Die wichtigsten SAS/STAT-Prozeduren oder: Welches methodische Verfahren berechne ich mit welcher SAS-Prozedur?

Carina Ortseifen
Universität Heidelberg
Universitätsrechenzentrum
Im Neuenheimer Feld 293
69120 Heidelberg
carina.ortseifen@urz.uni-heidelberg.de

Zusammenfassung

Beim Blick in die Online-Hilfe des Statistikmoduls SAS/STAT (oder das zweibändige Handbuch) verliert man leicht den Überblick bei der Fülle an Prozeduren. In Version 9.4 beginnt die Liste beispielsweise mit der Prozedur ACECLUS und endet nach 98 weiteren Prozeduren mit der Prozedur VARIOGRAM. Nicht nur der Einsteiger kann sich bei diesem Angebot überfordert fühlen.
Wie soll man die richtige Auswahl treffen bzw. wo fängt man mit dem Lernen, aber auch Lehren, an? Welches sind die wichtigen Prozeduren, mit denen ich mich unbedingt beschäftigen sollte, welches die Prozeduren für Spezialgebiete?
In diesem Beitrag wird dem Lernenden und Einsteiger, aber auch dem Dozenten eine Orientierungshilfe angeboten und ein Weg aufgezeigt, wie man sich dieser Fülle nähern kann. Dazu werden Programmbeispiele zu den SAS-Prozeduren für die Berechnung von grundlegenden statistischen Methoden aufgezeigt und auf die Stellen im Output hingewiesen, an denen die zugehörigen Ergebnisse erscheinen. Ausgangspunkte der Betrachtungen werden neben den SAS-Prozeduren die den methodischen Verfahren zugrundeliegenden Kenngrößen sein: (a) Anzahl der abhängigen (AV) und unabhängigen (UV) Variablen, (b) Typ der Verteilung und/oder Zahl der Ausprägungen der AV und UV, sowie (c) Unabhängigkeit bzw. Verbundenheit der den Daten zugrundeliegenden Populationen und Stichproben.
Der Beitrag wendet sich an Einsteiger in das SAS System, die mit der Programmsyntax von Prozedur- und Datenschritten vertraut sind, und über grundlegende Statistikkenntnisse verfügen. Letztere sind wichtig, da im Beitrag die methodischen Verfahren nicht in der Tiefe behandelt werden und auch die Interpretationen der Ergebnisse dem Zuhörer überlassen bleiben. Und spätestens nachdem man 5-6 Statistikprozeduren kennengelernt hat, wird man als Anwender bemerken, dass nicht die SAS-Programmierkenntnisse das Problem darstellen als vielmehr die Kenntnis der statistischen Methoden.

Schlüsselwörter: SAS/STAT, Statistik-Prozeduren, PROC TTEST, PROC FREQ, PROC REG

1 Einleitung

Beim Blick in die Online-Hilfe des Statistikmoduls SAS/STAT (oder das zweibändige Handbuch) verliert man leicht den Überblick bei der Fülle an Prozeduren. In Version

9.4 beginnt die Liste beispielsweise mit der Prozedur ACECLUS und endet nach 98 weiteren Prozeduren mit der Prozedur VARIOGRAM. Nicht nur der Einsteiger kann sich bei diesem Angebot überfordert fühlen.
Wie soll man die richtige Auswahl treffen bzw. wo fängt man mit dem Lernen, aber auch Lehren, an? Welches sind die wichtigen Prozeduren, mit denen ich mich unbedingt beschäftigen sollte, welches die Prozeduren für Spezialgebiete?
Neben den SAS-Prozeduren selbst müssen auch die methodischen Verfahren betrachtet werden, die bei der statistischen Analyse zum Einsatz kommen (können), wie beispielsweise

- t-Test, Varianzanalyse
- Chiquadrat-Test, Mc Nemar Test
- Wilcoxon-Rangsummentest (Mann-Whitney-U-Test), Kruskal-Wallis-Test
- Regression, Logistische Regression, Überlebenszeitanalyse

Will man diese Verfahren anwenden, muss man die den methodischen Verfahren zugrundeliegenden Kenngrößen berücksichtigen, wie

- Anzahl der abhängigen (AV, auch Zielvariablen genannt) und unabhängigen (UV) Variablen (Einflussfaktoren)
- Typ der Verteilung und Skalenniveau, sowie Zahl der Variablen und Zahl der Ausprägungen der AV und UV
- Unabhängigkeit bzw. Verbundenheit der den Daten zugrundeliegenden Populationen und Stichproben

Das deutsche Ärzteblatt [1] unterteilt die Verfahren für eine AV in solche mit kategoriellen und solche mit stetigen Zielgrößen (Abbildungen 1 und 2).

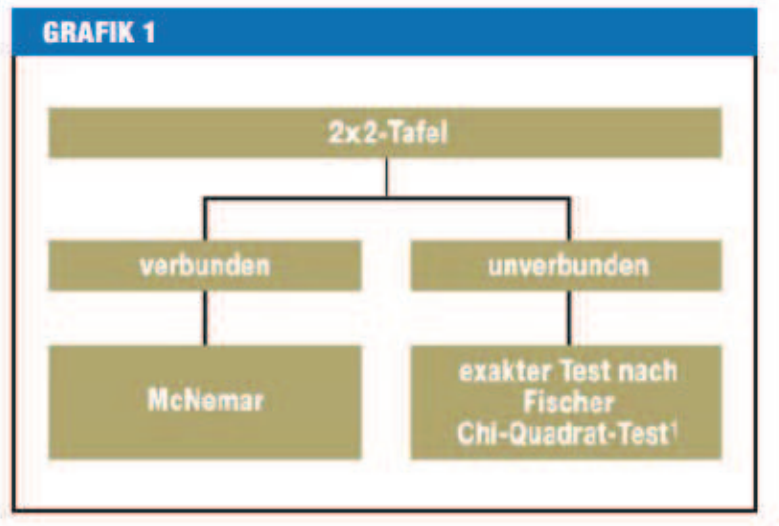

Abbildung 1: Tests für kategorielle AV (Zielgrößen)

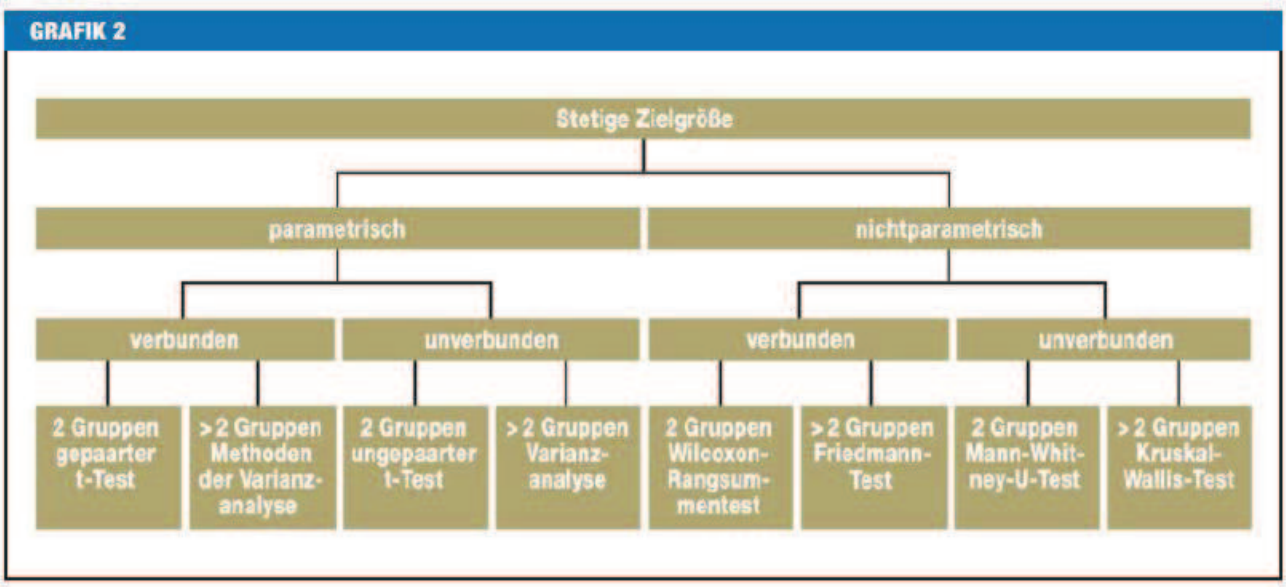

Abbildung 2: Tests für stetige AV (Zielgrößen)

Eine erweiterte Darstellung, ebenfalls für eine AV, findet sich in Wittenberg et.al. [2]:

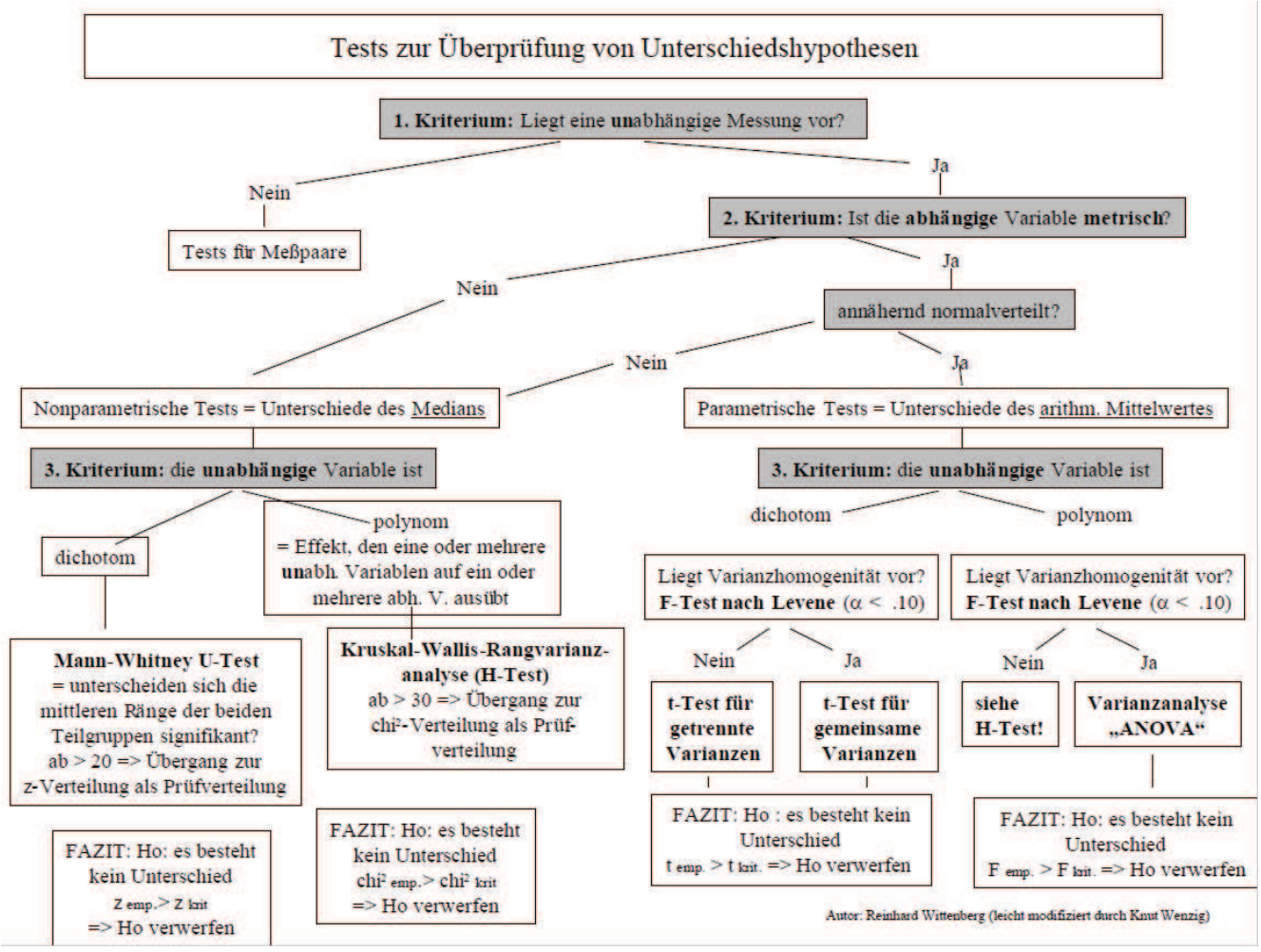

Abbildung 3: Tests zur Überprüfung von Unterschiedshypothesen

Hierbei werden der Reihe nach die Kriterien Unabhängigkeit der Messung, Skalenniveau der AV und Anzahl der Werte der UV betrachtet.

Eine Verfahrensübersicht auch für Situationen, in denen mehrere abhängige Variablen berücksichtigt werden, beschreibt das Institut for Digital Research and Education (IDRE) der University of California in Los Angeles (UCLA) [3].

Number of Dependent Variables	Nature of Independent Variables	Nature of Dependent Variable(s)	Test(s)	How to SAS	H t S
	0 IVs (1 population)	interval & normal	one-sample t-test	SAS	S
		ordinal or interval	one-sample median	SAS	S
		categorical (2 categories)	binomial test	SAS	S
		categorical	Chi-square goodness-of-fit	SAS	S
		interval & normal	2 independent sample t-test	SAS	S
		ordinal or	Wilcoxon-Mann		
...					
	1 or more interval IVs and/or 1 or more categorical IVs	interval & normal	analysis of covariance	SAS	
		categorical	multiple logistic regression	SAS	
			discriminant analysis	SAS	
2+	1 IV with 2 or more levels (independent groups)	interval & normal	one-way MANOVA	SAS	
	2+	interval & normal	multivariate multiple linear regression	SAS	
	0	interval & normal	factor analysis	SAS	
2 sets of 2+	0	interval & normal	canonical correlation	SAS	

Abbildung 4: Empfehlungen des IDRE der UCLA

Während die IDRE-Tabelle auch Umsetzungen in SPSS, Stata und R ausgibt, beschränkt sich die adaptierte Tabelle 1 auf die SAS-Prozeduren.

Tabelle 1: Geeignete methodische Verfahren und SAS-Prozeduren

# AV	Art der UV	AV metrisch u. normal-verteilt	SAS	AV ordinal oder metrisch	SAS	AV kategorisch	SAS
1	0 UV (1 Gruppe)	Ein-Stich-proben-t-Test	TTEST	Ein-Stichpr.-Mediantest	UNI-VARIATE	Binomialtest Chi-Quadrat-Anpassungst.	FREQ FREQ
1	1 UV, 2 Stufen (unabh. Gruppen)	t-Test	TTEST	Wilcoxon-RS-Test / Mann-Whitney U	NPAR1WAY	Chi-Quadrat. Fishers exakter Test	FREQ FREQ
1	1 UV, 2+ Stufen (unabh. Gr.)	ANOVA	GLM	Kruskal-Wallis-Test	NPAR1WAY	Chi-Quadrat-Test	FREQ
1	1 UV, 2 Stufen (abh. Gruppen)	Gepaarter t-Test	TTEST	Wilcoxon-Vorzeichen-Rangtest	UNIVARIATE	Mc Nemar-Test	FREQ
1	1 UV, 2+ Stufen (abhängige Gruppen)	ANOVA mit Messwieder-holung	GLM	Friedman-Test	FREQ	Logist. Regress. mit Messwied.	GEN MOD
1	2+ UV (unabh. Gruppen)	Faktorielle ANOVA	GLM	"Ordered" Logist. Regr.	LOGISTIC	Faktorielle logist. Regr.	LOGI STIC
1	1 metrische UV	Korrelation Lineare Regression	CORR REG	Nichtpara-metrische Korrelation	CORR	Logistische Regression	LOGI STIC
1	1+ metr. UV und/oder 1+ kategorielle UV	Multiple Regr. Kovarianz-analyse	REG GLM			Multiple Log. Regr. Diskrimi-nanzanalyse	LOGI STIC DISC RIM
2+	1 UV, 2+ Stufen (unabh. Gr.)	Einweg-MANOVA	GLM				
2+	2+	Multivariate multiple lin. Regression	REG				
2+	0	Faktoren-analyse	FACTOR				
2 set of 2+	0	Kanonische Korrelation	CANCOR				

Die Tabelle wurde dabei so umstrukturiert, dass für die jeweilige Anzahl an AV und die Art der UV für jeweils metrische und normalverteilte, ordinale oder metrische sowie kategorische/nominale UV ein empfohlenes Verfahren sowie eine SAS-Prozedur benannt werden.

Damit können aus den knapp 100 Statistikprozeduren fürs Erste 12 Prozeduren ausgewählt werden, die ein breites Spektrum abdecken können, wobei es auch unter den beschriebenen Szenarien weitere Verfahren und Prozeduren geben mag, die besser geeignet sind als die hier beschriebenen. Betrachten Sie daher diese Tabelle eher als Vorschlag denn als Bibel!

2 Anwendungen

An drei Beispielen wird nun die Verwendung der Tabelle demonstriert. Eine vollständige Liste mit Programmbeispielen finden Sie in [3] sowie – hoffentlich - in naher Zukunft auch im deutschsprachigen SAS Wiki [4], wenn der Autorin eine zündende Idee für eine geschickte Darstellung eingefallen ist.

2.1 Beispieldatei High School-Studenten (HSB2)

Für die Beispiele wird die aus IBM/SPSS Statistics importierte SAS-Tabelle HSB2 verwendet (http://stats.idre.ucla.edu/wp-content/uploads/2016/02/hsb2-3.sav).
Diese Tabelle enthält eine Stichprobe von 200 High School-Studierenden mit demographischen Merkmalen, wie Geschlecht, sozioökonomischer Status, ethnischer Hintergrund sowie Bewertungen für Lesen, Schreiben, Rechnen, Naturwissenschaften und Gemeinschaftskunde. Die ersten 10 Beobachtungen mit Variablennamen und Werten sind in Abbildung 5 zu sehen.

Obs	id	female	race	ses	schtyp	prog	read	write	math	science	socst
1	70	male	white	low	public	general	57	52	41	47	57
2	121	female	white	middle	public	vocation	68	59	53	63	61
3	86	male	white	high	public	general	44	33	54	58	31
4	141	male	white	high	public	vocation	63	44	47	53	56
5	172	male	white	middle	public	academic	47	52	57	53	61
6	113	male	white	middle	public	academic	44	52	51	63	61
7	50	male	african-amer	middle	public	general	50	59	42	53	61
8	11	male	hispanic	middle	public	academic	34	46	45	39	36
9	84	male	white	middle	public	general	63	57	54	58	51
10	48	male	african-amer	middle	public	academic	57	55	52	50	51

Abbildung 5: Ausschnitt aus dem Listenbericht der SAS-Tabelle HSB2

2.2 Ein-Stichproben-t-Test mit PROC TTEST

Für ein erstes Beispiel betrachten wir die Bewertung aller Studierenden im Fach Schreiben (`write`) und stellen die Frage: Beträgt der mittlere Wert im Schreiben 50 oder ist der Wert ungleich 50?
Da wir die gesamte Stichprobe betrachten, also keine Unterscheidung nach Gruppen vornehmen, gibt es keine unabhängige Variable und lediglich eine abhängige Variable, nämlich `write`.
Für diese Situation entnehmen wir Tabelle 1 als methodisches Verfahren den „Ein-Stichproben-t-Test“ und als geeignete SAS-Prozedur „TTEST“.

Die Nullhypothese für den Test lautet demnach:

H0: Der mittlere Wert für die Variable `write` beträgt 50.

Mit der Prozedur PROC TTEST lässt sich der Test mit der folgenden Syntax berechnen:

```
Proc TTEST Data=my.hsb2 H0=50;
   Var write;
Run;
```

Das Ergebnis der Prozedur TTEST erscheint in folgender Form:

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
52.7750	51.4533	54.0967	9.4786	8.6318	10.5110

DF	t Value	Pr > |t|
199	4.14	<.0001

Abbildung 6: Ausgabe der Prozedur TTEST

Die obere Tabelle zeigt den Mittelwert (Mean), hier: 52,775, sowie das 95% Konfidenzintervall (95% CL Mean) für den Mittelwert, 51,45 als untere Grenze, 54,10 als obere Grenze. Der Wert 50 wird davon nicht eingeschlossen, woraus wir bereits schließen können, dass die Nullhypothese abgelehnt werden kann. Die gleiche Aussage kann man auch mit dem t-Wert und der zugehörigen Wahrscheinlichkeit dieses t-Testes unter der Nullhypothese treffen (untere Tabelle): Die Studierenden haben einen signifikant höheren Wert im Schreiben als 50.

Die Prozedur TTEST liefert – mit der SAS-Version 9.4 TS1M3 unter Windows – noch zwei Grafiken, die hier nicht wiedergegeben werden.

Man könnte für diese Fragestellung auch die Prozedur MEANS verwenden. Da diese aber nur gegen den Wert 0 testen kann, müsste man zunächst die Werte transformieren (Subtraktion mit 50), um gegen den Schreibwert von 50 testen zu können.

```
Data hsb2_neu;
   Set hsb2;
   write_n=write-50;
Run;
Proc MEANS Data=hsb2_neu Means Lclm Uclm t probt;
   Var write;
Run;
```

2.3 χ^2-Test mit PROC FREQ

Mit der zweiten Anwendung soll untersucht werden, ob es einen Zusammenhang zwischen den Variablen Schultyp (`schtyp`) und Geschlecht (`gender`) gibt, oder, anders formuliert, ob der Schultyp vom Geschlecht abhängig ist. Es liegt damit eine unabhängige Variable vor mit zwei Stufen (Geschlecht, `male` und `female`) und eine abhängige Variable (Schultyp, auch mit zwei Stufen, `public` und `private`).
Die Nullhypothese lautet, dass die beiden Variablen Geschlecht und Schultyp voneinander unabhängig sind.
Mit der Prozedur FREQ und der Option CHISQ in der TABLES-Anweisung kann der χ^2-Test gerechnet werden:

```
Proc FREQ Data=my.hsb2;
   Tables schtyp * gender / Chisq Norow Nocol;
Run;
```

Die Option NOROW und NOCOL werden verwendet, um die Ausgabe in Abbildung 7 schlanker zu gestalten.

Der Testwert (Value) sowie dessen Wahrscheinlichkeit unter der Nullhypothese (Prob) werden in der zweiten Tabelle angezeigt (neben weiteren Statistiken, die hier ausgeblendet wurden). Die Nullhypothese kann in diesem Fall nicht abgelehnt werden, d.h. es gibt keinen Hinweis für einen Zusammenhang zwischen den beiden Variablen Schultyp und Geschlecht.

In dem Fall der Vierfeldertafel, also beim Vorliegen von zwei dichotomen Variablen, liefert die Prozedur zusätzlich noch automatisch Fishers exakten Test, der ebenfalls die Nullhypothese nicht ablehnt (p=0,8492). Bei größeren Kreuztabellen muss dieser Test explizit angefordert werden mit der Option EXACT.

Table of schtyp by gender			
schtyp(type of school)	gender		
Frequency Percent	male	female	Total
public	77 38.50	91 45.50	168 84.00
private	14 7.00	18 9.00	32 16.00
Total	91 45.50	109 54.50	200 100.00

Statistic	DF	Value	Prob
Chi-Square	1	0.0470	0.8283
…	..	..	..

Fisher's Exact Test	
…	…
Two-sided Pr <= P	0.8492

Abbildung 7: Ausgabe der Prozedur FREQ

2.4 Lineare Regression mit PROC REG

Mit einer linearen Regression kann die Beziehung zwischen einer normalverteilten, intervallskalierten Prädiktorvariablen (UV) und einer ebenso verteilten Zielvariable (AV) untersucht werden. Für unser Anwendungsbeispiel untersuchen wir die Beziehung zwischen Schreiben und Lesen bei den 200 Studierenden. Oder genauer: Kann die Bewertung für Schreiben aus der Bewertung für Lesen abgeleitet werden.

Dazu verwenden wir die SAS-Prozedur REG. Mit der Model-Anweisung beschreiben wir die Beziehung, rechts vom Gleichheitszeichen die UV, links vom Gleichheitszeichen die AV. Und die Option STB (nach dem „/“) liefert zusätzlich die standardisierten Regressionskoeffizienten.

```
Proc REG Data=my.hsb2;
   Model write=read / Stb;
Run;
```

In Abbildung 8 wird die Anpassungsgerade mitsamt ihrer Konfidenzbänder und der beobachteten Punkt grafisch veranschaulicht.

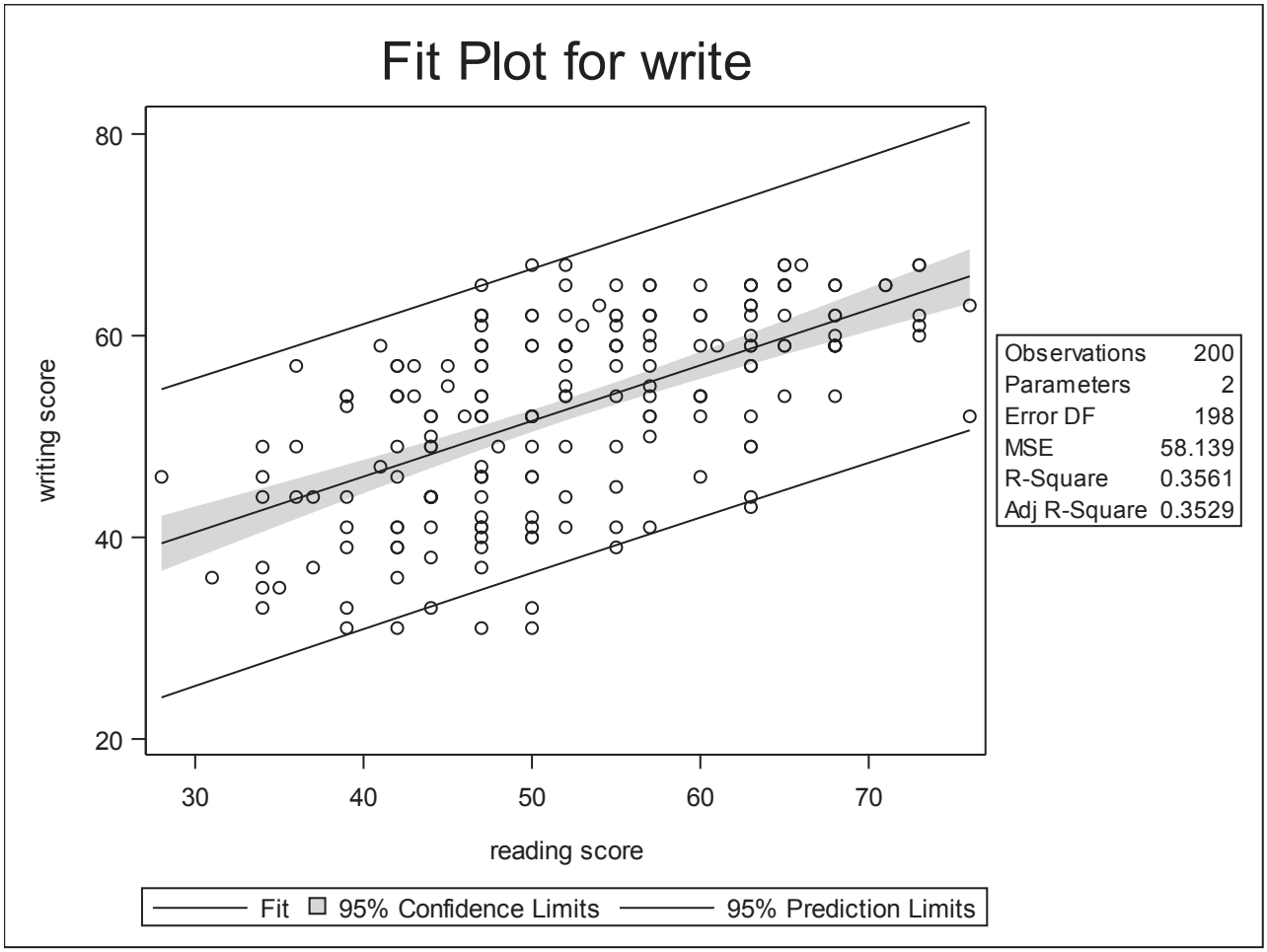

Abbildung 8: Graphische Ausgabe der Prozedur REG (Ausschnitt)

Die Parameterschätzer zur genaueren Beschreibung der Beziehung können der untersten Tabelle von Abbildung 9 entnommen werden.

Der Zusammenhang lässt sich damit dann wie folgt beschreiben:

Lesen = 23.96 + 0.55 Schreiben

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6367.42127	6367.42127	109.52	<.0001
Error	198	11511	58.13866		
Corrected Total	199	17879			

Root MSE	7.62487	R-Square	0.3561
Dependent Mean	52.77500	Adj R-Sq	0.3529
Coeff Var	14.44788		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	23.95944	2.80574	8.54	<.0001	0
read	reading score	1	0.55171	0.05272	10.47	<.0001	0.59678

Abbildung 9: Tabellarische Ausgabe der Prozedur REG

3 Fazit

Beginnt man sich mit der SAS Software zu beschäftigen, dann begegnet man – neben dem Datenschritt – und den Prozeduren PRINT und SORT ganz schnell folgenden Prozeduren:

- MEANS
- UNIVARIATE
- FREQ
- CORR

Damit kann man bereits Deskriptive Statistiken ermitteln und erste Tests rechnen. Ergänzt man diese dann noch um

- TTEST
- NPAR1WAY
- GLM
- REG und eventuell
- LOGISTIC

ist man gut gerüstet, um mit SAS statistische Auswertungen bearbeiten zu können. Je nach Arbeitsgebiet können dann natürlich weitere Prozeduren hinzukommen, z.B. für Überlebenszeitanalysen.

Der Schwerpunkt liegt dann, zumindest nach den Erfahrungen der Autorin, nicht mehr ausschließlich in der Anwendung von SAS, da SAS-Grundlagen bereits vorliegen und das Aneignen von neuen Prozeduren mithilfe der Beispiele in der Online-Hilfe gut unterstützt wird. Gute methodische Kenntnisse sind dann bei weitem mehr gefragt und das Schritthalten mit neueren Entwicklungen in der Statistik, die zunehmend auch wieder in SAS-Prozeduren umgesetzt werden.

Literatur

[1] Deutsches Ärzteblatt, 2010 (Int 107(19): 343-8): Auswahl statistischer Testverfahren. https://www.aerzteblatt.de/archiv/74880 [01.03.2017]

[2] R. Wittenberg et.al.: Datenanalyse mit IBM SPSS Statistics. utb, 2014, S. 232.

[3] Choosing the correct statistical test in SAS, Stata, SPSS and R. From: http://stats.idre.ucla.edu/other/mult-pkg/whatstat/ [01.03.2017]

[4] Deutschsprachiges SAS-Wiki, http://de.saswiki.org [01.03.2017]