Abstract

Many statistical models rely on assumptions towards the distribution of the data and on the type of influence of the covariates. These assumptions are useful to simplify the estimation of the model. However, these assumptions regularly restrict the applicability of a model. Therefore, statistical models with more flexibility and relaxed assumptions are available.

A rather arbitrary assumption is that all covariates influence the response linearly. So all models discussed in the following relax this assumption and allow for any smooth function as a relationship between covariates and the response. These models are called additive models or models with semiparametric predictors. Another restriction is the distribution of the response variable. In the classical linear model Gaussian data are assumed, but several approaches exist that relax this assumption.

First of all, generalized linear or additive models are not restricted to the Gaussian distribution, but accept responses whose distribution is a member of the exponential family. The choice of the correct distribution is often evident by the structure of the data. Additionally, these models require the assumption of a response function, which is in applications often chosen without any proof, even though severe bias on the estimates occurs if the wrong response function is chosen. Therefore, I propose in a contribution to this thesis a method where the response function is estimated jointly with the semiparametric predictors. The approach remains in the maximum likelihood framework and provides inference for the estimates.

Alternatively, the assumption of any parametric distribution can be avoided with expectile regression. There the estimation is based on least asymmetric weighted squares approaches and does not depend on the assumption of a distribution nor on homoscedastic error terms. Thus, in combination with smooth predictors, this results in very flexible models. Nevertheless, as in any other model, the choice of the right covariates is important. Including non-informative covariates results in predictions with a high variance, while missing informative covariates will result in biased estimates. Therefore, model selection is also a major issue in expectile regression and will be considered in this thesis. To account for the dependence of expectile regression on the asymmetry parameter, two possibilities to specify the model are presented. Either the models are selected for each asymmetry parameter separately, or jointly for the whole distribution. Additionally, our methods are able to decide whether a smooth effect is necessary, or a linear effect would be sufficient. The model selection itself is either performed by goodness-of-fit criteria or shrinkage approaches. Besides the correct choice of single covariates interactions between covariates can be considered to improve the model fit. Thus, tensor products of one-dimensional P-splines provide convenient solutions. Moreover, the separation of main effects and interaction effects usually results in improved interpretations. An important example, where high-dimensional interactions are often applied, are temporal variations of spatial effects. However, these models usually depend on the correct specification of a distribution. To avoid this assumption I propose expectile regression with spatio-temporal predictors in an article contributing to this dissertation.

All methods suggested in this thesis are accompanied with real-world examples from economics, biology or meteorology. Furthermore, R-code to apply the models on new data sets is also provided.

Zusammenfassung

Statistische Modelle basieren im Allgemeinen auf Annahmen über die Verteilung der Daten und die Art des Einflusses der Kovariablen. Diese Annahmen sind sinnvoll um Modelle schätzen zu können, aber sie schränken regelmäßig die Anwendbarkeit ein. Deshalb wurden statische Modelle mit weniger Annahmen und mehr Flexibilität entwickelt.

Eine stark einschränkende Annahme ist, dass alle Kovariablen einen linearen Einfluss auf die Zielgröße haben. Daher werden im Folgenden nur Modelle betrachtet deren Einfluss auf die Zielgröße eine beliebige glatte Funktion ist. Diese Modelle werden additive Modelle oder Modelle mit semiparametrischen Prädiktor genannt. Eine weitere Annahme betrifft die Verteilung der Zielgröße. Im klassischen linearen Modell werden normalverteilte Zielvariablen betrachtet, aber auch hierfür existieren allgemeiner gültige Modelle.

Einerseits gibt es Generalisierte Lineare bzw. Additive Modelle, bei denen die Daten nicht mehr nur normalverteilt sein müssen, sondern aus einer beliebigen Verteilung aus der Exponentialfamile kommen dürfen. Jedoch benötigen diese Modelle die Annahme einer Responsefunktion, welche oftmals ohne Nachweis der Gültigkeit verwendet wird. Falls die falsche Responsefunktion verwendet wird, treten verzerrte Schätzungen auf. Daher wird in einem der Beiträge zu dieser Dissertation ein Modell vorgeschlagen, bei dem die Responsefunktion gemeinsam mit den glatten Kovariableneffekten geschätzt wird. Da der Ansatz in die Maximum Likelihood Theorie eingebettet ist, ergeben sich auch Konfidenzintervalle für die Schätzer.

Andererseits existieren Modellen, bei denen keine exakte Verteilung der Daten angenommen wird. So wird bei der Expektilregression das Modell mittels asymmetrisch gewichteten kleinsten Quadraten geschätzt und ist somit frei von der Annahme einer Verteilung und erlaubt sogar die Modellierung heteroskedastischer Fehler. Daher ist dieses Modell, zusammen mit semiparametrischen Prädiktoren, sehr flexibel, wobei auch hier die Auswahl der richtigen Kovariablen entscheidend ist. So führen unnötige Kovariablen zu unsicheren Vorhersagen, während fehlende Kovariablen zu Verzerrungen führen. Daher ist auch bei der Expektilregression die Modellwahl entscheidend und wird in einem Beitrag genauer vorgestellt. Durch die Abhängigkeit der Expektilregression von den Asymmetrieparametern ergeben sich zwei Möglichkeiten die Modelle zu wählen. Entweder wird für jeden Asymmetrieparameter eine separate Modellwahl durchgeführt, oder das Modell für die gesamte Verteilung optimiert. Zusätzlich stellen wir einen Ansatz vor, bei dem durch Modellwahl entschieden wird, ob eine Kovariable als glatter Effekt, oder als linearer Effekt am besten modelliert wird. Die Modellwahl an sich wird entweder über Gütekriterien, oder über Shrinkage ausgeführt. Neben der korrekten Auswahl einzelner Kovariablen können oft auch Interaktionen zwischen den Kovariablen hilfreich sein. Dafür bieten sich Tensorprodukte von eindimensionalen P-Splines an. Bei diesen Interaktionen ergeben sich sinnvollere Interpretationen, falls zusätzlich die Haupteffekte von der eigentlichen Interaktion getrennt werden. Ein wichtiges Beispiel für mehrdimensionale Interaktionen stellen zeitlich variierende räumliche Effekte dar. In der bestehenden Literatur basieren diese Modelle aber immer auf der korrekten Spezifikation der Verteilung, weshalb in einem Beitrag räumlich-zeitliche Modelle für Expektile, also ohne Verteilungsannahme, vorgestellt werden.

Alle Modelle, die diese Dissertation beinhaltet, werden anhand von echten Daten aus den Wirtschaftswissenschaften, der Biologie und der Meteorologie erklärt. Außerdem wird `R`-Code zur Anwendung dieser Methoden auf neue Daten zur Verfügung gestellt.