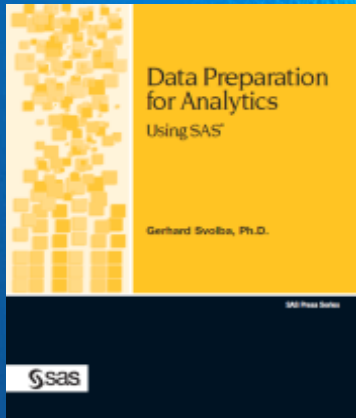


Die Vortragsfolien sind bereits online!
Google: *gerhard sas samples*

MEHR ALS LINEAR ODER LOGISTISCH – AUSGEWÄHLTE MÖGLICHKEITEN NEUER REGRESSIONSMETHODEN IN SAS

DR. GERHARD SVOLBA
COMPETENCE CENTER ANALYTICS
HANNOVER, 27. MÄRZ 2015

(In Zusammenarbeit mit
Dr. Mihai Paunescu)



EINLEITUNG DAS ERWARTET SIE IN MEINEM VORTRAG

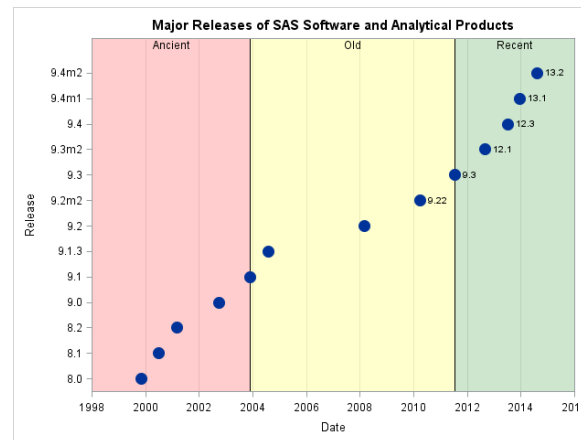
- SAS® STAT Procedures und High-Performance Procedures für die Regressionsanalyse
- Variablenselektion in linearen und nicht-linearen Modellen mit PROC GLMSELECT und HPGENSELECT
- Quantils-Regression mit PROC QUANTREG, QUANTSELECT und HPQUANTSELECT
- Multivariate Adaptive Regression Splines mit PROC ADAPTIVEREG

SAS® STAT PROCEDURES UND HIGH-PERFORMANCE PROCEDURES FÜR DIE REGRESSIONSANALYSE



86	(klassische) Procedures
11	High-Performance-Analytic Procedures
9429	Seiten Dokumentation

SAS Release	SAS/STAT Release	Release Date
9.4m2	13.2	05AUG14
9.4m1	13.1	15DEC13
9.4	12.3	10JUL13
9.3m2	12.1	29AUG12
9.3	9.3	12JUL11
9.2m2	9.22	01APR10
9.2		01MAR08
9.1.3		01AUG04
9.1		01DEC03
9.0		01OCT02
8.2		01MAR01
8.1		01JUL00
8.0		01NOV99



Kopiert und adaptiert aus:

<http://blogs.sas.com/content/iml/2013/08/02/how-old-is-your-version-of-sas-release-dates-for-sas-software/>

ADAPTIVEREG

fits multivariate adaptive regression spline models. This is a nonparametric regression technique that combines both regression splines and model selection methods. PROC ADAPTIVEREG produces parsimonious models that do not overfit the data and thus have good predictive power. PROC ADAPTIVEREG supports CLASS variables. For more information, see [Chapter 25: The ADAPTIVEREG Procedure](#).

CATMOD

analyzes data that can be represented by a contingency table. PROC CATMOD fits linear models to functions of response frequencies, and it can be used for linear and logistic regression. PROC CATMOD supports CLASS variables. For more information, see [Chapter 8: Introduction to Categorical Data Analysis Procedures](#), and [Chapter 32: The CATMOD Procedure](#).

GAM

fits generalized additive models. Generalized additive models are nonparametric in that the usual assumption of linear predictors is relaxed. Generalized additive models consist of additive, smooth functions of the regression variables. PROC GAM can fit additive models to nonnormal data. PROC GAM supports CLASS variables. For more information, see [Chapter 41: The GAM Procedure](#).

GENMOD

fits generalized linear models. PROC GENMOD is especially suited for responses that have discrete outcomes, and it performs logistic regression and Poisson regression in addition to fitting generalized estimating equations for repeated measures data. PROC GENMOD supports CLASS variables and provides Bayesian analysis capabilities. For more information, see [Chapter 8: Introduction to Categorical Data Analysis Procedures](#), and [Chapter 43: The GENMOD Procedure](#).

GLIMMIX

uses likelihood-based methods to fit generalized linear mixed models. PROC GLIMMIX can perform simple, multiple, polynomial, and weighted regression, in addition to many other analyses. PROC GLIMMIX can fit linear mixed models, which have random effects, and models that do not have random effects. PROC GLIMMIX supports CLASS variables. For more information, see [Chapter 44: The GLIMMIX Procedure](#).

GLM

uses the method of least squares to fit general linear models. PROC GLM can perform simple, multiple, polynomial, and weighted regression in addition to many other analyses. PROC GLM has many of the same input/output capabilities as PROC REG, but it does not provide as many diagnostic tools or allow interactive changes in the model or data. PROC GLM supports CLASS variables. For more information, see [Chapter 5: Introduction to Analysis of Variance Procedures](#), and [Chapter 45: The GLM Procedure](#).

GLMSELECT

performs variable selection in the framework of general linear models. PROC GLMSELECT supports CLASS variables (like PROC GLM) and model selection (like PROC REG). A variety of model selection methods are available, including forward, backward, stepwise, LASSO, and least angle regression. PROC GLMSELECT provides a variety of selection and stopping criteria. For more information, see [Chapter 48: The GLMSELECT Procedure](#).

LIFEREG

fits parametric models to failure-time data that might be right-censored. These types of models are commonly used in survival analysis. PROC LIFEREG supports CLASS variables and provides Bayesian analysis capabilities. For more information, see [Chapter 13: Introduction to Survival Analysis Procedures](#), and [Chapter 57: The LIFEREG Procedure](#).

LOESS

uses a local regression method to fit nonparametric models. PROC LOESS is suitable for modeling regression surfaces in which the underlying parametric form is unknown and for which robustness in the presence of outliers is required. For more information, see [Chapter 59: The LOESS Procedure](#).

LOGISTIC

fits logistic models for binomial and ordinal outcomes. PROC LOGISTIC provides a wide variety of model selection methods and computes numerous regression diagnostics. PROC LOGISTIC supports CLASS variables. For more information, see [Chapter 8: Introduction to Categorical Data Analysis Procedures](#), and [Chapter 60: The LOGISTIC Procedure](#).

MIXED

uses likelihood-based techniques to fit linear mixed models. PROC MIXED can perform simple, multiple, polynomial, and weighted regression, in addition to many other analyses. PROC MIXED can fit linear mixed models, which have random effects, and models that do not have random effects. PROC MIXED supports CLASS variables. For more information, see [Chapter 65: The MIXED Procedure](#).

NLIN

uses the method of nonlinear least squares to fit general nonlinear regression models. Several different iterative methods are available. For more information, see [Chapter 69: The NLIN Procedure](#).

NLMIXED

uses the method of maximum likelihood to fit general nonlinear mixed regression models. PROC NLMIXED enables you to specify a custom objective function for parameter estimation and to fit models with or without random effects. For more information, see [Chapter 70: The NLMIXED Procedure](#).

ORTHOREG

uses the Gentleman-Givens computational method to perform regression. For ill-conditioned data, PROC ORTHOREG can produce more-accurate parameter estimates than procedures such as PROC GLM and PROC REG. PROC ORTHOREG supports CLASS variables. For more information, see [Chapter 72: The ORTHOREG Procedure](#).

PHREG

fits Cox proportional hazards regression models to survival data. PROC PHREG supports CLASS variables and provides Bayesian analysis capabilities. For more information, see [Chapter 13: Introduction to Survival Analysis Procedures](#), and [Chapter 73: The PHREG Procedure](#).

PLS

performs partial least squares regression, principal component regression, and reduced rank regression, along with cross validation for the number of components. PROC PLS supports CLASS variables. For more information, see [Chapter 76: The PLS Procedure](#).

PROBIT

performs probit regression in addition to logistic regression and ordinal logistic regression. PROC PROBIT is useful when the dependent variable is either dichotomous or polychotomous and the independent variables are continuous. PROC PROBIT supports CLASS variables. For more information, see [Chapter 81: The PROBIT Procedure](#).

QUANTREG

uses quantile regression to model the effects of covariates on the conditional quantiles of a response variable. PROC QUANTREG supports CLASS variables. For more information, see [Chapter 83: The QUANTREG Procedure](#).

PROCEDURES IN SAS® STAT

SAS® PROCEDURES FÜR REGRESSIONSANALYSEN (FORTS.)

QUANTSELECT	performs multiple regression analysis for multivariate time series dependent variables by using current and past vectors of dependent and independent variables as predictors, with vector autoregressive moving-average errors, and with modeling of time-varying heteroscedasticity. For more information, see Chapter 35: The VARMAX Procedure in <i>SAS/ETS 13.2 User's Guide</i> . provides variable selection for quantile regression models. Selection methods include forward, backward, stepwise, and LASSO. The procedure provides a variety of selection and stopping criteria. PROC QUANTSELECT supports CLASS variables. For more information, see Chapter 84: The QUANTSELECT Procedure .
REG	performs linear regression with many diagnostic capabilities. PROC REG produces fit, residual, and diagnostic plots; heat maps; and many other types of graphs. PROC REG enables you to select models by using any one of nine methods, and you can interactively change both the regression model and the data that are used to fit the model. For more information, see Chapter 85: The REG Procedure .
ROBUSTREG	uses Huber M estimation and high breakdown value estimation to perform robust regression. PROC ROBUSTREG is suitable for detecting outliers and providing resistant (stable) results in the presence of outliers. PROC ROBUSTREG supports CLASS variables. For more information, see Chapter 86: The ROBUSTREG Procedure .
RSREG	builds quadratic response-surface regression models. PROC RSREG analyzes the fitted response surface to determine the factor levels of optimum response and performs a ridge analysis to search for the region of optimum response. For more information, see Chapter 87: The RSREG Procedure .
SURVEYLOGISTIC	uses the method of maximum likelihood to fit logistic models for binary and ordinal outcomes to survey data. PROC SURVEYLOGISTIC supports CLASS variables. For more information, see Chapter 14: Introduction to Survey Sampling and Analysis Procedures , and Chapter 98: The SURVEYLOGISTIC Procedure .
SURVEYPHREG	fits proportional hazards models for survey data by maximizing a partial pseudo-likelihood function that incorporates the sampling weights. The SURVEYPHREG procedure provides design-based variance estimates, confidence intervals, and tests for the estimated proportional hazards regression coefficients. PROC SURVEYPHREG supports CLASS variables. For more information, see Chapter 14: Introduction to Survey Sampling and Analysis Procedures , Chapter 13: Introduction to Survival Analysis Procedures , and Chapter 100: The SURVEYPHREG Procedure .
SURVEYREG	uses elementwise regression to fit linear regression models to survey data by generalized least squares. PROC SURVEYREG supports CLASS variables. For more information, see Chapter 14: Introduction to Survey Sampling and Analysis Procedures , and Chapter 101: The SURVEYREG Procedure .
TPSPLINE	uses penalized least squares to fit nonparametric regression models. PROC TPSPLINE makes no assumptions of a parametric form for the model. For more information, see Chapter 103: The TPSPLINE Procedure .
TRANSREG	fits univariate and multivariate linear models, optionally with spline, Box-Cox, and other nonlinear transformations. Models include regression and ANOVA, conjoint analysis, preference mapping, redundancy analysis, canonical correlation, and penalized B-spline regression. PROC TRANSREG supports CLASS variables. For more information, see Chapter 104: The TRANSREG Procedure .
Several SAS/ETS procedures also perform regression. The following procedures are documented in the <i>SAS/ETS User's Guide</i> :	
ARIMA	uses autoregressive moving-average errors to perform multiple regression analysis. For more information, see Chapter 7: The ARIMA Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
AUTOREG	implements regression models that use time series data in which the errors are autocorrelated. For more information, see Chapter 8: The AUTOREG Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
COUNTREG	analyzes regression models in which the dependent variable takes nonnegative integer or count values. For more information, see Chapter 11: The COUNTREG Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
MDC	fits conditional logit, mixed logit, heteroscedastic extreme value, nested logit, and multinomial probit models to discrete choice data. For more information, see Chapter 18: The MDC Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
MODEL	handles nonlinear simultaneous systems of equations, such as econometric models. For more information, see Chapter 19: The MODEL Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
PANEL	analyzes a class of linear econometric models that commonly arise when time series and cross-sectional data are combined. For more information, see Chapter 20: The PANEL Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
PDLREG	fits polynomial distributed lag regression models. For more information, see Chapter 21: The PDLREG Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
QLIM	analyzes limited dependent variable models in which dependent variables take discrete values or are observed only in a limited range of values. For more information, see Chapter 22: The QLIM Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
SYSLIN	handles linear simultaneous systems of equations, such as econometric models. For more information, see Chapter 29: The SYSLIN Procedure in <i>SAS/ETS 13.2 User's Guide</i> .
VARMAX	












PROCEDURES IN SAS® STAT

HIGH PERFORMANCE PROCEDURES FÜR SAS® STAT SIND AB VERSION 9.4 (STAT 12.1) OHNE AUFPREIS INKLUDIERT

SAS Release	SAS/STAT Release	Release Date
9.4m2	13.2	05AUG14
9.4m1	13.1	15DEC13
9.4	12.3	10JUL13
9.3m2	12.1	29AUG12
9.3	9.3	12JUL11
9.2m2	9.22	01APR10
9.2		01MAR08
9.1.3		01AUG04
9.1		01DEC03
9.0		01OCT02
8.2		01MAR01
8.1		01JUL00
8.0		01NOV99



Ab dieser Version sind in SAS® STAT
die High-Performance-Analytics
Procedures im Single Server Modus
ohne Aufpreis inkludiert

- +  [The HPCANDISC Procedure](#)
- +  [The HPFMM Procedure](#)
- +  [The HPGENSELECT Procedure](#)
- +  [The HPLMIXED Procedure](#)
- +  [The HPLOGISTIC Procedure](#)
- +  [The HPNLMOD Procedure](#)
- +  [The HPPLS Procedure](#)
- +  [The HPPRINCOMP Procedure](#)
- +  [The HPQUANTSELECT Procedure](#)
- +  [The HPREG Procedure](#)
- +  [The HPSPLIT Procedure](#)

Single Server Modus: SAS mit Desktop Lizenz, SAS mit klassischem
Server (Windows, Unix, Linux, ...).

Nicht: Verteiltes System mit In-Memory Computing → eigene Lizenz

Table 5.1: Overview over Regression Procedures for a Linear and Quantile Regression

Approach	Linear Regression	Quantile Regression
Rich set of parameters for analysis and output, limited or no possibility to perform variable selection	REG and GLM	QUANTREG
Various variable selection methods, few options to parameterize the analysis and the output	GLMSELECT	QUANTSELECT
High Performance Analytics, can deal with large datasets at quick runtimes, variable selection, few options to parameterize the analysis and the output	HPGENSELECT HPREG	HPQUANTSELECT

Entnommen aus: *My Favorite Analytic Case Studies*, Gerhard Svolba, SAS Press, expected 2016/2017

VARIABLENSELEKTION IN LINEAREN UND NICHT-LINEAREN MODELLEN MIT PROC GLMSELECT UND HPGENSELECT



[Paper 401-2013: High-Performance Statistical Modeling](#)

[Paper 259-2009: Applications of the GLMSELECT Procedure for
Megamodel Selection](#)

- Variablen-Selektion mit der Stepwise Option im MODEL-Statement: z.B: in
 - PROC REG, PROC LOGISTIC, PROC GLMSELECT, PROC PHREG
 - Nicht vorhanden z.B: in PROC GENMOD, PROC GLM, PROC QUANTREG
- Hinweis: die HP-Procedures haben ein STEPWISE-Statement und keine STEPWISE Option im MODEL-Statement
 - Z.B: PROC HPREG, PROC HPLOGISTIC, PROC HPGENSELECT

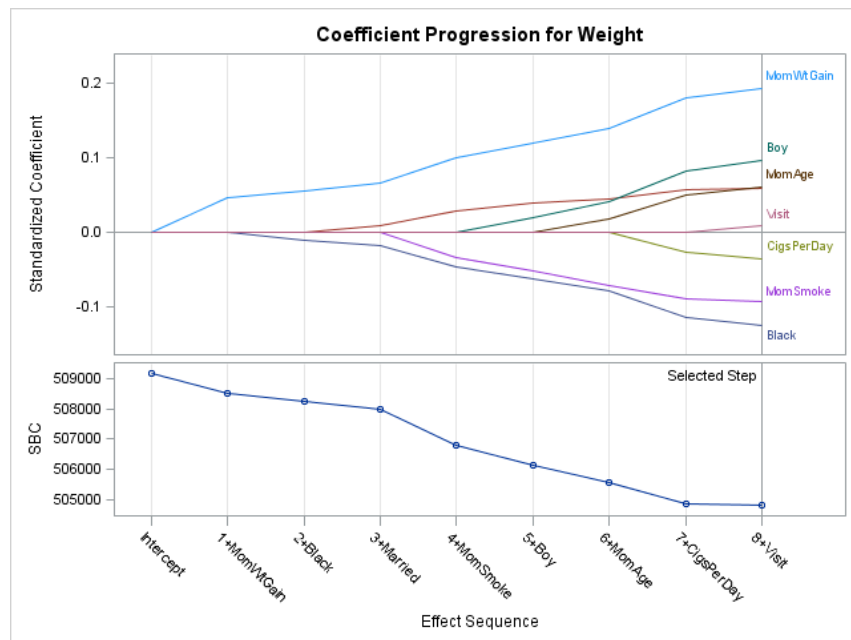
- PROC GLMSELECT
 - General Linear Models
 - Methoden: FORWARD, BACKWARD, STEPWISE, LAR, LASSO, ELASTICNET
 - Ähnlich zu PROC REG und PROC GLM
- PROC HPGENSELECT
 - Generalized Linear Models
 - Z.B: "Exponentialfamilie" NORMAL, POISSON, TWEEDIE, ZERO-INFLATED POISSON, NEGATIVE BINOMIAL
 - Methoden: FORWARD, BACKWARD, STEPWISE
 - Ähnlich zu PROC GENMOD

VARIABLEN- SELEKTION

VARIABLENSELEKTION NACH DER LAR (LEAST-ANGLE-REGRESSION) METHODE

```
proc glmselect data=bweight_train
    testdata= bweight_test
    plots=all;
    model weight = black married boy momage momsmoke
        cigsperday momwtgain visit momedlevel
        /selection=lar;
run;
```

LAR Selection Summary					
Step	Effect Entered	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	508270.928	313932.272	310818.509
3	Married	4	507985.483	311626.299	308411.865
4	MomSmoke	5	506802.888	302502.624	298567.810
5	Boy	6	506153.166	297568.850	293358.267
6	MomAge	7	505545.290	293020.737	288701.909
7	CigsPerDay	8	504864.137	288016.077	283586.450
8	Visit	9	504806.347*	287525.959	283140.546
* Optimal Value of Criterion					



Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	3294.159400
Black	1	-192.171885
Married	1	73.339732
Boy	1	109.250094
MomAge	1	6.042609
MomSmoke	1	-158.232646
CigsPerDay	1	-4.376877
MomWtGain	1	8.545629
Visit	1	7.464571

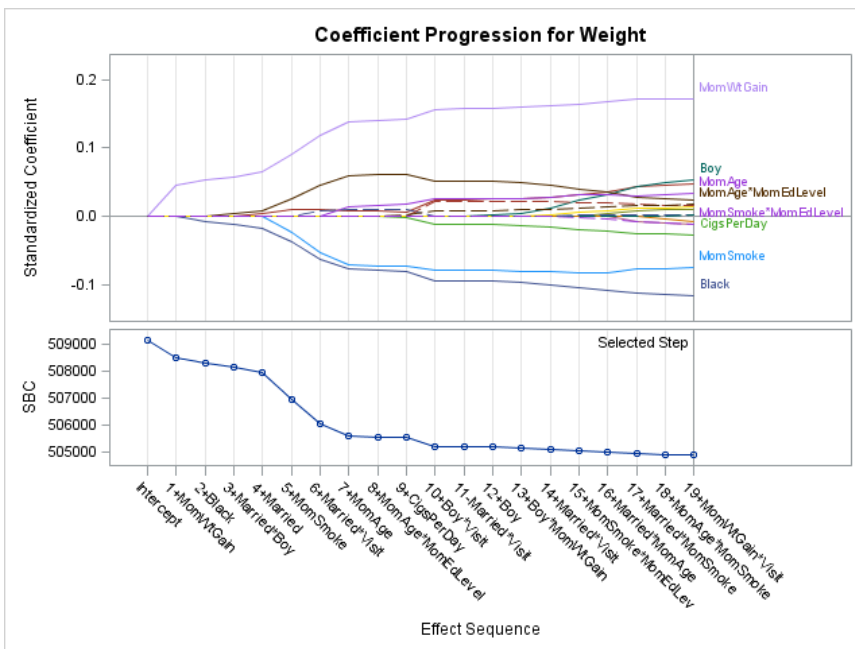
VARIABLEN-SELEKTION

VARIABLENSELEKTION NACH DER LASSO-METHODEN

```
proc glmselect data=bweight_train
    testdata= bweight_test
    plots=all;

    model weight = black|married|boy|momage|momsmoke|
        cigsperday|momwtgain|visit|momedlevel @2
        /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

Definition der Interaktionen

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	3321.643301
Black	1	-178.808648
Married	1	59.780012
Boy	1	61.004800
Married*Boy	1	28.275716
MomAge	1	3.332014
Married*MomAge	1	1.279508
MomSmoke	1	-127.950643
Married*MomSmoke	1	-25.507518
MomAge*MomSmoke	1	-2.212552
CigsPerDay	1	-3.311760
MomWtGain	1	7.633594
Boy*MomWtGain	1	0.876675
Married*Visit	1	1.148892
Boy*Visit	1	6.336393
MomWtGain*Visit	1	0.024404
MomAge*MomEdLevel	1	1.015330
MomSmoke*MomEdLevel	1	-10.941775

VARIABLEN- SELEKTION

STEPWISE POISSON-REGRESSION FÜR DIE ANZAHL ADVERSE EVENTS

```
proc hpgenselct 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

Model Information	
Data Source	WORK.PATIENTS_XT
Response Variable	CNT_AES
Class Parameterization	GLM
Distribution	Poisson
Link Function	Log
Optimization Technique	Newton-Raphson with Ridging

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.600631	0.334799	22.8567	<.0001
BRESLOW	1	-0.050654	0.016453	9.4783	0.0021
CentNr 1	1	0.316211	0.337624	0.8772	0.3490
CentNr 2	1	-0.556209	0.342332	2.6399	0.1042
CentNr 3	1	-0.271410	0.352837	0.5917	0.4418
CentNr 4	1	0.276425	0.340515	0.6590	0.4169
CentNr 5	1	0.202505	0.344037	0.3465	0.5561
CentNr 6	1	-0.073692	0.343588	0.0460	0.8302
CentNr 7	1	0.247420	0.344125	0.5169	0.4722
CentNr 9	1	0.039550	0.378986	0.0109	0.9169
CentNr 10	1	-0.118548	0.375270	0.0998	0.7521
CentNr 11	1	-0.697271	0.366435	3.6208	0.0571
CentNr 12	1	-0.000642	0.398991	0.0000	0.9987
CentNr 20	1	-2.143663	1.054227	4.1347	0.0420
CentNr 21	1	-0.582011	0.604982	0.9255	0.3360
CentNr 22	1	-0.220395	0.460197	0.2294	0.6320
CentNr 24	1	1.435491	0.404654	12.5844	0.0004
CentNr 25	1	-0.200264	0.531164	0.1421	0.7062
CentNr 26	1	-0.335495	0.382973	0.7674	0.3810
CentNr 27	1	-0.477202	0.486136	0.9636	0.3263
CentNr 31	1	-0.064334	0.506866	0.0161	0.8990
CentNr 40	0	0	.	.	.

QUANTILS-REGRESSION MIT PROC QUANTREG, QUANTSELECT UND HPQUANTSELECT



QUANTILS- REGRESSION

PROC QUANTREG UND DER MEDIAN

```
proc univariate data=dat;  
  ods select moments quantiles;  
  var sales;  
run;
```

Basic Statistical Measures			
Location		Variability	
Mean	241825.8	Std Deviation	823581
Median	75875.1	Variance	6,78E+16
Mode	.	Range	18605198
		Interquartile Range	168416

```
proc quantreg data=dat;  
  model Sales= / quantile=(0.5);  
run;
```

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > t
Intercept	1	75875.10	2.654.220	70.670.691	81.079.509	28.59	<.0001

QUANTILS- REGRESSION

PROC QUANTREG MIT EINER KLASSIFIKATIONSVARIABLE

```
proc quantreg data=dat ci=resampling ;
  class cust_grp;
  model Sales=cust_grp / quantile=(0.5) seed=12345;
run;
```

Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > t
Intercept		1	47335.60	3.071.371	41.313.237	53.357.963	15.41	<.0001
Cust_grp	I	1	88078.60	6.903.957	74.541.278	101615.92	12.76	<.0001
Cust_grp	O	1	21870.35	5.324.410	11.430.214	32.310.486	4.11	<.0001
Cust_grp	N	0	0.0000	0.0000	0.0000	0.0000	.	.

```
proc means data=dat P1 median P90;
  var sales;
  class Cust_grp;
run;
```

Analysis Variable : Sales		
Cust_grp	N Obs	Median
I	793	135414.2
O	1024	69062.08
N	1004	47323.3

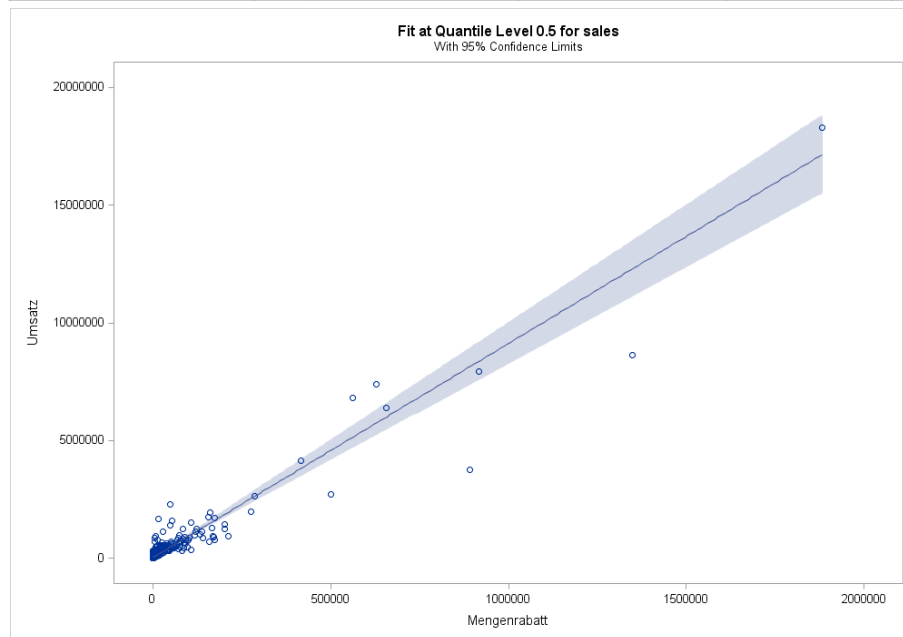
$$47335.6 + 88078.6 = 135414.2$$

QUANTILS- REGRESSION

PROC QUANTREG MIT EINER INTERVALLVARIABLE

```
proc quantreg data=dat ci=resampling plots=(fitplot);  
  where cust_grp='n';  
  model Sales=Rabatt_Menge / quantile=(0.5) seed=12345;  
run;
```

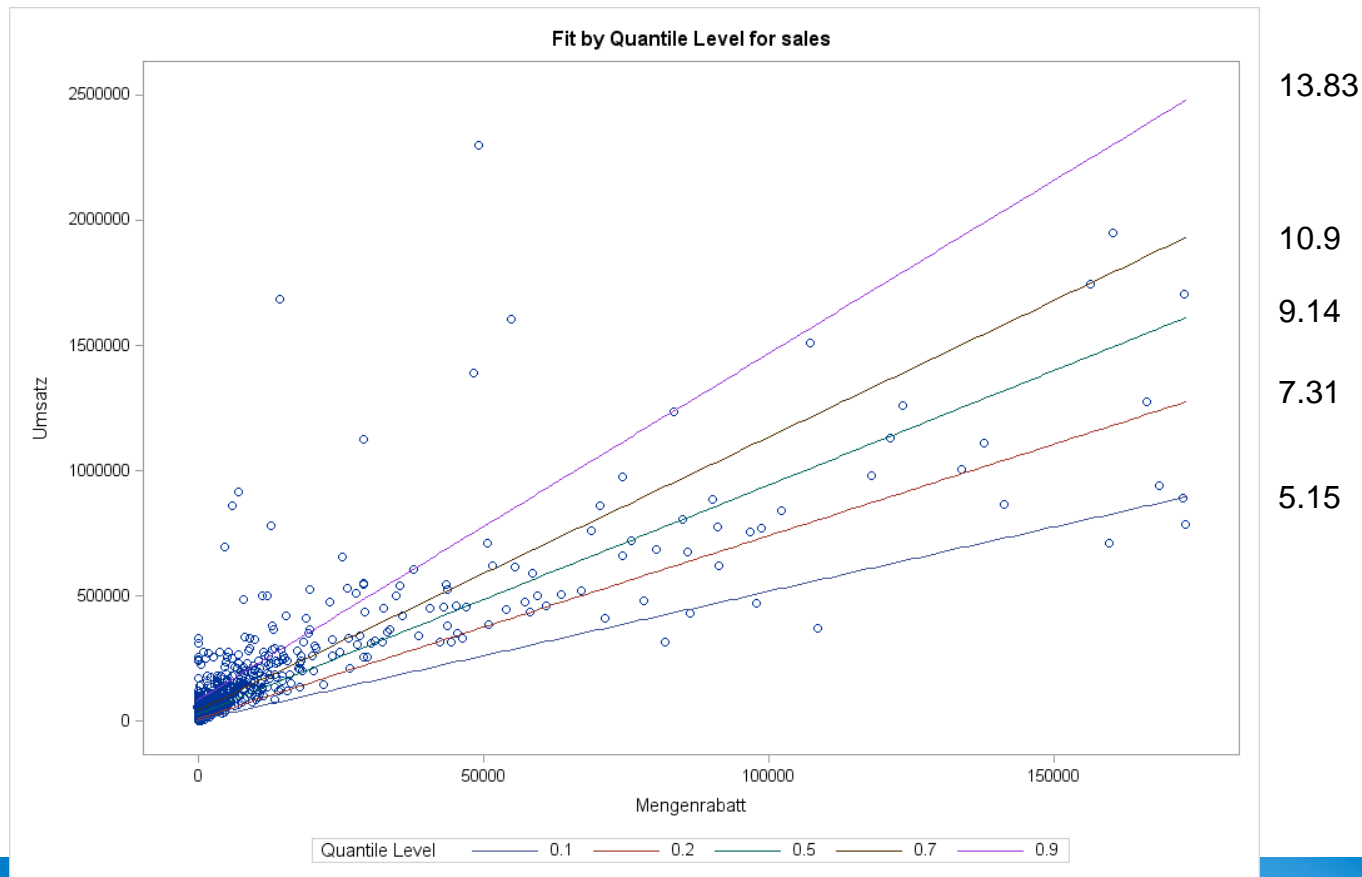
Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > t
Intercept	1	28552.1	5873.371	17026.581	40077.615	4.86	<.0001
Rabatt_Menge	1	9.1	0.4522	8.2082	9.9829	20.12	<.0001



QUANTILS- REGRESSION

PROC QUANTREG FÜR MEHRERE QUANTILE

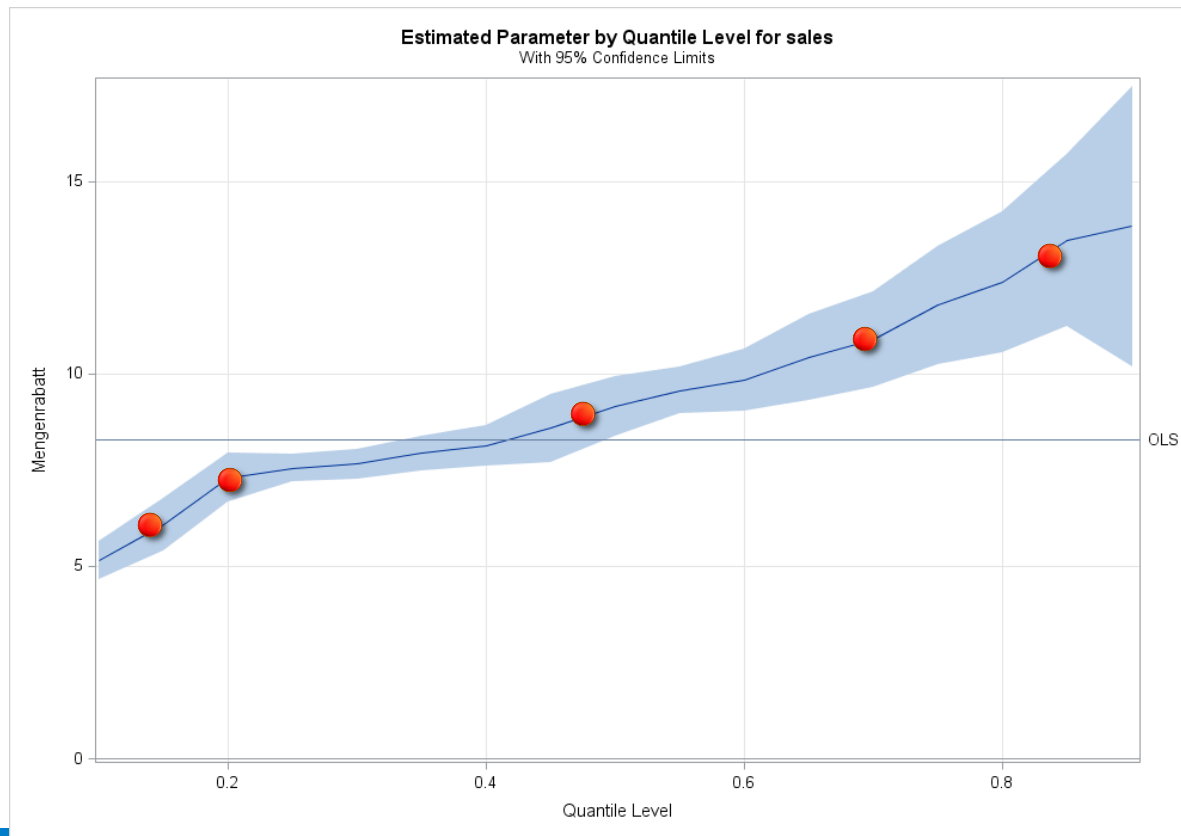
```
proc quantreg data=dat ci=resampling plots=(fitplot);  
ods select fitplot;  
where cust_grp='n' and Rabatt_Menge < 200000;  
model Sales=Rabatt_Menge / quantile=(0.1 0.2 0.5 0.7 0.9) seed=12345;  
run;
```



QUANTILS- REGRESSION

DER QUANTILE PROCESS PLOT

```
proc quantreg data=dat ci=resampling;  
ods select quantplot;  
where cust_grp='n' and Rabatt_Menge < 200000;  
model Sales=Rabatt_Menge  
/quantile=(0.1 to 0.9 by 0.05) plot=(quantplot /unpack ols) seed=1268 ;  
run;
```



13.83

10.9

9.14

7.31

5.15

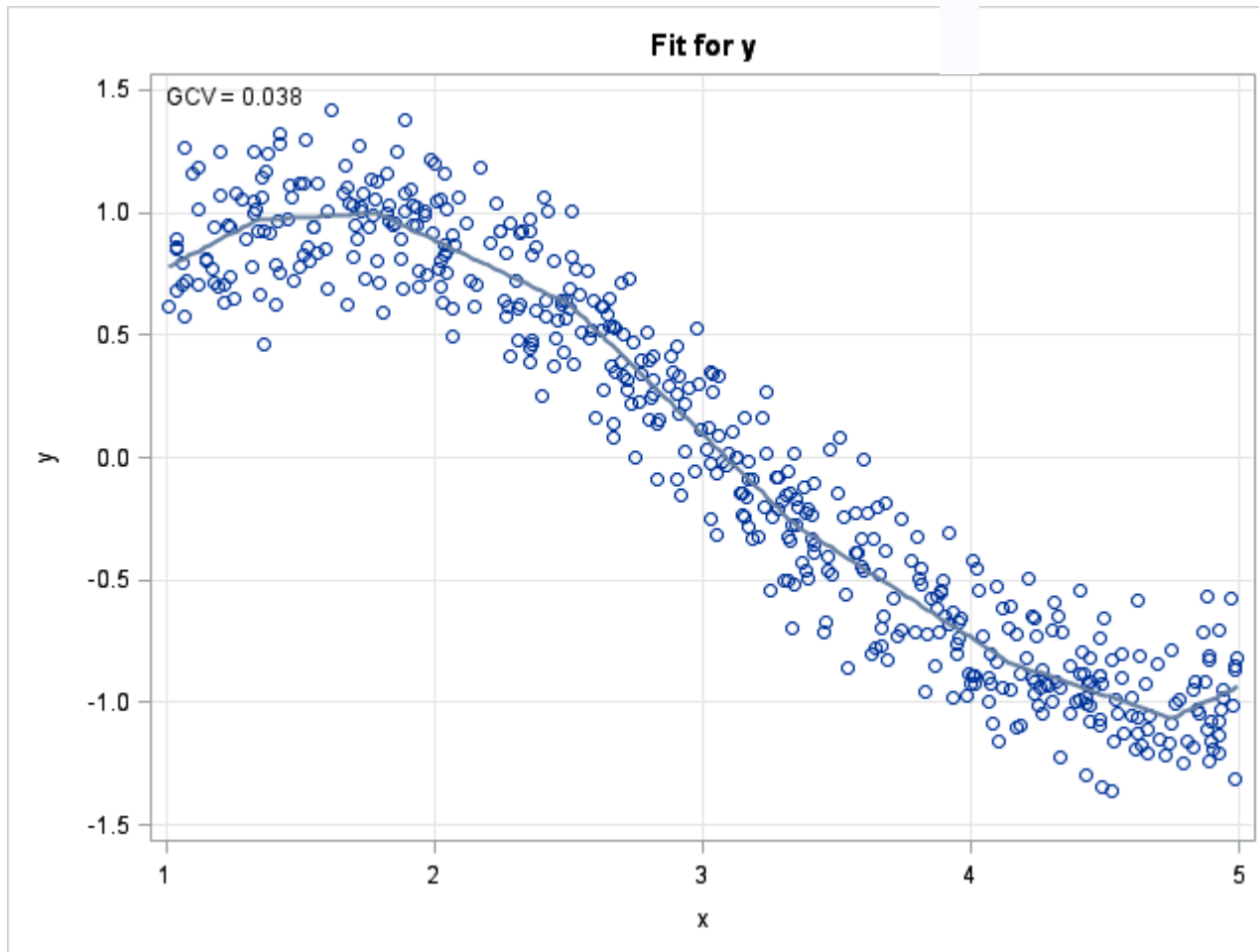
- *Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile (Chen, 2005)*
- Verwendung:
 - Zusammenhänge entdecken für extreme Bereiche der Zielvariable
 - Robuste Medianschätzung gegenüber Ausreißer, ohne Verteilungsannahmen
- Hinweise:
 - Ist nicht äquivalent zu linearen Regressionen für Segmente von Beobachtungen
 - HPQUANTSELECT ab SAS/STAT 13.2 → deutliche Performance-Verbesserung
- [Paper 213-30: An Introduction to Quantile Regression and the QUANTREG Procedure](#)

MULTIVARIATE ADAPTIVE REGRESSION SPLINES MIT PROC ADAPTIVEREG



ADAPTIVEREG

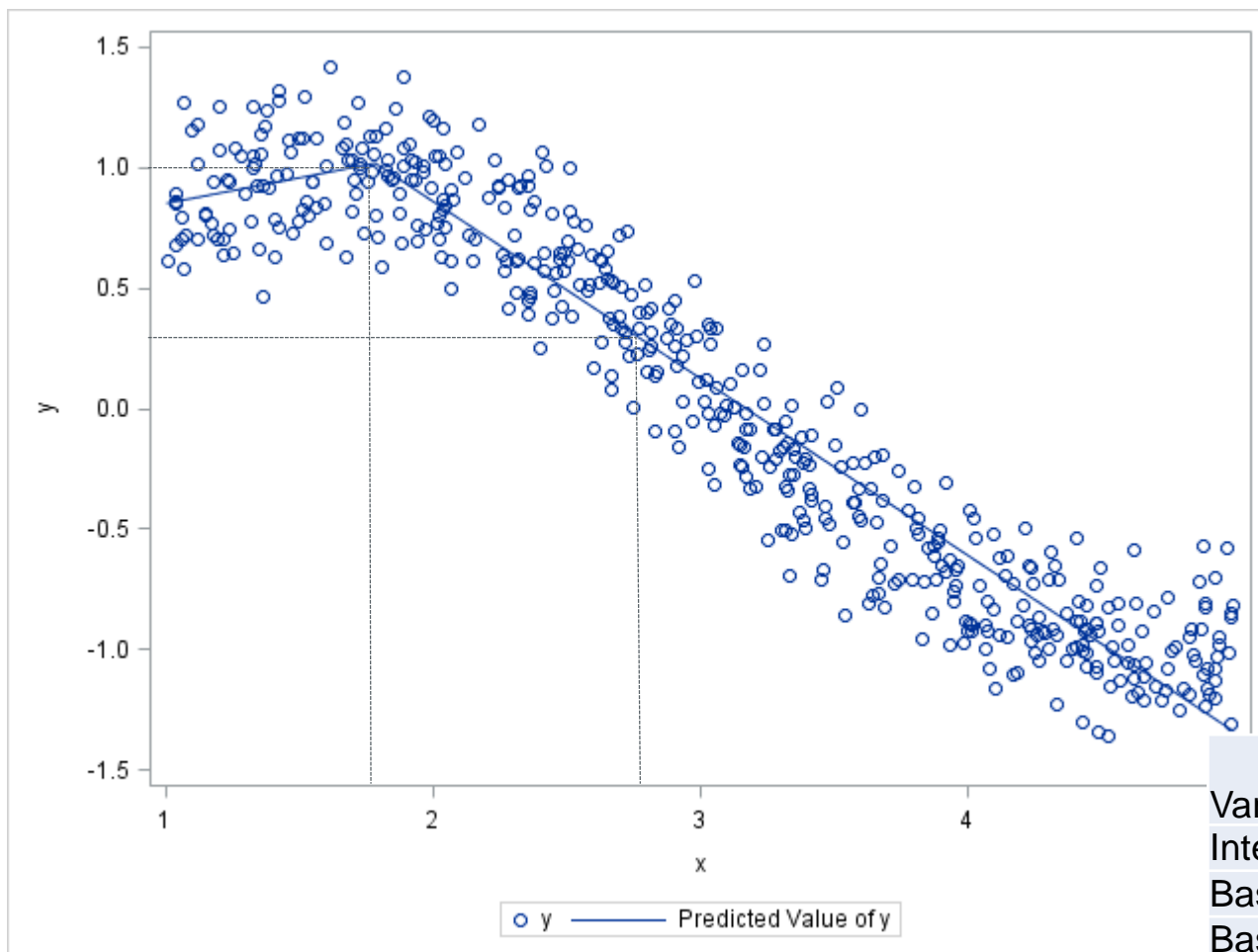
MULTIVARIATE ADAPTIVE REGRESSION SPLINES



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

ADAPTIVEREG GRUNDLAGEN

AUTOMATISCHE SELEKTION VON GEEIGNETEN STELLEN FÜR KNOTEN



Basis1	$\text{MAX}(x - 1.8, 0)$
--------	--------------------------

Basis2	$\text{MAX}(1.8 - x, 0)$
--------	--------------------------

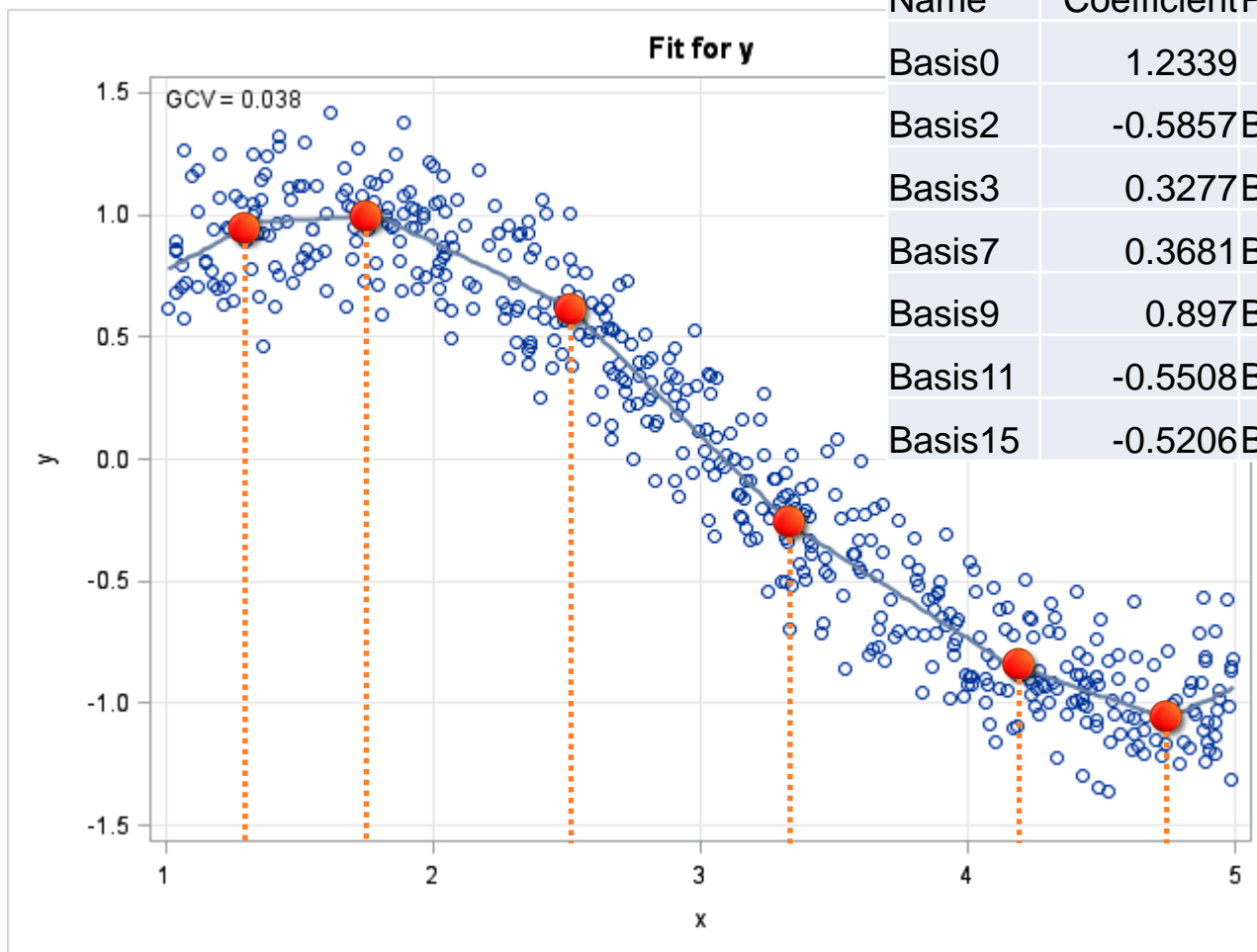
```
data ds2;
set ds;
Basis1 = max(x-1.8, 0);
Basis2 = max(1.8 - x, 0);
run;

proc reg data=ds2;
model y = basis1 basis2;
run;
```

Variable	Parameter Estimate	t Value	Pr > t
Intercept	1.02	52.04	<.0001
Basis1	-0.73	-67.82	<.0001
Basis2	-0.21	-3.31	0.001

ADAPTIVEREG GRUNDLAGEN

REDUKTION DES MODELS DURCH SELEKTIONSMETHODEN



Regression Spline Model after Backward Selection

Name	Coefficient	Parent	Variable	Knot
Basis0	1.2339		Intercept	
Basis2	-0.5857	Basis0	x	1.7865
Basis3	0.3277	Basis0	x	4.1447
Basis7	0.3681	Basis0	x	3.3424
Basis9	0.897	Basis0	x	4.7489
Basis11	-0.5508	Basis0	x	2.5061
Basis15	-0.5206	Basis0	x	1.3345

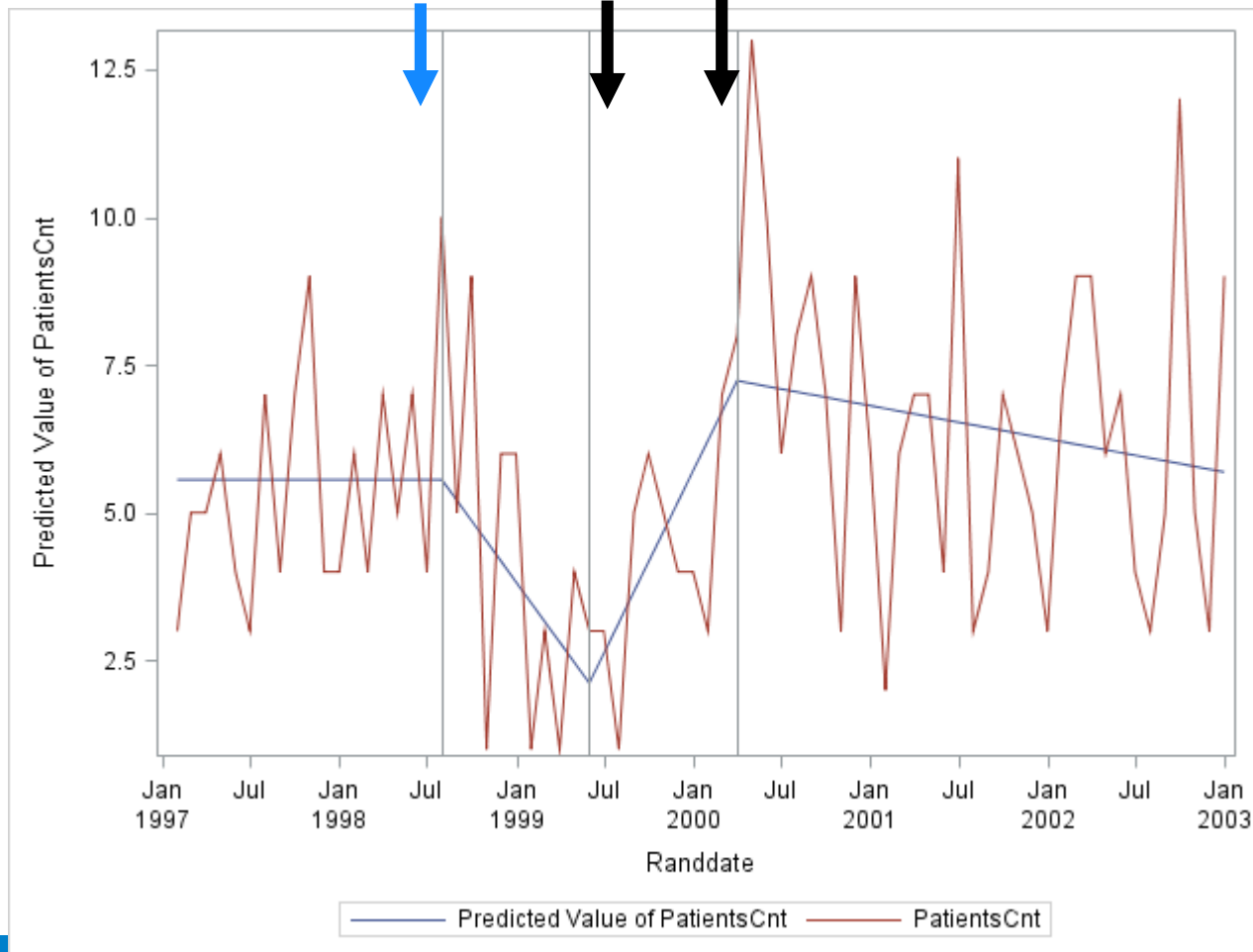
ADAPTIVEREG ANWENDUNGS- BEISPIEL

RECRUITMENT-MONITORING IN EINER LANGFRISTIGEN KLINISCHEN STUDIE BREAKPOINT DETECTION

Ein Patient
erkrankt schwer.
Möglicher Zusammenhang mit der
Medikation wird diskutiert

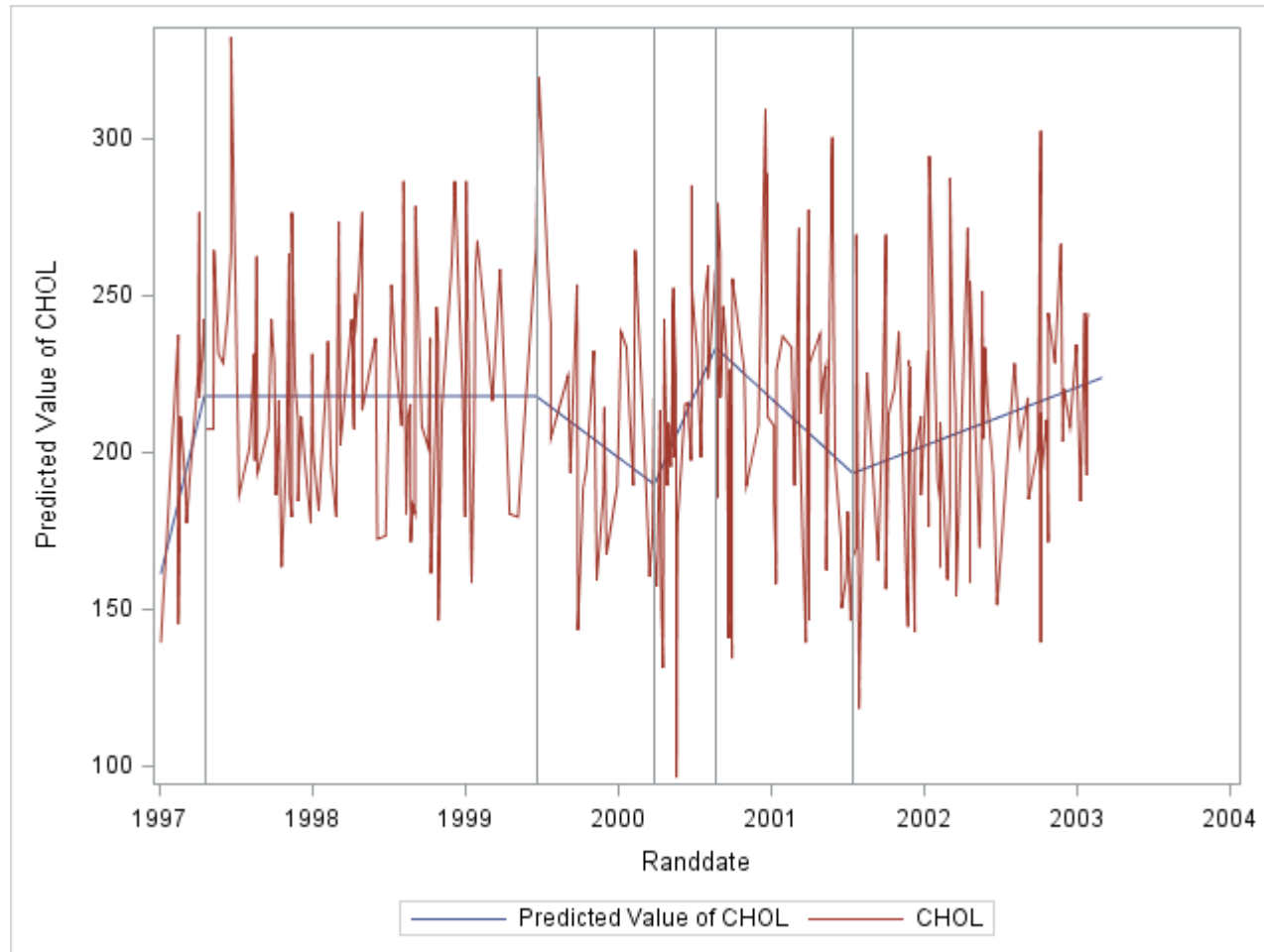
Studientreffen:
Aufruf zur
Priorisierung
der Studie

Rekrutierung
kommt
wieder in Fahrt



ADAPTIVEREG ANWENDUNGS- BEISPIEL

MONITORING DER BASELINE CHARAKTERISTIKA IN EINER LANGFRISTIGEN KLINISCHEN STUDIE



CODING TIPP

AUTOMATISCHES ANZEIGEN DER VERTIKALEN REFERENZ-LINIEN BEI DEN JEWEILIGEN BREAKPOINTS (3 SCHRITTE)

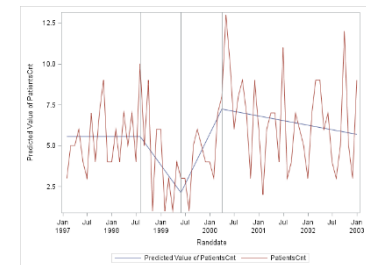
```
proc adaptivereg data=patients_1997_2002
                    plots=all details=bases ;
    format randdate date9.;
    model PatientsCnt = randdate / maxbasis=100;
    output out=recruit_adpt predicted=pred;
    ods output BWDPParams=BWDPParams;
run;
```

```
filename reflines 'c:/tmp/reflines.sas';
data _NULL_;
    set bwdparams;
    where upcase(variable) eq upcase('randdate');
    format knot 8.;
    file reflines;
    put @04 "refline " knot " / axis = x;";
run;
```

```
proc sgplot data=recruit_adpt;
    series x=randdate y=pred;
    series x=randdate y=PatientsCnt;
    %include reflines;
run;
```

	Name	Coefficient	Parent	Variable	Knot
1	Basis0	5.5580		Intercept	—
2	Basis1	0.02808	Basis0	Randdate	14396
3	Basis3	-0.01830	Basis0	Randdate	14701
4	Basis5	-0.01131	Basis0	Randdate	14092

```
refline 14396 / axis = x;
refline 14701 / axis = x;
refline 14092 / axis = x;
```



ADAPTIVEREG ZUSAMMENFASSUNG

- Erweitert lineare Modelle für die Analyse nicht-linearer Abhängigkeiten
- Nicht-parametrische Regressions-Technik die Regressions-Splines und Variablenselektion kombiniert.
 - Knotenpunkte werden automatisch ermittelt
 - Modelle werden durch Selektionstechniken reduziert
- Weitere Möglichkeiten Nicht-Parametrischer Regression in SAS/STAT: PROC GAM, PROC LOESS
- [Paper 457-2013: Introducing the New ADAPTIVEREG Procedure](#)

ZUSAMMENFASSUNG



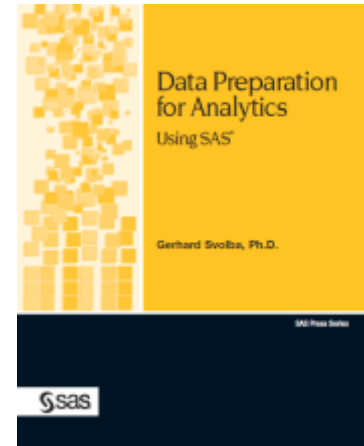
ZUSAMMENFASSUNG Links und Referenzen

- [Paper 401-2013: High-Performance Statistical Modeling](#)
- [Paper 259-2009: Applications of the GLMSELECT Procedure for Megamodel Selection](#)
- [Paper 213-30: An Introduction to Quantile Regression and the QUANTREG Procedure](#)
- [Paper 457-2013: Introducing the New ADAPTIVEREG Procedure](#)
- [Blogbeitrag](#) zur KSFE im Mehr-Wissen Blog von SAS
- Svolba, My Favorite Business Case Studie With SAS, SAS Press, expected 2016/2017
- http://www.sascommunity.org/wiki/New_Features_in_SAS_STAT_9.2
- [WBS-Seminar 2012](#)



Data Quality for Analytics Using SAS SAS Press 2012

http://www.sascommunity.org/wiki/Data_Quality_for_Analytics



Data Preparation for Analytics Using SAS SAS Press 2006

http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics



Gerhard Svolba

Analytic Solution Architect
SAS-Austria

Gerhard.svolba@sas.com

http://www.sascommunity.org/wiki/Gerhard_Svolba

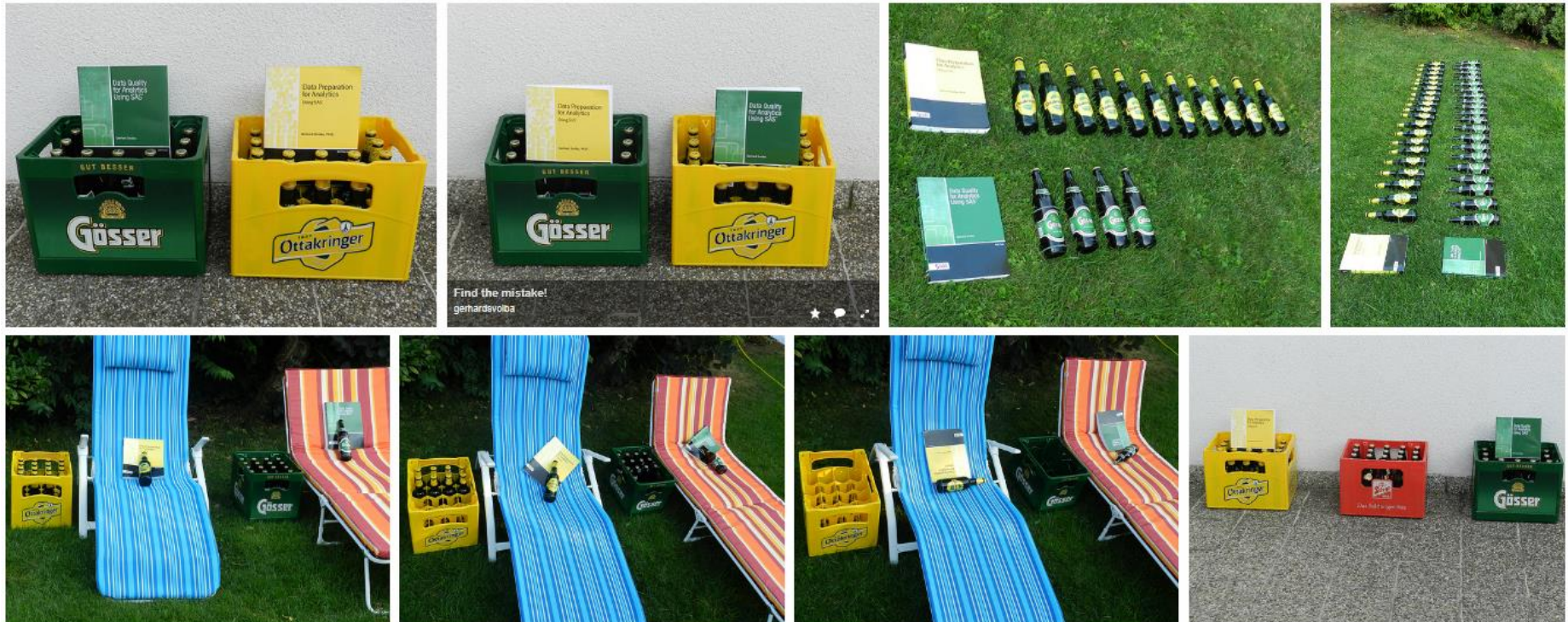
[LinkedIn](#) – [XING](#) – [PictureBlog](#)

ZUSAMMENFASSUNG

Picture Blog mit Downloads

"Data Preparation" and "Data Quality" on the Road

- [http://www.sascommunity.org/wiki/%22Data Preparation for Analytics%22 and %22Data Quality for Analytics%22 on the Road](http://www.sascommunity.org/wiki/%22Data_Preparation_for_Analytics%22_and_%22Data_Quality_for_Analytics%22_on_the_Road)



SAS1440-2015

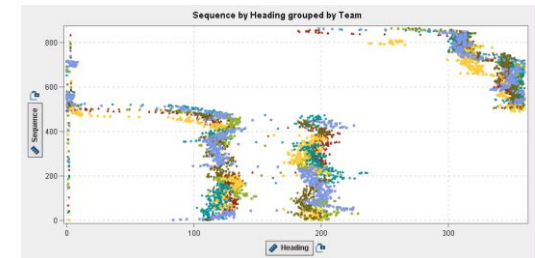
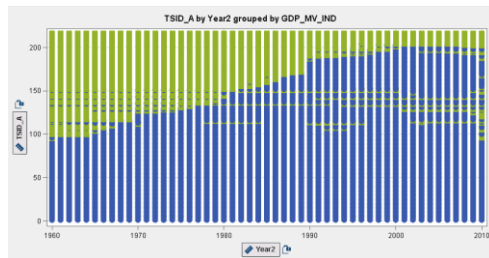
Want an Early Picture of the Data Quality of Your Analysis Data? SAS® Visual Analytics Shows You How

Gerhard Svolba, SAS Institute Inc. - Austria

ABSTRACT

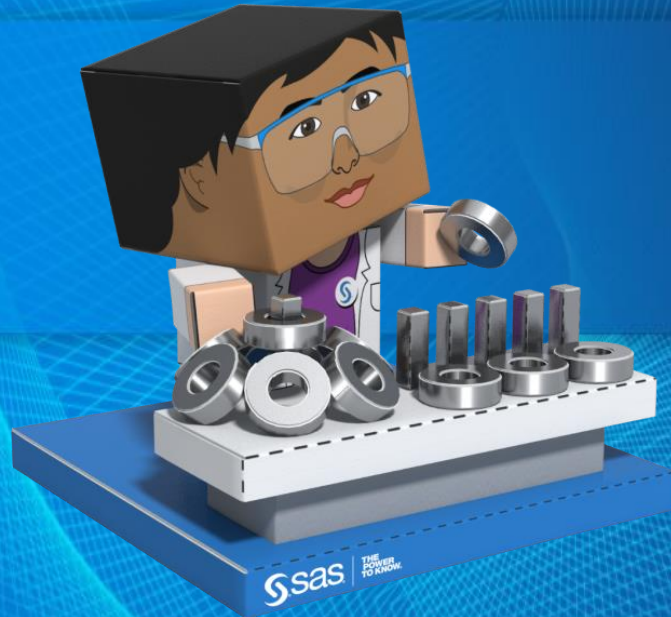
When you are analyzing your data and building your models, you often find out that the data cannot be used in the intended way. Systematic pattern, incomplete data, and inconsistencies from a business point of view are often the reason. You wish you could get a complete picture of the quality status of your data much earlier in the analytic lifecycle. SAS® analytics tools like SAS® Visual Analytics help you to profile and visualize the quality status of your data in an easy and powerful way. In this session, you learn advanced methods for analytic data quality profiling.

You see case studies based on real life data, where you look at time series data from a bird's-eye-view, and interactively profile GPS trackpoint data from a sail race.



DATA SCIENTIST

WWW.SAS.DE/DS



**THE
POWER
TO KNOW®**

**BESUCHEN SIE UNS
AM SAS STAND**