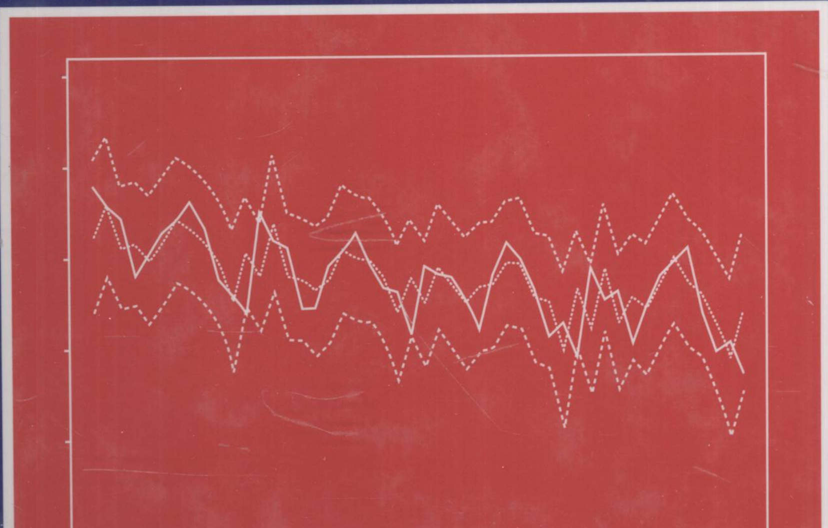
WILEY

Regression Models for Time Series Analysis



Benjamin Kedem
Konstantinos Fokianos

WILEY SERIES IN PROBABILITY AND STATISTICS

Regression Models for Time Series Analysis

BENJAMIN KEDEM
University of Maryland

KONSTANTINOS FOKIANOS
University of Cyprus

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

This text is printed on acid-free paper. ♾

Published simultaneously in Canada.

For ordering and customer service, call 1-800-CALL WILEY.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 0-471-36355-3

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Regression Models for Time Series Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

To Carmella
B. K.

To my mother and to the memory of my beloved grandmother
K. F.

Preface

Regression methods have been an integral part of time series analysis for a long time, dating back at least one hundred years to the work of Schuster (1898) [379]. Schuster's work on sinusoidal regression was applied in the estimation of "hidden periodicities" and led to the invention of the periodogram. Structural regression models for time series have been around for many years and have figured prominently in the econometrics and business literature. Treated rigorously by Anderson (1971) [20], and Fuller (1996) [161] among others, these structural models have been used for years in forecasting and decomposition of time series into "trend", "seasonal", and "irregular" components. Another distinctive example is the class of autoregressive integrated moving average models that came to be associated with Box and Jenkins (1976) [61] but has its roots in the pioneering work of E.E. Slutski and G.U. Yule in the 1920s, and of H.O. Wold in the 1930s. Most of the aforementioned work deals with linear models for time series assuming continuous values. However, there are many instances in practice where the data are not continuous and a linear model is not appropriate. This points to the necessity for alternative modeling.

This book introduces the reader to relatively newer developments and somewhat more diverse regression models and methods for time series analysis. It has been written against the backdrop of a vast modern literature on regression methods for time series and related topics as is apparent from the long list of references.

A relatively recent statistical development is the important class of models known as *generalized linear models* (GLM) that was introduced by Nelder and Wedderburn (1972) [336], and which provides under some conditions a unified regression theory suitable for continuous, categorical, and count data. The theory of GLM was originally intended for independent data, but it can be extended to dependent data under various assumptions. In the first four chapters of this book the GLM methodology is extended systematically to time series where the primary and covariate data are both random and stochastically dependent. There are three notions which enable this [152], [395]. The notion of an increasing sequence of histories relative to an observer, the notion of *partial likelihood* introduced by Cox (1975) [105] and further elaborated on by Wong (1986) [439], and the notion of martingale with respect to a sequence of histories. The latter, under suitable conditions, is applied in asymptotic inference including goodness of fit.

After a general introduction to time series that follow generalized linear models in Chapter 1, Chapters 2, 3, and 4 specialize to regression models for binary, categorical, and count time series, respectively. Chapter 5 is an introduction to various regression models developed during the last thirty years or so, particularly regression models for integer valued time series including hidden Markov models. Chapter 6 summarizes classical and more recent results concerning state space models. The last chapter, Chapter 7, presents a Bayesian approach to prediction and interpolation in spatial data adapted to time series that may be short and/or observed irregularly. We also describe a specially designed software for the implementation of the Bayesian prediction method. A brief introduction to stationary processes can be found in the Appendix. Throughout the book there are quite a few real data applications and further results presented by means of problems and complements.

Parts of the book were taught at the University of Maryland to a mixed audience of beginning and more advanced graduate students. Based on our experience, the book should be accessible to anyone who is familiar with basic modern concepts of statistical inference, corresponding roughly with the master's degree level. A basic course in applied stochastic processes consistent with the level of Parzen (1962) [343] is helpful.

We are very grateful to V. De Oliveira, L. Fahrmeir, R. Gagnon, N.O. Jeffries, C. Kedem, D.E.K. Martin, M. Nerlove, J. Picka, E. Russek-Cohen, T.J. Santner, and others who read parts of the book and provided very helpful suggestions. Special thanks are due to Amy Hendrickson of TeXnology Inc. for many useful LaTeX tips. We would like to acknowledge with thanks the travel support from the University of Cyprus and the grant support, over many years, from the National Aeronautics and Space Administration (NASA). Finally we thank our families for their unlimited support and patience throughout this project.

BENJAMIN KEDEM, College Park, Maryland
KONSTANTINOS FOKIANOS, Nicosia, Cyprus

Contents

1

Time Series Following Generalized Linear Models

In ordinary linear regression, a most useful and much dealt with statistical tool, the problem is to relate the mean response of a variable of interest to a set of explanatory variables by means of a linear equation. In many cases this is done under the assumption that the data are normal and independent. There are situations however, regarding non-normal observations such as binary and count data, when ordinary linear regression leads to certain inconsistencies, some of which are resolved very elegantly and successfully by generalized linear models. Emboldened by this success, we wish to import ideas from generalized linear models in modeling time series data. The question then is how to extend the generalized linear models methodology to time series where the data are dependent and the covariates and perhaps even the auxiliary data are time dependent and also random. As we shall see, by using partial likelihood we can transport quite straightforwardly the main inferential features appropriate for independent data to time series, not necessarily stationary, following generalized linear models. An essential component of this is that partial likelihood allows for temporal or sequential conditional inference with respect to a filtration generated by all that is known to the observer at the time of observation. This enables very flexible conditional inference that can easily accommodate autoregressive components, functions of past covariates, and all sorts of interactions among covariates.

In this chapter we provide the necessary background and an overview of generalized linear models by discussing their theoretical underpinnings, having in mind dependent time series data. Specifically, we define what we mean by time series following generalized linear models, introduce the notion of partial likelihood, and discuss in some detail the statistical properties–including large sample results–of the maximum partial likelihood estimator. Examples of special cases are presented at the end of the chapter and in subsequent chapters.

1.1 PARTIAL LIKELIHOOD

The likelihood, defined as the joint distribution of the data as a function of the unknown parameters, lies at the core of statistical theory and practice and its importance cannot be exaggerated. When the data are independent or when the dependence in the data is limited, the likelihood is readily available under appropriate assumptions on the factors in terms of which the joint distribution is expressed. In practice, however, things tend to be more complicated as the nature of dependence is not always known or even understood and consequently the likelihood is not within an easy reach. This gives the impetus for seeking suitable modifications usually by means of clever conditioning. Partial likelihood is an example of such a modification.

To motivate partial likelihood, consider a time series $\{Y_t\}$, $t = 1, \ldots, N$, with a joint density $f_{\boldsymbol{\theta}}(y_1, \ldots, y_N)$ parametrized by a vector parameter $\boldsymbol{\theta}$. In addition, suppose there is some auxiliary information AI known throughout the period of observation. Then the likelihood is a function of $\boldsymbol{\theta}$ defined by the equation

$$f_{\boldsymbol{\theta}}(y_1, \ldots, y_N|\mathrm{AI}) = f_{\boldsymbol{\theta}}(y_1|\mathrm{AI}) \prod_{t=2}^{N} f_{\boldsymbol{\theta}}(y_t \mid y_1, y_2, \ldots, y_{t-1}, \mathrm{AI}). \tag{1.1}$$

When auxiliary information is not available or is not relevant, it can be dropped from the equation as we shall do forthwith to simplify the notation to

$$f_{\boldsymbol{\theta}}(y_1, \ldots, y_N) = f_{\boldsymbol{\theta}}(y_1) \prod_{t=2}^{N} f_{\boldsymbol{\theta}}(y_t \mid y_1, y_2, \ldots, y_{t-1}). \tag{1.2}$$

The main difficulty with (1.2) is that quite generally, if no additional assumptions are made, as the series size N increases so does the size of $\boldsymbol{\theta}$. Hence, instead of getting more and more information about a fixed set of parameters, we obtain information but about an increasing number of parameters, a fact which raises consistency as well as modeling problems. This is rectified when the conditional dependence in the data is limited and the increased amount of information obtained by a growing time series size concerns a fixed set of parameters.

Appropriate assumptions and modifications of the general likelihood (1.2) are called for to accommodate dependent time series data. Helpful clues in the search for a successful definition of "likelihood" can be obtained from Markovian time series, and the notion of partial likelihood advanced by Cox [104], [105].

Markov dependence of some order typifies what we mean by conditional limited dependence. As an example, suppose we observe a first order stationary Markov process, $\{Y_t\}$, at $t = 1, \ldots, N$, and that $f_{\boldsymbol{\theta}}(y_1, \ldots, y_N)$ is the joint density of the observations where $\boldsymbol{\theta}$ is a fixed vector parameter. Due to the Markov assumption the joint density can be factored as

$$f_{\boldsymbol{\theta}}(y_1, \ldots, y_N) = f_{\boldsymbol{\theta}}(y_1) \prod_{t=2}^{N} f_{\boldsymbol{\theta}}(y_t \mid y_{t-1}). \tag{1.3}$$

Ignoring the first factor $f_{\boldsymbol{\theta}}(y_1)$, as it is independent of N, inference regarding $\boldsymbol{\theta}$ can be based on the product term in (1.3). This is an example of *conditional likelihood* resulting from dependent observations expressed as a product of conditional densities. The factorization (1.3), without $f_{\boldsymbol{\theta}}(y_1)$, has some desirable properties worth keeping in mind, such as the fact that the dimension of the factors, as well as that of $\boldsymbol{\theta}$, is fixed regardless of N, and that the derivative with respect to $\boldsymbol{\theta}$ of the logarithm of (1.3) is a zero mean square integrable martingale (see [191].) The latter is useful when studying the asymptotic properties of the resulting maximum likelihood estimator. Important early references where the martingale property was recognized and applied in statistical inference are [50] and [390]. In [50], a central limit theorem for martingales was proved and applied in asymptotic large sample theory.

Next we turn to an idea due to Cox [105] who suggested using only a part of (1.2) such as a factorization that consists only of the odd numbered conditional densities. This suggests an inference based on *partial likelihood*. More precisely, consider an occasion when a time series is observed jointly with some *random time dependent covariates*. Thus, suppose we observe a pair of jointly distributed time series, (X_t, Y_t), $t = 1, \ldots, N$, where $\{Y_t\}$ is a *response* series and $\{X_t\}$ is a *time dependent random covariate*. Employing the rules of conditional probability, as was done in (1.2) and (1.1), the joint density of all the X, Y observations can be expressed as

$$f_{\boldsymbol{\theta}}(x_1, y_1, \ldots, x_N, y_N) = f_{\boldsymbol{\theta}}(x_1) \left[\prod_{t=2}^{N} f_{\boldsymbol{\theta}}(x_t \mid d_t)\right] \left[\prod_{t=1}^{N} f_{\boldsymbol{\theta}}(y_t \mid c_t)\right], \tag{1.4}$$

where $d_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1})$ and $c_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1}, x_t)$. The second product on the right hand side of (1.4) constitutes a partial likelihood according to [105] and can be used for inference. Clearly, there is information about $\boldsymbol{\theta}$ in the first product as well, and a question arises as to what happens when this factor is ignored. It turns out that under some reasonable conditions the loss of information due to the ignored factor is small, and in exchange the remaining factor is a simplified yet useful likelihood function. The adjective "partial" also refers to the fact that the remaining factor does not specify the full joint distribution of the response and the covariate data.

The previous discussion points to the potentially useful idea of forming certain likelihood functions by taking products of conditional densities, where the densities depend on a fixed parameter and where the formed products do not necessarily give complete joint or full likelihood information. This motivates the following definition of partial likelihood with respect to a nested sequence of conditioning histories.

Definition 1.1.1 Let $\mathcal{F}_t$, $t = 0, 1, \ldots$ be an increasing sequence of σ-fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \ldots$, and let $Y_1, Y_2, \ldots$ be a sequence of random variables on some common probability space such that Y_t is $\mathcal{F}_t$ measurable. Denote the density of Y_t, given $\mathcal{F}_{t-1}$, by $f_t(y_t;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in R^p$ is a fixed parameter. The partial likelihood (PL) function relative to $\boldsymbol{\theta}$, $\mathcal{F}_t$, and the data $Y_1, Y_2, \ldots, Y_N$, is given by the product

$$\mathrm{PL}(\boldsymbol{\theta}; y_1, \ldots, y_N) = \prod_{t=1}^{N} f_t(y_t;\boldsymbol{\theta}). \tag{1.5}$$

According to Definition 1.1.1, the notion of partial likelihood generalizes both concepts of likelihood and conditional likelihood. Indeed, partial likelihood simplifies to ordinary likelihood when auxiliary information is absent and the data are independent while it becomes a conditional likelihood if the covariate process is deterministic, that is, known throughout the period of observation. Partial likelihood takes into account only what is known to the observer up to the time of actual observation, that is, it allows for *sequential conditional inference*. Closely associated with this is the martingale property alluded to earlier; it manifests itself in Section 1.4.2 on large sample results for generalized linear models. Evidently, partial likelihood does not require full knowledge of the joint distribution–that is, joint statistical dynamics–of the response and the covariates. This enables conditional inference for a fairly large class of "transition" or "transitional" non-Markovian processes where the response depends on its past values *and* on past values of the covariates. See Remark 1.2.1 and compare with [75], [123, Ch. 10].

The vector $\boldsymbol{\theta}$ that maximizes equation (1.5) is called the maximum partial likelihood estimator (MPLE). Its theoretical properties, including consistency, asymptotic normality, and efficiency, have been studied extensively in [439].

Definition 1.1.1 has been extended to continuous time stochastic processes in connection with survival analysis in [392], [393]. Additional references that treat theoretical properties of partial likelihood processes include [224] and [225]. Ramifications of partial likelihood have been considered by several authors. The notion of *marginal partial likelihood* was introduced in [174], and that of *projected partial likelihood* for modeling longitudinal data with covariates subject to drop-out is studied in [333]. Other types of pseudo-likelihoods have been considered by a fairly large number of authors of which we mention the pseudo-likelihood introduced in [45], [46] for spatial data analysis, and the notion of *empirical likelihood* introduced in [341] for nonparametric inference. See [183] for a general treatment of *pseudo-likelihood* and [342, Ch. 4] for a survey of pseudo-likelihoods including *profile* and empirical likelihoods.

1.2 GENERALIZED LINEAR MODELS AND TIME SERIES

Let $\{Y_t\}$ be a time series of interest, called the *response*, and with an eye toward prediction, let

$$\mathbf{Z}_{t-1} = (Z_{(t-1)1}, \ldots, Z_{(t-1)p})',$$