A First Course on Time Series Analysis with SAS - Ein Open-Source Projekt

Michael Falk
Lehrstuhl für Statistik,
Universität Würzburg
Am Hubland
97074 Würzburg
falk@mathematik.uni-wuerzburg.de

Frank Marohn
Lehrstuhl für Statistik,
Universität Würzburg
Am Hubland
97074 Würzburg
marohn@mathematik.uni-wuerzburg.de

Bernward Tewes
Universitätsrechenzentrum,
Katholische Universität Eichstätt
Ostenstraße 24
85072 Eichstätt
bernward.tewes@ku-eichstaett.de

Daniel Hofmann
Lehrstuhl für Statistik,
Universität Würzburg
Am Hubland
97074 Würzburg
daniel.c.hofmann@web.de

Maria Macke
Lehrstuhl für Statistik,
Universität Würzburg
Am Hubland
97074 Würzburg
maria.macke@gmx.de

Peter Dinges
Lehrstuhl für Statistik,
Universität Würzburg
Am Hubland
97074 Würzburg
me@elwedgo.de

René Michel
Lehrstuhl für Statistik,
Universität Würzburg
Am Hubland
97074 Würzburg
michel@mathematik.uni-wuerzburg.de

Zusammenfassung

Das Projekt „A First Course on Time Series Analysis with SAS“ hat die Erstellung eines Open-Source-Lehrbuches zur Zeitreihenanalyse zum Ziel. Dieses verbindet theoretische Elemente der Zeitreihenanalyse mit einer Auswahl von statistischen Verfahren, die wiederum mittels geeigneter SAS-Prozeduren auf reale Datensätze angewendet werden. Zentrale Stelle dieses Projektes ist eine Internetseite

(http://statistik.mathematik.uni-wuerzburg.de/timeseries),

auf der sämtliche das Projekt betreffenden Elemente wie der Buchtext und LaTeX-Quellcode, die Datensätze und SAS-Programme kostenlos zum Download bereit gestellt werden.

Schlüsselworte: Zeitreihenanalyse, ETS-Modul, Open-Source, GNU Free Documentation License

1 Idee und Motivation

1995 erschien das Buch „Angewandte Statistik mit SAS“ von M. Falk, F. Marohn und R. Becker und 2002 die englische Version „Foundations of Statistical Analysis and Applications with SAS“ von M. Falk, F. Marohn und B. Tewes.

Diese Bücher sind eine Verbindung von Theorie und Anwendung grundlegender statistischer Verfahren wie Varianzanalyse, Regressionsanalyse, Clusteranalyse, Faktorenanalyse etc. Die Anwendung wird dabei mittels SAS dargestellt und an realen Datensätzen vorgenommen.

Die Idee ist nun ein ähnliches Buch zur Zeitreihenanalyse zu erstellen, jedoch in einer anderen Form und unter gezieltem Einsatz der Neuen Medien.

Dieses Buch soll zusammen mit dem LaTeX-Quellcode, den Datensätzen und SAS-Programmen im Internet unter

http://statistik.mathematik.uni-wuerzburg.de/timeseries

veröffentlicht werden, so dass Interessierte das Buch kostenlos verwenden und gegebenenfalls durch eigene Ideen, Verbesserungsvorschläge, Datensätze, neue Themengebiete etc. verbessern und ergänzen können.

Im Gegensatz zu einem konventionellen Buch, das bis zum Erscheinen der nächsten Auflage statisch bleibt, bietet ein Open-Source Projekt unter der GNU Free Documentation License die Gelegenheit ein dynamisches Buch zu erzeugen, das sich ständig an Neuerungen anpassen kann. So können die neuesten Verfahren und aktuelle Datensätze jederzeit eingefügt werden. Außerdem kann das Buch immer mit der aktuellsten SAS-Version abgeglichen werden. Die genauen Bedingungen, unter denen das Buch zu kommerziellen und nicht-kommerziellen Zwecken weitergegeben werden darf, finden sich als Anhang im Buchtext sowie im Internet[1].

Wir würden uns freuen, wenn vorgenommene Änderungen auch an die E-Mail-Adresse des Projektes

timeseries@statistik.mathematik.uni-wuerzburg.de

geschickt werden, denn in regelmäßigen Abständen soll eine neue „offizielle“ Version mit den vorgeschlagenen Änderungen auf der Internetseite des Projektes erscheinen.

2 Anwendungsbereich und Zielgruppe

Das Buch kann verwendet werden als Grundlage für eine zweisemestrige Vorlesung über Zeitreihenanalyse.

Ergänzend dazu ist es möglich ein Seminar zu veranstalten, im Rahmen dessen man versucht, das Buch durch neue Inhalte, Datensätze oder SAS-Programme weiterzuentwickeln.

Ebenfalls ist das Buch zum Selbststudium geeignet.

Zielgruppe sind somit Studenten der Mathematik/Statistik, sowie Studenten anderer Fachrichtungen (Wirtschaftswissenschaften, Ingenieurwissenschaften etc.), bei denen Statistik zur akademischen Ausbildung gehört und in der beruflichen Praxis große Anwendung findet.

Ebenso geeignet ist das Buch für Praktiker, die sich über die einfache Anwendung hinaus für die mathematischen Hintergründe interessieren.

[1] http://www.gnu.org/licenses/fdl.html

Als Vorkenntnisse sollte man Grundlagen der Stochastik (inklusive der Testtheorie) mitbringen.

Vorkenntnisse in SAS sind nicht nötig, aber (wie immer) hilfreich. Das Buch kann auch als eine erste Einführung in SAS dienen.

3 Inhalte

Die folgenden Themenbereiche sind bisher abgedeckt:

Elemente der explorativen Zeitreihenanalyse (additives Modell, lineares Filtern, Autokovarianzen und Autokorrelationen, ...).
Modelle der Zeitreihenanalyse (AR-, MA-, ARMA-, ARIMA-Prozesse, sowie ARCH- und GARCH-Prozesse, dazu das Box-Jenkins-Programm und State-Space-Modelle,...).
Die Analyse im Frequenzbereich (Harmonische Wellen, Fourier-Frequenzen, Periodogramm, ...)
Die Analyse des Spektrums eines stationären Prozesses (Spektraldichte, Powerfunktion, Tiefpass-, Hochpassfilter,...)
Statistische Analyse im Frequenzbereich (Test auf weißes Rauschen, Schätzung der Spektraldichte,...)

Zu allen diesen Bereichen sind ausführliche Übungsaufgaben vorhanden, die auch in einer entsprechenden Vorlesung verwendet werden können.

Für die Auswertung der statistischen Verfahren und Programme wird hauptsächlich das Modul ETS verwendet.

Ein gewünschter Inhalt, der bisher noch nicht realisiert werden konnte, ist eine Fallstudie an einem geeigneten realen Datensatz, bei der die vorher vorgestellten statistischen Verfahren mittels SAS durchgeführt werden und eine detaillierte Analyse des Datensatzes erstellt wird.

4 Gründe für den Einsatz von SAS

Die Verwendung von SAS zur Darstellung der Anwendungen geschieht hauptsächlich aus den folgenden Gründen:

- Aufgrund der weiten Verbreitung in Wirtschaft und Industrie ist das Buch in der Praxis einsetzbar und dazu geeignet, Studenten auf das spätere Arbeiten im Beruf vorzubereiten.
- Durch die Eingabe eines Programmcodes kann verwirrendes „Herumklicken“ vermieden werden und so eine gute Dokumentation eines Projektes erstellt werden. Dies erleichtert auch die Darstellung der SAS-Elemente im Buchtext erheblich.
- Die SAS-Dokumentation ist mathematisch orientiert und bietet somit einen guten Übergang von den theoretischen Hintergründen zu den praktischen Anwendungen. Sie ist nach der Lektüre des Buches „verstehbar“.

5 Vorgehensweise des Buches

Wir wollen nun die Vorgehensweise des Buches anhand der Einführung der ARCH- und GARCH-Prozesse demonstrieren.

Zunächst werden die ARCH-Prozesse definiert. Danach wird der Zusammenhang mit den bereits bekannten AR-Prozessen hergestellt. Anschließend werden noch die GARCH-Prozesse eingeführt. Dieser komplette Theorieteil kann aus Platzgründen hier nicht abgedruckt werden, deshalb nur der Anfang mit der Hinführung:

ARCH and GARCH-Processes

In particular the monitoring of stock prices gave rise to the idea that the *volatility* of a time series (Y_t) might not be a constant but rather a random variable, which depends on preceding realizations. The following approach to model such a change in volatility is due to Engle (1982).

We assume the multiplicative model

$$Y_t = \sigma_t Z_t, \qquad t \in \mathbb{Z},$$

where the Z_t are independent and identically distributed random variables with

$$\mathrm{E}(Z_t) = 0 \text{ and } \mathrm{E}(Z_t^2) = 1, \qquad t \in \mathbb{Z}.$$

The *scale* σ_t is supposed to be a function of the past p values of the series:

$$\sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j Y_{t-j}^2, \qquad t \in \mathbb{Z},$$

Nach den theoretischen Überlegungen kommt eine praktische Anwendung an einem realen Datensatz. Dieser besteht aus den täglichen Logreturns des HangSeng-Index zwischen dem 16. Juli 1981 und dem 30. September 1983. Begonnen wird mit der Erklärung des Datensatzes.

2.2.13 Example (Hongkong Data). The daily Hang Seng closing index was recorded between July 16th, 1981 and September 30th, 1983, leading to a total amount of 552 observations p_t. The daily *log returns* are defined as

$$y_t := \log(p_t) - \log(p_{t-1}),$$

where we now have a total of $n = 551$ observations. The expansion $\log(1+\varepsilon) \approx \varepsilon$ implies that

$$y_t = \log\left(1 + \frac{p_t - p_{t-1}}{p_{t-1}}\right) \approx \frac{p_t - p_{t-1}}{p_{t-1}},$$

provided that p_{t-1} and p_t are close to each other. In this case we can interpret the return as the difference of indices on subsequent days, relative to the initial one.

Danach werden der Datensatz und seine Quadrate visualisiert.

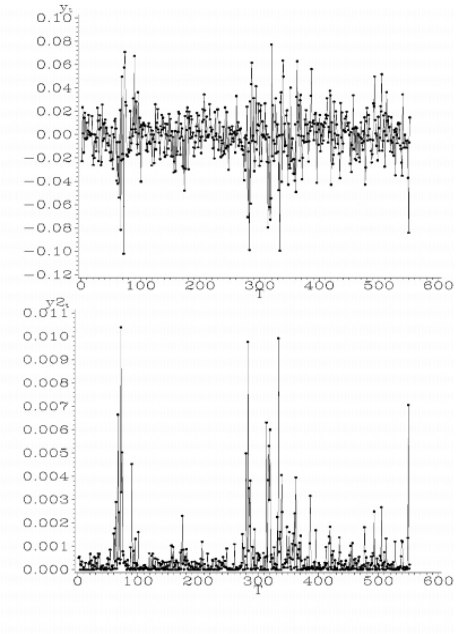


Figure 2.2.8. Log returns of Hang Seng index and their squares.

Zu jeder SAS-Ausgabe erfolgt im Buch die Angabe des SAS-Programmcodes, mit dem der Output (in diesem Fall die vorhergehende Graphik) erzeugt wurde.

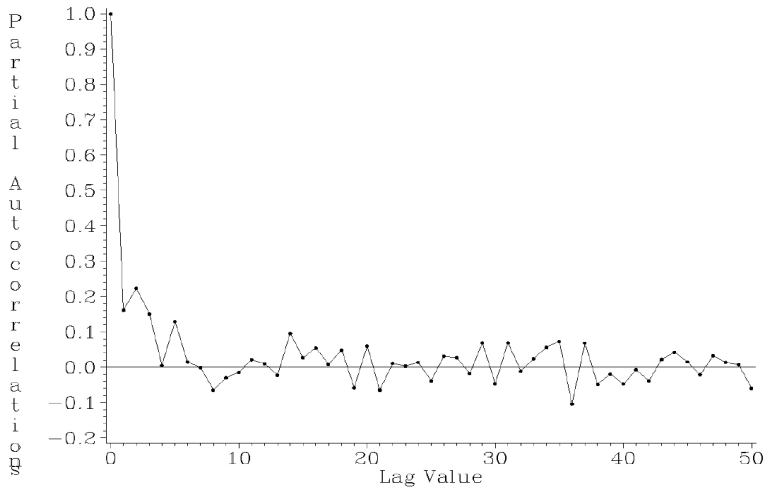
```
***     Program 2_2_8     ***;
TITLE1 'Daily log returns and their squares';
TITLE2 'Hongkong Data ';

DATA data1;
   INFILE 'c:\data\hongkong.txt';
   INPUT p;
   t=_N_;
   y=DIF(LOG(p));
   y2=y**2;

SYMBOL1 C=RED V=DOT H=0.5 I=JOIN L=1;
AXIS1 LABEL=('y' H=1 't') ORDER=(-.12 TO .10 BY .02);
AXIS2 LABEL=('y2' H=1 't');
GOPTIONS NODISPLAY;
PROC GPLOT DATA=data1 GOUT=abb;
   PLOT y*t / VAXIS=AXIS1;
   PLOT y2*t / VAXIS=AXIS2;
RUN;

GOPTIONS DISPLAY;
PROC GREPLAY NOFS IGOUT=abb TC=SASHELP.TEMPLT;
   TEMPLATE=V2;
   TREPLAY 1:GPLOT 2:GPLOT1;
RUN; DELETE _ALL_; QUIT;
```

Die einzelnen Schritte des Programmes werden im Anschluss gesondert erläutert.

In the `DATA` step the observed values of the Hang Seng closing index are read into the variable `p` from an external file. The time index variable `t` uses the SAS-variable `_N_`, and the log transformed and differenced values of the index are stored in the variable `y`, their squared values in `y2`.

After defining different axis labels, two plots are generated by two `PLOT` statements in `PROC GPLOT`, but they are not displayed. By means of `PROC GREPLAY` the plots are merged vertically in one graphic.

Nachdem ein erster visueller Eindruck der Daten gegeben wurde, soll nun eine Analyse der Daten mit PROC ARIMA und PROC AUTOREG erfolgen. Wiederum wird zunächst der Output angegeben



Autoreg Procedure

Dependent Variable = Y

Ordinary Least Squares Estimates

SSE	0.265971	DFE	551
MSE	0.000483	Root MSE	0.021971
SBC	-2643.82	AIC	-2643.82
Reg Rsq	0.0000	Total Rsq	0.0000
Durbin-Watson	1.8540		

NOTE: No intercept term is used. R-squares are redefined.

GARCH Estimates

SSE	0.265971	OBS	551
MSE	0.000483	UVAR	0.000515
Log L	1706.532	Total Rsq	0.0000
SBC	-3381.5	AIC	-3403.06
Normality Test	119.7698	Prob>Chi-Sq	0.0001

Variable	DF	B Value	Std Error	t Ratio	Approx Prob
ARCH0	1	0.000214	0.000039	5.444	0.0001
ARCH1	1	0.147593	0.0667	2.213	0.0269
ARCH2	1	0.278166	0.0846	3.287	0.0010
ARCH3	1	0.157807	0.0608	2.594	0.0095
TDFI	1	0.178074	0.0465	3.833	0.0001

Figure 2.2.9. Partial autocorrelations of squares of log returns of Hang Seng index and parameter estimates in the $ARCH(3)$ model for stock returns.

und danach das erzeugende Programm mit einer ausführlichen Erläuterung.

```
***     Program 2_2_9     ***;
TITLE1 'ARCH(3)-model';
TITLE2 'Hongkong Data';
* Note that this program needs data1 generated by program 2_2_8;

PROC ARIMA DATA=data1;
   IDENTIFY VAR=y2 NLAG=50 OUTCOV=data2;

SYMBOL1 C=RED V=DOT H=0.5 I=JOIN;
PROC GPLOT DATA=data2;
   PLOT partcorr*lag / VREF=0;
RUN;

PROC AUTOREG DATA=data1;
   MODEL y = / NOINT GARCH=(q=3) DIST=T;
RUN;
```

To identify the order of a possibly underlying ARCH process for the daily log returns of the Hang Seng closing index, the empirical partial autocorrelations of their squared values, which are stored in the variable `y2` of the data set `data1` in Program 2_2_8, are calculated by means of `PROC ARIMA` and the `IDENTIFY` statement. The subsequent procedure `GPLOT` displays these partial autocorrelations. A horizontal reference line helps to decide whether a value is substantially different from 0.

`PROC AUTOREG` is used to analyze the ARCH(3) model for the daily log returns. The `MODEL` statement specifies the dependent variable `y`. The option `NOINT` suppresses an intercept parameter, `GARCH=(q=3)` selects the ARCH(3) model and `DIST=T` determines a t distribution for the innovations Z_t in the model equation. Note that, in contrast to our notation, SAS uses the letter `q` for the ARCH model order.

Am Ende gibt es noch eine abschließende Erklärung und Analyse der von SAS erzeugten Graphiken und Outputs und die Herstellung der Verbindung mit dem anfänglichen Theorieteil.

> We use an $ARCH(3)$ model for the generation of y_t, which seems to be a plausible choice by the partial autocorrelations plot. If one assumes t-distributed innovations Z_t, SAS estimates the distribution's degrees of freedom and displays the reciprocal in the `TDFI`-line, here $m = 1/0.1780 = 5.61$ degrees of freedom. Following we obtain the estimates $a_0 = 0.000214$, $a_1 = 0.147593$, $a_2 = 0.278166$ and $a_3 = 0.157807$. The SAS output also contains some general regression model information from an ordinary least squares estimation approach, some specific information for the (G)ARCH approach and as mentioned above the estimates for the ARCH model parameters in combination with t ratios and approximated p-values.

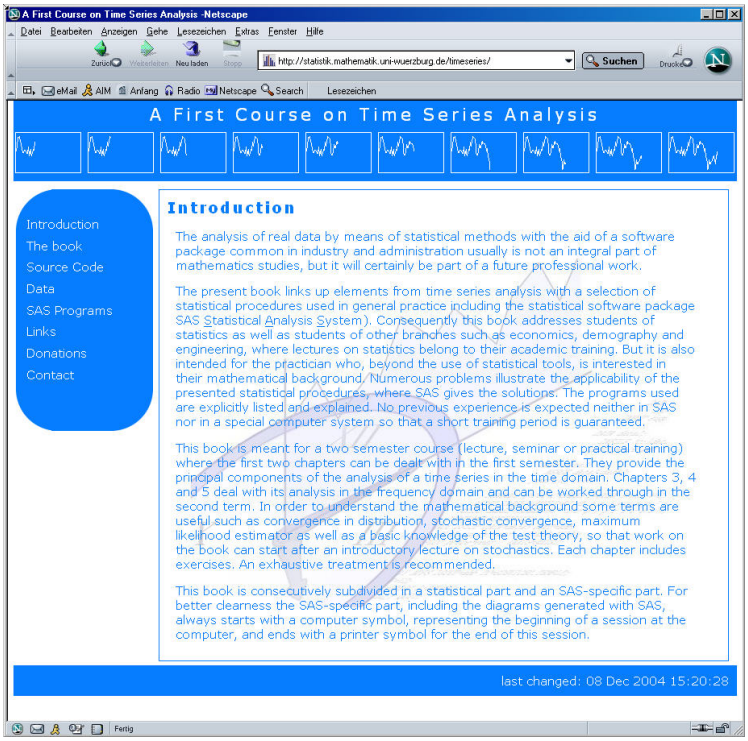
Diese ist obligatorisch, da neben dem Programmieren ein wichtiger Aspekt des Umgangs mit SAS die Auswertung des Outputs und die Beantwortung der ursprünglich gestellten Fragen ist. Dies soll immer dadurch geschehen, dass die im theoretischen Teil eingeführten Größen im SAS-Output identifiziert werden.

6 Der Internetauftritt

Zum Abschluß soll noch kurz die Internetseite des Projektes vorgestellt werden. Sie hat die Adresse

http://statistik.mathematik.uni-wuerzburg.de/timeseries

Auf der Homepage des Projektes findet sich zunächst eine Einführung, sowohl in die Open-Source-Struktur des Projektes als auch inhaltlich.



Unter den weiteren Menüpunkten kann das Buch (komplett oder in einzelne Kapitel getrennt) heruntergeladen werden, genauso wie der LaTeX-Quellcode mit den Abbildungen und Graphiken.

Die zu den Beispielen gehörenden Datensätze finden sich ebenfalls zum Download bereit (wieder komplett oder einzeln). Der Benutzer kann dabei zwischen dem txt-Format und dem SAS-Datenformat wählen. Zu allen Datensätzen finden sich Übersichten mit der Herkunft der Daten und weiteren Erläuterungen sowie der Output von PROC CONTENTS.

Genauso stehen die in den Beispielen des Buches verwendeten Programme zum Download (wieder komplett oder einzeln) bereit. Auch hier gibt es wieder Einzelheiten zu den Programmen, an vorderster Stelle die Erläuterungen zur Funktionsweise.

Neben einer Linksammlung zum Thema Zeitreihenanalyse und SAS sowie der Kontaktadresse

timeseries@statistik.mathematik.uni-wuerzburg.de

gibt es auch die Möglichkeit dieses Projekt durch finanzielle Spenden[2] zu unterstützen. Diese würden uns sehr helfen das Projekt für längere Zeit gesichert fortzuführen.

Literatur

[1] M. Falk, F. Marohn, R. Becker (1995) – Angewandte Statistik mit SAS - Springer

[2] M. Falk, F. Marohn, B. Tewes (2002) – Foundations of Statistical Analysis and Applications with SAS - Birkhäuser

[2] Kontonummer: 743 015 40, Empfänger: Staatsoberkasse Bayern in Landshut, Geldinstitut: Bundesbank Regensburg, BLZ: 750 000 00, Verwendungszweck (wichtig, bitte unbedingt angeben!): 1517010/824030-1