Journal of Econometrics 177 (2013) 357–373

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Complete subset regressions☆

Graham Elliott[a], Antonio Gargano[b], Allan Timmermann[a,*]

[a] UC San Diego, United States

[b] University of Melbourne, Australia

ARTICLE INFO

Article history:
Available online 17 April 2013

Keywords:
Subset regression
Forecast combination
Shrinkage

ABSTRACT

This paper proposes a new method for combining forecasts based on complete subset regressions. For a given set of potential predictor variables we combine forecasts from all possible linear regression models that keep the number of predictors fixed. We explore how the choice of model complexity, as measured by the number of included predictor variables, can be used to trade off the bias and variance of the forecast errors, generating a setup akin to the efficient frontier known from modern portfolio theory. In an application to predictability of stock returns, we find that combinations of subset regressions can produce more accurate forecasts than conventional approaches based on equal-weighted forecasts (which fail to account for the dimensionality of the underlying models), combinations of univariate forecasts, or forecasts generated by methods such as bagging, ridge regression or Bayesian Model Averaging.

1. Introduction

Methods for controlling estimation error in forecasting problems that involve small sample sizes and many potential predictor variables have been the subject of much recent research.[1] One lesson learned from this literature is that a strategy of including all possible variables is typically too profligate; given the relatively short data samples typically available to estimate the parameters of economic forecasting models, it is important to limit the number of parameters that have to be estimated or in other ways reduce the effect of parameter estimation error. This has led to the preponderance of forecast methods such as shrinkage or ridge regression (Hoerl and Kennard, 1970), model averaging (Bates and Granger, 1969; Raftery et al., 1997), bagging (Breiman, 1996), and the Lasso (Tibshirani, 1996), which accomplish this in different ways.

This paper proposes a new method for combining forecasts based on complete subset regressions. For a given set of potential predictor variables we combine forecasts from all possible linear regression models that keep the number of predictors fixed. For example, with K possible predictors, there are K unique univariate models and $n_{k,K} = K!/((K-k)!k!)$ different k-variate models for $k \leq K$. We refer to the set of models for a fixed value of k as a complete subset and propose to use equal-weighted combinations of the forecasts from all models within these subsets indexed by k. Moreover, we show that an optimal value of k can be determined from the covariance matrix of the potential regressors and so lends itself to being selected recursively in time.

Special cases of subset regression combinations have appeared in the empirical literature. For example, Rapach et al. (2010) consider equal-weighted combinations of all possible univariate equity premium models and find that they produce better forecasts of stock returns than a simple no-predictability model. This corresponds to setting $k = 1$ in our context. Papers such as Aiolfi and Favero (2003) consider equal-weighted combinations of forecasts of stock returns from all possible 2^K models. While their combination scheme is not directly nested by our approach, this can nevertheless be obtained from a combination of the individual subset regression forecasts.

From a theoretical perspective, we show that subset regression combinations are akin to a complex version of shrinkage which, in general, does not reduce to shrinking the Ordinary Least Squares (OLS) estimates coefficient by coefficient. Rather, the adjustment to the coefficients depends on all least squares estimates and is a function of both k, the number of variables included in the model, and K, the total number of potential predictors. Only in the special case where the covariance matrix of the predictors is orthonormal does subset regression reduce to ridge regression or, equivalently, to a Bayes estimator with a specific prior distribution. For this special case we derive the exact degree of shrinkage implied by different values of k and thus formalize how k, the number of parameters in the conditional mean equation, is equivalent to other measures of model complexity that have previously been proposed in the literature.

☆ We thank the Editor, Herman van Dijk, and two anonymous referees for many constructive and helpful comments.

* Correspondence to: UC San Diego, Rady School of Management, 9500 Gilman Drive, La Jolla, CA 92093-0553, United States.

E-mail address: atimmermann@ucsd.edu (A. Timmermann).

[1] See, e.g., Stock and Watson (2006) for a review of the literature.

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.04.017

We also show that the weights implied by subset regression reflect omitted variable bias in a way that can be useful for forecasting. This holds particularly in situations with strongly positively correlated regressors since the subset regression estimates account for the omitted predictors.

An attractive property of the proposed method is that, unlike the ridge estimator and conventional Bayesian estimators, it does not impose the same amount of shrinkage on each coefficient. Unlike model selection methods, it also does not assign binary zero–one weights to the OLS coefficients. Other approaches that apply flexible weighting to individual predictors include bagging (Breiman, 1996) which applies differential shrinkage weights to each coefficient, the adaptive Lasso (Zou, 2006) which applies variable-specific weights to the individual predictors in a data-dependent adaptive manner, the Elastic Net (Zou and Hastie, 2005; Zou and Zhang, 2009) which introduces extra parameters to control the penalty for inclusion of additional variables, and Bayesian methods such as adaptive Monte Carlo (Lamnisos et al., 2012).

To illustrate the subset regression approach empirically we consider, like many previous studies, predictability of US stock returns. In particular, following Rapach et al. (2010), we study quarterly data on US stock returns in an application that has 12 potential predictor variables and so generates subset regressions with $k = 1, 2, \ldots, 12$ predictor variables. We find that subset regression combinations that use $k = 2$, 3, or 4 predictors produce the lowest out-of-sample mean squared error (MSE)-values. Moreover, these subset models generate superior predictive accuracy relative to the equal-weighted average computed across all possible models, a benchmark that is well-known to be difficult to beat, see Clemen (1989). We also find that the value of k in the subset regression approach can be chosen recursively (in pseudo "real time") in such a manner that the approach produces forecasts with lower out-of-sample MSE-values than those produced by recursive versions of Bayesian Model Averaging, ridge regression, Lasso, or bagging.

The outline of the paper is as follows. Section 2 introduces the subset regression approach and characterizes its theoretical properties, Section 3 presents a Monte Carlo simulation study, Section 4 conducts the empirical analysis of US stock returns, while Section 5 concludes.

2. Theoretical results

This section presents the setup for the analysis and derives theoretical results for the proposed complete subset regression method.

2.1. Setup

Suppose we are interested in predicting the univariate (scalar) variable y_{T+1} using a linear regression model based on observing K predictors $x_T \in \mathbb{R}^K$, and a history of data, $\{y_{t+1}, x_t\}_{t=0}^{T-1}$. Let $E[x_t x_t'] = \Sigma_X$ for all t and, without loss of generality, assume that $E[x_t] = 0$ for all t. To focus on regressions that include only a subset of the predictors, define β to be a $K \times 1$ vector with slope coefficients in the rows representing included regressors and zeros in the rows of the excluded variables. Let β_0 be the pseudo true value for β, the population value of the projection of y on X, where $y = (y_1, \ldots, y_T)$ is a $T \times 1$ vector and $X = (x_0, x_1, \ldots, x_{T-1})'$ stacks the x observations into a $T \times K$ matrix. Denote the generalized inverse of a matrix A by A^-. Let S_i be a $K \times K$ matrix with zeros everywhere except for ones in the diagonal cells corresponding to included variables, so that if the $[j, j]$ element of S_i is one, the jth regressor is included, while if this element is zero, the jth regressor is excluded. Sums over i are sums over all permutations of S_i.

We propose an estimation method that uses equal-weighted combinations of forecasts based on all possible models that include a particular subset of the predictor variables. Each subset is defined by the set of regression models that include a fixed (given) number of regressors, $k \leq K$. Specifically, we run the 'short' regression of y_t on a particular subset of the regressors, then average the results across all k dimensional subsets of the regressors to provide an estimator, $\hat{\beta}$, for forecasting, where $k \leq K$. With K regressors in the full model and k regressors chosen for each of the short models, there will be subset regressions to average over. In turn, each regressor gets included a total of $n_{k-1,K-1}$ times.

As an illustration, consider the univariate case, $k = 1$, which has $n_{1,K} = K$ short regressions, each with a single variable. Here all elements of $\hat{\beta}_i$ are zero except for the least squares estimate of y_t on x_{it} in the ith row. The equal-weighted combination of forecasts from the individual models is then

$$\hat{y}_{T+1} = \frac{1}{K} \sum_{i=1}^{K} x_T' \hat{\beta}_i. \tag{1}$$

Following common practice, our analysis assumes quadratic or mean square error (MSE) loss. For any estimator, we have

$$\begin{aligned} & E\left[\left(y_{T+1} - \hat{\beta}_T' x_T\right)^2\right] \\ &= E\left[\left(y_{T+1} - \beta_0' x_T + (\beta_0 - \hat{\beta}_T)' x_T\right)^2\right] \\ &= E\left[\left(\varepsilon_{T+1} + (\beta_0 - \hat{\beta}_T)' x_T\right)^2\right] \\ &= \sigma_\varepsilon^2 \left(1 + T^{-1} \sigma_\varepsilon^{-2} E\left[T(\hat{\beta}_T - \beta_0)' x_T x_T' (\hat{\beta}_T - \beta_0)\right]\right). \end{aligned} \tag{2}$$

Here ε_{T+1} is the residual from the population projection of y_{T+1} on x_T and σ_ε^2 is its variance. We concentrate on the last term since the first term does not depend on $\hat{\beta}$. Hence, we are interested in examining $\sigma_\varepsilon^{-2} E\left[(\hat{\beta}_T - \beta)' x_T x_T' (\hat{\beta}_T - \beta)\right]$.

2.2. Complete subset regressions

Subset regression coefficients can be computed as averages over least squares estimates of the subset regressions. When the covariates are correlated, the individual regressions will be affected by omitted variable bias. However, as we next show, the subset regression estimators are themselves approximately a weighted average of the components of the full regression OLS estimator, $\hat{\beta}_{\text{OLS}}$.

Theorem 1. *Assume that as the sample size gets large $\hat{\beta}_{\text{OLS}} \to^p \beta_0$ for some β_0 and $T^{-1}X'X \to^p \Sigma_X$. Then, for fixed K, the estimator for the complete subset regression, $\hat{\beta}_{k,K}$, can be written as*

$$\hat{\beta}_{k,K} = \Lambda_{k,K} \hat{\beta}_{\text{OLS}} + o_p(1),$$

where

$$\Lambda_{k,K} \equiv \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \left(S_i' \Sigma_X S_i\right)^- (S_i' \Sigma_X).$$

A proof of this result is contained in the Appendix.

This result on the relationship between $\hat{\beta}_{k,K}$ and the OLS estimator makes use of high level assumptions that hold under very general conditions on the data; see White (2001, Chapter 3) for a set of sufficient conditions. Effectively, any assumptions on the model that result in the OLS estimators being consistent for their population values and asymptotically normal will suffice. For