Umsetzung robuster Regressionsverfahren in SAS

Hans-Peter Altenburg
Siemens Healthcare Diagnostics
Emil-von-Behring-Str. 76
35041 Marburg
hans-peter.altenburg@siemens.com

Zusammenfassung

Es wird gezeigt, welche Möglichkeiten SAS bietet, um robuste Verfahren standardmäßig umzusetzen, und wie man sich beispielsweise behelfen kann, wenn andere, nicht in SAS realisierte Verfahren verwendet werden sollen. Im Einzelnen werden folgende Punkte beschrieben: Eine kurze Beschreibung der Prozeduren UNIVARIATE bzw. CAPABILITY und ihrer Optionen, welche die Bestimmung robuster Lage- bzw. Streuungsparameter erlauben. Die SAS Prozedur ROBUSTREG für robuste lineare Regression, deren Möglichkeiten und was bei der Anwendung dieser Prozedur beachtet werden sollte. Robuste nichtlineare Regression bei Ausreißern in Y-Richtung (IRLS-Algorithmus).

Schlüsselwörter: Robuste Verfahren, lineare und nichtlineare Regression, nichtnormale Fehlerverteilung

1 Einleitung

Robuste Verfahren sind heutzutage aus der statistischen Anwendungspraxis nicht mehr wegzudenken. Viele Guidelines für die pharmazeutische Industrie verlangen sogar explizit die Anwendung robuster Verfahren bei der Datenanalyse. Die Bezeichnung „*ein robustes Verfahren*“ kann verschiedene Bedeutungen haben, je nachdem welche statistische Fragestellung vorliegt. Viele Annahmen in statistischen Verfahren sind oft nur Approximationen der Realität. In der Realität kann es schon öfters Werte geben, die weit ab von der Masse der restlichen Daten liegen. Ein robustes Verfahren bei der Bestimmung von Lage- oder Streuungskennzahlen kann demnach einfach nur eine Rechenvorschrift sein, welche solche extremen Werte (Ausreißer?) in der Stichprobe bei der Berechnung weniger berücksichtigt oder gar ausschließt. Robust kann aber auch ein Verfahren sein, wenn es Abweichungen von den Voraussetzungen für ein statistisches Verfahren toleriert, z.B. ist der t-Test robust gegen Abweichungen von der Annahme „die Daten sind normal verteilt“, solange die Verteilung der Stichprobe symmetrisch bleibt. In diesem Sinne gibt es viele Varianten für den Begriff „Robustheit“, die sich jeweils an der vorliegenden Fragestellung orientieren. Nicht alle Parameter einer Fragestellung müssen mit robusten Verfahren bearbeitet werden. Viele Modelle haben einen empirischen Charakter und entsprechen nur näherungsweise dem theoretischen Modell.
Im nächsten Kapitel soll zunächst gezeigt werden, wie man mit Hilfe von SAS robuste Kennzahlen für Lage- und Streuungsparameter über die Prozeduren UNIVARIATE

bzw. CAPABILITY erhält. In den nachfolgenden Abschnitten werden verschiedene Varianten für robuste Regressionsverfahren skizziert.

2 Robuste Kennzahlen

Die wichtigsten Kennzahlen zur Repräsentation einer Lage sind Mittelwert, Median oder Perzentile. Entsprechende robuste Lagemaße wären Median, Perzentile, winsorisierte oder getrimmte Mittelwerte. Mit Hilfe der Prozeduren `UNIVARIATE` oder `CAPABILITY` erhält man diese Varianten über die Schlüsselworte `MEDIAN`, `P1`, `P5`, `P10`, `P90`, `P95` und `P99` im OUTPUT-Statement. (Bei der Prozedur `MEANS` müssen diese Schlüsselworte als Option im Prozeduraufruf angegeben werden.)

Winsorisierte oder getrimmte Mittelwerte können mit den beiden Prozeduren `UNIVARIATE` oder `CAPABILITY` über die Optionen

`TRIM=k (TYPE=TWOSIDED)` für getrimmte, bzw.
`WINSOR=k (TYPE=TWOSIDED)` für winsorisierte Werte

bestimmt werden.

Getrimmte Mittelwerte werden berechnet, nachdem die k größten und kleinsten Werte der Stichprobe eliminiert wurden. Winsorisieren heißt, die k „extremen“ Werte der Stichprobe durch den darauf in der Stichprobe folgenden Wert zu ersetzen und anschließend den Mittelwert zu berechnen. `TYPE=TWOSIDED` kann dabei durch `TYPE=ONESIDED` ersetzt werden.

In SAS verfügbare robuste Streuungsmaße sind Interquartile Range, Gini’s Mean Difference, Median absolute Abweichung (MAD) sowie die Streuungskennzahlen von Rousseeuw und Croux (Rousseeuw and Croux (1993)) welche ohne ein Lagemaß auskommen. Sie alle werden über die Option `ROBUSTSCALE` in den Prozeduren `UNIVARIATE` oder `CAPABILITY` angefordert.

3 Robuste Lineare Regression

Ziel eines robusten Regressionsverfahren ist es in der Regel, stabile Ergebnisse zu erhalten, auch wenn Ausreißer die zugrundeliegende funktionale Beziehung verfälschen. Historisch gesehen gibt es drei Klassen von Ausreißern:

- Ausreißer in Y-Richtung
- Multivariate Ausreißer im Kovariablenraum (X-Raum), sog. Leverage-Punkte, und
- Ausreißer in beiden Richtungen (X-Richtung und Y-Richtung).

Für lineare Regressionsbeziehungen steht die Prozedur `ROBUSTREG` zur Verfügung, welche in ihrer Syntax stark an die Prozedur `REG` angelehnt ist, aber in ihren Optionen bei weitem nicht so viele Möglichkeiten bietet und sich nur auf robuste Methoden konzentriert. Die Prozedur `ROBUSTREG` unterstützt vier robuste Methoden:

- M-Schätzung (Huber, 1973), gut geeignet für Kontamination in Y-Richtung, dagegen weniger gut bei Vorliegen von Leverage-Punkten,
- Least Trimmed Squares (LTS) (Rousseeuw, 1984), ein Verfahren mit einem hohen Breakdown-Punkt, wurde von Rousseeuw and Van Driessen (1999) verbessert: FAST-LTS,
- S-Schätzung (Rousseeuw and Yohai, 1984), gleicher Breakdown-Punkt, aber höhere Effizienz, und
- MM Schätzung (Yohai, 1987), kombiniert Breakdown-Verfahren mit M-Schätzung, hat die High-Breakdown-Eigenschaft und eine höhere statistische Effizienz als die S-Schätzung.

Beispiel `ROBUSTREG` Syntax:

```
PROC ROBUSTREG DATA=dataset ;
MODEL objective = variable list
/ DIAGNOSTICS LEVERAGE ;
OUTPUT OUT=RobOut r=resid sr=stdres;
RUN ;
```

liefert eine M-Schätzung der Parameter „`variable list`“ mit der Default-Gewichtsfunktion Bisquare. Der Prozedur-Aufruf

```
PROC ROBUSTREG METHOD=M WF=Huber DATA=dataset ;
```

verwendet die Gewichtsfunktion `Huber`. Zahlreiche weitere Gewichtsfunktionen können für eine M-Schätzung verwendet werden:

Option Gewichtsfunktion (**Weight function**) für eine M-Schätzung:

`WF=`	`BISQUARE`	Default Parameter: 4.685(default)
	`ANDREWS`	Default Parameter: 1.339
	`CAUCHY`	Default Parameter: 2.385
	`FAIR`	Default Parameter: 1.4
	`HAMPEL`	Default Parameter: 2, 4, 8
	`HUBER`	Default Parameter: 1.345
	`LOGISTIC`	Default Parameter: 1.205
	`MEDIAN`	Default Parameter: 0.01
	`TALWORTH`	Default Parameter: 2.795
	`WELSCH`	Default Parameter: 2.985

Alle Gewichtsfunktionen benutzen standardmäßig einen Parameter (siehe obige Liste), der aber bei Bedarf geändert werden kann. Sollen in der Variablenliste Potenzen der

unabhängigen Variablen verwendet werden, wie etwa x^2, x^3, usw. so müssen diese in einem vorausgehenden DATA-Step neu gebildet werden, z.B.:

```
DATA new ;
SET old ;
x_2 = x*x ;
x_3 = x_2*x ;
…
RUN ;
```

Die anderen robusten Methoden können über die **Option**

```
METHOD =  M     (default)
          LTS
          S
          MM
```

angefordert werden. LTS und S verwenden Subsampling Methoden. Es macht keinen Sinn diese Verfahren bei unabhängigen Variablen zu verwenden, welche nur wenige Werteausprägungen (ungleich null) besitzen.
Eine weitere wichtige Option ist der zu verwendende Skalierungsparameter:

Option:

`SCALE=`	`MED`	(default)
	`TUKEY`	Default Parameter: 2.5
	`HUBER`	Default Parameter: 2.5
	`fix`	feste Zahl *k*

Die Skalierungsverfahren Tukey und Huber erfordern ebenfalls die Angabe eines Parameters, welcher bei Bedarf an die Situation angepasst werden kann.
Wichtige und nützliche Optionen im MODEL-Statement sind:

`DIAGNOSTICS`	für Ausreißer-Diagnostiken und
`LEVERAGE`	für die Analyse von Hebepunkten.

Die Option `OUTPUT` verwendet Schlüsselworte analog zur Prozedur `REG` wie z.B. `OUT=`*`newdata`* für eine Ausgabedatei. Die Option

`OUTLIER`	Variable „markiert“ Ausreißer
`LEVERAGE`	Variable „markiert“ Hebelpunkte

Vorsicht ist bei der Verwendung dieser Optionen geboten, da die Ausgabedatei dann mehr Beobachtungen enthält als die Eingangsdatei: Neben den „alten“ Daten werden neue Werte für die gefundenen Ausreißer hinzugefügt.

4 Robuste nichtlineare Schätzverfahren

Bevor die Vorgehensweise für robuste, nichtlineare Schätzverfahren skizziert wird, soll kurz das Prinzip einfacher robuster Verfahren (M-Schätzung) erläutert werden. Bei der Kleinste-Quadrate-Schätzung (OLS) werden die Residuenquadrate r_i^2 minimiert:

$$Q_{LS} = (1/2)\sum_{i=1}^{n} r_i^2 \rightarrow \min,$$

mit einem Skalierungsparameter

$$\sigma_{LS} = \sqrt{[(1/(n-p))Q_{LS}]}.$$

p ist dabei die Anzahl zu schätzender Parameter.

Hiervon ausgehend minimieren Huber-Typ-Schätzungen eine Funktion $\rho(.)$ der skalierten Residuen:

$$Q = \sum_{i=1}^{n} \rho(r_i/\sigma) \rightarrow \min .$$

Die Funktion $\rho(.)$ sollte dabei konvex sein. Die Parameterschätzungen erhält man als Lösung eines Systems von p Gleichungen der Form:

$$\sum_{i=1}^{n} \psi (r_i/\sigma) x_{ij} =0, j=1, \ldots, p$$

Die Lösung dieses Systems liefert ein IRLS-Verfahren (IRLS: iteratively reweighted Least Squares) mit einer Gewichtsfunktion $w(x)$, $w(x)=\psi(x)/x$. Die Funktion $\psi(x)$ ist die Ableitung der Funktion $\rho(x)$. Allgemein (wie in der folgenden Problemstellung) verwendet man Quasi-Likelihood-Verfahren (siehe z.B. Wedderburn (1974) bzw. McCullagh (1983)), um die Parameter der Funktion und den Skalierungsparameter gemeinsam zu schätzen.
Mit dieser Vorgehensweise lassen sich auch für zahlreiche nichtlineare Funktionen M-Schätzer finden, wie z.B. Funktionen der Art

$$y= A_{lower} + (A_{upper} - A_{lower}) * G(x)$$

wobei A_{lower} bzw. A_{upper} zu schätzende Konstanten sind und $G(x)$ eine Funktion der Art

$$G(x)=1/[1+EXP(ax+b))]$$

oder

$$G(x) = \Phi(ax+b)$$

oder aber auch y sich als eine Bateman-Funktion darstellen lässt. Wichtig wäre stets nur, dass die Ausreißer nur in Y-Richtung vorkommen können. Ein Algorithmus für eine M-

Schätzung der Parameter einer solchen Funktion könnte dann folgendermaßen aussehen:

1. Start mit einer normalen OLS-Schätzung,
2. Schätze die Gewichtsfunktion bzw. den Skalierungsparameter auf der Basis der vorherigen Schätzwerte,
3. Schätze die Kurvenparameter über eine gewichtete Kleinste-Quadrate-Methode
4. Iteriere die beiden Schritte 2 und 3.

In der Regel konvergiert dieser Algorithmus schon nach wenigen Iterationen.

Lineare Modelle sowie multiple lineare Regressionsmodelle, die nichtnormale Fehlerverteilungen aufweisen, können mit Hilfe der SAS-Prozedur NLMIXED robust geschätzt werden. Es können dabei Vorteile gegenüber einer M-Schätzung durchaus vorliegen. Für Details sei auf das Paper von Gilbert (2007) verwiesen.

Literatur

[1] Gilbert, S.A. and Chen, L.C. (2007), Using SAS Proc NLMIXED for Robust Regression. SAS Global Forum 2007, Orlando, *Proceedings*

[2] Huber, P.J. (1973), Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Stat.*, **1**, 799-821

[3] McCullagh, P. (1983), Quasi-Likelihood functions. *Ann. Stat.* **11**, 59-67.

[4] Rousseeuw, P. J. and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*. **88**, 1273 - 1283.

[5] Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, **79**, 871-880.

[6] Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*. Wiley-Interscience, New York

[7] Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223

[8] Rousseeuw, P.J. and Yohai, V. (1984), "Robust Regression by Means of S Estimators," in *Robust and Nonlinear Time Series Analysis*, ed. J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics, 26, New York: Springer-Verlag, 256-274.

[9] Wedderburn, R.W.M. (1974), Quasi-Likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439-47.

[10] Yohai V.J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *Annals of Statistics*, **15**, 642-656.