nicht erörtert werden. Glücklicherweise erkennt man multiple Ausreisser meistens auch relativ gut anhand von Scatterplots und partiellen Regressionsplots.

Ein extremes Beispiel für multiple Ausreisser findet sich in einer Arbeit von Freitag (2000). Freitag untersuchte den Zusammenhang zwischen Sozialkapital und Arbeitslosenquoten in Schweizer Kantonen. Die empirischen Ergebnisse deuten auf einen – gemäss Theorie erwarteten – starken negativen Zusammenhang: Je höher das Sozialkapital in einem Kanton, desto geringer die Arbeitslosenquote (Abbildung 8a).

Bei genauerer Betrachtung des Streudiagramms erkennt man jedoch, dass der Zusammenhang hauptsächlich durch den Unterschied zwischen zwei Clustern von Datenpunkten zustande kommt. Die beiden Cluster repräsentieren nun gerade die Kantone der (überwiegend) „lateini-

Tabelle 4: Regressionsmodelle zum Einfluss des Sozialkapitals auf die Arbeitslosenquote

	Modell 1		Modell 2		Modell 3 (ohne BS)	
	$\hat{\beta}$	<i>t</i> -Wert	$\hat{\beta}$	<i>t</i> -Wert	$\hat{\beta}$	<i>t</i> -Wert
Sozialkapital 1992	-5.32	-4.39	-1.08	-0.86	0.27	0.20
Lateinische Schweiz (0/1)			2.77	4.78	3.33	5.54
Konstante	3.95	15.70	3.10	12.13	2.83	10.52
$R^2$	0.45		0.72		0.76	
<i>N</i>	26		26		25	

Quelle: Daten aus Freitag (2000), eigene Berechnungen.

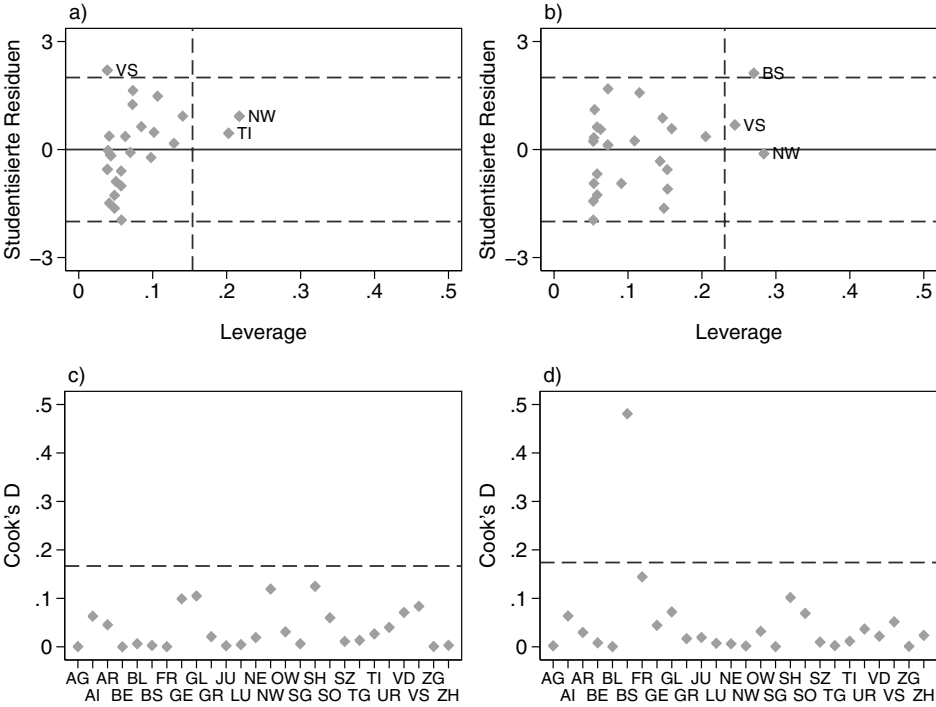
schen“ Schweiz (FR, GE, JU, NE, TI, VD, VS) und der Deutschschweiz (einzige Ausnahme ist Basel-Stadt). Die Annahme liegt also nahe, dass ein klassisches Drittvariablenproblem vorliegt und nicht das Sozialkapital eine Auswirkung auf die Arbeitslosenquote hat, sondern beide Variablen durch gemeinsame, mit der Sprachregion zusammenhängende Faktoren bedingt sind. Wird das Modell durch Aufnahme eines Indikators für die Sprachregion erweitert, verschwindet denn auch der Zusammenhang zwischen Sozialkapital und Arbeitslosenquote fast vollständig, wie man am partiellen Regressionsplot unter Kontrolle der Sprachregion (Abbildung 8b) und den Regressionsmodellen in Tabelle 4 erkennen kann (bzw. in beiden Sprachregionen ist der Effekt des Sozialkapitals praktisch null). Im Modell mit der Kontrollvariable „Sprachregion“ wird Basel-Stadt (BS) als einflussreicher Datenpunkt identifiziert (vgl. Abbildung 9) und bei Ausschluss von Basel-Stadt wird der Effekt sogar positiv (Modell 3 in Tabelle 4). Das heisst aber natürlich noch lange nicht, dass das Modell ohne Basel-Stadt zutreffender ist. Zur Verteidigung von Freitag (2000) muss zudem gesagt werden, dass mit der Sprach-Dummy eigentlich nicht viel erklärt wird. Es ist lediglich ein Hinweis darauf, dass weitere, mit dem sprachkulturellen Hintergrund zusammenhängende Faktoren von Bedeutung sind. Welche Faktoren das sind, kann mit den vorliegenden Daten nicht geklärt werden.

#### 4 Reputationseffekte in Internet-Auktionen: Ein Anwendungsbeispiel

Im Folgenden sollen die vorgestellten Methoden der Regressionsdiagnostik an den Daten einer Studie von Diekmann und Wyder (2002) nochmals veranschaulicht werden. Diekmann und Wyder untersuchten die Auswirkungen von Reputation in Transaktionen zwischen anonymen Akteuren anhand von Internet-Auktionen. Die Daten stammen aus der Beobachtung der Auktionen von neuen Nokia-Mobiltelefonen (Nokia 8310) bei Ricardo.ch im Zeitraum Oktober 2001 bis Januar 2002. Unter anderem interessierten sich Diekmann und Wyder für den Effekt der Reputation auf den realisierten Verkaufspreis und schätzten entsprechende Regressionsmodelle. Die Hypothese war, dass eine gute Reputation den Verkaufspreis im Sinne einer Risikoprämie erhöht. Die Hypothese konnte in der Studie weitgehend bestätigt werden.

Für die folgende Reanalyse des Reputationseffektes wurden die Daten zuerst einer Bereinigung anhand des Rohdatenmaterials (gespeicherte Screenshots der Auktionen) unterzogen. Neben der Korrektur einiger kleinerer Übertragungsfehler musste eine Reihe von Fällen ausgeschlossen werden, weil es sich (a) um doppelt erfasste Auktionen handelte, (b) die Rohdaten unvollständig oder (c) die Auktionen nicht vergleichbar waren („neuwertige“ anstatt „neue“

Abbildung 9: Diagnostische Plots für den Zusammenhang zwischen Arbeitslosenquote und Sozialkapital (Diagramme a und c beziehen sich auf das bivariate Modell; Diagramme b und d auf ein Modell mit der Kontrollvariable Sprachregion)



Quelle: Daten aus Freitag (2000), eigene Berechnungen.

Produkte, Mehrfachangebote etc.). Zur Analyse der Verkaufspreise verbleiben nach der Bereinigung 84 Fälle.

Diekmann und Wyder verwendeten als Regressoren zur Erklärung des Verkaufspreises (in Franken) die Reputation (operationalisiert als die Anzahl Bewertungen oder die logarithmierte Anzahl Bewertungen; es wurden zudem nur Fälle mit durchschnittlich positiver Reputation berücksichtigt), den Mindestpreis (das vom Anbieter festgelegte Startgebot), die veranschlagten Versandkosten, die Dauer der Auktion, die vom Verkäufer festgelegte Mindesthöhung (Mindestschrittweite der Gebote) und die Anzahl der Gebote. Für die Reputation und die Mindesthöhung wurde je ein signifikanter positiver Effekt festgestellt, für die Versandkosten ein stark negativer Effekt. Der Mindestpreis und die Anzahl Gebote hatten keinen Effekt.

Aus Gründen der Übersichtlichkeit werde ich mich hier mit einem vereinfachten Modell ohne die Variablen „Dauer der Auktion“ und „Anzahl der Gebote“ befassen und als abhängige Variable den Bruttopreis (Verkaufspreis plus Versandkosten) verwenden. Als zusätzliche Variable wird jedoch das Startdatum der Auktion (Kalenderzeit) aufgenommen, da es sich um einen sehr wichtigen Prädiktor für den Verkaufspreis handelt. Die Resultate der Modellschätzung sind in Tabelle 5 (Spalte 1) abgebildet (die Bootstrap-Standardfehler in eckigen Klammern sollen vorerst ausser Acht gelassen werden). Wie bei Diekmann und Wyder (2002) zeigt sich auch in dem modifizierten Modell ein signifikanter Reputationseffekt (mit einem  $t$ -Wert von 2.83) und ein überproportionaler Effekt der Mindesthöhung (mehr als 3 Franken Zusatzerlös bei einer Erhöhung der Schrittweite der Gebote um einen Franken;  $t = 5.90$ ).<sup>9</sup> Im Vergleich mit den Resultaten von Diekmann und Wyder (2002: Tabelle 3) fällt die deutlich verbesserte Erklärungskraft des Modells auf, was zum Teil mit der oben angesprochenen Bereinigung des Datenmaterials zusammenhängt, in der Hauptsache jedoch auf den starken Effekt der Kalenderzeit zurückzuführen ist.

Es stellt sich nun die Frage, ob die Resultate einem genaueren diagnostischen Check standhalten. Bei der Betrachtung der bivariaten Scatterplots zwischen dem Verkaufspreis und den Regressoren (Abbildung 10) sticht ein Ausreisser mit einem sehr hohen Verkaufspreis von rund 800 Franken sofort ins Auge (Nr. 71).<sup>10</sup> In allen vier Diagrammen liegt der Punkt weit weg vom Rest der Daten. Bezüglich der Variable „Mindesthöhung“ scheint der Ausreisser besonders problematisch zu sein, weil er auch in der Richtung der  $X$ -Achse deutlich von der Punktwolke abweicht und somit sehr starken Einfluss auf die Steigung der Regressionsgeraden nehmen kann. Im Diagramm, das sich auf den Zusammenhang zwischen dem Verkaufspreis und der Reputation bezieht, fällt weiterhin Punkt 73 als extremer  $X$ -Wert mit einer relativ grossen Abweichung von der Regressionsgerade auf. Zudem gruppiert sich der Grossteil der Punkte in einem Bereich von 0 bis ca. 15 Bewertungen (ca. 85 % der Punkte liegen in diesem Bereich; in 32 % der Fälle ist die Anzahl Bewertungen gleich null) und es besteht die Gefahr, dass der Zusammenhang von den wenigen, lose verteilten Punkten im Bereich von 20 bis 100 Bewertungen dominiert wird.

Das Ergebnis von Punkt 71 als starkem Ausreisser wird durch die partiellen Regressionsplots (Abbildung 11), die im Falle einer multiplen Regression ein verlässlicheres Bild ergeben, bestätigt. Im Plot für die Mindesthöhung sticht der Punkt nun sogar noch stärker hervor. Eben-

---

<sup>9</sup>Bei der Datenbereinigung wurde festgestellt, dass einige Anbieter mit mehreren Auktionen in der Stichprobe vertreten sind. Dies hat zur Folge, dass die Beobachtungen genau genommen nicht unabhängig sind, was bei der Berechnung der Standardfehler und somit der  $t$ -Werte zu berücksichtigen wäre. Ich verzichte jedoch hier auf diese Korrektur, weil sie sich kaum auf die Resultate auswirkt.

<sup>10</sup>Bei den markierten Punkten Nr. 9, 18, 32, 71 und 73 handelt es sich diejenigen Fälle, die nach Cooks  $D$  und  $DFITS$  den grössten Gesamteinfluss auf die Regressionschätzung nehmen (mehr dazu weiter unten).

Tabelle 5: Regressionsmodelle zur Erklärung des Verkaufspreises bei Internet-Auktionen

	OLS1	OLS2	OLS3	RREG
Reputation	0.615 (0.217) [0.296]	0.611 (0.188) [0.270]	0.836 (0.215) [0.260]	0.892 (0.180) [0.357]
Startpreis	-0.019 (0.024) [0.027]	0.000 (0.021) [0.022]	-0.001 (0.020) [0.023]	0.009 (0.020) [0.023]
Mindestserhöhung	3.401 (0.577) [1.339]	1.492 (0.621) [0.669]	2.282 (0.579) [0.580]	1.997 (0.593) [0.761]
Kalenderzeit	-0.885 (0.152) [0.155]	-0.941 (0.133) [0.136]	-0.947 (0.118) [0.119]	-0.991 (0.127) [0.151]
Konstante	518.609 (9.558) [10.280]	524.102 (8.356) [9.710]	516.510 (7.501) [8.360]	512.363 (7.988) [9.874]
$R^2$	0.510	0.526	0.622	
$N$	84	83	79	83

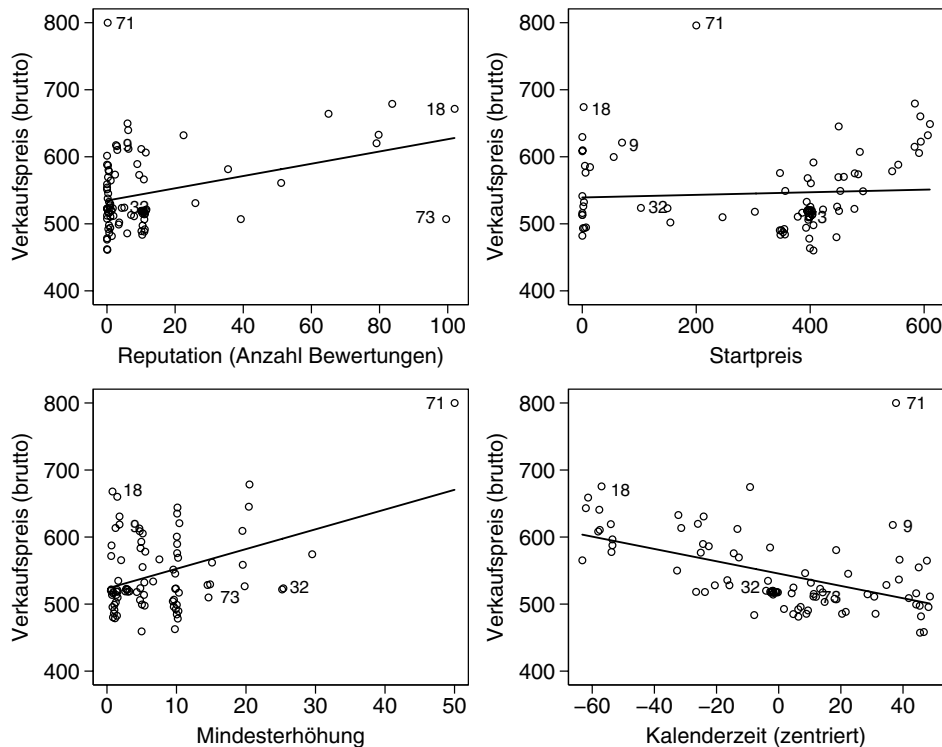
*Quelle:* Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

*Anmerkungen:* OLS-Standardfehler in runden Klammern; Bootstrap-Standardfehler in eckigen Klammern (1000 Replikationen; vgl. z. B. Mooney und Duval 1993).

*Modelle:* OLS1: Lineare Regression mit allen Fällen; OLS2: Lineare Regression ohne Fall 71; OLS3: Lineare Regression ohne die Ausreisser 9, 18, 32, 71 und 73; RREG: Robuste Regression nach Hamilton (1991). Die abhängige Variable ist in allen Modellen der Bruttoverkaufspreis in Franken.



Abbildung 10: Bivariate Scatterplots für das Reputationsmodell



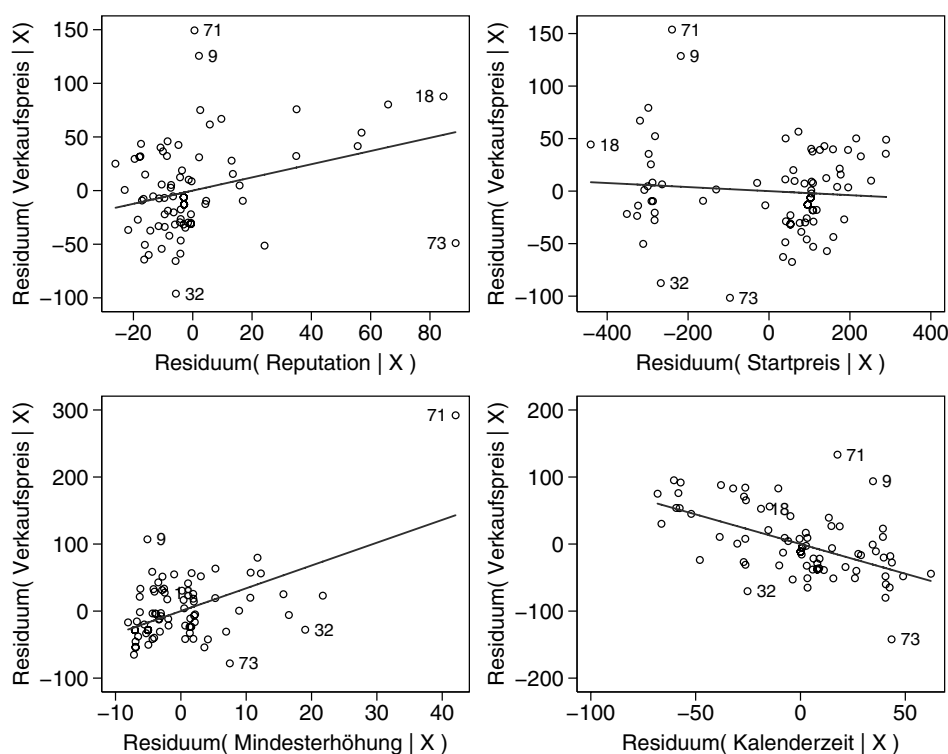
Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

falls erscheint der bereits identifizierte Punkt 73 im Plot für die Reputation deutlich als einflussreicher Ausreisser. Weiterhin gibt es offensichtlich einige bedeutsame Fälle, die mit den bivariaten Streudiagrammen nicht identifiziert werden konnten. Zum Beispiel hebt sich Punkt 9 in allen vier partiellen Regressionsplots mit einem relativ grossen Residuum von der Punktwolke ab. Zudem tritt Punkt 73 nun auch in den Plots bezüglich des Startpreises, der Mindesterhöhung und der Kalenderzeit in Erscheinung. Weitere bedeutsame Punkte sind Nr. 32 mit einem eher grossen Residuum in allen vier Diagrammen und Nr. 18 als Ausreisser bezüglich der X-Achse in den Plots für die Reputation und den Startpreis.

Ein eindrückliches Bild des Effektes, den Punkt 71 auf den Koeffizienten für den Mindestpreis hat, liefert der partielle Residuenplot in Abbildung 12. Die eingezeichnete Lowess-Kurve verläuft bis zu einer Schrittweite der Gebote von rund 30 Franken relativ flach und steigt danach steil an. Bei Ausschluss von Punkt 71 würde man eine Regressionsgerade erwarten, die eher dem ersten Teil der Lowess-Kurve entspricht.<sup>11</sup> Die partiellen Residuenplots weisen auf zwei weitere Aspekte hin, die hier aus Platzgründen jedoch nicht weiterverfolgt werden sollen: Eine eher unerwartete Zusammenhangsform wird für die Variable „Startpreis“ angezeigt. Dies hängt damit zusammen, dass der Verkaufspreis durch den Startpreis von unten her zentriert ist (was man übrigens auch im bivariaten Scatterplot deutlich erkennt): Es können nur Gebote abgegeben werden, die über diesem Mindestpreis liegen. Das Problem kann mit einem

<sup>11</sup>Man könnte, nebenbei bemerkt, natürlich auch versuchen, die durch die Lowess-Kurve angezeigte Zusammenhangsform mit einem nichtlinearen (hier z. B. exponentiellen) Modell zu approximieren. Die Wahl eines solchen Zusammenhangs dürfte hier aber inhaltlich relativ schwierig zu begründen sein und würde zudem nur auf dem Ausschereen eines einzigen Punktes beruhen.

Abbildung 11: Partielle Regressionsplots für das Reputationsmodell



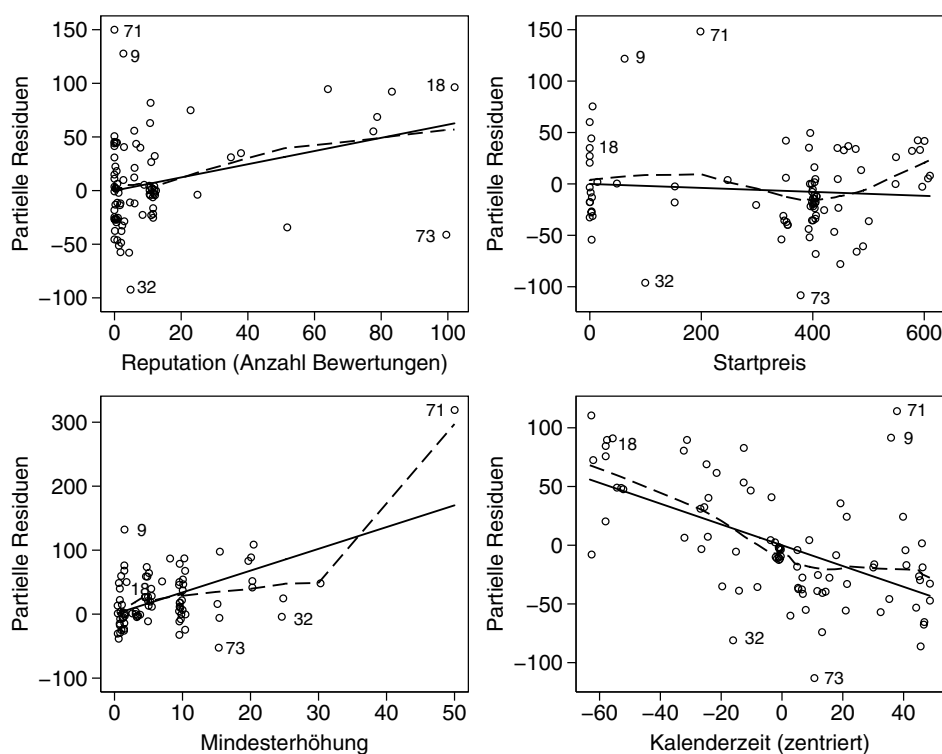
Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

zensierten Regressionsmodell angegangen werden (siehe dazu z. B. Maddala 1983 oder Breen 1996).<sup>12</sup> Weiterhin bestehen für die Variable „Kalenderzeit“ Hinweise auf einen nichtlinearen Zusammenhang. In der Tat erhält man einen (knapp) signifikanten Koeffizienten, wenn man die quadrierte Kalenderzeit als zusätzliche Variable in das Modell aufnimmt (und somit einen parabolischen Zusammenhang modelliert), was auch theoretisch Sinn macht (Verringerung des Wertes über die Zeit mit abnehmender Rate).

In Abbildung 13 sind die Werte einiger Masszahlen für den Einfluss einzelner Datenpunkte auf die Regressionsergebnisse mit Hilfe von Indexplots grafisch dargestellt. Die Masse Cooks  $D$  und  $DFFITS$  beziehen sich auf den Einfluss auf die Koeffizienten. Die bisherigen Ergebnisse werden bestätigt: Es gibt fünf Punkte (Nr. 9, 18, 32, 71 und 73), die die jeweilige Schranke zur Identifikation problematischer Fälle spürbar überschreiten. Mit einer Ausnahme haben diese Punkte aufgrund ihrer grossen studentisierten Residuen alle einen negativen Effekt auf die Präzision der Schätzung ( $COVRATIO < 1$ ). Punkt 18 hingegen erhöht die Schätzgenauigkeit, weil sein Residuum relativ klein ist, die Leverage aber eher gross (in Abbildung 11 erkennt man, dass Punkt 18 immer relativ nahe an der Regressionsgerade liegt, jedoch bezüglich der Reputation und des Startpreises Extremwerte annimmt). Recht informativ ist der letzte Plot in Abbildung 13, in dem die studentisierten Residuen gegen die Leverage abgetragen wird und die Grösse der

<sup>12</sup>Solche Modelle haben allerdings den Nachteil das die Ergebnisse relativ stark von der Verteilungsannahme über die Residuen abhängen (eine Alternative können u. U. semi-parametrische Modelle sein; vgl. z. B. Chay und Powell 2001, wobei sich die dort verwendeten Modelle nicht direkt übertragen lassen). Zu vermerken ist ferner, dass die Verkaufspreise zusätzlich von oben her zensiert sind, da einige Anbieter einen „Sofortpreis“ angegeben haben. Gibt ein Käufer ein Gebot in der Höhe des Sofortpreises ab, wird die Auktion gestoppt und das Produkt zu diesem Preis verkauft.

Abbildung 12: Partielle Residuenplots für das Reputationsmodell



Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

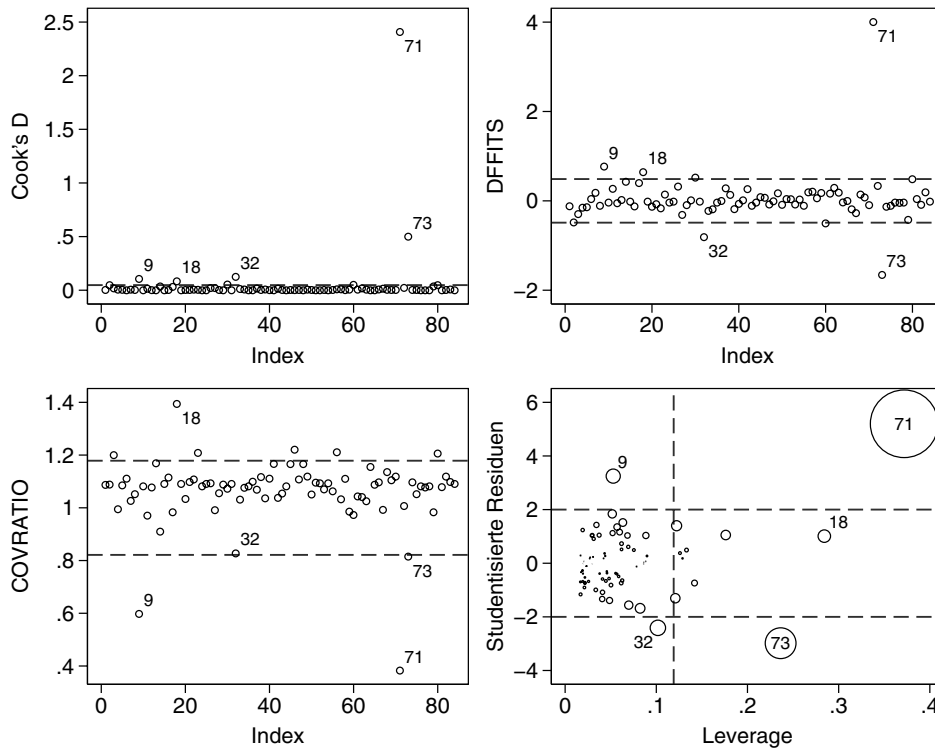
Symbole vom Wert von Cooks  $D$  abhängt.<sup>13</sup> Man sieht hier sehr gut, wie die Kombination aus Residuum und Leverage zu den Werten für die Einflussstatistiken führt. Ein differenziertes Bild der Effekte von Ausreißern auf die einzelnen Koeffizienten geben zudem die in Abbildung 14 enthaltenen Indexplots der  $DFBETAS$ -Statistiken. Man erkennt zum Beispiel, dass sich Fall 71 stark positiv auf den Effekt der Mindestserhöhung auswirkt und eher negativ auf den Effekt des Startpreises. Ferner vermindert Punkt 73 den Koeffizienten für die Reputation deutlich.

Als Ergänzung oder als Ersatz für die Grafiken kann es sinnvoll sein, eine Tabelle mit den Ausreißern zu erstellen, die die Werte der verschiedenen Einflussstatistiken und ggf. weiterer Variablen aufführt. Dies erlaubt es, die Extremfälle einzeln bezüglich verschiedener Grössen zu untersuchen. Tabelle 6 listet alle 20 Fälle der Reputationsmodells auf, die für mindestens eine Einflussstatistik gemäss den Definitionen in Tabelle 3 im kritischen Bereich liegen. Die entsprechenden Werte sind in der Tabelle hervorgehoben. Man erkennt, dass die bereits identifizierten Punkte 9, 18, 32, 71 und 73 jeweils auf einer ganzen Reihe von Massen ausschlagen. Die weiteren Punkte erscheinen jedoch eher unproblematisch.

Interessant ist nun zu betrachten, inwieweit sich die Regressionsergebnisse durch die Eliminierung einzelner Datenpunkte tatsächlich inhaltlich verändern. In Tabelle 5 ist neben dem Ausgangsmodell mit allen 84 Fällen (OLS1) ein Modell enthalten, bei dessen Schätzung der stärkste Ausreisser (Fall 71) ausgeschlossen wurde (OLS2). Man erkennt, dass der überproportionale Effekt der Mindestserhöhung durch Ausschluss dieser einen Beobachtung stark schrumpft und nicht mehr weit vom theoretisch erwarteten Wert 1 entfernt ist. Interessant ist auch noch ein

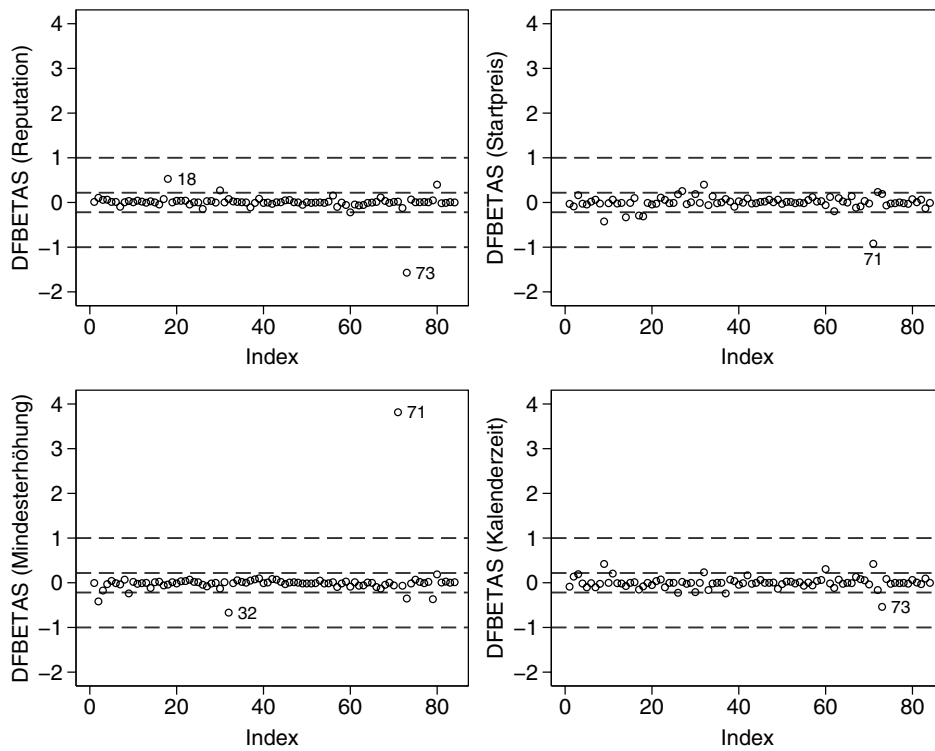
<sup>13</sup>Ein weiteres Beispiel für einen Plot, der die Information aus verschiedenen Masszahlen verdichtet darstellt, gibt Gray (1989b).

Abbildung 13: Verschiedene Einflussstatistiken für das Reputationsmodell



Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

Abbildung 14: Einfluss auf einzelne Koeffizienten des Reputationsmodells



Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

Tabelle 6: Ausreisser im Reputationsmodell

Nr.	Preis	Studen- tisierte Residuen	Leverage	Cooks <i>D</i>	<i>DFFITs</i>	<i>COVRATIO</i>	<i>DFBETAs</i>			
							Reputation	Startpreis	Mindest- erhöhung	Kalender- zeit
2	575	-1.308	0.121	0.047	-0.485	1.088	0.108	-0.089	<b>-0.419</b>	0.138
3	610	-0.733	0.142	0.018	-0.298	<b>1.200</b>	0.059	0.162	-0.172	0.192
9	617	3.246	0.053	<b>0.105</b>	<b>0.767</b>	<b>0.597</b>	0.036	<b>-0.424</b>	<b>-0.236</b>	<b>0.423</b>
14	613	1.838	0.052	0.036	0.429	0.909	0.025	<b>-0.328</b>	-0.115	-0.073
17	628	1.522	0.063	0.031	0.396	0.983	0.079	<b>-0.293</b>	-0.061	-0.151
18	671	1.016	0.284	<b>0.082</b>	<b>0.640</b>	<b>1.394</b>	<b>0.529</b>	<b>-0.309</b>	-0.041	-0.083
23	648	0.375	0.126	0.004	0.142	<b>1.208</b>	-0.048	0.059	0.068	-0.100
26	615	1.032	0.089	0.021	0.322	1.093	-0.147	0.182	-0.043	<b>-0.226</b>
27	495	-1.391	0.049	0.020	-0.315	0.991	0.031	<b>0.259</b>	-0.082	0.019
30	660	1.400	0.122	<b>0.054</b>	<b>0.523</b>	1.073	<b>0.272</b>	0.188	-0.128	-0.207
32	525	-2.414	0.102	<b>0.124</b>	<b>-0.812</b>	0.828	0.075	<b>0.397</b>	<b>-0.671</b>	<b>0.235</b>
37	640	1.034	0.068	0.016	0.280	1.069	-0.109	0.079	0.046	<b>-0.236</b>
46	620	0.186	0.129	0.001	0.071	<b>1.221</b>	0.058	0.020	0.005	-0.000
56	630	0.489	0.133	0.007	0.191	<b>1.211</b>	0.156	0.054	0.014	0.005
60	565	-1.677	0.082	<b>0.049</b>	<b>-0.503</b>	0.973	<b>-0.221</b>	-0.058	-0.087	<b>0.308</b>
71	800	5.201	0.372	<b>2.408</b>	<b>4.001</b>	<b>0.383</b>	0.018	<b>-0.919</b>	<b>3.816</b>	<b>0.423</b>
72	610	1.351	0.057	0.022	0.332	1.007	-0.127	<b>0.235</b>	-0.066	-0.164
73	510	-2.978	0.236	<b>0.499</b>	<b>-1.657</b>	<b>0.814</b>	<b>-1.573</b>	0.193	<b>-0.353</b>	<b>-0.542</b>
79	525	-1.558	0.070	0.036	-0.427	0.983	0.049	-0.033	<b>-0.370</b>	-0.020
80	676	1.051	0.176	0.047	0.486	<b>1.206</b>	<b>0.397</b>	0.073	0.188	0.062

Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

anderes Ergebnis. Der Standardfehler des Koeffizienten wird im vollen Modell verglichen mit einem mittels des Bootstrap-Verfahrens gewonnenen Standardfehler durch die OLS-Methode stark unterschätzt. Dies ist höchst plausibel. Bootstrap-Standardfehler werden mit Hilfe der Varianz der geschätzten Koeffizienten ermittelt, die sich ergibt, wenn die Modellschätzung anhand einer Reihe von Stichproben aus den vorliegenden Daten repetiert wird (die Bootstrap-Stichproben haben jeweils den gleichen Umfang wie die Originalstichprobe, werden jedoch mit Zurücklegen gezogen und unterscheiden sich deshalb voneinander). Die einzelnen Bootstrap-Koeffizienten für die Mindesterrhöhung hängen nun stark davon ab, ob sich Fall 71 gerade in der Stichprobe befindet oder nicht, was zu starker Variabilität des Koeffizienten führt. Ein grosser Bootstrap-Standardfehler verglichen mit dem gewöhnlichen OLS-Schätzer kann somit ein Indiz für die Präsenz von einflussreichen Ausreissern sein.

Tabelle 5 enthält weiterhin ein Modell, bei dem zusätzlich die Punkte 9, 18, 32, und 73 ausgeschlossen wurden (OLS3), sowie eine „robuste“ Regression (RREG; es handelt sich um einen so genannten *M*-Schätzer, bei dem verschiedene Gewichtungsfunktionen zur Anwendung kommen; zu den Details vgl. Hamilton 1991). Die Resultate sind ähnlich, wobei der Effekt der Mindesterrhöhung in diesen Modellen wieder etwas stärker ist. Ebenfalls leicht zugenommen hat in diesen Modellen der Effekt der Reputation, wobei zu berücksichtigen ist, dass auch hier der OLS-Schätzer für den Standardfehler verglichen mit dem Bootstrap-Schätzer jeweils etwas tief ausfällt. Dies hat mit der bereits angesprochenen starken Massierung der Datenpunkte im unteren Bereich der Reputationsskala (0–15 Bewertungen) und dem damit verbundenen relativ starken Einfluss der weit gestreuten restlichen Punkte zu tun (der Effekt der Reputation hängt stark davon ab, welche Fälle mit hoher Reputation in einem Bootstrap-Sample enthalten sind). Der Reputationseffekt ist also aufgrund der schiefen Verteilung der *X*-Werte weniger stabil als man auf den ersten Blick annehmen könnte und ist gemessen am Bootstrap-Standardfehler nur noch knapp signifikant. Allerdings bleibt der Koeffizient von ca. 0.6 auch dann noch bestehen, wenn alle Fälle mit mehr als 15 Bewertungen von den Analysen ausgeschlossen werden, was wiederum als Indiz für einen Reputationseffekt gewertet werden könnte (der Standardfehler nimmt jedoch stark zu; Modell nicht dargestellt).

Zusammenfassend lässt sich als Resultat der Ausreisseranalyse festhalten, dass erstens der Effekt der Mindesterrhöhung in dem Modell aufgrund eines einzelnen Extremwerts stark überschätzt wird und zweitens der Reputationseffekt etwas weniger verlässlich ist, als es durch eine einfache OLS-Schätzung suggeriert wird. Trotzdem muss die Schlussfolgerung von Diekmann und Wyder (2002) bezüglich der Hypothese eines positiven Effektes von Reputation auf den Auktionserlös nicht revidiert werden. Der Reputationseffekt ist in allen Modellen signifikant und liegt jeweils in einem Bereich von 0.6 oder höher, das heisst, pro zusätzliche (positive) Bewertung ist gemäss den Daten mit einer Erhöhung des Preises um gut 0.6 Franken zu rechnen (bei Diekmann und Wyder 2002 betrug der Effekt 0.67 Franken). Drittens wurden für die Kalenderzeit Hinweise auf einen nichtlinearen Effekt gefunden und es wurde ein Zensierungsproblem bezüglich des Startpreises identifiziert.<sup>14</sup>

## 5 Diskussion

Wie alle statistischen Verfahren, die zum Ziel haben, gegebene Daten auf eine überschaubare Menge an Kennzahlen zu reduzieren, erfassen auch Regressionsschätzungen nur einen Teil der

---

<sup>14</sup>Anzumerken ist allerdings, dass auch Diekmann und Wyder (2002) das Zensierungsproblem erkannt und behandelt hatten.

Eigenschaften und spiegeln die Daten nicht immer in angemessener Weise wider. Insbesondere in kleinen Stichproben besteht die Gefahr, dass die gewonnenen Resultate durch einige wenige Beobachtungen dominiert werden, was Zweifel an der Verlässlichkeit einer Regressions-schätzung und den daraus abgeleiteten Schlussfolgerungen aufkommen lassen kann.

Die in diesem Beitrag vorgestellten grafischen und formalen Instrumente sind einfach zu berechnen und stehen in den üblichen Statistikprogrammen mehrheitlich zur Verfügung. Sie sollten zumindest bei kleinen bis mittleren Stichproben, in denen einzelne „atypische“ Beobachtungen einen besonders starken Einfluss ausüben können, im Prozess der Datenanalyse standardmässig angewendet werden. Es bleibt jedoch darauf hinzuweisen, dass durchaus auch bei grossen Stichproben übermässig einflussreiche Datenpunkte auftreten können. Diese Gefahr besteht besonders bei komplexen Modellen mit vielen Parametern und z. B. Interaktionseffekten. Auch bei grösseren Stichproben lassen sich die Instrumente also unter Umständen gewinnbringend einsetzen.

Trotz der Wichtigkeit der Regressionsdiagnostik muss man bei der Interpretation der diagnostischen Ergebnisse auch eine gewisse Grosszügigkeit walten lassen. Bei kleineren Stichproben ist die Wahrscheinlichkeit gross, dass vermeintlich systematische Muster entstehen, die in Wahrheit aber nur auf Zufall beruhen (der Mensch ist bekanntlich äusserst anfällig dafür, in Zufallsprozessen eine Systematik zu erkennen; vgl. Schnell 1994: 213f.). Bei kleineren Unregelmässigkeiten besteht sicherlich kein Anlass zu Besorgnis und man muss aufpassen, dass man die Daten nicht „überinterpretiert“. Erst deutliche Abweichungen bedürfen einer genaueren Betrachtung. Im Allgemeinen sind auch die Faustregeln zur Identifikation bedeutender Datenpunkte ziemlich sensitiv, das heisst, selbst unter normalen Bedingungen ist zu erwarten, dass ein Teil der Punkte im kritischen Bereich liegt.

Es stellt sich schliesslich noch die Frage, was zu tun ist, wenn Ausreisser und einflussreiche Datenpunkte identifiziert worden sind. Die Ausreisser sollten (1) einer detaillierten Betrachtung unterzogen werden. Ratsam ist beispielsweise, für alle Punkte, die auf mindestens einer der Einfluss-Masszahlen extreme Werte erzielen, die  $Y$ - und  $X$ -Werte sowie die Ausreisserstatistiken zu tabellieren. Man sollte sich auch nicht scheuen, das Rohdatenmaterial nochmals zu sichten, da Ausreisser unter Umständen auf Übertragungsfehler zurückzuführen sind oder die Rohdaten zusätzliche Informationen enthalten, die die Ausreisser erklären können. Im Beispiel mit den Internet-Auktionen (Abschnitt 4) konnte so für den kritischen Punkt Nr. 71 ausfindig gemacht werden, dass im Lieferumfang einige Zusatzaccessoires enthalten waren, in der Beschreibung des Produkts ein sehr hoher Neupreis von 900 Franken angegeben wurde (der Ladenpreis des Geräts betrug ca. 700 Franken), und der realisierte Verkaufspreis von 800 Franken dem Sofortpreis entsprach. Einerseits scheint der Wert der gelieferten Ware also tatsächlich etwas höher als üblich gewesen zu sein. Andererseits hatten unter Umständen der angegebene Neupreis sowie die Sofortpreis-Option eine zusätzliche Signalwirkung die zur Realisierung des hohen Verkaufspreises beitrugen.

Ausreisser, bei denen es sich um offensichtliche Messfehler handelt, können (2) getrost eliminiert bzw., falls möglich, korrigiert werden. Bei dem Ausschluss von Messfehlern sollte man allerdings nicht voreilig handeln, da die Klassifizierung als fehlerhafte Beobachtung nicht immer eindeutig ist und der Ausschluss von gültigen Daten unter Umständen zu Fehlschlüssen führen kann. Eine berühmte Kontroverse um den Einbezug von vermeintlich fehlerhaftem Datenmaterial bezieht sich auf eine Studie von Jasso (1985), in der die Entwicklung der ehelichen Koitusfrequenz in Abhängigkeit der Ehedauer untersucht wurde. Kahn und Udry (1986) warfen Jasso vor, vier Fälle mit Wert 88 (!) für die monatliche Koitusfrequenz, bei denen es sich um fehlcodierte fehlende Werte handle (der Code für fehlende Werte war 99), übersehen zu ha-

ben und dass dadurch ihre Regressionsschätzungen substantiell verzerrt seien. Zwar mag Jasso (1986) damit Recht haben, dass durch Ausschluss der höchsten Werte und die damit verbundene Stützung der beobachteten Häufigkeitsverteilung ebenfalls eine Verzerrung der Regressionsresultate herbeigeführt werden könne, ihre verteidigenden Argumente für die Gültigkeit der vier in Frage stehenden Werte (z. B. „... a figure of ‘88’ represents a mean weakly figure of 22, which, in turn, represents three daily plus a weekly lagniappe ...“, S. 741) klingen allerdings nicht sonderlich überzeugend.<sup>15</sup> Jasso hätte zumindest auf das Vorhandensein der Extremwerte hinweisen müssen.

Allgemein empfiehlt es sich (3) einige „robuste“ Modelle zu schätzen, bei denen Ausreisser und einflussreiche Punkte ausgeschlossen werden. Ziel dieses Verfahrens kann allerdings nicht sein, einfach die Ausreisser wegzulassen und Modellschätzungen ohne die fraglichen Datenpunkte zu berichten. Oftmals geben ja gerade die unüblichen Datenpunkte wichtige Hinweise über die zu untersuchenden Sachverhalte und es wäre fahrlässig, diese Punkte zu ignorieren. Überdies kann die Entscheidung, welche Ausreisser im Detail auszuschliessen sind, eine höchst subjektive Angelegenheit sein (dies demonstrieren z. B. Collett und Lewis 1976), und es lassen sich durch den gezielten Ausschluss von Punkten, „die nicht ins Bild passen“, fast beliebige Resultate erzielen. Oder allgemeiner: Eliminiert man alle Befunde, die gegen eine bestimmte Vorstellung (z. B. eine spezifische Theorie) sprechen, wird man die Vorstellung auch eher bestätigen können. Der endgültige Ausschluss einzelner Fälle ist somit nicht zu empfehlen bzw. sollte ggf. genau dokumentiert und stichhaltig begründet werden.

Der Zweck von Modellschätzung unter Ausschluss von Ausreissern ist vielmehr zu ermitteln, inwieweit sich die Regressionsergebnisse tatsächlich inhaltlich verändern. Bleiben die Schätzergebnisse im Grossen und Ganzen stabil, ist man auf der sicheren Seite und kann davon ausgehen, dass von den Ausreissern keine grundlegende Verzerrung ausgeht. Sind die Schätzergebnisse jedoch stark abhängig von geringen Änderungen der Daten, dann sollte dies bei der Interpretation der Resultate verdeutlicht werden. Auf Ergebnisse instabiler Modelle ist kein Verlass und sie sollten in Folgestudien mit neuen Daten überprüft werden.

Eine Alternative zur Schätzung von Modellen unter Ausschluss von Ausreissern wird manchmal auch in der Anwendung von Methoden der robusten Regression gesehen. Dazu ist zu bemerken, dass mit der klassischen robusten Regression, die auf der Anwendung von Gewichtungsfunktionen beruht (sog. *M*-Schätzer wie z. B. die Median-Regression; vgl. Huber 1981, 1996, Hampel et al. 1986, sowie einführend Berk 1990), das Ziel verfolgt wird, im Falle von nicht normalverteilten Fehlern eine grössere Effizienz zu erreichen als mit der OLS-Regression. Die Verfahren sind somit zwar robust gegen *Y*-Ausreisser, jedoch nicht gegen Datenpunkte mit hoher Leverage (*X*-Ausreisser). Man spricht in diesem Zusammenhang davon, dass diese Schätzer genau gleich wie die OLS-Regression einen tiefen „Breakdown-Point“ haben: durch die willkürliche Platzierung eines einzigen (!) Datenpunktes kann die Regressionsschätzung beliebig stark verändert werden. Zur Begrenzung des Einflusses von Punkten mit hoher Leverage wurden zwar so genannte Bounded-Influence- bzw. Generalized-*M*-Schätzer (z. B. Welsch 1980; Krasker und Welsch 1982) oder High-Breakdown-Schätzer (Rousseeuw 1984; Rousseeuw und Leroy 1987) vorgeschlagen (für Übersichten siehe auch Davies 1993, Chave und

---

<sup>15</sup>Zudem ist Jassos Behauptung, dass ihre Schätzungen auch bei tatsächlicher Fehlerhaftigkeit der vier Werte erwartungstreu seien, falsch. Jasso macht geltend, dass gemäss statistischer Theorie die Schätzer auch dann unverzerrt sind, wenn die abhängige Variable mit *zufälligen* Messfehlern behaftet ist. Die Definition eines Zufallsfehlers in dieser Theorie ist jedoch nicht kompatibel mit der vorliegenden Situation.



Thomson 2003 und Lawrence 2003); diese Schätzer sind aber leider bis jetzt noch kaum praxistauglich.<sup>16</sup>

Ausreisser und einflussreiche Datenpunkte können schliesslich darauf hinweisen, dass ein Modell falsch spezifiziert ist und zum Beispiel wichtige zusätzliche Regressoren oder Interaktionseffekte übersehen oder eine unpassende Zusammenhangsform modelliert wurde (ein illustratives Beispiel gibt etwa Gray 1989a). Insbesondere das Vorliegen ganzer Cluster von Ausreissern weist auf das Vorhandensein starker Dritteinflüsse hin (vgl. dazu das Beispiel zum Zusammenhang zwischen Sozialkapital und Arbeitslosenquoten in Abschnitt 3.5). Man kann also (4) versuchen das Zusammenhangsmodell neu spezifizieren, wobei dies möglichst theoriegeleitet geschehen sollte. Eine rein explorative Anpassung des Modells an die Daten ist nicht zu empfehlen, weil man so mit grosser Wahrscheinlichkeit Eigenschaften modelliert, die zwar in der spezifischen Stichprobe erkennbar sind, jedoch in der Grundgesamtheit kein Gegenüber finden. So kann beispielsweise die passende Zusammenhangsform ebenfalls hauptsächlich von einigen wenigen einflussreichen Datenpunkten abhängen (vgl. die Literaturhinweise in Abschnitt 3.5). Oft besteht ferner das Problem, dass zwar Ausreisser relativ deutliche Hinweise über mögliche weitere Einflussfaktoren geben, entsprechende Daten aber nicht zur Verfügung stehen oder sich die Faktoren auf Phänomene beziehen, die in der Stichprobe zu selten auftreten, um noch statistisch behandelt werden zu können. In beiden Fällen wird man zur Klärung der genauen Zusammenhänge auf Folgestudien angewiesen sein.

## Literatur

- Andrews, David F., und Daryl Pregibon, 1978: Finding the Outliers that Matter. Journal of the Royal Statistical Society (Series B) 40: 85–93.*
- Anscombe, F. J., 1973: Graphs in Statistical Analysis. The American Statistician 27: 17–21.*
- Atkinson, A. C., 1981: Two Graphical Displays for Outlying and Influential Observations in Regression. Biometrika 68: 13–20.*
- Atkinson, A. C., 1982: Regression Diagnostics, Transformations and Constructed Variables. Journal of the Royal Statistical Society (Series B) 44: 1–36.*
- Atkinson, A. C., 1983: Diagnostic Regression Analysis and Shifted Power Transformations. Technometrics 25: 23–33.*
- Atkinson, A. C., 1986: Diagnostic Tests for Transformations. Technometrics 28: 29–37.*
- Barnett, Vic, und Toby Lewis, 1995: Outliers in Statistical Data, 3. Aufl. New York: John Wiley & Sons.*
- Beckman, R. J., und R. D. Cook, 1983: Outlier.....s. Technometrics 25: 119–149.*
- Belsley, David A., Edwin Kuh und Roy E. Welsch, 1980: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons.*
- Berk, Richard A., 1990: A Primer on Robust Regression. S. 292–324 in: John Fox und J. Scott Long (Hg.): Modern Methods of Data Analysis. Newbury Park, CA: Sage.*
- Bollen, Kenneth A., und Robert W. Jackman, 1990: Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. S. 257–291 in: John Fox und J. Scott Long (Hg.): Modern Methods of Data Analysis. Newbury Park, CA: Sage.*

---

<sup>16</sup>Zum Beispiel liessen sich für das Reputationsmodell in Abschnitt 4 keine einigermaßen stabilen High-Breakdown-Schätzer ermitteln (verwendet wurde für die Berechnungen das lqs-Modul von Brian Ripley für R; vgl. Venables und Ripley 2002: 156ff.).

- Brüderl, Josef*, 2000: Regressionsverfahren in der Bevölkerungswissenschaft. S. 589–642 in: *Ulrich Mueller, Bernhard Nauck und Andreas Diekmann* (Hg.): Handbuch der Demographie, Band 1. Berlin: Springer.
- Breen, Richard*, 1996: Regression Models. Censored, Sample Selected, or Truncated Data. Newbury Park, CA: Sage.
- Cain, Kevin C.*, und *Nicholas T. Lange*, 1984: Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data. *Biometrics* 40: 493–499.
- Carroll, R. J.*, und *D. Ruppert*, 1987: Diagnostics and Robust Estimation When Transforming the Regression Model and the Response. *Technometrics* 29: 287–299.
- Chatterjee, Samprit*, und *Ali S. Hadi*, 1986: Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science* 1: 379–393.
- Chatterjee, Samprit*, und *Ali S. Hadi*, 1988: Sensitivity Analysis in Linear Regression. New York: John Wiley & Sons.
- Chave, Alan D.*, und *David J. Thomson*, 2003: A Bounded Influence Regression Estimator Based on the Statistics of the Hat Matrix. *Applied Statistics* 52: 307–322.
- Chay, Kenneth Y.*, und *James L. Powell*, 2001: Semiparametric Censored Regression Models. *Journal of Economic Perspectives* 15: 29–42.
- Cleveland, William S.*, 1979: Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74: 829–836.
- Collett, D.*, und *T. Lewis*, 1976: The Subjective Nature of Outlier Rejection Procedures. *Applied Statistics* 25: 228–237.
- Cook, R. D.*, und *P. C. Wang*, 1983: Transformations and Influential Cases in Regression. *Technometrics* 25: 337–343.
- Cook, R. Dennis*, 1977: Detection of Influential Observation in Linear Regression. *Technometrics* 19: 15–18.
- Cook, R. Dennis*, 1979: Influential Observations in Linear Regression. *Journal of the American Statistical Association* 74: 169–174.
- Cook, R. Dennis*, und *Sanford Weisberg*, 1980: Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics* 22: 495–508.
- Cook, R. Dennis*, und *Sanford Weisberg*, 1982: Residuals and Influence in Regression. New York: Chapman and Hall.
- Cook, R. Dennis*, und *Sanford Weisberg*, 1982: Criticism and Influence Analysis in Regression. *Sociological Methodology* 13: 313–361.
- Cook, R. Dennis*, und *Sanford Weisberg*, 1999: Applied Regression Including Computing and Graphics. New York: John Wiley & Sons.
- Crawford, Kelly D.*, *Daniel J. Vasicek* und *Roger L. Wainwright*, 1995: Detecting Multiple Outliers in Regression Data Using Genetic Algorithms. Proceedings of the 1995 ACM symposium on applied computing, February 26–28, 1995, Nashville, Tennessee, United States.
- Dalton, Russell J.*, 2004: The Greening of the Globe?: Cross-national Levels of Environmental Group Membership. University of California.
- Davies, P. L.*, 1993: Aspects of Robust Linear Regression. *The Annals of Statistics* 21: 1843–1899.
- Diekmann, Andreas*, und *David Wyder*, 2002: Vertrauen und Reputationseffekte bei Internet-Auktionen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 54: 674–693.
- Draper, N. R.*, und *J. A. John*, 1981: Influential Observations and Outliers in Regression. *Technometrics* 23: 21–26.
- Fowlkes, Edward B.*, 1987: Some Diagnostics for Binary Logistic Regression via Smoothing.

- Biometrika 74: 503–515.
- Fox, John*, 1991: *Regression Diagnostics*. Newbury Park, CA: Sage.
- Fox, John*, 1997: *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.
- Freitag, Markus*, 2000: Soziales Kapital und Arbeitslosigkeit. Eine empirische Analyse zu den Schweizer Kantonen. *Zeitschrift für Soziologie* 29: 186–201.
- Gray, J. Brian*, 1986: A Simple Graphic for Assessing Influence in Regression. *Journal of Statistical Computation and Simulation* 24: 121–134.
- Gray, J. Brian*, 1989: On the Use of Regression Diagnostics. *The Statistician* 38: 97–105.
- Gray, J. Brian*, 1989: The Four-Measure Influence Plot. *Computational Statistics & Data Analysis* 8: 135–225.
- Gray, J. Brian*, und *Robert F. Ling*, 1984: *K*-Clustering as a Detection Tool for Influential Subsets in Regression. *Technometrics* 26: 305–318.
- Greene, William H.*, 2003: *Econometric Analysis*, 5. Aufl. Upper Saddle River, NJ: Pearson Education.
- Hamilton, Lawrence C.*, 1991: How Robust is Robust Regression? *Stata Technical Bulletin*: 21–26.
- Hampel, Frank R.*, *Elvezio M. Ronchetti*, *Peter J. Rousseeuw* und *Werner A. Stahel*, 1986: *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley & Sons.
- Hennig, Christian*, 2003: Clusters, Outliers, and Regression: Fixed Point Clusters. *Journal of Multivariate Analysis* 86: 183–212.
- Hinkley, D. V.*, und *S. Wang*, 1988: More About Transformations and Influential Cases in Regression. *Technometrics* 30: 435–440.
- Hoaglin, David C.*, und *Roy E. Welsch*, 1978: The Hat Matrix in Regression and ANOVA. *The American Statistician* 32: 17–22.
- Hocking, R. R.*, 1983: Developments in Linear Regression Methodology: 1959–1982. *Technometrics* 25: 219–230.
- Hosmer, David W., Jr.*, und *Stanley Lemeshow*, 1999: *Applied Survival Analysis. Regression Modeling of Time to Event Data*. New York: John Wiley & Sons.
- Hosmer, David W., Jr.*, und *Stanley Lemeshow*, 2000: *Applied Logistic Regression*, 2. Aufl. New York: John Wiley & Sons.
- Huber, Peter J.*, 1981: *Robust Statistics*. New York: John Wiley & Sons.
- Huber, Peter J.*, 1996 (1977): *Robust Statistical Procedures*, 2. Aufl. Philadelphia: Society for Industrial and Applied Mathematics.
- Jasso, Guillermina*, 1985: Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. *American Sociological Review* 50: 224–241.
- Jasso, Guillermina*, 1986: Is It Outlier Deletion or Is It Sample Truncation? *Notes on Science and Sexuality. American Sociological Review* 51: 738–742.
- Jennings, Dennis E.*, 1986: Outliers and Residual Distributions in Logistic Regression. *Journal of the American Statistical Association* 81: 987–990.
- Kahn, Joan R.*, und *J. Richard Udry*, 1986: Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. *American Sociological Review* 51: 734–737.
- Kohler, Ulrich*, und *Frauke Kreuter*, 2001: *Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. München: Oldenbourg.
- Krasker, William S.*, und *Roy E. Welsch*, 1982: Efficient Bounded-Influence Regression Estima-

- tion. *Journal of the American Statistical Association* 77: 595–604.
- Kruskal, William H.*, 1960: Some Remarks on Wild Observations. *Technometrics* 2: 1–3.
- Landwehr, James M., Daryl Pregibon und Anne C. Shoemaker*, 1984: Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association* 79: 61–71.
- Lawrence, David E.*, 2003: Cluster-Based Bounded Influence Regression. Blacksburg, Virginia: Virginia Polytechnic Institute and State University.
- Lesaffre, E.*, und *A. Albert*, 1989: Multiple-Group Logistic Regression Diagnostics. *Applied Statistics* 38: 425–440.
- Maddala, G. S.*, 1983: Limited Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.
- Mallows, C. L.*, 1986: Augmented Partial Residuals. *Technometrics* 28: 313–319.
- Marchette, David J.*, und *Jeffrey L. Solka*, 2003: Using data images for outlier detection. *Computational Statistics & Data Analysis* 43: 541–552.
- McCullagh, P.*, und *J. A. Nelder*, 1989: *Generalized Linear Models*, 2. Aufl. London: Chapman & Hall.
- Menard, Scott*, 2002: *Applied Logistic Regression Analysis*, 2. Aufl. Newbury Park, CA: Sage.
- Mooney, Christopher Z.*, und *Robert D. Duval*, 1993: Bootstrapping. A Nonparametric Approach to Statistical Inference. Newbury Park, CA: Sage.
- O’Hara Hines, R. J.*, und *E. M. Carter*, 1993: Improved Added Variable and Partial Residual Plots for the Detection of Influential Observations in Generalized Linear Models. *Applied Statistics* 42: 3–20.
- Peña, Daniel*, und *Francisco J. Prieto*, 2001: Multivariate Outlier Detection and Robust Covariance Matrix Estimation. *Technometrics* 43: 286–310.
- Pena, Daniel*, und *Victor J. Yohai*, 1995: The Detection of Influential Subsets in Linear Regression by using an Influence Matrix. *Journal of the Royal Statistical Society (Series B)* 57: 145–156.
- Pettitt, A. N.*, und *I. Bin Daud*, 1989: Case-Weighted Measures of Influence for Proportional Hazards Regression. *Applied Statistics* 38: 51–67.
- Polasek, Wolfgang*, 1984: Regression Diagnostics for General Linear Regression Models. *Journal of the American Statistical Association* 79: 336–340.
- Pregibon, Daryl*, 1981: Logistic Regression Diagnostics. *The Annals of Statistics* 9: 705–724.
- Ritschard, Gilbert*, und *Gerard Antille*, 1992: A Robust Look at the Use of Regression Diagnostics. *The Statistician* 41: 41–53.
- Rousseeuw, Peter J.*, 1984: Least Median of Squares Regression. *Journal of the American Statistical Association* 79: 871–880.
- Rousseeuw, Peter J.*, und *Annick M. Leroy*, 1987: *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.
- Rousseeuw, Peter J.*, und *Bert C. van Zomeren*, 1990: Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association* 85: 633–639.
- Schnell, Rainer*, 1994: *Graphisch gestützte Datenanalyse*. München: Oldenbourg.
- Velleman, Paul F.*, und *Roy E. Welsch*, 1981: Efficient Computing of Regression Diagnostics. *The American Statistician* 35: 234–242.
- Venables, William N.*, und *Brian D. Ripley*, 2002: *Modern Applied Statistics With S*, 4. Aufl. New York: Springer.
- Vos, Paul W.*, 1991: A Geometric Approach to Detecting Influential Cases. *The Annals of Statistics* 19: 1570–1581.

- Wang, P. C., 1985: Adding a Variable in Generalized Linear Models. *Technometrics* 27: 273–276.
- Weisberg, Sanford, 1985: *Applied Linear Regression*, 2. Aufl. New York: John Wiley & Sons.
- Welsch, Roy E., 1980: Regression Sensitivity Analysis and Bounded-Influence Estimation. S. 153–167 in: *Jan Kmenta und James Bernard Ramsey (Hg.): Evaluation of Econometric Models*. New York: Academic Press.
- Welsch, Roy E., 1982: Influence Functions and Regression Diagnostics. S. 149–169 in: *Robert L. Launer und Andrew F. Siegel (Hg.): Modern Data Analysis*. New York: Academic Press.
- Wooldridge, Jeffrey M., 2003: *Introductory Econometrics. A Modern Approach*, 2. Aufl. Mason, Ohio: Thomson, South-Western.