Ein Vergleich von SAS und S-Plus

Axel Benner
Biostatistik, DKFZ
Telefon: (06221) 42 2390
Email: benner@dkfz-heidelberg.de

Was ist S-PLUS ?

- *S-PLUS* ist die kommerzielle Variante der *S*-Language, vertrieben durch MathSoft Inc., Seattle.
- *S(-PLUS)* ist eine Umgebung von Software-Werkzeugen zur Datenmanipulation, Datenanalyse und grafischer Darstellung.
- *S(-PLUS)* ist objekt-orientiert, also gibt es Klassen und generische Funktionen
- *S(-PLUS)* arbeitet mit Datenobjekten, die verschiedene Arten der Daten“sammlung“ beschreiben. Diese Sammlung kann ein Feld von numerischen Werten, Zeichen oder logischen Werten sein, eine Liste oder ein ``Data Frame''. Eine Liste ist eine Verknüpfung von Datenobjekten, die wieder Listen oder Dataframes enthalten kann. Ein Dataframe ist wie eine Datenmatrix, wobei die Spalten unterschiedliche Typen haben können. Zeilen und Spalten können direkt durch Namen adressiert werden.
- *S(-PLUS)* speichert Objekte temporär in Frames und permanent in ″Data Directories″. Permanente Objekte werden durch Zuweisung erzeugt

  x <- 5

  und werden einzeln in Dateien gespeichert.
- Neue Programme werden in *S(-PLUS)* als Objekte des Typs `function` erzeugt. Eine Funktion ist definiert durch

```
name <- function(arg.1, arg.2, ...) expression
```

Ein Aufruf der Funktion hat die Gestalt

```
name(arg.1, arg.2,...).
```

Beispiele:

1. Binomialkoeffizienten

```
cbinom <- function(n)
          {
           lfact <- c(0, cumsum(log(1:n)))
           exp(lfact[n + 1] - lfact - lfact[(n + 1):1])
          }
```

2. Konfidenzlimits fuer den Korrelationskoeffizienten

```
cor.confint <- function(x,y,conf.level=.95)
                {
                 z <- atanh(cor(x,y))
                 b <- qnorm((1-conf.level)/2)/sqrt(length(x)-3)
                 ci.z <- c(z+b,z-b)
                 return(cor(x,y), tanh(ci.z))
                }
```

3. Logistisches Regressionsmodell (Aufruf und Attribute)

```
fit <- glm(ami.who ~ bmi+alter+l.ldh+l.tnt+l.ck,
             data=tnt, family=binomial(link=logit))
summary(fit)
plot(fit)

> attributes(fit)
$names:
 [1] "coefficients"      "residuals"         "fitted.values"
 [4] "effects"           "R"                 "rank"
 [7] "assign"            "df.residual"       "weights"
[10] "family"            "linear.predictors" "deviance"
[13] "null.deviance"     "call"              "iter"
[16] "y"                 "contrasts"         "terms"
[19] "formula"

$class:
[1] "glm" "lm"
```

Vergleich der Anwendung von SAS und S-Plus an einem Beispiel

SAS	S-Plus

`PROC MEANS; VAR AGE;`	`mean(age)`
	`sum(age<mean(age))`
	`mean(age<mean(age))`

`DATA NEW; SET OLD;`
`IF AGE < 16 THEN GROUP ='young';`
`ELSE GROUP = 'old';`
	`group <- ifelse(age<16,'young','old')`

Missing values ergeben 'young'!	Missing values ergeben missing values!

(Nicht vollständiger) Vergleich der Eigenschaften von SAS und S-Plus

Allgemein	SAS	S-Plus
Variablennamen	Bis 8 Zeichen Sonderzeichen:_ Keine Unterscheidung: Gross/Kleinschreibung	Beliebige Länge Sonderzeichen: . Unterscheidung zwischen Gross/Kleinschreibung
Benutzerdefinierte Attribute	Nicht vorhanden	Attribute eines Objekts können beliebig sein: 1. Vektor x: `comment(x) <- 'Corr:1.4.98'` 2. Elemente des Vektors x: `is.imputed(x)`
Label von Faktor-stufen	PROC FORMAT; getrennt von Daten	Wesentliches Attribut von Faktorvariablen
Datenverarbeitung	Record für Record	Vektor- bzw. Matrixweise
Merging	Allgemeine effiziente Verwendung	Allgemeine Verwendung, aber langsam
Verarbeitung grosser Datensätze	Nur beschränkt durch Plattenplatz	Beschränkung durch Arbeitsspeicher
Geschwindigkeit	Linear abnehmend mit Umfang	(Exponentiell) Abnehmend mit Umfang
Service	Online: Elementar; Internet: Guter Support (FAQ)	Online: Gutes Hilfesystem; Internet: allgemeine Texte

Fehlende Werte	SAS	S-Plus
Standard	Symbol: . Prüfung: x=.	Symbol: `NA` Prüfung: `is.na(x)`
Speziell	Spezielle Symbole: .A,...,.Z	Benutzerdefinierte Attribute
Anwendung in Ausdrücken	Wird als kleinster Wert verwendet Logische Operationen fehlerhaft	Korrekte logische Operationen, 1. `NA < 50` ergibt `NA` 2. `NA | True` ergibt `True`

Der SAS-Aufruf

```
data;
input x y @@;
if x < y;
cards;
1 .  2 2  . 3  4 5
;
proc print;
```

ergibt als Resultat

```
OBS     X     Y

1       .     3
2       4     5
```

Verarbeitung	SAS	S-Plus
Processing	Getrennte DATA und PROC Arbeitsschritte	Interpreter; beliebiges Mischen von Datenverarbeitung und Analysen
BY-Processing	PROC SORT, dann BY-Statement im Prozeduraufruf	tapply und verwandte Funktionen Schleifenweise Verarbeitung (langsam)
Matrizenrechnung	PROC IML; getrennt von anderen Prozeduren	Wesentliches Element der Sprache
Nachbearbeitung des Outputs	Output-Datensatz enthält oft nicht alle Print-Resultate, Schwierige Weiterverarbeitung	Alle Resultate sind in dem "Resultat"-Objekt gespeichert, das von einer Funktionen erzeugt wird
Benutzererweiterungen	PROC IML, aber keine Mischung mit Standard PROCs; Macro-Sprache (umständlich)	Funktionen entsprechend der S language; Einfaches Verknüpfen mit Fortran oder C Routinen
Quelltext	Nicht vorhanden	Viele Funktionen sind (zumindest teilweise) offen Alle sind aber veränderbar (!)
Kategoriale Variablen	Nur wenige Prozeduren bieten ein CLASS-Statement	Dummy-Variablen werden automatisch generiert
Nichtlineare Effekte	In der Regel keine Generierung nichtlinearer Komponenten im Modell	Alle Modelle erlauben beliebige Transformation und Funktionen von Prädiktoren im Aufruf Smoothing Splines in GAM: `gam(y ~ s(x))`

Wechselwirkungen	In der Regel keine Generierung von Wechselwirkungsfaktoren	Bestandteil der Modellformulierung
		`lm(y~x1+x2+x1:x2);`
		`coxph(y~(x1+x2+x3)^2)`

Man stellt fest:

- SAS Prozeduren sind orientiert auf den gedruckten Output
  - umfangreich, strukturiert
  - mühsame Weiterverarbeitung
    ```
    PROC SORT; BY SUBSTANZ;
    PROC NPAR1WAY WILCOXON; BY SUBSTANZ;
    CLASS GRUPPE;
    ```
- S-Plus zielt auf die Weiterverwendung des Ouptuts als Objekt
  - knapper gedruckter Output
  - einfache Weiterverarbeitung
    ```
    for(i in unique(Substanz))
    print(kruskal.test(Wert[Substanz==i],Gruppe[Substanz==i]))
    ```

Beispiel: Kruskal-Wallis-Test

SAS

```
SUBSTANZ=Albumin

N P A R 1 W A Y  P R O C E D U R E

Wilcoxon Scores (Rank Sums) for Variable WERT
Classified by Variable GRUPPE


                              Sum of        Expected         Std Dev           Mean
  GRUPPE          N           Scores        Under H0        Under H0          Score

  Gesund         42       3724.50000          2520.0      179.824291     88.6785714
  NAIV           29       1298.50000          1740.0      161.546890     44.7758621
  ART            36       1511.50000          2160.0      172.849706     41.9861111
  ART.mit.       12        605.50000           720.0      113.308075     50.4583333
Average Scores Used for Ties


Kruskal-Wallis Test (Chi-Square Approximation)
CHISQ =  45.416          DF =  3          Prob > CHISQ = 0.0001
```

S-Plus

```
[1] "Albumin"

        Kruskal-Wallis rank sum test

data:  Wert[Substanz == i] and Gruppe[Substanz == i]
Kruskal-Wallis chi-square = 45.4157, df = 3, p-value = 0
alternative hypothesis: two.sided
```

Grafiken	SAS	S-Plus
Präsentationen	Umständlich und zeitaufwendig	Einfach und komfortabel
Modellbeschreibung/ -diagnose	Wenige Methoden vorhanden	Viele vorgegebene Grafiken, vgl. plot.glm; plot.lme Erweiterbar
Besonderheiten	PROC INSIGHT	TRELLIS Displays

Trellis Displays:

- Grafische Darstellungen, die multiple Panele enthalten, die in einer regelmäßigen gitterartigen Struktur angeordnet sind.
- Analyse des Zusammenhangs zweier Variablen in Abhängigkeit von anderen Variablen
- Darstellungen bedingter Verteilungen (bzgl. mehrerer bedingender Variablen):
  1D, 2D und 3D Darstellungen in Dreiweg-Wiedergaben von Tafeln
  Histogramme, Dotplots, Streudiagramme, Kontourplots, 3D-Plots,...
- Zusätzlich: ``Interaktive'' Manipulation der Darstellung

Besonderheiten	SAS	S-Plus
Gemischte Effekte	PROC MIXED für lineare Modelle	Einige Funktionen verfügbar: aov, lme oder nlme
CART-Verfahren	Noch nicht vorhanden	In S-Plus tree; Ergänzung rpart
Generalisierte Additive Modelle	Nicht vorhanden	Funktion gam
Nichtparametrische Glättungsverfahren	Nur PROC INSIGHT, nicht in Standard-prozeduren	Eine Vielzahl eingebauter Glättungsverfahren
Zufallszahlen/ Verteilungen	``Standard''-Verteilungen; auch nichtzentral	Viele Verteilungen (ca. 100 Funktionen) i.w. nur zentral
Exakte Verfahren	Sign, Fisher's, McNemar, Rangtests, LR, Mantel-Haenszel, Pearson, Spearman, Cochran-Armitage, ...	Wilcoxon (ohne Bindungen!)

Exemplarischer Vergleich von Prozeduren und Funktionen

SAS Prozeduren	S-Plus Funktionen
ANOVA	Aov
REG, GLM	lm, (ols)
NLIN	nls, ms, smooth
MIXED	lme, nlme
GENMOD	glm, gam, (gee)
CATMOD, LOGISTIC	(lrm)
LIFETEST	surv.diff, survfit

LIFEREG	survreg, (psm)
PHREG	coxph, (cph)
TTEST	t.test
NPAR1WAY	wilcox.test, kruskal.test
FREQ, CORR	table, crosstabs, mantelhaen.test, fisher.test, chisq.test, mcnemar.test, friedman.test, cor.test, cor
UNIVARIATE, MEANS	median, quantile, mean, var
...	...

Zwei Beispiele:

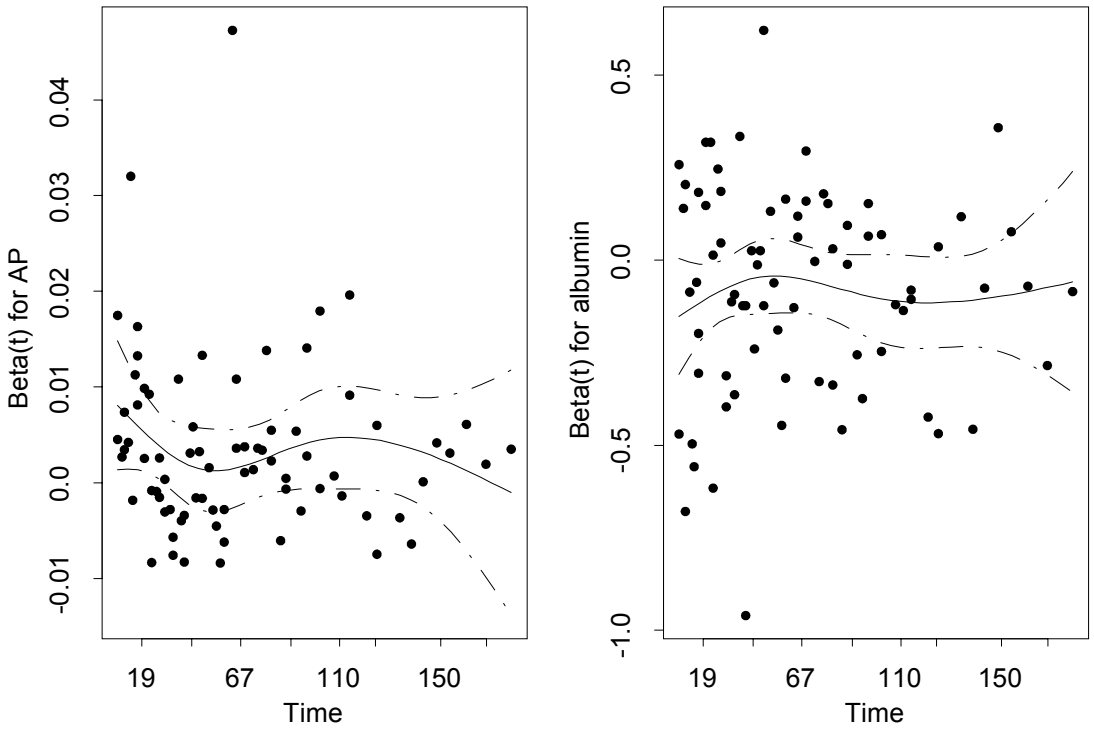
1. Analyse von Überlebenszeiten

Vorteile von S-Plus

- Andersen-Gill Modell (Andersen & Gill, 1982)
  Multiple Events/Person; Zeitabh. Strata
- Grafische Diagnose von Cox-Modellen
- Allgemeine Modellformulierung
- Buckley-James Modell (Ergänzung; Buckley & James, 1979)
- CART für Überlebenszeiten (Ergänzung)

Grafischer Test der Proportional Hazards Annahme (Grambsch & Therneau, 1994)

```
fit <- coxph(Surv(Survival, status) ~ AP + albumin)
zph <- cox.zph(fit)
par(mfrow=c(1,2))
plot(zph)
```



2. Logistische Regression

SAS

5 Prozeduren: LOGISTIC, PROBIT, CATMOD, GENMOD, LIFEREG

- LOGISTIC
  - Dies ist die primäre Prozedur für binäre und ordinale logistische Regression, aber
  - Kein CLASS-Statement für kategorielle Einflußvariablen
  - Keine Wechselwirkungen
  - DESCENDING Option nötig für Modellierung von P(Y=1)
- PROBIT: CLASS Statement
- GENMOD
  - Logistische Regression im Rahmen verallgemeinerter linearer Modelle, Responseverteilung: Binomial, Link: logit
  - CLASS Statement
- CATMOD
  - Keine quantitativen Einflußfaktoren
  - CLASS-Statement
- LIFEREG
  - Logistische Regression für P(Y=1)
  - CLASS-Statement; mehrere MODEL-Statements möglich

S-Plus

2 Funktionen: glm, lrm

- glm
  - Standardfunktion in S-Plus
  - Nur binäre logistische Regression
  - Entspricht GENMOD
- lrm (Erweiterung)
  - Im Rahmen des Pakets Design (letzte Fassung: Harrell, 1998)
  - Binäre und ordinale logistische Regression
  - Grafische Diagnose, Validierung

<u>Beispiel:</u> Ordinale logistische Regression (Bender & Benner, 1999)

Hier zur Demonstration nur 2 Einflussfaktoren:

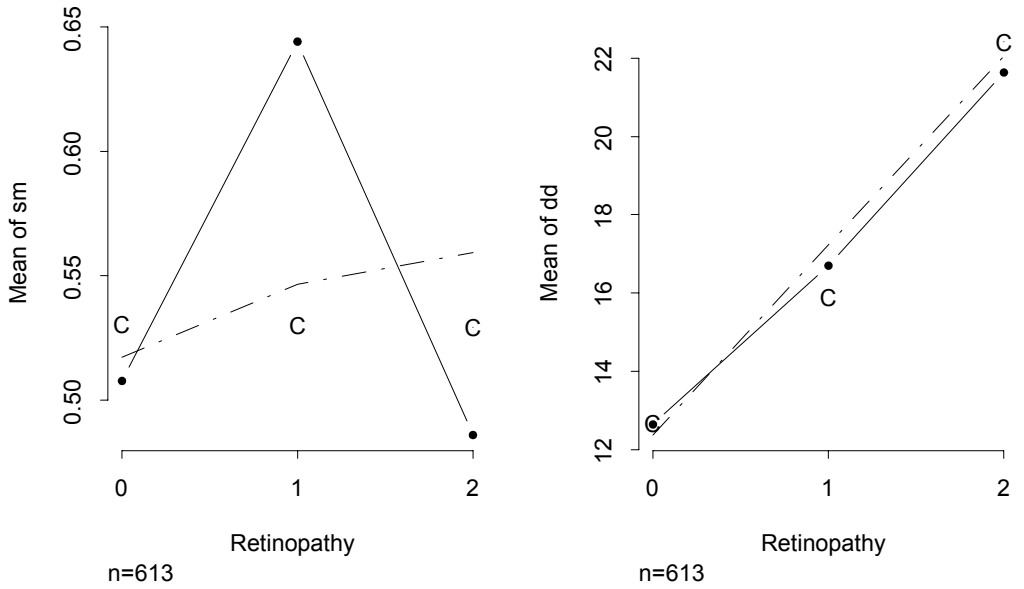
sm: smoking (yes/no), **dd:** duration diabetes

Ordinale Zielgrösse ist der Grad der Netzhauterkrankung:

Retinopathy (Levels 0,1,2)

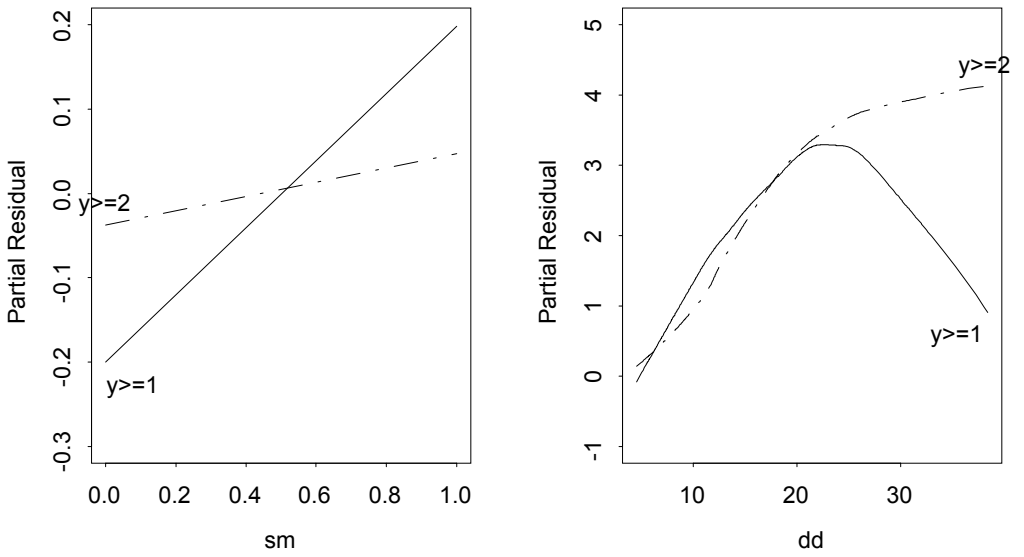
1. Plot der bedingten Mittelwerte von sm und dd als erster Eindruck

```
> plot.xmean.ordinaly(Retinopathy ~ sm + dd, cr=T)
```



2. Plot der partiellen Residuen zur Modelldiagnose

```
> fit <- lrm(Retinopathy ~ sm + dd, x=T, y=T)
> resid(fit, 'partial', pl=T)
```



Zusammenfassung

Vorteile von S-Plus

- Erweiterbarkeit/Flexibilität
  - Mischen von Funktionsaufrufen und Datenmanipulationsschritten
  - Einbindung eigener C und FORTRAN Subroutinen
- Grafik
  - Qualität der Präsentation (Publikation)
  - Datenanalytische Grafiken
- Schnelle Verfügbarkeit neuer statistischer Verfahren

Vorteile von SAS

- Routineaufgaben/Standardverfahren
  - Exakte Tests
- Geschwindigkeit
  - Analyse großer Datensätze
  - Simulationen
- Technical Support

Kommentar

- S-Plus ist einfacher zu lernen als SAS.
- SAS-Benutzer mit IML-Kenntnissen haben einen leichteren Einstieg in S-Plus. **Die Kombination von SAS und S-Plus ist beiden einzelnen Paketen DEUTLICH überlegen, da S-Plus dort schwach ist, wo SAS stark ist und SAS dort schwach ist, wo S-Plus stark ist.**
- S-Plus unterstützt die Verknüpfung mit SAS durch Funktionen wie sas.get.

Hinweis

Es wurden SAS Version 6.12 und S-Plus Version 3.4 Rel. 1, Sun Solaris 2.6 bzw. 2.5.1 verglichen. Viele Fakten und S-Funktionen sind von Frank Harrell Jr. übernommen, der in der Einfuehrung zu seinen S-Plus-Ergänzungen *Hmisc* und *Design* auch einen ausführlichen Vergleich von SAS und S-Plus präsentiert. Der vollständige Text ist auf der Webseite

http://hesweb1.med.virginia.edu/biostat/s/index.html

zu finden.

Das Dokument selbst,

An Introduction to S-Plus and the Hmisc and Design Libraries; CF Alzola and FE Harrell (16Feb99, PDF, 323 pages)

ist dort im Adobe Acrobat 3.0 format (pdf) gespeichert,

http://hesweb1.med.virginia.edu/biostat/s/doc/splus.pdf

Literatur

Bender, R. und Benner, A. (1999). Calculating Ordinal Regression Models in SAS and S-Plus. In Vorbereitung.

Grambsch, P.M. und Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81*, 515-526.

Harrell F.E. (1998). Design: S functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, and prediction.
Programs available from lib.stat.cmu.edu.