Linear Regression and Applications of MIP for Optimal Selection of Variables

Filiz Ersoz & Stan Uryasev

Department of Industrial Engineering, Karabuk University
Risk Management and Financial Engineering Lab, University of Florida

STRUCTURE OF THE PRESENTATION

- **THE INTRODUCTION**
  - Objectives of the Study
  - Background: Linear Regression
  - Linear Regression with Mean Square Error
  - R-Square
  - Adjusted R-Square
  - Background: Stepwise Multiple Regression
  - Pros and Cons of Stepwise Regression
- **THE MAIN BODY**
  - Mixed Integer Second-Order Cone Programming (MISOCP): Variable Selection
  - MISOCP : Problem statement
  - MISOCP : Numerical Experiments
  - Data
  - MISOCP: Numerical Results
  - Commercial Packages
  - Calculation Results
- **CONCLUSION**

Objective

- The purpose of this paper is to compare several approaches for selecting the best subset of explanatory variables in a multiple linear regression model.

Linear Regression

- The regression model

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_K x_{i,K} + e_i$$

$x_{i,k}$ = value of k^{th} predictor

b_0 = regression constant

b_k = coefficient on the k^{th} predictor

K = total number of predictors

y_i = predictand

e_i = error term

Linear Regression with Mean Square Error

- Prediction equation

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} + \cdots + \hat{b}_K x_{i,K}$$

- Residuals

$$\hat{e}_i = y_i - \hat{y}_i$$

y_i = observed value

$\hat{y}_i$ = predicted value

$s^2 = \dfrac{\sum e_i^2}{n - k - 1}$ = Adjusted mean-squared error (or Adjusted MSE)

Linear Regression with Mean Square Error (Cont'd)

- Standard notations:

$$SSE = \sum_{i=1}^{n} \hat{e}_i^{\,2} \qquad \text{sum of squares, error}$$

$$SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \qquad \text{sum of squares, total}$$

$$SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 \qquad \text{sum of squares, regression}$$

- Partition of variation $SST = SSR + SSE$

R-Squared

- Coefficient of multiple determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- $0 \le R^2 \le 1$

- Values closer to 1 represent better fits

- Adding of factors increases (in-sample) R^2

Adjusted R-Squared

- The adjusted coefficient of determination (R^2) is the percentage of the variability of the dependent variable that is explained by the variation of the independent variables after accounting for the intercept and number of independent variables $k \leq p$ (p= the number of all candidate variables).

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2/(n-k-1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)}$$

Background: Stepwise Multiple Regression

- In multiple regression, the *"best"* factors (i.e., subset of factors) should be identified that have the strongest relationship to a dependent variable.

- Three heuristic statistical techniques are often used for variables selection: forward selection, backward elimination, and *stepwise regression.*

Background: Stepwise Multiple Regression (Cont'd)

- In stepwise multiple regression, *the independent variables are entered* according to their statistical contribution in explaining the variance in the dependent variable.

- Variables are added to the regression equation one at a time, by maximizing the *adjusted* R^2.

Pros and Cons of Stepwise Regression

Pro's:

- Easy to code
- Fast
- Provide near optimal (and frequently optimal) solutions

Pros and Cons of Stepwise Regression (Cont'd)

Con's:

- Stepwise regression is a greedy-type heuristic and do not always provide the best subset of explanatory variables.
- Unstable (small perturbation of data can lead to very different results), may miss 'best' model.
- Is not applicable in statistical problems with constraints (e.g., portfolio replication problem with budget constraint)

Mixed Integer Second-Order Cone Programming (MISOCP) : Variable Selection

- Miyashiroa and Takano [1] suggested to select subset of explanatory variables with mixed-integer second-order cone programming formulations.

- Goodness-of-fit measures: Adjusted R^2, AIC and BIC.

[1] http://www.me.titech.ac.jp/technicalreport/h25/2013-7.pdf

MISOCP : Problem Statement [1]

- Variable selection problem by maximizing adjusted R^2

p = number of candidate variables
n = number of data points is much larger than p

$$
\begin{aligned}
\underset{\substack{a,b,\varepsilon,f,\\ g,k,z}}{\text{minimize}} \quad & f \\
\text{subject to} \quad & \varepsilon_i = y_i - \left(b + \sum_{j=1}^{p} a_j x_{ij} \right) \qquad (i = 1, 2, \ldots, n), \\
& \sum_{i=1}^{n} \varepsilon_i^2 \le f \cdot g, \\
& g = n - k - 1, \\
& -M z_j \le a_j \le M z_j \quad (j = 1, 2, \ldots, p), \\
& \sum_{j=1}^{p} z_j = k, \\
& z_j \in \{0, 1\} \qquad (j = 1, 2, \ldots, p).
\end{aligned}
$$

MISOCP : Numerical Experiments [1]

- All computations for solving: a Dell Precision T5500 PC3 (For each instances, 16GB memory, eight threads and up to 10,000 seconds were assigned for a branch-and-bound procedure)
- Stepwise regression function implemented in the statistics toolbox of MATLAB R2012b on a NEC Mate J PC.
- Numerical experiments to assess the effectiveness of the proposed MISOCP formulations (CPLEX 12.5 solver)

Data

- Downloaded from the UCI Machine Learning Repository
- 10 different datasets with different sample sizes and number of factors (*n=sample size, p= number of candidate variables*)

Abbreviation	n	p	Dataset
Housing	506	13	Housing
Servo	167	19	Servo
AutoMPG	392	25	AutoMPG
SolarFlareC	1066	26	Solar Flare (C-class flares production)
SolarFlareM	1066	26	Solar Flare (M-class flares production)
SolarFlareX	1066	26	Solar Flare (X-class flares production)
BreastCancer	194	32	Breast Cancer Wisconsin
ForestFires	517	63	Forest Fires
Automobile	159	65	Automobile
Crime	1933	100	Communities and Crime

MISOCP: Numerical Results [1]

instance	n	p	method	R^2	k	time (s)
Housing	506	13	SWR_{const}	**0.7348**	11	1.12
			SWR_{all}	0.7338	13	0.18
			MISOCP	**0.7348**	11	9.03
Servo	167	19	SWR_{const}	**0.7419**	10	1.75
			SWR_{all}	0.7348	15	0.52
			MISOCP	**0.7419**	10	4.43
AutoMPG	392	25	SWR_{const}	0.8683	17	3.39
			SWR_{all}	0.8669	22	0.71
			MISOCP	**0.8686**	16	29.83
SolarFlareC	1066	26	SWR_{const}	**0.1869**	11	3.28
			SWR_{all}	0.1818	20	1.34
			MISOCP	**0.1869**	11	184.97
SolarFlareM	1066	26	SWR_{const}	**0.0955**	9	2.80
			SWR_{all}	0.0873	20	1.24
			MISOCP	**0.0955**	9	95.61
SolarFlareX	1066	26	SWR_{const}	$0.1295^{\dagger}$	6	1.90
			SWR_{all}	0.1195	20	1.30
			MISOCP	$\mathbf{0.1295}^{\ddagger}$	6	19.03
BreastCancer	194	32	SWR_{const}	0.2305	11	3.41
			SWR_{all}	0.1999	32	0.51
			MISOCP	**0.2494**	16	3211.08
ForestFires	517	63	SWR_{const}	0.1006	22	15.65
			SWR_{all}	0.0558	60	3.72
			MISOCP	**0.1024**	26	$>$ 10000
Automobile	159	65	SWR_{const}	0.9656	43	24.02
			SWR_{all}	0.9630	55	5.71
			MISOCP	**0.9674**	35	$>$ 10000
Crime	1933	100	SWR_{const}	0.6839	65	104.63
			SWR_{all}	0.6796	100	8.89
			MISOCP	**0.6841**	53	$>$ 10000

- SWR_{const} = Stepwise regression starting with no explanatory variables
- SWR_{all} = Stepwise regression starting with all candidate variables

MISOCP: Numerical Results [1] (Cont'd)

- MISOCP provides quickly an optimal solution for small instances. For medium-sized instances, the method often generates a better subset of variables than stepwise regression.

- MISOCP solving times are much longer than the stepwise regression

IBM SPSS Statistics, IBM SPSS Modeler, Minitab

Benchmarking computations are performed with IBM SPSS Statistics, IBM SPSS Modeler, Minitab, and Portfolio Safeguard (PSG) solver.

- SPSS (*Statistical package* for the Social Sciences). The newest version of SPSS is called "IBM SPSS Statistics 20". IBM purchased SPSS in 2009.

- IBM SPSS Modeler is a set of *data mining tools* to develop predictive models using business expertise and deploy them into business operations to improve decision making. SPSS Modeler supports the entire data mining process, from data to better business results. SPSS Modeler offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics.

- Minitab *statistical package* developed at Pennsylvania State University. Minitab is often used in conjunction with the implementation of Six sigma and other statistics-based process improvement methods.

Portfolio Safeguard Package (PSG)

- PSG is an *advanced mixed-integer optimization package* tuned to solve large stochastic optimization problems
- The PSG meta-codes, data, and solutions for various case studies: http://www.aorda.com/aod/psg.action
- PSG code (simplified) for minimizing mean-squared error with cardinality constraint:

```
Problem: problem_h, type = minimize
Objective:
meansquare (matrix_s)
Constraint:  upper_bound = 12
cardn (0.000001, matrix_1)
```

Calculation Results

Datadet	Miyashiroa and Takano	Portfolio Safeguard	Data mining IBM SPSS Modeler	Statistics IBM SPSS Statistics	Minitab
Housing	0.7348 (11)	0,7344(11)(van) 0.7344(11)(car)	0.7348	0.7348	0.7215
Servo	0.7419 (10)	0.7371(10)(car) 0.7339(10)(van)	0.7404	0.7404	0.7182
AutoMPG	0.8686 (16)	0.8686(16)(car) 0.8671(16)(van)	0.8668	0.8668	0.8685
Crime	0.6841(53)	0.6735(53)(van)	0.6664	0.6664	0.6798
ForestFires	0.1024 (26)	0.1004(26)(van) 0.0941(26)(car)	0.0745	0.0745	0.0928
BreastCanser	0.2494 (16)	0.2401(16)(car) 0.2446(16)(van)	0.1949	0.1949	0.2265
SolarC	0.1869 (11)	0.1855(11)(car) 0.1846(11)(van)	0.1830	0.1830	0.1521
SolarM	0.0955 (9)	0.0905(9)(car) 0.0938(9)(van)	0.0887	0.0887	0.0949
SolarX	0.1295 (6)	0.1162(6)(car) 0.1290(6)(van)	0.1282	0.1282	0.1283

Calculation Results (Cont'd)

- Stepwise heuristic provides near optimal solutions (in many cases solutions are optimal).
- Miyashiroa and Takano [1] MISOCP algorithm gives optimal solutions, but may be slow for large dimensions. Flexible tool: can be used with additional constraints.
- IBM SPSS Statistics, IBM SPSS Modeler, Minitab sometimes provide solutions which are quite far from optimality.
- Portfolio Safeguard gives near optimal solutions. Fast and flexible: can optimize problems with additional constraints.

THANK YOU

- **For further questions, please contact us at:**

fersoz@karabuk.edu.tr

uryasev@ufl.edu