Mehr als linear oder logistisch – ausgewählte Möglichkeiten neuer Regressionsmethoden in SAS

Gerhard Svolba
SAS Austria
Mariahilfer Straße 116
A-1070 Wien
gerhard.svolba@sas.com

Zusammenfassung

Methoden zu Regressionsanalysen haben bei SAS eine lange Tradition. Genauso lang ist die Liste möglicher SAS Prozeduren für die Regressionsanalyse; REG, GLM, GENMOD, LOGISTIC, MIXED sind gute alte Bekannte für viele SAS User. SAS erweitert aber nach wie vor sein Angebot an Regressionsmethoden. So sind in den letzten Versionen einige neue Prozeduren hinzugekommen:
QUANTREG, QUANTSELECT und HPQUANTSELECT für die Quantile Regression; ADAPTIVEREG für multivariate adaptive Regression Splines, GLMSELECT für die Variablenselektion in linearen Modellen und HPGENSELECT für die Variablenselektion bei verallgemeinerten linearen Modellen, um nur einige zu nennen.
Dieser Beitrag gibt einen Überblick über die neuen Möglichkeiten und zeigt Anwendungsbeispiele mit SAS Code.

Schlüsselwörter: SAS/STAT, Variablenselektion, Multivariate Adpative Regression Splines, Quantile Regression, High Performance Analytics, PROC HPGENSELECT, PROC HPQUANTSELECT

1 Die Weiterentwicklung der SAS/STAT Software

1.1 Verfügbarkeit der SAS High Performance Prozeduren

Mit der SAS/STAT Version 13.2, die mit SAS 9.4M2 verfügbar ist, bietet SAS seinen Anwendern bereits 97 analytische Prozeduren an. Diese 97 Prozeduren setzen sich aus 86 „klassischen“ SAS Prozeduren zusammen, die im Laufe der Jahre in das SAS/STAT Modul hinzugefügt wurde. Weiters haben SAS Anwender seit der SAS/STAT Version 12.3, die ab SAS 9.4 verfügbar ist, Zugriff auf zusätzliche 11 High Performance Analytic Prozeduren (HPA-Prozeduren).

Diese HPA-Prozeduren sind grundsätzlich Teil von „SAS High Performance Statistics“ und bieten in diesem Modul die Möglich von enorm hoher Performance in einem verteilten Rechnersystem mit entsprechender Anzahl von CPUs und zugeteilten Hauptspeicher. Um die Verfügbarkeit der statistischen Methoden in diesen HPA-Prozeduren für

eine bereite Gruppe von Anwendern zu ermöglichen, sind diese HPA-Prozeduren auch in SAS/STAT verfügbar, sofern SAS/STAT auf einem Rechner im Single-Server Mode ausgeführt wird. Mit Single-Server Mode sind all jene Systeme gemeint, die kein verteiltes In-Memory System sind, also SAS Installation auf Desktop Workstations und Laptops bzw. auf klassischen Servern.

Die HPA-Prozeduren bieten auch auf einem nicht-verteilten System große Performancegewinne. Tabelle 1 zeigt für die lineare Regression und für die Quantil-Regression ausgewählte Prozeduren und welche Ansätze mit den jeweiligen Prozeduren verfolgt werden.

Tabelle 1: Überblick über Modellierungsmöglichkeiten ausgewählter Prozeduren

Ansatz	**Linear Regression**	**Quantile Regression**
Vielzahl von Parametern für die Analyse, Modelanpassung, Post-Hoc Analyse von Modellen, keine oder nur wenige Möglichkeiten der Variablenselektion	REG und GLM	QUANTREG
Vielzahl von Variablenselektionsmöglichkeiten, geringere Anzahl von Modell-Parametrisierungen	GLMSELECT	QUANTSELECT
High Performance Analytics, Analyse für Big Data mit hoher Performance, In-Memory Computing, Variablenselektion, weniger Optionen für Analyse und Ergebnisdarstellung	HPGENSELECT HPREG	HPQUANTSELECT

1.2 Erweitertes Methodenspektrum

Die Weiterentwicklung der SAS Software fokussiert sich aber nicht nur auf die Steigerung der Performance durch In-Memory Computing. SAS hat in den letzten Jahren jährlich eine neue Analytik-Version auf den Markt gebracht, in die eine Reihe neuer Features und neuer SAS Prozeduren eingeflossen sind.

Die Entwicklungsabteilung von SAS in Cary, North Carolina fokussiert sich in der Weiterentwicklung der SAS/STAT Software somit auf die Bereiche Performancesteigerung und Erweiterung des Methodenangebots.

Dieser Beitrag fokussiert sich auf drei Aspekte des erweiterten Methodenangebots in SAS/STAT.

- Es werden erweiterte Möglichkeiten in der Variablenselektion gezeigt.
- Die Quantil-Regression wird vorgestellt.
- Die Ideen und Beispiele der Multiple Adaptive Splines Regression werden gezeigt.

2 Variablenselektion in linearen und nicht-linearen Modellen mit PROC GLMSELECT und HPGENSELECT

2.1 Überblick

Prozeduren in SAS/STAT bieten folgende Möglichkeiten für die Variablenselektion:

- **Variablen-Selektion mit der Stepwise Option im MODEL-Statement**
  die STEPWISE Option ist zum Beispiel in folgenden Prozeduren verfügbar: PROC REG, PROC LOGISTIC, PROC PHREG
  Die Stepwise Option ist aber z.B. nicht für General Linear Models (PROC GLM) oder verallgemeinerte lineare Modell (PROC GENMOD) und auch nicht in der Quantil-Regression (PROC QUANTREG) verfügbar.

- **Variablen-Selektion mit den <name>SELECT Prozeduren**
  Die GLMSELECT Prozedur bietet hier Möglichkeit der Variablenselektion in General Linear Models. Weiters steht die PROC QUANTSELECT zur Verfügung.

- **Variablen Selektion mit den HP-Prozeduren**
  Diese haben ein STEPWISE-Statement und keine STEPWISE Option im MODEL-Statement und sind z.B. in folgenden Prozeduren verfügbar: PROC HPREG, PROC HPLOGISTIC, PROC HPGENSELECT

2.2 Vergleich: PROC GLMSELECT und PROC HPGENSELECT

Aufgrund der Namensgebung der SAS Prozeduren ist es wichtig zwischen der Bedeutung von „General“ und „Generalized“ zu unterscheiden.

PROC GLMSELECT modelliert “**General Linear Models**” und ist vom her Ansatz mit PROC GLM und PROC REG zu vergleichen. Als Methoden für die Variablenselektion werden folgende Verfahren angeboten: FORWARD, BACKWARD, STEPWISE, LAR, LASSO, ELASTICNET.

PROC HPGENSELECT modelliert “**Generalized Linear Models**” und ist mit PROC GENMOD vergleichbar. PROC HPGENSELECT passt Standardmodelle der "Exponentialfamilie" an. Das sind z.B. folgende Verteilungen: NORMAL, POISSON, TWEEDIE, ZERO-INFLATED POISSON, NEGATIVE BINOMIAL.
PROC HPGENSELCT ermöglicht auch Variablenselektion nach der FORWARD, BACKWARD, STEPWISE Methode für diese Verfahren.

2.3 Beispiel für Variablenselektion mit der LASSO-Methode

Folgendes Beispielprogramm zeigt den Aufruf von PROC GLMSELECT für die LASSO-Methode. Ausgewählte Ausgabeobjekte sind in Abbildung 1 und 2 zu sehen.

```
PROC GLMSELECT DATA=BWEIGHT_TRAIN
               TESTDATA= BWEIGHT_TEST
               PLOTS=ALL;
 MODEL WEIGHT = BLACK MARRIED BOY MOMAGE MOMSMOKE
                CIGSPERDAY MOMWTGAIN VISIT MOMEDLEVEL
                /SELECTION=LASSO;
RUN;
```

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE
0	Intercept		1	509180.253	321292.085	318744.413
1	MomWtGain		2	508517.489	315949.216	312924.447
2	Black		3	508318.158	314301.734	311204.622
3	Married*Boy		4	508190.410	313220.718	310115.449
4	Married		5	507944.144	311223.473	308085.366
5	MomSmoke		6	506974.792	303720.226	300218.628
6	Married*Visit		7	506070.315	296877.131	292997.973
7	MomAge		8	505596.488	293317.140	289360.746
8	MomAge*MomEdLevel		9	505552.279	292917.048	288949.324
9	CigsPerDay		10	505521.947	292618.614	288645.859
10	Boy*Visit		11	505196.346	290178.864	286157.756
11		Married*Visit	10	505174.947	290100.840	286078.634
12	Boy		11	505171.544	289999.686	285970.724
13	Boy*MomWtGain		12	505146.398	289741.641	285690.779
14	Married*Visit		13	505093.522	289283.986	285191.412
15	MomSmoke*MomEdLevel		14	505021.888	288692.160	284539.951
16	Married*MomAge		15	504976.642	288290.933	284094.552
17	Married*MomSmoke		16	504913.118	287759.250	283529.499
18	MomAge*MomSmoke		17	504893.668	287543.983	283265.446
19	MomWtGain*Visit		18	504881.736*	287382.677	283058.251

* Optimal Value of Criterion

Abbildung 1: Selection Summary für Variablenselektion nach der LASSO-Methode mit PROC GLMSELECT

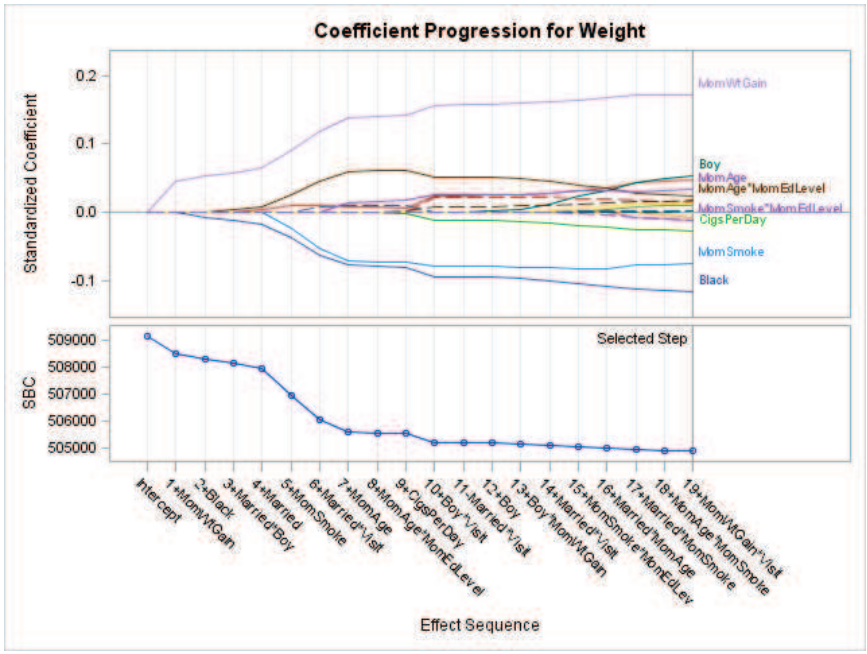

Abbildung 2: Graphische Darstellung der Veränderung der Koeffizienten in jedem Selektionsschritt

2.4 Beispiel für Variablenselektion in der Poisson-Regression

Folgendes Syntax-Beispiel zeigt den Aufruf der HPGENSELECT Prozedur für eine Stepwise Variablenselektion in der Poisson-Regression. Der Output ist in Abbildung 3 dargestellt.

```
proc hpgenselect data=patients_xt;
 class centnr treatment;
 model cnt_aes = treatment age breslow weight
                 stage secsurgyn centnr
                 /link=log distribution=poisson;
 selection method= stepwise;
run;
```

Selection Summary

Step	Effect Entered	Number Effects In	p Value
0	Intercept	1	.
1	CentNr	2	<.0001
2	BRESLOW	3	0.0021

Abbildung 3: Ergebnis der Stepwise Variablenselektion für die Poisson-Regression

3 Quantil-Regression

3.1 Idee der Quantil-Regression

Folgendes Zitat aus dem Paper von Chen [3] beschreibt die Idee der Quantil-Regression:

> *Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile.*

Wenn die Veränderung des Zusammenhangs der untersucht wird und durch den Regressionskoeffizienten ausgedrückt wird, vom Quantil abhängig ist, macht die Verwendung der Quantil-Regression Sinn.

Die Quantil-Regression eignet sich dort, wo Zusammenhänge für extreme Bereiche der Zielvariable analysiert werden sollen. Die Quantil-Regression bietet auch eine robuste Medianschätzung bei Ausreißern, ohne dass Verteilungsannahmen gemacht werden müssen. Die Quantil-Regression ist aber nicht äquivalent zu einer linearen Regression für bestimmte Segmente der Beobachtungen. Im diesem Fall würde nur eine Submenge an Beobachtungen mit deren Mittelwert modelliert werden. Bei der Quantil-Regression werden alle Beobachtungen verwendet und das jeweilige bedingte Quantil wird modelliert.

Abbildung 4 zeigt ein Beispiel der Quantil-Regression für unterschiedliche Quantile des Body-Mass-Index (BMI) über das Lebensalter. Es ist ersichtlich, dass die unterschiedlichen BMI-Quantile unterschiedliche Zusammenhangsverläufe über das Lebensalter haben.

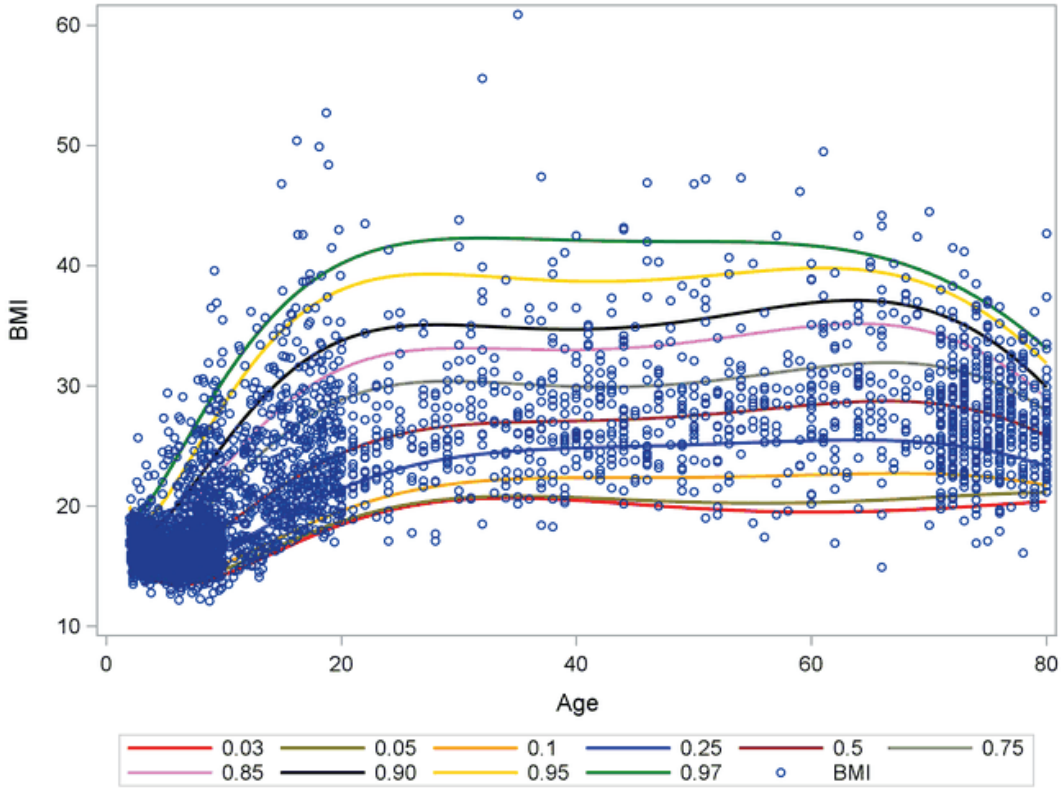

Abbildung 4: Quantil-Regression für unterschiedliche Quantile des Body-Mass-Index

3.2 Verfügbare Prozeduren für die Quantil-Regression in SAS

Folgendes Code Beispiel, mit dem auch die Analyse für die Graphik in Abbildung 4 erstellt wurde, ist dem Beispiel 13.2 der Online-Hilfe zur HPQUANTSELECT Prozedur in SAS/STAT entnommen. Die Quelldaten enthalten BMI und AGE. Zusätzlich werden folgende abgeleitete Variablen erstellt.

```
SqrtAge = sqrt(Age);
InveAge = 1/Age;
LogBMI  = log(BMI);
```

Die Analyse erfolgt mit der Prozedur HPQUANTSELECT.

```
proc hpquantselect data=BMIMen;
   model logBMI = InveAge SqrtAge Age SqrtAge*Age Age*Age
                  Age*Age*Age  / quantile=0.03 0.05 0.1 0.25 0.5
                                          0.75 0.85 0.90 0.95 0.97;
   code file='bmicode.sas';
   output out=Bmiout copyvars=(BMI Age) pred=P_LogBMI;
run;
```

Die Modell-Anweisung enthält Interaktionen und quadratische und kubische Terme von AGE. Beachten Sie, dass die gewünschten Quantile als Option in der Model-Anweisung angegeben werden.

Die Code-Anweisung erlaubt es, die Parameter des finalen Modell-Ergebnisses als Datastep Scorecode in ein SAS-File zu schreiben. Beachten Sie auch, dass bei den HP-Prozeduren die Variablen, welche im OUTPUT Dataset enthalten sein sollen, mit der COPYVARS Option explizit spezifiziert werden müssen.

Während die HPQUANTREG Prozedur High-Performance Analyse und Variablenselektion erlaubt, bietet die QUANTREG Prozedur detailliertere Analysemöglichkeiten (vgl. Tabelle 1). Folgendes Syntax-Beispiel zeigt, wie ein Quantil-Plot, wie in Abbildung 5 dargestellt, erstellt werden kann.

```
proc quantreg data=dat ci=resampling;
ods select quantplot;
model Sales=Rabatt_Menge
             /quantile=(0.1 to 0.9 by 0.05)
              plot=(quantplot /unpack ols)
              seed=1268 ;
run;
```

Die gewünschten Quantile werden mit der QUANTILE= Option spezifiziert. Für den Quantil-Plot (QUANTPLOT) wurde zusätzlich die Option OLS angegeben. Dies erlaubt einen Vergleich der Koeffizienten der Quantil-Regressionen mit dem Kleinst-Quadrate-Schätzer. In Abbildung 5 ist der OLS-Schätzer als horizontale Linie dargestellt. Es ist zu erkennen, dass die Koeffizienten der Quantil-Regression sich über den Quantil-Bereich unterschiedlich sind.

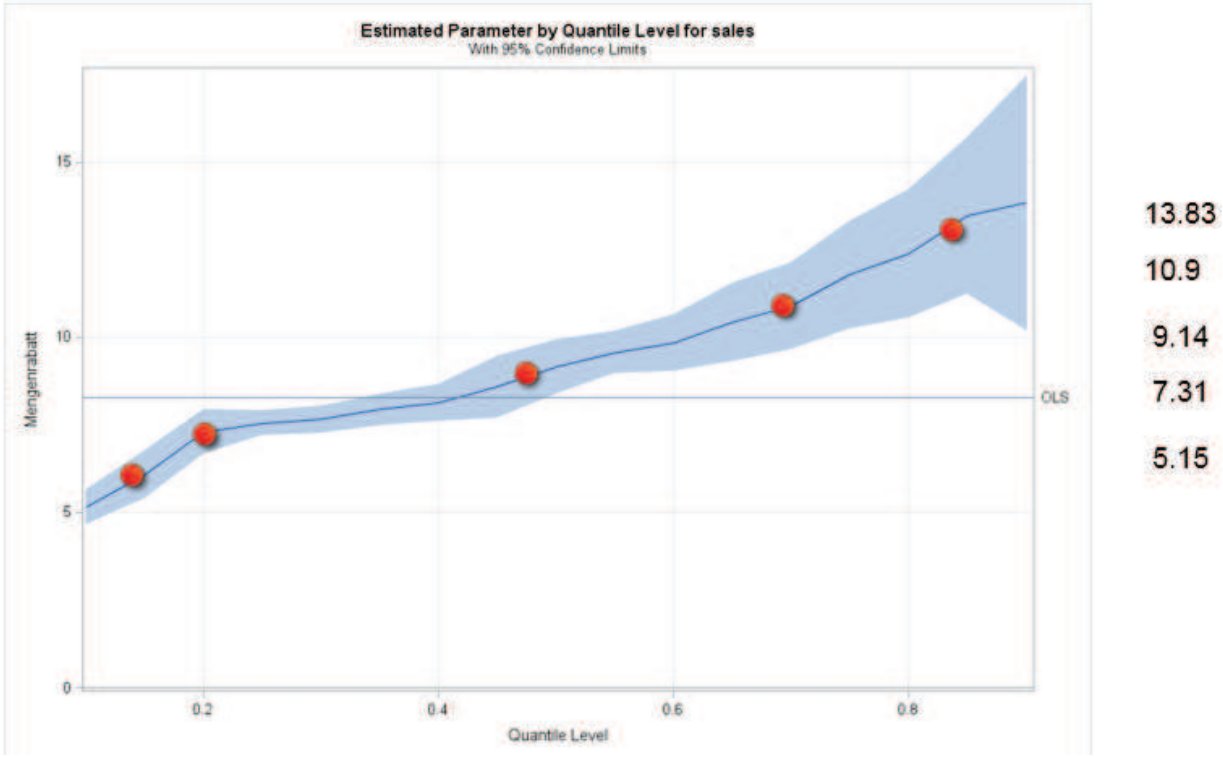

Abbildung 5: Quantilplot mit dem Kleinste-Quadrate-Schätzer

4 Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines erweitern lineare Modelle für die Analyse nicht-linearer Abhängigkeiten. Dabei werden die nicht-parametrische Regressions-Technik „Regressions-Splines“ und die Variablenselektion kombiniert.

Im ersten Schritt werden aus den Daten die Knotenpunkte automatisch ermittelt und die jeweiligen Modell-Parameter geschätzt. Im zweiten Schritt werden die Modelle durch Selektionstechniken reduziert.

Multivariate Adaptive Regression Splines stehen in SAS mit der ADAPTIVEREG Prozedur zur Verfügung. Weitere Möglichkeiten Nicht-Parametrischer Regression in SAS/STAT bieten die PROC GAM und die PROC LOESS.

Mit folgenden Code Beispiel kann ein Multivariate Adaptive Regression Splines Modell in SAS erstellt werden.

```
proc adaptivereg plots=all details=bases;
   model y = x;
run;
```

Abbildung 6 zeigt, wie die x/y Punktwolke durch die jeweiligen linearen Abschnitte modelliert wird.

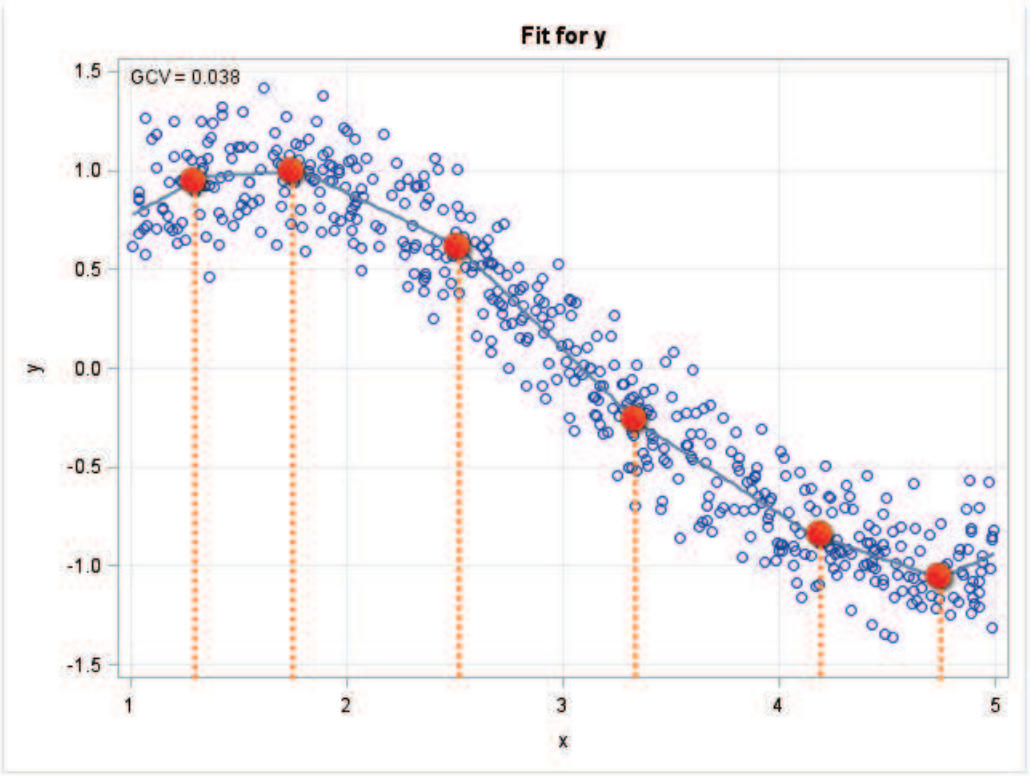

Abbildung 6: Knotenpunkte als Output der AdaptiveReg Prozedur.

Die Knotenpunkte werden von der AdaptiveReg Prozedur automatisch gefunden, durch eine Variablenselektion reduziert und die linearen Zusammenhänge in jedem Abschnitt entsprechend geschätzt.

Aufgrund der Tatsache, dass die Knotenpunkte nicht vorab spezifiziert werden müssen, eignet sich die AdpativeReg Prozedur auch für das Aufdecken von strukturellen Veränderungen in Verlaufsdaten.

5 Zusammenfassung

SAS erweitert laufend sein Angebot an analytischen Prozeduren. In der Version SAS 9.4M2 stehen mit SAS/STAT 13.2 bereits 97 statistische Prozeduren bereit. Neue Features beinhalten zum Beispiel die Multivariate Adaptive Regression Splines, die Quantil-Regression und die Variablenselektion für verallgemeinerte lineare Modelle.

Danksagung

Mihai Paunescu für die Bereitstellung der Inhalte seines Vortrags vom SAS Club 2014.

Literatur

[1] Cohen R., Rodriguez R.: High-Performance Statistical Modeling; SAS Global Forum 2013, Paper 401.
http://support.sas.com/resources/papers/proceedings13/401-2013.pdf

[2] Cohen R.: Applications of the GLMSELECT Procedure for Megamodel Selection; SAS Global Forum 2009, Paper 259:
https://support.sas.com/resources/papers/proceedings09/259-2009.pdf

[3] Chen C.: An Introduction to Quantile Regression and the QUANTREG Procedure; SUGI 2005, Paper 213: http://www2.sas.com/proceedings/sugi30/213-30.pdf

[4] Kuhfeld W., Cai W.: Introducing the New ADAPTIVEREG Procedure for Adaptive Regression; SAS Global Forum 2013,
Paper 457: http://support.sas.com/resources/papers/proceedings13/457-2013.pdf

Bei Fragen und Kommentaren können Sie sich gerne an den Autor wenden:
sastools.by.gerhard@gmx.at , bzw. die Seite
http://www.sascommunity.org/wiki/Gerhard_Svolba besuchen.