Thesis
Essays on Semiparametric Model Selection and Model Averaging

Arihiro Yoshimura

Abstract

Econometric methods and the resulting implications are strongly affected by the specification of the models. The omitted variable bias, for instance, is one of the problems caused by the misspecification. If one uses the misspecified models, the resulting economic implications are no longer reasonable because the analysis based on the misspecified models could lead to biased estimates for the true parameters of the models. Since most economic phenomena are quite complex, applied researchers need to pay attention to the model specification.

When considering the specification of the models, one can consider the nonparametric models and techniques. Nonparametric methods are quite robust to the model specification and then useful to deal with the complex phenomena. However, the problem of "curse of dimensionality," for example, arisen in the fully nonparametric methods, results in the slower rates of convergence of the estimators. Due to the difficulty of efficient aggregation of the information, requirements that are too loose regarding model specification can result in an inaccurate estimation.

On the other hand, the semiparametric models and techniques are quite useful in econometric analysis because of the model specification flexibility. Here, the term "semiparametric" in this thesis refers to the middle approach of the parametric and the fully nonparametric methods. In particular, this thesis mainly considers the semiparametric models that include explicitly some unknown functions to be estimated (or approximated) in a part of the models. In semiparametric setup, one can allow a part of the model to be parametric without specifying a functional form of an unknown part of the model. By incorporating the information about the known structures of the model as much as possible, the semiparametric approach provides more efficient estimation for the parameter of interest than the fully nonparametric methods. This flexibility and the efficiency of the semiparametric methods are quite reasonable and thus suited to the econometric analysis.

When approximating unknown functions by semiparametric methods, one usually needs to introduce some additional smoothing parameters. The question is how to choose these "nuisance" parameters in practice, even though the asymptotic theory justifies the validity of the semiparametric methods. For example, if one approximates (or estimates) an unknown function by the kernel method, the estimator of the parameter of interest and its finite sample distribution are practically affected by the selection of the bandwidth. Another notable example is the approximation of an unknown function by the nonparametric series method. In that case, any model to be used in practice is at most an "approximation" of the true model, that is, this is intrinsically misspecified. Hence, one needs to select one particular "wrong" model, while one still focuses on the true parameters as the target of the analysis.

Developing the model selection procedures for the semiparametric models is one of the most important challenges in econometrics. An interesting feature of model selection is that it

basically seeks for the “good” or “best” model rather than the true model. In fact, the spirit behind the model selection is that the true (or bigger) model is not necessarily the best model, even if one intends to estimate the true parameters in the true model. In model selection, the objective (target) of the analysis is crucial. For the given purposes of the analysis, the sense of adequacy of the model differs.

In addition, investigating the model averaging methods for semiparametric models is also an important research topic. Model averaging is useful alternative to model selection and it combines the candidate models rather than select one particular model. Model averaging has several properties that are better than model selection. For example, it can often improve the estimation risk versus model selection and this point is independently important. The developments of the model averaging method for semiparametric models are also worth considering.

This thesis addresses some issues regarding the semiparametric model selection and model averaging. The outline of this thesis is organized as follows.

In Chapter 2, we briefly review the existing model selection criteria and their properties. The purpose of this review is to summarize the motivations and relationships between the existing procedures. In particular, we focus on the AIC, FIC, and Mallows' C_p because these model selection criteria play important roles throughout this thesis. The contents in this review improve the understanding of our proposed model selection method.

In Chapter 3, we propose a new focused information criterion for variable selection in semiparametric partially linear models. Our criterion is designed to select an optimal model for estimating a focus parameter, which is a parameter of interest. We estimate the model by the series method and jointly select the variables in the linear part and the number of series terms in the nonparametric part. A Monte Carlo simulation illustrates that the proposed criterion successfully selects the model that has a relatively small mean squared error of the estimator for the focus parameter. This study has been conducted as a collaborative research with Naoya Sueishi.

In Chapter 4, we provide a brief review of the recent model averaging methods. The purpose of this review is to summarize the principles and the properties of the model averaging. In particular, we focus on the study of model averaging for regression models that have recently been studied intensively. In these studies, we see that the optimality results of model selection criteria are carried over into the model averaging procedures. The content in this review leads to our subsequent studies on model averaging methods provided in Chapters 5 and 6.

In Chapter 5, we propose a method of averaging generalized least squares estimators for linear regression models with heteroskedastic errors. The averaging weights are chosen to minimize Mallows' C_p-like criterion. We show that the weighting vector selected by our procedure is asymptotically optimal. It is also shown that the optimality holds even when the variances of the error terms are estimated and the feasible generalized least squares estimators are averaged. The variances can be estimated not only parametrically but also nonparametrically. Monte Carlo simulation results support our procedure and an empirical example illustrates that the proposed method is useful for predicting a measure of firms' performance. This study has been conducted as a joint research with Qingfeng Liu and Ryo Okui.

In Chapter 6, we examine a method of averaging series-based least squares estimators for semiparametric partially linear varying coefficient models. In this study, our major aim is to improve the estimation risk for the varying coefficient functions. We derive a Mallows' C_p-like criterion by calculating an estimate of the mean squared errors of estimation of the nonparametric part. The averaging weights are chosen to minimize this criterion. We show that this method achieves optimality in the sense that the mean squared error evaluated at the chosen weights is asymptotically equivalent to the optimal risk. Monte Carlo simulation illustrates that our model averaging procedure works well compared with existing selection methods for estimating the varying coefficient functions.

Mathematical proofs regarding Chapters 5 and 6 are provided in the Appendix.