Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computers & Industrial Engineering 49 (2005) 155–167
computers & industrial engineering
www.elsevier.com/locate/dsw

Subset selection in multiple linear regression: a new mathematical programming approach☆

Burak Eksioglu*,[a], Riza Demirer[b], Ismail Capar[a]

[a]Department of Industrial Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[b]Department of Economics and Finance, Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA

Received 2 December 2004; revised 8 March 2005; accepted 28 March 2005
Available online 12 July 2005

Abstract

A new mathematical programming model is proposed to address the subset selection problem in multiple linear regression where the objective is to select a minimal subset of predictor variables without sacrificing any explanatory power. A parametric solution of this model yields a number of efficient subsets. To obtain this solution, an optimal or one of two heuristic algorithms is repeatedly used. The subsets generated are compared to ones generated by several standard procedures. The results suggest that the new approach finds subsets that compare favorably against the standard procedures in terms of the generally accepted measure: adjusted R^2.

Keywords: Mathematical programming; Heuristics; Multivariate statistics; Regression; Lagragian relaxation; GRASP

1. Introduction

A regression analyst is commonly challenged to select the best subset from a set of predictor variables using some specified criterion. Historically, when there are many predictor variables, one or more subsets with fewer predictor variables are generated using a method of the analyst's choice. Given data of the form $\{\{y_i, x_{1i},\ldots,x_{ki}\}, i=1,\ldots,n\}$, the subset selection problem involves selecting a subset M of N ($M \subseteq N$) where $N=\{1,\ldots,k\}$ is the index set of the predictor variables $\{\mathbf{X}_1,\ldots,\mathbf{X}_\mathbf{k}\}$ such that some measure of the model's explanatory power is maximized.

☆ This manuscript was processed by Area Editor Elsayed A. Elsayed.

* Corresponding author. Tel.: +1 662 325 7625; fax: +1 662 325 7618.
E-mail address: beksioglu@ie.msstate.edu (B. Eksioglu).

0360-8352/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cie.2005.03.004

The main motivation for subset-selection seems to be parsimony: "if 3 regressors can 'explain' or 'satisfactorily fit' a response **Y**, why use 4?" as Mandel (1989) notes. Some of the reasons for using only a subset of the available predictor variables (given by Miller, 1984) are

- to estimate or predict at a lower cost by reducing the number of variables on which data are to be collected;
- to predict more accurately by eliminating uninformative variables;
- to describe multivariate data sets parsimoniously; and
- to estimate regression coefficients with smaller standard errors (particularly when some of the predictors are highly correlated).

A number of studies in the statistical literature discuss the problem of selecting the best subset of predictor variables in regression. Such studies focus on subset selection methodologies, selection criteria, or a combination of both. The traditional selection methodologies can be enumerative (e.g. all subsets and best subsets procedures), sequential (e.g. forward selection, backward elimination, stepwise regression, and stagewise regression procedures), and screening-based (e.g. ridge regression and principal components analysis). Standard texts like Draper and Smith (1998) and Montgomery and Peck (1992) provide clear descriptions of these methodologies. New methodologies, such as Broersen's (1986) stepwise directed search and Breiman's (1995) nonnegative garrote were developed recently. Mitchell and Beauchamp (1988) create a parallel approach to the subset selection problem using a Bayesian perspective. With respect to the selection criteria, a number of measures have been proposed such as adjusted R^2, Mallow's Cp, and Akaike's AIC. Once again, Draper and Smith (1998) and Montgomery and Peck (1992) offer adequate explanations on this topic.

Several papers can help understand the state-of-the-art in subset selection research. Hocking's (1976) early work provides a detailed overview of the field until the mid-70 s. At about the same time, Berk (1978) reports a computational comparison of various selection procedures, and Thompson (1978a, b) details both a review and an evaluation of selection procedures and criteria. Subsequently, Miller (1984) offers a comprehensive survey of selection methods and criteria and discusses the potential pitfalls an analyst faces in using subset selection. Grechanovsky (1987) provides a somewhat similar account, though in a limited way. Sparks, Zucchini, and Coutsourides (1985) examine the same issues, but for the case when there are multiple Y variables. Hoerl, Schuenemeyer, and Hoerl (1986) report a computational study involving ridge regression, sequential and screening-based subset selection. Cavalier and Melloy (1991) use a mathematical programming approach to solve the n-dimensional linear Euclidean regression problem. Recently, Kashid and Kulkarni (2002) propose a new criterion called Sp-criterion for subset selection in multiple linear regression.

Opinions regarding the advantages and disadvantages of the various procedures clearly differ, and no final word seems to be forthcoming. We propose a new mathematical programming based approach for subset selection that is similar to "all subsets" and "best subsets" procedures because it concerns itself with the selection of good subsets. However, unlike the "all subsets" procedure, it identifies only a limited number of subsets, and, unlike the "best subsets" procedure, it uses a non-traditional selection criterion. The criterion used is based on the intuition that, in a good model, the correlations between the **Y** variable and the **X** variables (**Y–X** correlations) should be high and those between the **X** variables (**X–X** correlations) should be low.