Helmut Bludszuweit

Die Sensitivitätsfrage in der Regressionsrechnung, illustriert an kleinen Stichproben

Datensätze zur Regressionsrechnung weisen oft merkwürdige Datenpunkte auf, die einer sorgfältigeren Betrachtung unterzogen werden sollten. Das sind einmal solche, die nicht in das vom unterstellten Modell her bestimmte Bild der Verteilungen passen . Sie provozieren die Entscheidungsfrage: Sind diese Punkte als 'Ausreißer' abzuqualifizieren und aus der weiteren Analyse zu elimnieren, oder muß das Modell in Frage gestellt und modifiziert werden. Zum anderen kann es Punkte geben mit einem relativ zu den anderen Punkten auffällig starken Einfluß auf wesentliche Seiten des Ergebnisbildes (Vorhersagewerte, Schätzungen der Modell-parameter, Streuungsparameter). Diese 'Influenzpunkte' sollten eine möglichst gut gesicherte Validität für die Untersuchungsfrage haben. Leider gibt es für diese Probleme keine einfache Patentlösung. Entsprechend reichhaltig ist das in der Fachliteratur angebotene Spektrum von Einzelkriterien. Auch ein so leistungsfähiges Statistikprogramm wie SAS (PROC REG) bietet dem Nutzer nur eine Auswahl solcher Kriterien, die ihm bei seiner Entscheidung helfen, diese ihm aber keineswegs ('automatisiert') abnehmen. Dieser Beitrag versucht, dem Nutzer dieser Kriterien eine theoretisch orientierte Verständnishilfe zu geben.

Wie kommen Influenzpunkte zustande?

Wir gehen vomStandardmodell der (linearen) Regressionsrechnung aus.
Eine zu erklärende bzw. vorherzusagende Variable Y wird als Linearkombination von 'unabhägigen Variablen' zuzüglich eines nicht erklärbaren Residualanteils e angesetzt:

$\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$ (n Beobachtungen in **Y**, k Parameter in **b**).

Unter Experimentalbedingungen werden die Werte der unabhängigen Variablen ('fehlerfrei') gesetzt. Diese Setzungen konstituieren die sog. Designmatrix **X** (n x k - Matrix). Die Gewichte **b** (Regressionskoeffizienten) sind so zu schätzen, daß die verbleibenden Residuen e möglichst (im Quadrat-Mittel) klein bleiben. Uber sie werden Veteilungsannahmen gemacht (Gauss-Markov-Bedingungen). Unter diesen Bedingungen liegt die ganze empirische Information in den Beobachtungen von **Y**, während **X** 'gesetzt' ist. **X** hat aber ganz wesentliche Einflüsse auf die hier in Rede stehenden Influenzen. Dazu sehen wir uns mal die 'Wirkspur' von X durch den Formalismus des Regressionsmodells an:

Die Kovarianzmatrix $\mathbf{Q} = \mathbf{(X'X)^{-1}}$

bestimmt wesentlich die Kovarianzen der Schätzungen der Regressionsparameter ($\mathbf{cov(b\hat{})} = s^2\mathbf{Q}$ mit $s^2 = \mathbf{e'e}/(n-k)$).

Die Schätzmatrix $\mathbf{A} = \mathbf{QX'}$

betimmt die Parameterschätzungen mittels der Beobachtungen **Y** ($\mathbf{b\hat{}} = \mathbf{AY}$).

Die Prädiktionsmatrix (engl.: 'hat matrix') $\mathbf{H} = \mathbf{XA}$

bestimmt die Vorhersage- bzw. Erklärungswerte $\mathbf{Y\hat{}} = \mathbf{HY}$, und schließlich die Residualmatrix (orthogonal zu **H**) $\mathbf{M} = \mathbf{I} - \mathbf{H}$

bestimmt die Residuenschätzungen $\mathbf{e\hat{}} = \mathbf{MY}$.

Für unsere Frage zentral ist die Prädiktionsmatrix **H**. Ihre Diagonalelemente sagen, mit welchem Gewicht (h_{ii}) die Beobachtung y_i (d.h. der i-te Meßpunkt im Versuchsplan) an der Bildung des Erklärungswertes $y_i\hat{}$ beteiligt ist. Dabei gilt: $0 \le h_{ii} \le 1$ und $\sum h_{ii} = k$. Ferner

wird die Anschaulichkeit der Einflußbeziehungen durch folgende Extremaussagen unterstützt:
Falls $h_{ii} = 0$, so sind alle h_{ij} in dieser Zeile = 0,
d.h., wenn der Beobachtungswert y_i keinerlei Erklärungswert für die (seine eigene) Vorhersage y_i^ hat, so haben auch alle anderen Meßwerte y_j keinen. (Null-Zeile in **H**).
Wenn $h_{ii} = 1$, d.h. wenn y_i seine Vorhersage y_i^ vollständig erklärt, haben alle anderen Meßwerte für y_i keinen ('zusätzlichen') Erklärungswert.
Verschärfung für Modelle mit additiver Konstanten: $\text{Min}(h_{ii}) = 1/n$.

Das Ideal $h_{ii} = k/n$ (gleiche Influenz für alle Meßpunkte) läßt sich (im 'fehlerfreien' Versuchsplan) realisieren, wenn alle n Meßpunkte im Raum der unabhängigen Variablen denselben Abstand zu ihrem Mittel haben. Je stärker die Abweichng von diesem Mittel, desto größer der Influenzwert. Am Ort des Mittels ist der Einflußwert minimal (0 bzw. 1/n).
Eine geometrische Veranschaulichung möge das theoretische Bild der Matrix **H** abrunden:
H ist eine Projektionsmatrix, die den n-dimensionalen Stichprobenraum der Beobachtungen **Y** auf einen k-dimensinalen Unterraum (Zahl der Parameter im Modell: k < n) orthogonal projiziert. Die komplementäre Matrix **M** = **I - H** ist dann die entsprechnede Orthogonalprojektion auf den n-k-dimensionalen 'Restraum' (der Residuen). Z.B. ergibt sich daraus eine für Modellprüfunen nützliche Eigenschaft: Die Residuen sind orthogonal zu den Vorhersagewerten.

Einige Beispiele

(1) Die einfache lineare Regression (k = 2) mit äquidistanten Meßpunkten x.
O.B.d.A. kann die Distanz zwischen benachbarten Meßpunkten gleich 1 gesetzt werden, da **H** invariant ist gegenüber linearen Transformationen der X-Skalen.
Für n = 3 erhalten wir :

$$X' = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

führt zu: $$H = \begin{pmatrix} 5/6 & 2/6 & -1/6 \\ 2/6 & 2/6 & 2/6 \\ -1/6 & 2/6 & 5/6 \end{pmatrix}$$

Allgemein für n = 2m+1 äquidistante Meßpunkte (Mittel bei Null) ergibt sich:

$$h_{ij} = \frac{1}{n} + \frac{3(m+1-i)(m+1-j)}{m(m+1)n}$$

Die 'Randpunkte' sind also von stärkerem Einfluß.Will man auch hier per Versuchsplan gleichen Einfluß erzwingen, muß man zwei Meßpunkte von jeweils gleicher Vielfachheit vorsehen, z.B.:

(2) $$X' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix}$$

führt zu: $H = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/1 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$

Allgemein: Plan mit 2 x m Meßpunkt-Wiederholungen:

$h_{ij} = 1/m$ für i,j ≤ m und für i,j > m (mithin für alle Diagonalelemente)

$hij = 0$ sonst.

(3) Bei mehr unabhängigen Variablen kann man die ‘gleiche Zentralität’ variabler gestalten, z.B. in der Ebene (k=3 Parameter) vier Meßpunkte auf einer ‘Kreisbahn’ um ihren (nicht besetzten) Zentralpunkt:

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ & & & \end{pmatrix} \quad \text{führt zu:} \quad H = \begin{pmatrix} 3/4 & 1/4 & 1/4 & -1/4 \\ 1/4 & 3/4 & -1/4 & 1/4 \\ 1/4 & -1/4 & 3/4 & 1/4 \\ -1/4 & 1/4 & 1/4 & 3/4 \end{pmatrix}$$

Realisierungen mit SAS/STAT

Wir realisieren SAS-Beisiele mit

```
PROC REG
        DATA= ...;
        MODEL ... /INFLUENCE P R;
RUN;
```

Außer der hier thematisierten Option ‘INFLUENCE’ nehmen wir die Optionen ‘P’ (‘Prediction’) und ‘R’ (Residuals’) hinzu, die unserem Thema bei einer Regressionsanalyse logisch vorangehen - man würde keine Influenzanalyse ohne Residuenanalyse machen.

Beispiel mit sieben gleichabständigen Meßpunkten:

```
libname disk "A:\";
data disk.ABSP1;
input y 2-3 x 5;
cards;
 35 1
 40 2
 37 3
 41 4
 39 5
 45 6
 49 7
;
proc reg data= disk.ABSP1;
        model Y= X/ INFLUENCE P R;
run;
```

Das liefert den Output:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	104.14286	104.14286	15.917	0.0104
Error	5	32.71429	6.54286		
C Total	6	136.85714			

Root MSE	2.55790	R-square	0.7610
Dep Mean	40.85714	Adj R-sq	0.7132
C.V.	6.26060		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	33.142857	2.16182085	15.331	0.0001
X	1	1.928571	0.48339784	3.990	0.0104

Obs	Dep Var Y	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	35.0000	35.0714	1.743	-0.0714	1.872	-0.038	| |	0.001
2	40.0000	37.0000	1.367	3.0000	2.162	1.388	| **	0.385
3	37.0000	38.9286	1.081	-1.9286	2.318	-0.832	* |	0.075
4	41.0000	40.8571	0.967	0.1429	2.368	0.060	| |	0.000
5	39.0000	42.7857	1.081	-3.7857	2.318	-1.633	*** |	0.290
6	45.0000	44.7143	1.367	0.2857	2.162	0.132	| |	0.003
7	49.0000	46.6429	1.743	2.3571	1.872	1.259	| **	0.687

Obs	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	X Dfbetas
1	-0.0341	0.4643	2.9150	-0.0318	-0.0315	0.0264
2	1.5829	0.2857	0.8270	1.0011	0.9498	-0.7079
3	-0.8016	0.1786	1.4121	-0.3738	-0.2990	0.1671
4	0.0540	0.1429	1.8203	0.0220	0.0099	0.0000
5	-2.1380	0.1786	0.4143	-0.9969	-0.0000	-0.4458
6	0.1184	0.2857	2.1722	0.0749	-0.0237	0.0530
7	1.3626	0.4643	1.3605	1.2685	-0.6294	1.0555

Sum of Residuals	0
Sum of Squared Residuals	32.7143
Predicted Resid SS (Press)	63.9579

Zu den Angaben der Residualanalyse bemerken wir hier lediglich:
Die praktisch wichtige Kenngröße ‘Student Residual’ (= r_i) ist in Wirklichkeit das Standardisierte Residuum (= ‘Residual’ / ‘Std Err Residual’).
Das studentisierte Residuum finden wir unter ‘Rstudent’ in der Tabelle zur Influenzanalyse (2.Tab.). Der formale Unterschied zum Standardisierten Residuum besteht darin, daß hier eine

Schätzung der Residuenvarianz bei jeweiligem Weglassen des betreffenden Meßpunktes i erfolgt:

$$'StdErr\ \mathrm{Re}\,sidual' = s\sqrt{1-h_{ii}}\ ,$$

aber in ‘RSTUDENT’ wird anstelle von s der jeweilige Wert $s_{(i)}$ gesetzt, wobei gilt:

$$s^2_{(i)} = \frac{n-k}{n-k-1}s^2 - \frac{e_i^2}{(n-k-1)(1-h_{ii})}\ ,$$ k = Zahl der Parameter im Modell.

Dieses heute generell bevorzugte Residuenmaß

$$\text{‘RSTUDENT’} = e^*_{(i)} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

kommt auch in den anderen Kenngrößen vor.
Die Leverage-Werte (Diagonalwerte aus H) sind die Werte der Spalte ‘Hat Diag H’.
‘COVRATIO’ ist das Verhältnis der Determinanten der geschätzten Kovarianzmatrizen für die Regressionsparameter b (ohne Meßpunkt i : mit allen Meßpunkten). Werte dicht bei 1 signali-sieren also ‘wenig Einfluß ‘ durch den entsprechenden Meßpunkt .
‘DFFITS’ beschreibt in standardisierter Form die Veränderung im Vorhersagewert bei Weglassen des Beobachtungspunktes i:

$$Dfit = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}}$$

Ein großer Wert signalisert starken Einfluß des Beobachtungspunktes.
Die ‘DFBETAS’ (für jeden Regresionskoeffizienten eine Spalte in der Tabelle) beschreiben in standardisierter Form den Effekt des Weglassens des Beobachtungspunktes i auf die Schätzung des jeweiligen Regressionskoeffizienten.
Diese letzten drei Einflußindizes akzentuieren gewissermaßen die ‘Blickrichtung’ beim Schauen auf den Einfluß des jeweiligen Beobachtungspunktes und sind damit eine zusätzliche Hilfe. Die folgenden Formeln mögen aber noch einmal zeigen, wie weitreichend sie durch die fundamen-taleren Informationen der Leverage-Werte h_{ii} und der studentisierten Residuen $r^*_{(i)}$ bestimmt sind:

COVRATIO: $$CR = \left(\frac{r_i^2}{r_{(i)}^{*2}}\right)^k \cdot \frac{h_{ii}}{1-h_{ii}}$$

DFFITS: $$Dfit = r^*_{(i)}\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

DFBETAS: $$Dbeta = r^*_{(i)} \cdot \frac{a_{ji}}{\sqrt{(1-h_{ii})q_{jj}}}$$

Beurteilungsfaustregeln (‘cutpoints’) für die Handhabung dieser Kriterien

Im Leverage auffällige Punkte
sind solche, die den Wert 2k/n überschreiten, d.h. deren Einfluß mehr als doppelt so groß ist als es im Idealfall ‘gleichmäßiger Einfluß für alle Meßpunkte’ der Fall wäre.

Im Standardisierten Residuum auffällige Punkte
sind solche, deren Wert-Betrag 2 überschreitet (sie sind t-verteilt, wenn die Gauß-Markov-Bedingungen für das Modell gültig sind).

Im Kovarianzverhältnis auffällige Punkte
sind solche, deren Verhältnis von 1 stärker abweicht als 3k/n.

Im DFFITS-Wert auffällige Punkte
sind solche, die den kritischen Wert $2\sqrt{\frac{k+1}{n-k-1}}$ überschreiten. Hierfür wurde ein schärferer kritischer Wert für h_{ii} von (k+1)/n gesetzt.

Im DFBETAS-Wert auffällig ist ein Punkt,
wenn er den kritischen Wert von $2/\sqrt{n}$ überschreitet.

Diese häufig empfohlenen Kriterien gehen auf Belsley u.a. (1980) zurück. Sie sollten als eine Hilfe verstanden werden, die Punkte in einer Regressionsanalyse herauszufinden, die aufgrund ihres überdurchschnittlichen Einflusses einer besonderen Aufmerksamkeit bedürfen. Ob sie nun für die Untersuchung besonders konstitutiv sind oder verfälschende Störungen mit sich bringen, die man eliminieren sollte, wird man i.a. ohne Zusatzinformationen nicht mit rein statistischen Mitteln sicher entscheiden können. Es kann immerhin passieren, das ‘schlechte Punkte’ maskiert werden und ‘gute Punkte’ als Ausreißer erscheinen, wie das Hawkins (wiedergegeben in Rousseeuw u.a., 1987) mit seinem konstruierten Testbeispiel demonstrierte. Er wollte damit die Vorzüge robuster Techniken demonstrieren, die die Verfälschungsgefahr umgehen, die vom Vorhandensein *mehrerer* ‘schlechter’ Punkte droht. Unsere Kriterien lassen nämlich immer nur *einen* Punkt aus, um dessen Wirkung auf das Analyseergebnis aufzuzeigen.

Wenden wir diese ‘Cutpoints’ auf unser Beispiel an:

Auffallige Beobachtungspunkte in der Stichprobe						
Beob.-Nr.	r*>2	h>0,57	CR>0,857	Dff>0,866	$Df\backslash_0$ / Df_1 > 0,756	
1	.	.	*	.	.	.
2	.	.	.	*	*	.
3	.	.	*	.	.	.
4	.	.	*	.	.	.
5	*	.	.	*	.	.
6	.	.	.	.	.	.
7	.	.	*	*	.	*

Es fällt hier auf: In den Elementargrößen ist nur ein Residuum zu markieren, aber die anderen signalisieren mehr. Demgemäß verdient vor allem der oberste Extrempunkt des Versuchsplanes erhöhte Aufmerksamkeit.

Literatur

Belsley,D.A., Kuh,E., and Welsch,R.E. (1980)
Regression Diagnostics / Wiley & Sons, New York
Chatterjee,S., and Hadi,A.S. (1988)
Sensitivity Analysis in Linear Regression / Wiley & Sons, New York
Hogalin,D.C., Mosteller,F., and Tukey,J.W. (1983)
Understanding Robust and Exploratory Data Analysis / Wiley & Sons, New York
Rousseuw, P.J., and Leroy, A.M. (1987)
Robust Regression and Outlier Detection / Wiley & Sons, New York
Sen,A., and Srivastava (1990)
Regression Analysis / Springer, Berlin
SAS/STAT User's Guide, Version 6, Fourth Edition ,
Kapitel PROC REG

Anschrift des Verfassers:
Dr.phil. Helmut Bludszuweit
Kernbergstr. 59
07749 Jena