12-2018

Budget-Constrained Regression Model Selection Using Mixed Integer Nonlinear Programming

Jingying Zhang
University of Arkansas, Fayetteville

Citation

Zhang, J. (2018). Budget-Constrained Regression Model Selection Using Mixed Integer Nonlinear Programming. *Graduate Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/2994

Budget-Constrained Regression Model Selection Using Mixed Integer Nonlinear Programming

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Engineering

by

Jingying Zhang
Henan Polytechnic University
Bachelor of Science in Electrical Engineering, 2000
Henan Polytechnic University
Master of Science in Electrical Engineering, 2007
University of Arkansas
Master of Science in Mathematics, 2010

December 2018
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Ronald L. Rardin, Ph.D
Dissertation Co-Director

Justin R. Chimka, Ph.D
Dissertation Co-Director

Ed Pohl, Ph.D
Committee Member

Giovanni Petris, Ph.D
Committee Member

Abstract

Regression analysis fits predictive models to data on a response variable and corresponding values for a set of explanatory variables. Often data on the explanatory variables come at a cost from commercial databases, so the available budget may limit which ones are used in the final model.

In this dissertation, two budget-constrained regression models are proposed for continuous and categorical variables respectively using Mixed Integer Nonlinear Programming (MINLP) to choose the explanatory variables to be included in solutions. First, we propose a budget-constrained linear regression model for continuous response variables. Properties such as solvability and global optimality of the proposed MINLP are established, and a data transformation is shown to significantly reduce needed big-Ms. Illustrative computational results on realistic retail store data sets indicate that the proposed MINLP outperforms the statistical software outputs in optimizing the objective function under a limit on the number of explanatory variables selected. Also our proposed MINLP is shown to be capable of selecting the optimal combination of explanatory variables under a budget limit covering cost of acquiring data sets.

A budget-constrained and /or count-constrained logistic regression MINLP model is also proposed for categorical response variables limited to two possible discrete values. Alternative transformations to reduce needed big-Ms are included to speed up the solving process. Computational results on realistic data sets indicate that the proposed optimization model is able to select the best choice for an exact number of explanatory variables in a modest amount of time, and these results frequently outperform standard heuristic methods in terms of minimizing the negative log-likelihood function. Results also show that the method can compute the best choice of explanatory variables affordable within a given budget. Further study adjusting the objective function to minimize the Bayesian Information Criterion **BIC** value instead of negative log-likelihood function proves that the new optimization model can also reduce the risk of over-fitting by introducing a penalty term to the objective function which grows with the number of parameters.

Finally we present two refinements in our proposed MINLP models with emphasis on multiple linear regression to speed branch and bound (B&B) convergence and extend the size range of instances that can be solved exactly. One adds cutting planes to the formulation, and the second develops warm start methods for computing a good starting solution. Extensive computational results indicate that our two proposed refinements significantly reduce the time for solving the budget constrained multiple linear regression model using a B&B algorithm, especially for larger data sets.

The dissertation concludes with a summary of main contributions and suggestions for extensions of all elements of the work in future research.

Acknowledgments

I would like to express my deepest and sincerest gratitude to Dr. Ronald L. Rardin and Dr. Justin Robert Chimka for guiding me through the Ph.D. study at University of Arkansas. Their wisdom, knowledge, personality, patience, and professional guidance were invaluable to me. I am also thankful to Dr. Ed Pohl and Dr. Giovanni Petris for offering valuable comments and suggestions as my committee members.

I also thank Dr. Shengfan Zhang and Dr. Manuel D. Rossetti for their help. I am also grateful to my colleagues: Kunlei Liang, Shuohao Wu, Fan Wang, Dong Xu, Aihong Wen, and Wayne Bolinger.

Finally, I would like to thank my parents for always believing in me and supporting me. I could not sustain myself through my Ph.D. study without their endless love and faith in me. I also want to give special thanks to my husband, Yueqing Li, and my son, Tianjian Li, without whom I would not have the courage to finsih my Ph.D program.

Dedication

This dissertation is dedicated to my parents, my son and my husband for their endless love and support.

Contents

List of Tables

List of Unpublished Manuscripts

Chapter 2. Zhang, J., Rardin, R. L., and Chimka, J. R. (2018). Budget Constrained Model Selection for Multiple Linear Regression. Quality Engineering, under review.

Chapter 3. Zhang, J., Rardin, R. L., and Chimka, J. R. (2018). Budget Constrained Model Selection for Logistic Regression. Journal of the Operational Research Society, under review.

Chapter 4. Zhang, J., Rardin, R. L., and Chimka, J. R. (2018). Computational Enhancements to Accelerate Budget Constrained Regression Model Selection by Mixed Integer Nonlinear Programming. Omega, to be submitted.

1. **Introduction**

Regression analysis is a well-known tool for understanding the relationship between a response variable and a set of explanatory variables. Linear regression and logistic regression are two commonly used regression models for continuous and categorical responsible variable, respectively. Variable selection is very important in model building to identify the best subset of available explanatory variables for predicting response values. Budget considerations arise because values of explanatory variables may be available only at cost and a budget limit may restrict the subset to be included in the fitted model. Although many different models and methods have been proposed for regression and variable selection, we know of none that has reported adding a budget constraint to the regression model while doing the variable selection.

The aim of this dissertation is to propose two budget-constrained regression models for continuous and categorical variables respectively using Mixed Integer Nonlinear Programming (MINLP) to choose the explanatory variables to be included in solutions. We first propose a budget-constrained and count-constrained linear regression model for continuous response variables. Then we propose a budget-constrained and count-constrained logistic regression model for categorical response variables limited to two possible discrete values. Finally, we refine the proposed MINLP Models by adding cutting planes and warm starts to facilitate solving bigger data sets.

Good variable selection or feature selection can lead to a clear relationship between the response variable and the selected variables and improve the model building effectiveness by filtering out less-significant features. Many feature selection methods such as forward selection, backward elimination and stepwise selection are well known and deeply studied for linear regression. In the past twenty years, hardware along with algorithm improvements have resulted in a dramatic speedup of solving optimization problems, and consequently, different optimization models have become practical for solving the classical variable selection problem. Specifically, Lasso regression, Ridge regression and a naive elastic net regression, are proposed for variable selection ((Tibshirani, 1996), (Rejchel, 2016), (Park and Klabjan, 2017), (Wu et al., 2018)). Bertsimas et al (Bertsimas et al., 2016) proposed a MINLP model for selecting the best fixed number p of features for linear regression models. Instead of fixing the number of selected features, Park (Park and Klabjan, 2017) proposed an optimization model for picking the best subset of variables in terms of minimizing mean absolute error (MAE) or mean squared error (MSE).

Feature selection methods are much less studied in logistic regression. Sato (Sato et al., 2016) proposed a Mixed Integer Optimization model and Lucadamo (Lucadamo and Simonetti, 2011) proposed the Disco Coefficient method to identify the significant variables for logistic regression. Bursac (Bursac et al., 2008) proposed a method called purposeful selection of co-variates in which an analyst makes a variable selection decision at each step of the modeling process.

Although many different models and methods have been proposed for regression and variable selection, to the best of our knowledge, none of those existing studies has considered budget constrained model selection in linear or logistic regression, and most do not guarantee an optimal choice of model. These are the main focus areas of this research.

The main body of this dissertation begins in Chapter 2, with a budget-constrained linear regression model for continuous response variables using MINLP. The objective function is constructed to minimize the sum of squared error and data standardization reduces the value of big-M coefficients to 1. Properties such as solvability and global optimality of the proposed MINLP are derived. Illustrative computational results on realistic store data sets indicate that the proposed MINLP outperforms standard statistical software outputs in optimizing the objective function under a limit on the number of explanatory variables selected. Also our proposed MINLP is shown to be capable of selecting the optimal combination of explanatory variables under a budget limit covering cost of acquiring data sets. This cannot be done through an exercise of the usual statistical software except by total enumeration of possible variable combinations.

In Chapter 3, we propose a corresponding MINLP to perform budget-constrained and /or count-constrained logistic regression modeling with categorical response variables limited to two possible discrete values. Instead of minimizing the sum of squared error, maximum likelihood through the logit transform function is used for constructing the objective function. Alternative transformations to reduce needed big-Ms are also proposed to speed up the B&B solving process. Computational results on realistic data sets indicate that the proposed optimization model is able to select the best choice for an exact number of variables in a modest amount of time, and these results frequently outperform standard heuristic methods in terms of minimizing the negative log-likelihood function. Studies of varying prices for variables and/or budget limits also demonstrate the new, optimization-based insights that can be available for analysis about what data sources to consider and how large a budget is needed to obtain satisfactory results. Further study adjusting the objective function to minimize the **BIC** value instead of negative log-likelihood function proves that the new optimization model can reduce the risk of over-fitting by introducing a penalty term to the objective function which grows with the number of parameters.

In Chapter 4, we propose two refinements to our Chapter 2 MINLP model for multiple linear regression to speed branch and bound convergence and extend the size range of instances that can be solved exactly. One part of the work considers adding cutting planes to the models. Noting that the budget constraint in our proposed models has the same form as 0–1 knapsack problems, minimal cover knapsack inequalities are proposed and tested on five realistic data sets under four different budget limits in comparison to cuts already available in GUROBI, a mathematical programming solver. A second set of enhancements investigate and test six candidate constructions to produce a good integer-feasible solutions as warm starts to the budget-constrained multiple linear regression MINLP models to speed branch and bound convergence. Extensive computational results indicate that our two proposed refinements significantly reduce the time for solving the budget constrained multiple linear regression model using B&B algorithm, especially for large data sets. Finally, Chapter 5 summarizes the results obtained in this dissertation and points out some directions for future research.

Reference

Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. Best subset selection via a modern optimization lens. The annals of statistics, 44(2):813–852, 2016.

Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. Purposeful selection of variables in logistic regression. Source code for biology and medicine, 3(1):17, 2008.

MA Efroymson. Multiple regression analysis. Mathematical methods for digital computers, pages 191–203, 1960.

Antonio Lucadamo and Biagio Simonetti. Variable selection in logistic regression. The publishing of this booklet is a part of the Tempus project “Master programme in applied statistics” MAS 511140-Tempus-1-2010-1-RS-Tempus-JPCR, page 42, 2011.

Young Woong Park and Diego Klabjan. Subset selection for multiple linear regression via optimization. arXiv preprint arXiv:1701.07920, 2017.

Wojciech Rejchel. Lasso with convex loss: Model selection consistency and estimation. Communications in Statistics-Theory and Methods, 45(7):1989–2004, 2016.

Toshiki Sato, Yuichi Takano, Ryuhei Miyashiro, and Akiko Yoshise. Feature subset selection for logistic regression via mixed integer optimization. Computational Optimization and Applications, 64(3):865–880, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Jian Wu, Liugen Xue, and Peixin Zhao. Quickly variable selection for varying coefficient models with missing response at random. Communications in Statistics-Theory and Methods, 47(10):2327–2336, 2018.

2. Budget Constrained Model Selection for Multiple Linear Regression

2.1 Introduction

Multiple linear regression is a well-known tool for understanding the variation in a response variable as a function of some explanatory variables. The regression model can be expressed as: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y}_{n\times 1}$ is a response vector, $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_p}] \in \mathbb{R}^{n\times p}$ is an explanatory variable matrix , $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is regression coefficient vector and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$ is error. Variable selection, also called feature subset selection, is choosing a subset of useful variables from many explanatory variables for model estimation. Currently variable selection seems even more critical in model construction for big data that were previously unwieldy. How we can effectively and accurately select a "best" subset of explanatory variables from a huge amount of data is very challenging. The three most commonly used variable selection methods are forward, backward and stepwise (Efroymson, 1960). Several shortcomings of these three methods have been documented (Bertsimas and King, 2015), and as a result, authors have proposed alternative methods, such as the Akaike's information criteria (AIC), the Bayesian information criterion (BIC) (Schwarz et al., 1978), the wrapper method (Kohavi and John, 1997), the supersaturated designs ((Parpoula et al., 2014), (Yamada, 2004)). However, all these alternatives use greedy search algorithms that add one variable at a time to the model to maximize the reduction in sum of squared of errors along with certain penalty terms and then drop variables from the model if they are redundant in terms of reducing sum of squared error. Therefore, there is no guarantee that a truly best subset of features will be selected.

In the past twenty years, hardware along with algorithm improvements have resulted in a dramatic speedup of solving optimization problems, and consequently, different optimization models have become practical for solving the classical variable selection problem. Tibshirani (Tibshirani, 1996) introduced an approach called Lasso to the variable selection problem and Rejchel (Rejchel, 2016) considered both Lasso and adaptive Lasso for variable selection. Lasso penalizes the least squares method by imposing an L_1-penalty on the regression coefficients.

$$\min_{(\beta_0,\beta)\in R^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in R^{(p+1)}} [\frac{1}{2n}\sum_{j=1}^{n}(y_j - \beta_0 - x^T\beta)^2 + \lambda P_\alpha(\beta)] \tag{1}$$

Where

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_{\ell_2}^2 + \alpha\|\beta\|_{\ell_1} \tag{2}$$

P_α is the penalty part. When $\alpha = 1$, the method is called lasso regression. When $\alpha = 0$, the method is called ridge-regression. When $\alpha \in (0,1)$, the method becomes a naive elastic net regression which combines the characteristics of both lasso and ridge regression. Lasso does both continuous shrinkage and variable selection at the same time, but ridge regression provides a better solution than lasso when the number of observations is greater than the number of explanatory variables and explanatory variables are highly correlated with each other. Still, none of these methods is easily adaptable to the problem of constrained model selection, where the otherwise best subset of explanatory variables may not be affordable. Recently, Mixed Integer Optimization (MIO) ((Bertsimas et al., 2016),(Park and Klabjan, 2017)) was proposed for feature subset selection. (Bertsimas et al., 2016) proposed a MIO model for selecting the best fixed number K of features for regression. A discrete extension of the modern first-order continuous optimization method was used to find high quality feasible solutions that can be used as warm starts to a MIO solver. Instead of fixing the number of selected features, Park and Klajan (Park and Klabjan, 2017) proposed an optimization model for picking the best subset of variables in terms of minimizing mean absolute error (MAE) or mean squared error (MSE). Wu, Xue, and Zhao (Wu et al., 2018) proposed a method by using basis function approximation with smooth-threshold estimating equations to achieve variables selection and coefficient estimation at the same time without solving a convex optimization problem. Li and Lin (Li and Lin, 2009) introduced a variable selection procedure via penalized least squares with the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (Fan and Li, 2001) for screening experiments.

Although many different models and methods have been proposed for forecasting and variable selection, we know of none that has reported adding a budget constraint to the forecasting model. Consider a sales forecasting problem. In order to get a useful forecast we may like to consider many explanatory variables such as unemployment rate, GDP, disposable income, profits, households, population, inflation, etc. from different websites. Some of the metrics may be free, but others may be very expensive. In reality, we would like to find a best subset of explanatory variables within the budget limit available while minimizing the regression sum of squared errors.

In this paper, we propose a constrained linear regression model to solve optimally this budget limited regression task. The model is discussed in the next section. Then, in Section 3, convexity of the constrained regression model is proved to assure efficient computation of optimal solutions in MIO search. Section 4 presents the computational results from tests on realistic retail store sales data sets. Our final conclusions are given in the last section.

2.2 Budget Constrained Regression Model for Continuous Response Variables

Inspired from the above lasso and ridge regression, and considering a realistic business application such as budget limits or constrained counts of explanatory variables, we first add the cost of predictors or count of predictors as a penalty to the objective function of ordinary least squares.

$$\min_{(\beta_0,\beta)\in R^{p+1}} (\sum_{j=1}^{n}(y_j - \beta_0 - x^T\beta)^2 + \sum_{i=1}^{p} c_i \cdot I_{(\beta_i\neq 0)})$$

or

$$+\sum_{i=1}^{p} I_{(\beta_i\neq 0)}) \tag{3}$$

where $c_i(i = 1, \cdots p)$ is the cost of ith predictor. $I_{(\beta_i\neq 0)}$ is the indicator function,

$$I = \begin{cases} 0, \text{ for } \beta_i = 0 \\ 1, \text{ for } \beta_i \neq 0 \end{cases} \tag{4}$$

There are two potential problems with the above optimization model: (1) The objective function is no longer a convex function of the parameters because of the indicator functions, and (2) we would like to find a best predictive model within the budget B or the count limit K instead of minimizing the budget or the count of predictors. To address the second problem, we move the cost of predictors or the count of predictors from the objective function to the constraints. After the transformation, the objective function becomes convex (see proof in the next section). However, the budget limit and count of predictors constraints are no longer convex functions due to the indicator functions. Also in the next section, a linear transformation of both budget and count of predictors constraints is introduced to address this last obstacle.

$$\begin{aligned}
\min \quad & \sum_{j=1}^{n}(y_j - \beta_0 - x^T\beta)^2 \\
s.t. \quad & \sum_{i=1}^{p} c_i \cdot I_{(\beta_i \neq 0)} \leq \mathrm{B} \\
& \sum_{i=1}^{p} I_{(\beta_i \neq 0)} \leq \mathrm{K} \\
& (\beta_0, \beta) \in R^{p+1}. \qquad (5)
\end{aligned}$$

2.3 MINLP Formulation of the Budget Constrained Linear Regression Model

2.3.1 Mixed Integer Nonlinear Programming

A Mixed Integer Nonlinear Programming (MINLP) is an optimization problem where some variables take integer values and some variables take continuous values. The objective function and constraints are described by nonlinear functions (Bussieck and Pruessner, 2003). The general form of a MINLP is defined as

$$\begin{aligned}
\min \quad & f(x,y) \\
s.t. \quad & g(x,y) \leq 0 \\
& x \in X \\
& y \in Y \text{integer} \qquad \text{(MINLP)}
\end{aligned}$$

Here, function $f(x,y)$ is the objective function and $\mathbf{g}(\mathbf{x},\mathbf{y})$ is the constraint function. Either of them can be nonlinear. Variables x, y are the decision variables, where y are restricted to integer values. MINLPs have been used in many areas such as engineering, management science, finance, and operations research(Grossmann and Sahinidis, 2003a,b; Bertsimas et al., 2016; Bertsimas and King, 2015).

The primary decision variables in the above constrained linear regression model are the coefficients of the predictors $\beta_i,\ i = 1, \ldots, p$ and the constant term β_0. In order to linearize both budget and count constraints in our proposed MINLP model, we first introduce non-negative deviation variables s_{i+}, s_{i-}, where $\beta_i = s_{i+} - s_{i-}$. Note that s_{i+}, s_{i-} cannot be both equal to 0 when $|\beta_i| \neq 0$. Secondly, we replace the indicator function $I_{(\beta_i \neq 0)}$ by adding binary variables $q_i, i = 1, \ldots p$ and two sets of "Big-M" constraints guaranteeing $q_i = 1$ when $|\beta_i| \neq 0$, otherwise $q_i = 0$. Here M is a constant upper bound on
$\max\{|\beta_i|\ i = 1, \ldots p\}$.
With these modifications, the optimization model for the above MINLP reduces to the following mixed-integer nonlinear program:

$$
\begin{aligned}
\min \quad & (\sum_{j=1}^{n} (y_j - \beta_0 - x^T \beta)^2 \\
s.t. \quad & \beta_i = s_{i+} - s_{i-}, \forall i \in 1, \ldots p \\
& s_{i+} \leq M \cdot q_i, \forall i \in 1, \ldots p \\
& s_{i-} \leq M \cdot q_i, \forall i \in 1, \ldots p \\
& \sum_{i=1}^{p} c_i \cdot q_i \leq \mathrm{B} && \text{(Budget constraint)} \\
& \sum_{i=1}^{p} q_i \leq \mathrm{K} && \text{(Number of variables constraint)} \\
& s_{i+} \geq 0, s_{i-} \geq 0, \forall i \in 1, \ldots p \\
& (\beta_0, \beta) \in R^{p+1} \\
& q_i = 0 \text{ or } 1, \forall i \in 1, \ldots p. && \text{(CLREG)}
\end{aligned}
$$

2.3.2 Data Transformation to Allow Choosing $M = 1$

Here the value of Big-M should be chosen carefully. First, M must be larger than $\max\{|\beta_i|\ i = 1, \ldots p\}$. If M is smaller than any estimated coefficient $\{|\beta_i|\ i = 1, \ldots p\}$, certain feasible solutions may be cutoff. However, if M is too big, the model may become numerically difficult to solve, and bounds from continuous relaxations may deteriorate severely.

Therefore, we standardize all predictors with mean 0 and standard deviation 1. This can be done by calculating the Z score for all predictors and the response variable. That is,

$$y' = \frac{(y - \hat{\mu}_y)}{\sigma_y}, x_1' = \frac{(x_1 - \hat{\mu}_{x_1})}{\sigma_{x_1}}, x_2' = \frac{(x_2 - \hat{\mu}_{x_2})}{\sigma_{x_2}}, \ldots, x_p' = \frac{(x_p - \hat{\mu}_{x_p})}{\sigma_{x_p}}$$

After standardizing all predictors and response variable, we build the regression model based on variables after standardization.

$$y' = \beta_1' x_1' + \beta_2' x_2' + \ldots + \beta_p' x_p' + \varepsilon$$

Here, $\beta_i', i \in 1, 2, \ldots p$ are the standardized regression coefficients. β_0' isn't included in the above regression model. That is because that $\beta_0' = \hat{\mu_y'} - \beta_1' \hat{\mu}_{x_1}' - \beta_2' \hat{\mu}_{x_2}' - \ldots - \beta_p' \hat{\mu}_{x_p}' = 0$.

$$\begin{aligned}
y \quad & = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \\
& \Rightarrow \\
y - \bar{y} \quad & = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon - \bar{y} \\
& = \bar{y} - \beta_1 \bar{x}_1 - \ldots - \beta_p \bar{x}_p + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon - \bar{y} \\
& = \beta_1 (x_1 - \bar{x}_1) + \ldots + \beta_p (x_p - \bar{x}_p) + \varepsilon \\
& = \beta_1 \sigma_{x_1} x_1' + \ldots + \beta_p \sigma_{x_p} x_p' + \varepsilon \\
& \Rightarrow \\
\frac{y - \bar{y}}{\sigma_y} = y' \quad & = \beta_1 \frac{\sigma_{x_1}}{\sigma_y} x_1' + \ldots + \beta_p \frac{\sigma_{x_p}}{\sigma_y} x_p' + \varepsilon'
\end{aligned}$$

Therefore, the standardized coefficients $\beta_i' = \beta_i \frac{\sigma_{x_i}}{\sigma_y} = \frac{\text{cov}(x_i,y)}{\sigma_{x_i}\sigma_y}$ which is same as the correlation between two vectors $r = \frac{\text{cov}(x_i,y)}{\sigma_{x_i}\sigma_y}$. The correlation between two vectors is definitely bounded between -1 and 1. This way, the absolute value of all estimated coefficients on the standardized variables are ≤ 1. The value of M is set to 1 in the model.

2.3.3 Tractability of MINLP

The transformed binary MINLP model can be addressed by applying the nonlinear form of branch-and-bound. The nonlinear branch-and-bound method starts with a continuous NLP problem formed by relaxing the binary constraints on discrete decision variables q_i from $q_i = 0$ or 1 to $0 \leq q_i \leq 1$. If the NLP relaxation problem is infeasible, then the MINLP is infeasible as well. If the solution of the NLP relaxation happens to be binary for all q_i, it also solves the MINLP. Otherwise, branch-and-bound may be used to systematically search more constrained candidate sub-problems of the current node to isolate an optimum, solving the corresponding relaxation at each step to find a bound on the objective value achievable. The best binary-feasible solution discovered in the search is retained as a provably global optimum for the full model.

Whether or not the transformed MINLP can be solved to global optimality in this way depends critically on the tractability of the NLP continuous relaxation. Many well-known improving search algorithms can produce a global optimum of an NLP as long as an NLP satisfies the following Definitions 2.3.1 and 2.3.2 (Rardin, 1998).

Definition 2.3.1. A function $f(x)$ is **convex** if

$$f(x^{(1)} + \lambda(x^{(2)} - x^{(1)})) \leq f(x^{(1)}) + \lambda(f(x^{(2)}) - f(x^{(1)}))$$

for every $x^{(1)}$ and $x^{(2)}$ in its domain and every step $\lambda \in [0, 1]$.

Similarly, $f(x)$ is **concave** if

$$f(x^{(1)} + \lambda(x^{(2)} - x^{(1)})) \geq f(x^{(1)}) + \lambda(f(x^{(2)}) - f(x^{(1)}))$$

for every $x^{(1)}$ and $x^{(2)}$ in its domain and every step $\lambda \in [0, 1]$.

Definition 2.3.2. A constrained nonlinear program in functional form

$$\begin{array}{lll} \max \ \mathit{or} \ \min & f(x) & \\ s.t. & g_i(x) \left\{ \begin{array}{c} \geq \\ \leq \\ = \end{array} \right\} b_i \ i = 1, \ldots m \end{array}$$

is a **convex program** if f is concave for maximize or convex for a minimize, each g_i of a $\geq$ constraint is concave, each g_i of a $\leq$ is convex, and each g_i of an $=$constraint is linear.

The continuous NLP relaxation of the above budget constrained regression model (CLREG) is indeed a convex program which can be proved as follows.
The continuous NLP relaxation of the budget constrained regression CLREG has all the same constraints as the full CLREG, except the binary constraints $q_i = 0$ or 1 for each binary variable i are replaced by $1 \geq q_i \geq 0$. All other constraints are linear, and the relaxation of q_i constraints is also linear. This assures that all constraints of the relaxed NLP model are linear and thus convex. What remains for the NLP relaxation to be a convex program is whether or not its objective function is convex for the minimization problem.

$$\min_{(\beta_0,\beta)\in R^(p+1)} \frac{1}{2n} \sum_{j=1}^{n} (y_i - \beta_0 - x^T\beta)^2$$

As we can see, the objective function is the sum of functions $f_j(\beta_0, \beta) = (y_j - \beta_0 - x_j^T\beta)^2$. It will be convex as long as each f_j is convex. Now, dropping the i subscripts, we examine

$$\begin{aligned} f_j(\beta_0, \beta) \quad &\triangleq [y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n]^2 \\ &= [|y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n|]^2 \\ &= [\max((y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n), -(y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n))]^2 \end{aligned}$$

Expressions $(y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n)$ and $-(y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n)$ are both linear and thus convex. Therefore,

$$h(\beta_0, \beta) = max((y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n), -(y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n))$$

is also convex since it is the maximum of convex functions. Finally, consider $s(y) \triangleq y^2$. Second derivative $s^{''}(y) = 2$ proves $s(y)$ is convex because $s^{''}(y)$ is the 1 by 1 Hessian matrix and positive definite. Over domain $y \geq 0, s(y) \triangleq y^2$ is also non-decreasing. Thus, by applying composition rule, we can conclude that
$f_j(\beta_0, \beta) \triangleq [y - \beta_0 - x_1\beta_1 - \cdots - x_n\beta_n]^2 = s(h(\beta_0, \beta))$ is convex.
This completes the argument for convexity of objective function of linear regression, and establishes that continuous CLREG relaxation is a **convex** program. As a result, our proposed budget constrained linear regression model for CLREG can be solved efficiently to global optimality via branch and bound.

2.4 Illustrative Computational Testing

In this section, we conduct illustrative computational experiments on realistic data sets to investigate the performance of the proposed model. All computational results in this section are performed using Knitro solver through AMPL on a desktop equipped with Intel core 2.70 GHz CPU, 8.00GB usable RAM and Microsoft Windows 7 Professional.

2.4.1 Realistic Test Sets

An entire data set derived from a real retail store forecasting includes 333 observations and continuous 53 variables. Here, the 53 variables are composed of one response variable (store annual sales) and 52 independent variables which have potential impact on store annual sales (e.g., associate engagement score, termination rate, population density, price gap, unemployment rate, household income). In order to better investigate the performance of the proposed optimization model with increasing number of variables as well as observations, we first separate the entire data set into nine different sub-data sets. The number of observations and variables of each sub-data set are shown in Table 2.1.

2.4.2 Model Application for Fixed Numbers of Explanatory Variables

The 52 independent variables which can be used for predicting sales are not free. First, we assume the cost of each variable is the same, and we can afford only eight out of the total 52 variables. As a result, our budget constraint is the same as a count constraint which determines a maximum number of independent variables allowed in the model. The optimization model detailed above was tested on all nine sub-data sets to select the best combination of eight variables which explain a majority of variance in sales. For each model, the value of the objective function, R^2 and CPU time are reported in Table 2.1. From the comparison of CPU time across the different data sets, we can see that the model run time does not increase too much when increasing the number of observations and keeping the number of variables fixed. However, the model run time is dramatically increased by increasing the number of predictor variables considered. The high R^2 values indicate that most variance of sales is explained by the selected eight independent variables. Since regular variable selection methods such as forward, backward and stepwise

Table 2.1: Sales Data Sets Test for Selecting Eight Explanatory Variables

Data set	Vars	Obs	Objective value	R^2	CPU time (sec)
1	13	111	11.84	0.89	0.48
2	26	111	8.42	0.93	128
3	52	111	7.71	0.93	8,416
4	13	222	24.85	0.88	0.79
5	26	222	17.59	0.92	56.9
6	52	222	16.04	0.92	21,910
7	13	333	38.81	0.88	0.70
8	26	333	27.83	0.91	68.95
9	52	333	24.00	0.93	9,863

are all heuristic one-step-ahead search algorithms, the only way to make sure that the exact best count of variables will be selected by such methods is fitting the regression model to all possible combinations of independent variables considered, in other words, by enumeration. The time spent on enumerating all possible combinations will exponentially increase with number of variables making it impractical for even medium-sized data sets. To further investigate the advantages of our optimization model, we compare the variable selection sequences and corresponding objective values obtained from statistical software R to our proposed optimization model on the first data-set which includes thirteen variables and 111 observations. The forward selection method is used while fitting the regression model in R. Table 2.2 shows the results.

From Table 2.2, we have the following observations.

- The predictor which tends to explain more variance in the response variable and results in the smaller objective value will get selected first. From the variable selection sequence, we can see that most of the time, our proposed optimization model and statistical software select the same combination of variables.

- Still the forward, backward and stepwise methods sometimes follow a suboptimal path and get stuck in a suboptimal area of the solution space. There is no guarantee

Table 2.2: Variable Selection Results

Count of selected variables	Objective value		Selected variables (R, Model)							
	R	Model	X1	X2	X3	X4	X5	X6	X7	X8
1	23.89	23.89	✓⋆							
2	18.47	18.47	✓⋆	✓⋆						
3	16.25	14.68	✓⋆	✓	✓⋆	⋆				
4	13.88	13.51	✓⋆	✓	✓⋆	✓⋆	⋆			
5	12.79	12.69	✓⋆	✓	✓⋆	✓⋆	✓⋆	⋆		
6	12.11	12.11	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆		
7	11.90	11.90	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	
8	11.84	11.84	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆

The check mark indicates the selected variables by statistical software R, star indicates the selected variables by our proposed optimization model.

that the best set of variables will be selected. Results in Table 2.2 when three, four and five variables are selected illustrate that our proposed optimization model selects a different set of variables compared to the variable selection results of statistical software R, with an optimal choice that improves on heuristic results.

2.4.3 Testing with Varying Prices and Budgets for Data Items

Cost was assumed to be the same for each variable in the above analysis. In reality, most likely that will not be the case. Hence, synthetic different costs for each variable are considered in the following analysis. To begin, a cost for each variable is fixed as shown in Table 2.3, and we suppose budget limits vary from \$300 to \$1800. Under different budget limits, the selected combination of variables, objective function values and corresponding budget utilization are shown.

From Table 2.3, we can see that, as would be expected, the best variable $X1$ with cost \$450 is no longer selected when we have only \$300 budget. The objective value obtained by the selected three variables $X4$, $X12$ and $X13$ within the budget limit is much worse than the objective value obtained by including only variable $X1$ which we can not afford. However, variable $X1$ is added to the model immediately as long as our budget limit is greater than \$450. Variables $X2$ and $X13$ are also selected with variable $X1$ while increasing the

Table 2.3: Variable Selection Results under Different Budget Limits

Variable	Cost ($)	Budget limits ($)			
		300	800	1200	1800
X1	450		⋆	⋆	⋆
X2	300		⋆		⋆
X3	250			⋆	⋆
X4	150	⋆		⋆	⋆
X5	200			⋆	⋆
X6	400				⋆
X7	350				
X8	300				
X9	250				
X10	200				
X11	150				
X12	100	⋆		⋆	
X13	50	⋆	⋆	⋆	⋆
Objective value		94.11	18.43	13.43	12.10
Budget usage		300	800	1200	1800

current budget from $300 to $800. The corresponding objective value is reduced from 94.11 to 18.43, but, from Table 2.2, we know that the best three variables in terms of minimizing the objective function value are $X1$, $X3$, and $X4$. The reason that they are not selected is because the total cost of those three variables ($850) exceeds our budget current limit $800. As the budget keeps increasing, additional significant variables are selected, and the corresponding objective value is further reduced.

To illustrate what happens when data sources become more expensive, we tested new scenarios that increase the cost of each variable by 20% and 100%. Variable selection results are shown in Table 2.4 and Table 2.5, respectively. As the variables become increasingly expensive, under the same budget limits, fewer and “less significant” variables are available for selection, and the objective function value of the optimization model increases. A valuable benefit of being able to test such optimization scenarios could be to

Table 2.4: Variable Selection Results at 20% Increment of Cost of Each Variable

Variable	Cost ($)	Budget limits ($)			
		300	800	1200	1800
X1	540		⋆	⋆	⋆
X2	360				
X3	300			⋆	⋆
X4	180		⋆	⋆	⋆
X5	240				⋆
X6	480				⋆
X7	420				
X8	360				
X9	300				
X10	240				
X11	180				
X12	120	⋆		⋆	
X13	60	⋆	⋆	⋆	⋆
Objective value		96.81	19.92	14.62	12.68
Budget usage		180	780	1200	1800

provide decision makers with information about the budget required to produce good results.

2.5 Conclusions and Extensions

One-step-ahead procedures, forward, backward and stepwise methods are commonly used for variable selection in multiple linear regression. However, as discussed in Section 4, common variable selection methods have no way to control the exact count of variables that will be selected. Moreover, all three variable selection methods are heuristic algorithms, that may follow a suboptimal path and converge to a suboptimal solution; there is no guarantee that a best subset of variables will be selected in terms of minimizing an objective function, even if the only constraint is variable count. Recently, a constrained linear regression optimization model has been proposed by Bertsimas et al. (Bertsimas

Table 2.5: Variable Selection Results at 100% Increment of Cost of Each Variable

Variable	Cost ($)	Budget limits ($)			
		300	800	1200	1800
X1	900			⋆	⋆
X2	600		⋆		
X3	500				⋆
X4	300			⋆	⋆
X5	400				
X6	800				
X7	700				
X8	600				
X9	500				
X10	400				
X11	300				
X12	200	⋆	⋆		
X13	100	⋆			⋆
Objective value		96.81	93.94	19.94	14.68
Budget usage		300	800	1200	1800

et al., 2016) to select an optimal subset of variables of a given size. To our knowledge, however, the more complicated task of choosing an optimal subset of variables under a budget constraint has not been reported even though such budget considerations are part of many applied data analytic environments.

- In this paper, we describe investigations into constrained linear regression models that add constraints for a budget limit and/or count limit to the regression task. Our model empowers the analyst to select a best set of variables without violating a budget or variable count limitation.

- Computational experiments on realistic data sets were conducted to investigate the performance of our approach. Computational results indicate that (i) the proposed optimization model enables us to select the best choice for an exact number of independent variables in a modest amount of time, and our results frequently out-perform standard heuristic methods in terms of minimizing squared regression error.

- Further studies varying prices of variables and/or budget limits also demonstrate the newly available, optimization-based insights into data analysis about what data sources to consider, and how large a budget is needed to obtain satisfactory forecasts or predictions.

From the results in Table 2.1 of computational experiments in Section 4, we also notice that the solution time rapidly increases as the number of variables increases. One natural extension of the current optimization approach is to exploit advanced integer programming techniques such as adding appropriate cutting planes to speed up the optimization model. That would permit us to solve even larger problems.

In this paper, we are focused on adding constraints to the linear regression model. Linear regression models are used when the response variable is continuous, and minimizing LSE can be used for parameters estimation. If the response variable is categorical (e.g. binary) and non-continuous, non-linear relationship should be considered through different link functions and distribution families(Agresti, 1996). Instead of minimizing LSE, a maximum likelihood method should be considered for parameter estimation. The objective function then becomes finding the values of parameters for a given statistic which makes the known likelihood distribution a maximum. In Section 3, we established that our proposed constrained linear regression model satisfies the convexity in both objective function and constraints that guarantee the global optimum of the corresponding continuous relaxations. Extending the optimization model framework proposed here to the Generalized Linear Models could be another useful step to pursue.

Reference

Alan Agresti. An introduction to categorical data analysis, volume 135. Wiley New York, 1996.

Dimitris Bertsimas and Angela King. Or forum—an algorithmic approach to linear regression. Operations Research, 64(1):2–16, 2015.

Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. Best subset selection via a modern optimization lens. The annals of statistics, 44(2):813–852, 2016.

Michael R Bussieck and Armin Pruessner. Mixed-integer nonlinear programming. SIAG/OPT Newsletter: Views & News, 14(1):19–22, 2003.

MA Efroymson. Multiple regression analysis. Mathematical methods for digital computers, pages 191–203, 1960.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360, 2001.

Ignacio E Grossmann and Nikolaos V Sahinidis. Special double issue on mixed-integer programming and its applications to engineering, part i, 2003a.

Ignacio E Grossmann and Nikolaos V Sahinidis. Special double issue on mixed-integer programming and its applications to engineering, part ii, 2003b.

Ron Kohavi and George H John. Wrappers for feature subset selection. Artificial intelligence, 97(1-2):273–324, 1997.

Runze Li and Dennis KJ Lin. Variable selection for screening experiments. Quality technology & quantitative management, 6(3):271–280, 2009.

Young Woong Park and Diego Klabjan. Subset selection for multiple linear regression via optimization. arXiv preprint arXiv:1701.07920, 2017.

Christina Parpoula, Krystallenia Drosou, Christos Koukouvinos, and Kalliopi Mylona. A new variable selection approach inspired by supersaturated designs given a large-dimensional dataset. Journal of Data Science, 12(1):35–52, 2014.

Ronald L Rardin. Optimization in operations research, volume 166. Prentice Hall Upper Saddle River, NJ, 1998.

Wojciech Rejchel. Lasso with convex loss: Model selection consistency and estimation. Communications in Statistics-Theory and Methods, 45(7):1989–2004, 2016.

Gideon Schwarz et al. Estimating the dimension of a model. The annals of statistics, 6(2): 461–464, 1978.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Jian Wu, Liugen Xue, and Peixin Zhao. Quickly variable selection for varying coefficient models with missing response at random. Communications in Statistics-Theory and Methods, 47(10):2327–2336, 2018.

Shu Yamada. Selection of active factors by stepwise regression in the data analysis of supersaturated design. Quality Engineering, 16(4):501–513, 2004.

3. **Budget Constrained Model Selection for Logistic Regression**

3.1 Introduction

In reality, there are many situations where we need to predict an output variable which is discrete (or categorical) instead of continuous such as given a set of input features to predict whether a breast tumor is benign or malignant. Linear regression is inappropriate for such modeling and classification problems because the response values are not measured on a continuous scale and the error term does not follow a normal distribution. The linear regression model can generate any real number ranging from negative to positive infinity as the predicted value, whereas a categorical variable might be restricted to discrete values such as ”Yes” or ”No”. Logistic regression and multinomial regression along with many classification techniques such as discriminant analysis, support vector machine (SVM), classification tree, random forest, naive Bayes classifier are useful for solving classification problems ((James et al., 2013), (Hosmer Jr et al., 2013), (Guyon et al., 2002), (Keerthi and Gilbert, 2002), (Friedman and Koller, 2003), (Friedman et al., 1997)). This paper focuses on the logistic regression model which is widely used for predicting a response variable with binary values.

Feature selection or model selection is very important in model building. It can lead to a clear relationship between the response variable and the selected features and improve the model prediction effectiveness by filtering out less-significant features. Many feature selection methods such as forward selection, backward elimination, stepwise selection, mixed integer optimization (MIO), Lasso regression, Ridge regression and a naive elastic net regression are well known and deeply studied for linear regression ((Efroymson, 1960), (Tibshirani, 1996), (Rejchel, 2016), (Bertsimas et al., 2016), (Park and Klabjan, 2017), (Wu et al., 2018)). Bertsimas et al (Bertsimas et al., 2016) proposed a MIO model for selecting the best fixed number p of features for linear regression models. Instead of fixing the number of selected features, Park (Park and Klabjan, 2017) proposed an optimization model for picking the best subset of variables in terms of minimizing mean absolute error (MAE) or mean squared error (MSE). Zhang (Zhang et al., 2018) proposed a mixed integer nonlinear programming model for selecting the best subset of independent variables under either count or budget constraints for linear regression.

Feature selection methods are much less studied in logistic regression. Sato (Sato et al., 2016) proposed a Mixed Integer Optimization model. Lucadamo (Lucadamo and Simonetti, 2011) proposed the Disco Coefficient method to identify the significant variables for logistic regression. Bursac (Bursac et al., 2008) proposed a method called purposeful selection of co-variates within which an analyst makes a variable selection decision at each step of the modeling process. To the best of our knowledge, none of those existing studies has considered budget constrained model selection in logistic regression, and most do not guarantee an optimal choice of model.

3.2 Budget Constrained Logistic Regression Model

In linear regression, OLS is used for estimating the parameters by minimizing the sum of squared errors. However, in logistic regression, least squares estimation is no longer appropriate for parameters estimation (Friedman et al., 2010). Instead, maximum likelihood estimation is used for estimating the parameters which best fit data. Let y represent the response variable having values in $(0, 1)$. The logistic regression model constructs the conditional probability of y as a function of a linear combination of the explanatory variables $\mathbf{x}$ through the logit transform function.

$$\text{logit}(\pi_j) = \log(\frac{\pi_j}{1-\pi_j}) = \beta_0 + x^T\beta \tag{1}$$

$$\pi_j = \frac{1}{1 + e^{-(\beta_0 + x^T \beta)}} \tag{2}$$

where $\pi_j = P(y = 1|x_j)$. With the logit transform, in other words, the log-odds of the probability of $(y = 1)$ is equal to the linear combination of explanatory variables. Unknown parameters β_0, β are estimated by using maximum likelihood which finds the estimations of parameters by maximizing the probability they could have generated the observed data. The joint conditional probability density function can be written as following

$$f(y|\beta) = \prod_{j=1}^{n} \pi_j^{y_j} (1 - \pi_j)^{1-y_j} \tag{3}$$

The joint conditional probability density function in (3) expresses the values of y as a function of known values of β. The likelihood function has the same form as the probability density function except that the parameters of functions are reversed: the likelihood function expresses the values of β in terms of known values of y. Thus,

$$L(\beta|y) = \prod_{j=1}^{n} \pi_j^{y_j} (1 - \pi_j)^{1-y_j} \tag{4}$$

The maximum likelihood finds the estimate of β which maximizes the likelihood function (4). By taking the first derivative of the likelihood function, we get the critical points which can be either maxima or minima. If the second derivative at that point is less than zero, then the critical point becomes a maximum. Thus, in order to find the maximum likelihood estimate of β, we need to take the first and second derivatives of the likelihood function. Taking the derivative of (4) is not easy due to the complexity of multiplicative terms. Actually for logistic regression, there is no closed form solution for MLE parameters. Since the logarithm is a monotonic function, any maximum of the likelihood function will be a maximum of the log likelihood function as well. After taking the natural log on both sides of function (4), it becomes

$$\begin{aligned}\ell(\beta|y) &= \sum_{j=1}^{N} y_j \mathrm{log}\pi_j + \sum_{j=1}^{N}(1-y_j)\mathrm{log}(1-\pi_j)\\ &= \sum_{j=1}^{N} \mathrm{log}(1-\pi_j) + \sum_{j=1}^{N} y_i \mathrm{log}\frac{\pi_i}{1-\pi_i}\\ &= \sum_{j=1}^{N} \mathrm{log}(1-\pi_j) + \sum_{j=1}^{N} y_i(\beta_0 + x^T\beta)\\ &= \sum_{j=1}^{N} -\mathrm{log}(1+e^{\beta_0+x^T\beta}) + \sum_{j=1}^{N} y_j(\beta_0 + x^T\beta) \qquad (5)\end{aligned}$$

We proposed a constrained linear regression model for selecting the best combination of independent variables within either count or budget constraints ((Zhang et al., 2018)). The objective function is defined as minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given data set and those predicted by the linear function. To be comparable to the linear regression model, instead of maximizing function (5), we minimize the negative log conditional likelihood. After adding both budget and count of predictors constraints, the constrained logistic regression model becomes

$$\begin{aligned}
\min \quad & \sum_{j=1}^{N} \log(1+e^{\beta_0+x^T\beta}) - \sum_{j=1}^{N} y_j(\beta_0+x^T\beta) \\
s.t. \quad & \sum_{i=1}^{p} c_i \cdot I_{(\beta_i \neq 0)} \leq \text{B} \\
& \sum_{i=1}^{p} I_{(\beta_i \neq 0)} \leq \text{K} \\
& (\beta_0, \beta) \in R^{p+1}. && (6)
\end{aligned}$$

Here, both budget and count of predictors constraints take the same form as our linear regression model (Zhang et al., 2018). The same transformation of the constraints using a constant "big-M" can be applied to obtain the appropriate model for the logistic case.

$$\begin{aligned}
\min \quad & \sum_{j=1}^{N} \log(1+e^{\beta_0+x^T\beta}) - \sum_{j=1}^{N} y_j(\beta_0+x^T\beta) \\
s.t. \quad & \beta_i = s_{i+} - s_{i-}, \forall i \in 1,\ldots p \\
& s_{i+} \leq M \cdot q_i, \forall i \in 1,\ldots p \\
& s_{i-} \leq M \cdot q_i, \forall i \in 1,\ldots p \\
& \sum_{i=1}^{p} c_i \cdot q_i \leq \text{B} && \text{(Budget constraint)} \\
& \sum_{i=1}^{p} q_i \leq \text{K} && \text{(Number of variables constraint)} \\
& s_{i+} \geq 0, s_{i-} \geq 0, \forall i \in 1,\ldots p \\
& (\beta_0, \beta) \in R^{p+1} \\
& q_i = 0 \text{ or } 1, \forall i \in 1,\ldots p. && \text{(CLOGREG)}
\end{aligned}$$

3.3 Tractability of MINLP

The above constrained logistic regression model (CLOGREG) can be solved to global optimum by branch and bound if its continuous NLP relaxation is a convex program. The continuous NLP relaxations of the budget constrained regression CLOGREG have the same forms constraints as the full CLOGREG except the binary constraints $q_i = 0$ or 1 are replaced by $1 \geq q_i \geq 0$. All other constraints are linear, and the relaxation of q_i constraints is also linear. This assures that all constraints of the relaxed NLP model are linear and thus convex. What remains for the NLP relaxation to be a convex program is whether or not its objective function is convex for this minimization problem.

$$\min \qquad \sum_{j=1}^{N} \log(1 + e^{\beta_0 + x^T\beta}) - \sum_{j=1}^{N} y_j(\beta_0 + x^T\beta)$$

Expression $y_j(\beta_0 + x^T\beta)$ is linear in β_0, β. Hence, it is a concave function and $-y_j(\beta_0 + x^T\beta)$ is thus convex. Exponential $e^{(\beta_0 + x^T\beta)}$ is a convex function of β_0, β, and $\log(x)$ is a non-decreasing single value function and thus convex. Finally since the non-negative weighted sum of convex functions is also convex, the full objective function of logistic regression model is convex. We may conclude that the full NLP relaxation of constrained logistic regression model (CLOGREG) is also a convex program. The convexity of the model guarantees branch and bound methods can (at least in principle) produce a global optimum to the MINLP(CLOGREG).

3.4 Value of Big-M

The value of Big-M in the above **CLOGREG** needs to be chosen carefully. First, M needs to be larger than $\max\{|\beta_i|\ i = 1, \ldots p\}$. If M is smaller than any estimated coefficient

$\{|\beta_i|\ i = 1, \ldots p\}$, then possible feasible solutions might be cutoff. On the other hand, if M is too big, the model may become numerically difficult to solve because bounds from continuous relaxations will be weak.

This issue can be addressed for the linear regression case by standardizing all predictor variables with their mean 0 and standard deviation 1 so that $M = 1$ suffices for all constraints of the MINLP (Zhang et al., 2018). The issue is much more complex for logistic regression, but this research will seek to find a suitable standardization permitting smaller values of M. Sections 3.4.1 and 3.4.2 describe methods to be considered.

3.4.1 Standardized Logistic Regression Coefficients

Menard (Menard, 2004) reviewed six different approaches for standardizing the logistic regression coefficients. Table 3.1 shows the explanation of all six. The first approach is dividing each unstandardized coefficient by its estimated standard deviation which was proposed by Goodman (Goodman, 1972). The second method suggested by Agresti (Agresti, 1996) and Menard (Menard, 2002) is to standardize only the predictors. The third approach proposed by Menard (Menard, 2002) is currently implemented in SAS statistical software. This procedure is to standardize both the predictors and the dependent variable. However, the same variance $\frac{\pi}{\sqrt{3}}$ is assumed for every dependent variable in every model while standardizing the dependent variable. The fourth approach was proposed by Long (Scott Long, 1997). The only difference between the third and fourth procedure is that $\frac{\pi}{\sqrt{3}} + 1$ is assumed as the constant variance of the dependent variable instead of $\frac{\pi}{\sqrt{3}}$. All of those four approaches are classified as partially standardized logistic regression coefficients since none of them really considers the empirical variation of the dependent variable.

The final two methods in Table 3.1 are variance-based fully standardized coefficients and information theoretic fully standardized coefficients proposed by Menard (Menard, 2002) and Soofi (Soofi, 1992) by taking into account the actual variation of the depend variable as well as the predictors.The information theoretic fully standardized coefficients is derived

Table 3.1: Standardized Logistic Regression Coefficients

Coefficients	Description	Type
$\beta^\star_G$	Standardize coefficients by dividing its std (Goodman)	Partial
$\beta^\star_A$	Standardize predictors only (Agresti, Menard)	Partial
$\beta^\star_S$	Standard logistic distribution (SAS)	Partial
$\beta^\star_L$	Standard logistic and normal distribution (Long)	Partial
$\beta^\star_M$	Variance-based fully standardized coefficients (Menard)	Fully
$\beta^\star_I$	Information theoretic fully standardized coefficients (Soofi)	Fully

from information theory through measuring the direct contribution of each predictor to the explained variance in the dependent variable. Menard (Menard, 2004) pointed out that the information theoretic fully standardized coefficients may be the best from a conceptual standpoint. But the practical application of this method is limited unless there is an appropriate algorithm to simplify this calculation.

3.4.2 Estimation of Variance-based Fully Standardized Coefficients

This research uses preferred variance-based fully standardized coefficients to construct needed big-M values. However, in the logistic regression, instead of directly modeling the relationship between the binary variable Y and the predictors, we model the logit transformed Y as the response variable. Therefore, in order to get the fully standardized coefficient $\beta^\star_M$, we must construct an appropriate estimation of the variance of logit(Y) instead of Y. Since, the value of logit transformed Y is from negative infinity to positive infinity, Menard (Menard, 2004) pointed out that it is impossible to directly calculate the standard deviation of logit(Y). However, it can be estimated indirectly by borrowing the formula from OLS:

$$\beta^\star = (\beta)(\frac{S_X}{S_Y})$$

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} \Rightarrow S_Y = \frac{S_{\hat{Y}}}{\sqrt{R^2}}$$

$$\beta^\star = (\beta)(S_X)\frac{\sqrt{R^2}}{S_{(\hat{Y})}}$$

Where β is the estimate of unstandardized linear regression coefficient, S_X is the sample standard deviation of the predictors X, and S_Y is the sample standard deviation of response variable Y. $S_{\hat{Y}}$ is the standard deviation of the predicted value of Y. In the parallel fashion, the estimation of $\beta^\star_M$ for the logistic regression can be written as following,

$$\beta_M{}^\star = (\beta)(\frac{S_X}{S_{logit(Y)}})$$

$$R^2 = \frac{S^2_{logit(\hat{Y}}}{S^2_{logit(Y)}}) \Rightarrow S_{logit(Y)} = \frac{S_{logit(\hat{Y})}}{\sqrt{R^2}}$$

$$\beta_M{}^\star = (\beta)(S_X)(\frac{\sqrt{R^2}}{S_{logit(\hat{Y})}})$$

$$(\beta)(S_X) = (\beta_M{}^\star) * (S_{logit(\hat{Y})})/\sqrt{R^2}$$

Agresti (Agresti, 1996) proved that $\beta^\star_A = (\beta)(S_X)$ can be easily obtained by standardizing only the predictors with mean 0 and variance 1. Menard (Menard, 2002) noted that the magnitude of the variance-based fully standardized logistic regression coefficients tend to be smaller than the magnitude of partially standardized coefficients. Especially, the magnitude of $\beta^{\star}_{M}$ is between -1 and +1 as long as there is no collinearity existing in the data set. Therefore, the value of $\beta^{\star}_{A}$ is between $-S_{logit(\hat{Y})}/\sqrt{R^2}$ and $S_{logit(\hat{Y})}/\sqrt{R^2}$. The value of M can be picked as any positive value greater than $S_{logit(\hat{Y})}/\sqrt{R^2}$ which can be obtained by fitting the non-constrained logistic regression.

3.5 Illustrative Computational Testing

In this section, we conduct illustrative computational experiments on real-world benchmark data sets to illustrate the performance of the proposed logistic regression model with constraints. All computational results in this section are performed using Knitro solver through AMPL on a desktop equipped with Intel core 2.70 GHz CPU, 8.00GB usable RAM and Microsoft Windows 7 Professional.

3.5.1 Data Introduction

A public benchmark data set is the *default of credit card clients* data set, obtained from the UCI repository(Dheeru and Karra Taniskidou, 2017). It is used for computational experiments. The entire data set is composed of 30000 observations, with one binary variable indicating whether the payment will default, and 23 explanatory variables such as the amount of the given credit in dollars, borrower gender, education, marital status, age, past 9 months payment history, the amount of bill statement and the amount of previous payment in dollars. In order to better investigate the performance of the proposed optimization model with increasing number of variables as well as observations, we split the entire data set into nine different subsets. The number of observations and variables in each subset are shown in Table 3.2.

Table 3.2: Credit Card Clients Data Sets Test for Selecting Three Explanatory Variables

Data set	Vars	Obervs	Upper bound of $\|\beta\|$	Obj(Mod)	Obj(R)	Mod CPU time (sec)
1	5	5000	2.667	2491.26	2491.26	0.42
2	5	10000	2.712	4926.31	4926.31	0.78
3	5	15000	2.704	7538.80	7538.80	1.26
4	10	5000	2.342	2279.77	2279.77	4.76
5	10	10000	2.338	4499.27	4502.59	6.71
6	10	15000	2.340	6828.17	6853.38	13.20
7	15	5000	2.808	2278.41	2283.20	18.78
8	15	10000	2.698	4497.75	4515.26	52.70
9	15	15000	2.512	6828.17	6841.83	89.47

Obj(Mod) is the objective of our proposed CLOGREG model, Obj(R) is the objective of backward selection logistic regression model in statistical software R. The last three variables kept by the backward selection procedure are picked as the best three variables in the R case.

3.5.2 Illustrative Testing for Fixed Numbers of Explanatory Variables

To begin we assume that the cost of each explanatory variable is the same and we can only afford three of them. Our proposed CLOGREG model is tested on the nine different data sets to pick the best combination of 3 variables which can help us better predict credit card default payment. First, we standardized all of the explanatory variables with mean 0 and standard deviation 1 (as explained in Section 3.4.1). Then we fitted the non-constrained logistic regression model and saved the predicted value of Y from the logistic regression model. Next we used the predicted value of Y to calculate R^2 and variance of $logit(\hat{Y})$. Finally, the upper bound of absolute value of estimated coefficients were calculated using formula $S_{logit(\hat{Y})}/\sqrt{R^2}$ and shown in Table 3.2. The value of M is set to three which is greater than the absolute value of all estimated coefficients on the standardized explanatory variables. For each model, the value of objective function of our proposed CLOGREG model and R backward selection, model CPU time are also reported in Table 3.2.

From the comparison of objective value between our proposed CLOGREG model and R backward selection model, we can see that for the smaller data sets backward selection method in R is able to pick the best combination of three variables, but when numbers of

Table 3.3: Variable Selection Results

Variable	Sequence of Selected Variables									
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
x1		⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
x2							⋆	⋆	⋆	⋆
x3										⋆
x4									⋆	⋆
x5				⋆	⋆	⋆	⋆	⋆	⋆	⋆
x6	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
x7			⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
x8						⋆	⋆	⋆	⋆	⋆
x9					⋆	⋆	⋆	⋆	⋆	⋆
x10								⋆	⋆	⋆
Objective	4562.6	4517.85	4499.27	4484.12	4476.99	4472.23	4468.61	4465.28	4462.04	4458.9

variables and observations increase, the backward selection method as one of heuristic one-step-ahead search algorithms tends to follow a wrong path and get stuck in a suboptimal area of the solution space. There is no guarantee that the best set of variables will be selected. From the CPU run time of nine data sets, we can see that the run time of our proposed model is more sensitive to the increment of number of variables compared with number of observations.

To further investigate the significance of each explanatory variable in terms of predicting credit card default payment, we construct the variable selection sequences using our proposed optimization model by selecting 1, 2,... up to all explanatory variables on data sets 5 through 10 variables and 10,000 observations. The results are shown in Table 3.3. From Table 3.3, we can see that variable x6 is the ”most significant”variable in terms of predicting credit card default payment, then variable x1 and the ”least significant” variable is x3. Also, the objective function (quality) of the fit improves each time more variables are allowed.

Table 3.4: Variable Selection Results under Different Budget Limits

Variable	Cost ($)	Budget limits B ($)			
		300	900	1200	1500
X1	500			⋆	⋆
X2	250				
X3	100				
X4	150			⋆	⋆
X5	400				
X6	550		⋆	⋆	⋆
X7	450				
X8	300	⋆	⋆		⋆
X9	350				
X10	200				
Objective value		4819.98	4535.64	4508.01	4490.5
Budget usage		300	850	1200	1500

3.5.3 Illustrative Testing with Varying Prices and Budgets for Data Items

The cost was assumed to be the same for each variable in the above analysis. In reality, that may not be the case. Hence, differing synthetic costs for each variable in data set 5 (see Table 3.4) are considered in the next analysis. We consider budget limits varying from $B = \$300$ to $B = \$1500$ in Table 3.4. Under different budget limits, the selected combination of variables, the objective values and the corresponding budget utilization for data set 5 are shown.

From Table 3.4, we can see that the best variable x6 with cost $550 is no longer selected when we have only $300 budget. The objective value obtained by the selected variable x8 within the budget limit is worse than the objective value obtained by variable x6 which we cannot afford anymore. However, variable x6 is added to the model immediately as long as the budget is greater than $550. There are also more variables selected when more budget is available.The objective value is further reduced as well.

What happens when data sources become more expensive? To illustrate, we tested new

Table 3.5: Variable Selection Results at 30% Increment of Cost of Each Variable

Variable	Cost ($)	Budget limits ($)			
		300	900	1200	1500
X1	650				⋆
X2	325		⋆		
X3	130				⋆
X4	195				
X5	520				
X6	715		⋆	⋆	⋆
X7	585				
X8	390			⋆	
X9	455				
X10	260	⋆			
Objective value		5085.38	4562.47	4535.64	4517.22
Budget usage		260	845	1105	1495

scenarios by increasing the cost of each variable by 30%. The variable selection results are shown in Table 3.5. As the variables become more expensive, under the same budget limits, fewer and ”less significant” variables can be afforded by the optimization model. The objective value of the optimization model becomes greater. A valuable benefit of being able to test such optimization scenarios could be to provide decision makers with information about the budget required for good results.

3.5.4 Illustrative Testing with Akaike and Bayesian Information Criterions for Over-fitting

The objective function of proposed CLOGREG model is minimizing the negative maximum log-likelihood function. It selects as many explanatory variables as it can as long as the budget allows. What happens if we have more than enough budget? We might end up building an unnecessarily complicated model, which has too many parameters to be estimated accurately on a given training data set. That potentially causes an over-fitting problem. The over-fitted model tends to memorize training data and therefore fail to fit additional data or predict unseen data unreliably.

Two most commonly used model selection criteria to address this challenge are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC and BIC are computed as follows (Fabozzi et al., 2014):

$$AIC = -2logL(\hat{\beta}) + 2k$$

$$BIC = -2logL(\hat{\beta}) + klog(n)$$

where

$$\begin{aligned} \hat{\beta} &= \text{the parameter values that maximize the likelihood function} \\ L(\hat{\beta}) &= \text{the maximized value of the likelihood function of the candidate model} \\ k &= \text{the number of parameters estimated by the model}(p+1) \\ n &= \text{the number of observations} \end{aligned}$$

The first component of both AIC and BIC is the log-likelihood function multiplied by -2. Ignoring the second component, the model with the minimum AIC or BIC is the one which maximizes the log-likelihood. However, a penalty term based on the number of estimated parameters is added to the first component for both AIC or BIC. The more parameters, the larger the penalty that will be added to the first component, increasing the value of either AIC or BIC. A difference between the AIC and BIC is that the larger penalty term imposed for number of parameters is added to BIC as compared to AIC. Also BIC as presented here is a function of n observations that do not affect AIC as presented here. In order to overcome the over-fitting problem and at the same time consider the budget limit, we can adjust the proposed CLOGREG model by changing the objective function from minimizing the negative log-likelihood function to minimize either the AIC or the BIC value. The adjusted objective functions for both AIC and BIC are defined as follows:

$$\min \quad 2(\sum_{j=1}^{N} \log(1 + e^{\beta_0 + x^T\beta}) - \sum_{j=1}^{N} y_j(\beta_0 + x^T\beta)) + 2(\sum_{i=1}^{p} q_i + 1) \tag{AIC}$$

$$\min \quad 2(\sum_{j=1}^{N} \log(1 + e^{\beta_0 + x^T\beta}) - \sum_{j=1}^{N} y_j(\beta_0 + x^T\beta)) + (\sum_{i=1}^{p} q_i + 1)\log n \tag{BIC}$$

The continuous NLP relaxations of adjusted AIC and BIC objective functions replace the binary variables $q_i = 0$ or 1 by $1 \geq q_i \geq 0$. The expressions of both penalty terms are linear and thus convex. Since all the constraints are still the same, the convexity of the adjusted MINLP(CLOGREG) guarantees that branch and bound methods can still produce a global optimum.

Suppose we have \$3500 budget which can cover the cost of all ten variables in data set 5. By applying the updated optimization models on data set 5, the selected variables, objective value and budget utilization are as shown in Table 3.6.

From the results in Table 3.6, we can see that AIC model selects all ten variables while the BIC model only selects six out of ten variables even though we have enough budget to cover each of the ten variables. The reason for AIC model selecting each of ten variables covered by budget is that the reduction in the first component is always greater than the

Table 3.6: Variable Selection Results based on AIC and BIC Values

Variable	Cost ($)	Budget: 3500 ($)	
		AIC	BIC
X1	500	$\star$	$\star$
X2	250	$\star$	
X3	100	$\star$	
X4	150	$\star$	
X5	400	$\star$	$\star$
X6	550	$\star$	$\star$
X7	450	$\star$	$\star$
X8	300	$\star$	$\star$
X9	350	$\star$	$\star$
X10	200	$\star$	
Objective value		8939.8	9008.9
Budget usage		3250	2550

penalty imposed by adding more parameters. However, for the BIC model, since there is greater penalty imposed for the number of parameters, the minimum BIC value is achieved by selecting the best six out of ten variables. If we select each of the ten variables, the BIC value would be increased from 9008.9 to 9019.1.

3.6 Conclusions and Extensions

One-step-ahead variable selection procedures, forward, backward and stepwise methods are commonly used for variable selection in logistic regression. However, as discussed in Section 3.5.2, the one-step-ahead variable selection has no way to control the exact count of variables to be selected. Moreover, all three of these variable selection methods are heuristic algorithms, sometimes following a wrong path and getting stuck in a suboptimal solution; there is no guarantee that the best set of variables will be selected in terms of minimizing the objective function even if the only limit is variable count. To our best knowledge, the more complicated task of choosing an optimal subset of variables under a budget constraint has not been addressed in any other journal paper even though such budget constraints are part of many data analytic environments.

- In this paper, we have investigated constrained logistic regression models that add constraints for a budget limit or a count limit to the logistic regression task. The proposed model is able to select the best set of variables without violating the budget or count limitation.

- Illustrative computational experiments on realistic data sets have been conducted to investigate the performance of the proposed approach. The computational results indicate that the proposed optimization model is able to select the best choice for an exact number of variables in a modest time, and that these results frequently out-perform standard heuristic methods in terms of minimizing the negative log-likelihood function.

- Studies varying prices of variables and/or budget limits also demonstrate the new, optimization-based insights that can be available for data analysis about what data sources to consider and how large a budget is needed to obtain satisfactory forecasts.

- Further study adjusting the objective function to minimize the BIC value instead of negative log-likelihood function proves that the new optimization model reduces the risk of over-fitting by introducing a penalty term to the objective function which grows with the number of parameters.

From the results in Table 3.2 of computational experiments in Section 3.5.2, we also notice that the time for solving the model increases rapidly with the number of variables. One natural extension of the current optimization approach is to exploit advanced integer programming techniques such as adding appropriate cutting planes to speed up the optimization model. The result would permit us to solve bigger data instances. Computation may also be reduced by starting the branch and bound with a strong starting solution.

In this paper, we are focused on adding constraints to the logistic regression model. Logistic regression models are used when the response variable has binary values. If the response variable is categorical with more than two values, multinomial or ordinal regression along with other machine learning techniques need to be considered. Extending the proposed optimization model to other Generalized Linear Models could be another useful step to pursue.

Reference

Alan Agresti. An introduction to categorical data analysis, volume 135. Wiley New York, 1996.

Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. Best subset selection via a modern optimization lens. The annals of statistics, 44(2):813–852, 2016.

Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. Purposeful selection of variables in logistic regression. Source code for biology and medicine, 3(1):17, 2008.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

MA Efroymson. Multiple regression analysis. Mathematical methods for digital computers, pages 191–203, 1960.

Frank J Fabozzi, Sergio M Focardi, Svetlozar T Rachev, and Bala G Arshanapalli. The basics of financial econometrics: Tools, concepts, and asset management applications. John Wiley & Sons, 2014.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010.

Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. Machine learning, 50(1-2):95–125, 2003.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. Machine learning, 29(2-3):131–163, 1997.

Leo A Goodman. A modified multiple regression approach to the analysis of dichotomous variables. American Sociological Review, pages 28–46, 1972.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3):389–422, 2002.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. Applied logistic regression, volume 398. John Wiley & Sons, 2013.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer, 2013.

S. Sathiya Keerthi and Elmer G Gilbert. Convergence of a generalized smo algorithm for svm classifier design. Machine Learning, 46(1-3):351–360, 2002.

Antonio Lucadamo and Biagio Simonetti. Variable selection in logistic regression. The publishing of this booklet is a part of the Tempus project “Master programme in applied statistics” MAS 511140-Tempus-1-2010-1-RS-Tempus-JPCR, page 42, 2011.

Scott Menard. Applied logistic regression analysis, volume 106. Sage, 2002.

Scott Menard. Six approaches to calculating standardized logistic regression coefficients. The American Statistician, 58(3):218–223, 2004.

Young Woong Park and Diego Klabjan. Subset selection for multiple linear regression via optimization. arXiv preprint arXiv:1701.07920, 2017.

Wojciech Rejchel. Lasso with convex loss: Model selection consistency and estimation. Communications in Statistics-Theory and Methods, 45(7):1989–2004, 2016.

Toshiki Sato, Yuichi Takano, Ryuhei Miyashiro, and Akiko Yoshise. Feature subset selection for logistic regression via mixed integer optimization. Computational Optimization and Applications, 64(3):865–880, 2016.

J Scott Long. Regression models for categorical and limited dependent variables. Advanced quantitative techniques in the social sciences, 7, 1997.

Ehsan S Soofi. A generalizable formulation of conditional logit with diagnostics. Journal of the American Statistical Association, 87(419):812–816, 1992.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Jian Wu, Liugen Xue, and Peixin Zhao. Quickly variable selection for varying coefficient models with missing response at random. Communications in Statistics-Theory and Methods, 47(10):2327–2336, 2018.

Jingying Zhang, Ronald L Rardin, and Justin R Chimka. Budget constrained model selection for multiple linear regression, 2018.

4. Computational Enhancements to Accelerate Budget Constrained Regression Model Selection by Mixed Integer Nonlinear Programming

4.1 Introduction

Branch and bound methods to solve the MINLP's developed in Chapters 2 and 3 over only a few predictor variables can compute exact optimal solutions in at most a few seconds. However, to deal with larger data sets, enhancements in branch and bound methods will be required. In this investigation, two classes of such enhancements are proposed, cutting planes and warm starts to strengthen continuous relaxations and speed branch and bound convergence. Computational experiments on five different data sets under different budget limits are conducted to illustrate their effectiveness.

The work will focus on the optimization model of Chapter 2 for budget constrained multiple linear regression is defined as the following mixed-integer nonlinear programming (CLREG):

$$
\begin{aligned}
\min \quad & (\sum_{j=1}^{n}(y_j - \beta_0 - x^T\beta)^2 \\
s.t. \quad & \beta_i = s_{i+} - s_{i-}, \forall i \in 1, \dots p \\
& s_{i+} \leq 1 * \cdot q_i, \forall i \in 1, \dots p \\
& s_{i-} \leq 1 * \cdot q_i, \forall i \in 1, \dots p \\
& \sum_{i=1}^{p} c_i \cdot q_i \leq \text{B} && \text{(Budget constraint)} \\
& s_{i+} \geq 0, s_{i-} \geq 0, \forall i \in 1, \dots p \\
& (\beta_0, \beta) \in R^{p+1} \\
& q_i = 0 \text{ or } 1, \forall i \in 1, \dots p. && \text{(CLREG)}
\end{aligned}
$$

Where, β_0 is the estimated constant term and $\beta_i,\ i = 1, \dots, p$ are estimated coefficients of the predictors. s_{i+}, s_{i-} are non-negative deviation variables. Here, s_{i+}, s_{i-} cannot be both equal to 0 if $|\beta_i| \neq 0$. $q_i, i = 1, \ldots p$ are binary variables. Two sets of "Big-M" constraints guarantee $q_i = 1$ when $|\beta_i| \neq 0$, otherwise $q_i = 0$. The value of M is replaced by 1 after standardizing both response variable and predictors with mean = 0 and standard deviation = 1.

4.2 Cutting Plane Enhancements

Cutting plane methods have been a very popular tool for solving larger integer/mixed integer programming (**IP/MIP**) models in recent years. The fundamental idea of cutting plane technique is to find inequalities that are valid (satisfied) for all feasible solutions to the underling **IPs** and **MIPs** but violated by some solutions to continuous relaxations. Including such cuts in the MINLP model sharpens the approximation provided by its continuous relaxation and thus improves bounds on integer solution values and makes integer-feasible solutions to the relaxations more likely.

There are general techniques for generating cutting planes for **IPs** and **MIPs** without considering the problem structure. Examples are Gomory's fractional cuts and rounding cuts ((Gomory et al., 1958), (Gomory, 1960a), (Gomory, 1960b), (Gomory, 1963)), simple disjunctive cuts ((Marchand, 1998), (Marchand and Wolsey, 2001)) and lift-and-project cuts ((Balas et al., 1993), (Lovász and Schrijver, 1991), (Sherali and Adams, 1990)). Indeed some such cutting planes are already available in the GUROBI solver which is going to be used for solving MINLPs in this chapter.

However, the cutting planes created by the general techniques can be quite inefficient in producing continuous relaxations that closely approximate the set of integer-feasible solutions to any model of interest. The budget constraint in our proposed Multiple Linear regression and Logistic regression optimization model has the exact same formulation as 0–1 knapsack problem – a single main constraint over binary decision variables. Here, ways are considered to obtain stronger inequalities by using such "local" structure.

4.2.1 Knapsacks and Minimal Cover Inequalities

Cover inequalities for 0–1 knapsack have been studied and used extensively in the literature to derive valid inequalities for **IP/MIP** sets. Generalizations of cover inequalities can be found in ((Balas, 1975), (Balas and Zemel, 1978), (Balas and Zemel, 1984), (Hammer et al., 1975), (Padberg, 1979), (Wolsey, 1975)) where the 0–1 knapsack set with generalized upper bounds constraints, the 0–1 knapsack with precedence constraints and the multiple 0–1 knapsack set are studied.

Consider the constraint set of a 0–1 knapsack problem

$$S = \{x \in B^n : \sum_{j \in N} a_j x_j \leq b\} \tag{1}$$

Where $N = \{1, \ldots, n\}$, $a_j \in Z^1_+$ for $j \in N$, and $b \in Z^1_+$. In other words, x belongs to binary sets $\{0, 1\}$ and $a_j \geq 0$ and $b \geq 0$, When $a_j > b$ implies that $x_j = 0$ for all $x \in S$. Thus it is assumed that $a_j \leq b$ for all $j \in N$. The set $C \subseteq N$ is a cover if

$$\lambda = \sum_{j \in C} a_j - b > 0 \tag{2}$$

The cover C is minimal if $a_j \geq \lambda$ for all $j \in C$.

Proposition 4.2.1. ((Nemhauser and Wolsey, 1988)) If $C \subseteq N$ is a minimal cover, then

$$\sum_{j \in C} x_j \leq |C| - 1 \tag{3}$$

is a valid inequality for S.

The extension $E(C)$ of a minimal cover set C is the set $C \cup \{k \in N \backslash C : a_k \geq a_j$ for all $j \in C\}$.

Proposition 4.2.2. If C is a minimal cover, then

$$\sum_{j \in E(C)} x_j \leq |C| - 1 \tag{4}$$

is a valid inequality for S.

Example 4.2.1. Consider the budget constraint in the multiple linear regression model

$$S = \{x \in B^8 : 100x_1 + 200x_2 + 250x_3 + 100x_4 + 150x_5 + 300x_6 + 400x_7 + 350x_8 \leq 600\}$$

$C = \{1, 2, 3, 4\}$ is a minimal cover for S because the sum of corresponding budget coefficients exceeds the limit, but removing any one leaves a subset that conforms to the budget. The corresponding minimal cover inequality

$$x_1 + x_2 + x_3 + x_4 \leq 3$$

The extension $E(C)$ of this minimal cover set C includes all other variables with knapsack coefficients as large as any in C to obtain $E(C) = C \cup \{6, 7, 8\}$ and inequality.

$$x_1 + x_2 + x_3 + x_4 + x_6 + x_7 + x_8 \leq 3$$

is a valid inequality for S.

4.2.2 Method for Generating Minimal Cover Inequalities

All extended minimal cover inequalities are generated and added at the root node to strengthen the constrained multiple linear regression models (CLREG). For a large data set with many predictor variables available, it is challenging to find them all since there are generally an enormous number of such constraints. The following simple example illustrates how all the possible minimal cover inequalities are found.

Table 4.1: Minimal Cover Inequalities Generation

Variable	Cost ($)	MinCover1	MinCover2	MinCover3
X1	1000	⋆		
X2	900	⋆	⋆	
X3	800		⋆	⋆
X4	700			⋆
X5	600			⋆
Total Cost		1900	1700	2100

Example 4.2.2. Suppose there is a data set which includes five predictor variables, $X1$ to $X5$. The cost for buying all those five variables is $4000. And there is only $1500 budget to spend. First all the predictors are ranked from high to low based on their costs, see Table 4.1. Secondly a valid minimal cover inequality is generated starting with the most expensive variables (MinCover1), lastly sequentially exclude the most expensive variable from the current minimal cover inequality and pick next expensive variables which are not included in the current inequality to generate another minimal cover inequality (MinCover2, MinCover3).

$$X1 + X2 \leq 1 \qquad (MinCover1)$$

$$X2 + X3 \leq 1 \qquad (MinCover2)$$

$$X3 + X4 + X5 \leq 2 \qquad (MinCover3)$$

According to Proposition 4.2.2, MinCover2 can be extended by including variable $X1$ with the cost that is higher than any variable in MinCover2. MinCover3 can be extended by including both variables $X1$ and $X2$ for the same reason. And the extended MinCover2 inequality dominates both MinCover1 and MinCover2 inequalities. Therefore, finally generated minimal cover inequalities are as following,

$$X1 + X2 + X3 \leq 1 \qquad (MinCover2\ Extension)$$

$$X1 + X2 + X3 + X4 + X5 \leq 2 \qquad (MinCover3\ Extension)$$

A similar routine will be followed while generating the minimal cover inequalities for a large data set. Those minimal cover inequalities cutting planes will be tested on different real data sets to evaluate whether including them has produced enough gains to be valuable.

4.3 Warm Starts for Constrained Linear Regression Models

The heart of any branch and bound search of an MINLP is comparing bounds computed from continuous relaxations to the objective value of the best known feasible solution to the full mixed-integer model. Having good feasible solutions, and finding them as quickly as possible, can be extremely valuable in the MIP search for lots of reasons. The better the objective value of a feasible solution, the more likely it is that the value of continuous relaxation will exceed it (in a minimization problem) and hence lead to a node being fathomed. Convergence of the process can be greatly accelerated if good warm-start feasible solutions are available. The second part of MINLP enhancement research in this chapter is to investigate and test different methods to produce such good feasible solutions as warm starts.

4.3.1 Processing Non-Integer Solutions as Knapsack Problems Over the Budget Constraint

The task of any such heuristic is to select which binary variables to make =1 (and thus which predictor variables to be included in the solution) while satisfying budget constraints. Focusing on the budget constraint, the task is to solve, at least approximately, a binary knapsack problem like the following to find a good feasible solution.

$$\begin{aligned} \max \quad & \sum_i v_i q_i \\ s.t. \quad & \sum_i c_i q_i \leq \text{Budget} && \text{(Budget Constraint)} \\ & q_i = 0 \text{ or } 1, \text{for every } i. \end{aligned}$$

Here v_i is some measure of the contribution to the overall regression solution of including predictor i in the warm start. For a smaller instance, this warm start knapsack problem can be solved exactly to the optimal. However for instances with relatively many candidate predictors, it will be preferable to only approximately solve the above knapsack after each continuous relaxations.There are two standard heuristic methods for approaching binary knapsacks. One is greedy algorithm by adding the remaining $q_i = 1$ with max v_i as long as budget permits. The other is to use "bang for buck" ratio v_i/c_i to rank variables i, iteratively fixing $q_i = 1$ in decreasing ratio sequence until the budget is filled, and taking the rest of the $q_i = 0$. Based on the number of candidate predictors that the testing data set has, either exact or approximate method will be used for solving warm start knapsack problems.

4.3.2 Continuous Relaxation Solutions as the Starting Point

A natural starting point for heuristics to produce good feasible solutions will be the continuous relaxation optima produced at at every iteration of branch and bound. Any such relaxation will produce relaxation optimal values $\{\bar{q}_i : i = 1, \ldots, p\}$ for the binary variables associated with each predictor variable. Of course if all such $\bar{q}_i$ are binary, the relaxation optimum is already feasible for the full model. But in the usual case where some or all of them are fractional, a heuristic is needed to choose which should be made =1 and which =0 in order to obtain a good feasible solution.

Table 4.2: Continuous Relaxation Solutions

Variable Description	Variable Cost (\$) (c_i)	$\bar{q}_i$	$\frac{\bar{q}_i}{c_i}$
Cylinders	300	0.19	0.000638
Displacement	350	0.22	0.000640
Horsepower	100	0.34	0.003382
Weight	600	0.78	0.001292
Acceleration	200	0.19	0.000935
Model_year	450	0.40	0.000881

Example 4.3.1. The data set used for this example is obtained from UCI repository((Dheeru and Karra Taniskidou, 2017)). The the original data set includes one response variable (consumption in miles per gallon), three multivalued discrete and five continuous predictors, and 392 observations after removing the missing values. In this example, six out of eight predictor variables are selected along with \$1000 budget limit to illustrate how the two proposed heuristics methods in Section 4.3.1 can be used to find a good warm start based on continuous relaxation solutions. We first solve a continuous NLP problem formed by relaxing the binary constraints of discrete decision variable q_i in the CLREG model for the root node of B&B. Variable description, variable cost, continuous relaxation solution $\bar{q}_i$, and ratio between continuous relaxation $\bar{q}_i$ and cost c_i are shown in Table 4.2.

The steps of the greedy algorithm to find a good warm start are: (1) sort the variables by $\bar{q}_i$ in descending order and (2) taking each j in turn, set the $q_i = 1$ if the corresponding cost fits within the remaining budget and $q_i = 0$ otherwise. The variables picked by this greedy algorithm are weight, horsepower and cylinders. The objective value of original CLREG model is 114.308 for the solution obtained by picking those three variables. Similar steps are followed to find a good warm start using the "bang for buck" algorithm. The only difference is sorting the variables based on $\frac{\bar{q}_i}{c_i}$ instead of $\bar{q}_i$. The variables picked by "bang for buck" algorithm are horsepower, weight, and acceleration. The objective value of original CLREG model is 114.807. The greedy algorithm outperforms the "bang for buck" algorithm in terms of the objective value minimization for this specific example.

4.3.3 Unconstrained Statistical Solutions as the Starting Point

In above warm start knapsack problem, v_i is defined as some measure of the contribution to the overall regression solution of including predictor i in the warm start. The non-constrained linear multiple regression describes the statistical relationship between predictor variables and the response variable. The p-value in linear regression output tests the null hypothesis that the estimated coefficient is equal to zero (no effect). A predictor that has a low p-value is likely to make a significant contribution to the regression model. Conversely, a larger p-value means that the predictors have no impact in the response. When p-value is very small, most statistical software tends to report the range value such as p-value <0.0002 instead of the exact value of p-value. However, the v_i measurement of contribution needs an exact value instead of range value. There is another metric called t-value reported in the regression analysis output. And p-value and t-value are inextricably linked as p-value is calculated from a t-test. The greater the magnitude of t-value (it can be either positive or negative), the smaller the p-value, and the greater the evidence that the predictor is highly important to the regression model. Therefore, the absolute t-value can be used as one of the choices for v_i in the warm start knapsack problem.

In linear regression statistical analysis, partial R^2 is another important metric for measuring the mutual relationship between response variable and explanatory variable x_i when other variables $x_j (j \neq i)$ are held constant. The partial R^2 is very useful in multiple linear regression, where it allows to directly estimate the proportion of unexplained variation of y that becomes explained with the addition of variable x_i to the model. Therefore, it can be used as another option of v_i in the warm start knapsack problem.

Both the absolute t-value and the partial R^2 value were obtained by fitting non constrained linear regression model in statistical software R. After that the warm start knapsack problem will be run with v_i replaced by either absolute t-value or partial R^2. Exact method is used for small data sets and two proposed heuristics methods in section 4.3.1 are used for large data sets to find a good feasible solution as a warm start.

4.4 Computational Experience

In this section, computational experiments are conducted on five real-world benchmark data sets to investigate the performance of the proposed knapsack cutting planes and warm starts methods. Computational results are reported in detail. All computational experiments in this chapter are performed using AMPL software with GUROBI solver on a desktop equipped with Intel core 2.70 GHz CPU, 8.00GB usable RAM and Microsoft Windows 7 Professional.

4.4.1 Real World Data Sets

Five benchmark databases, obtained from UCI repository((Dheeru and Karra Taniskidou, 2017)), are used for computational experiments. The descriptions of the five real-world benchmark data sets for budget constrained multiple linear regression are listed in Table 4.3. The first and fourth data sets are generated from Communities and Crime Data Set. The original data set includes 128 variables. 52 of them are selected for data set one and 99 variables are selected for data set four. The second and third data sets are generated from Residential Building Data Set. The original data set includes 105 variables. 52 of them are selected for both data sets two and three. The difference between data set two and three is the response variable. One is about the selling price and another one is about the construction cost. Data set five is generated from Blog Feedback Data Set. The original data set includes 281 variables and 60021 observations. 99 variables and 5000 observations are selected from the original data set for data set five. All the original data sets have no cost assigned for each variable. Therefore, we arbitrarily assigned a cost for each variable we picked. For the first three data sets, the cost of the selected variable starts with \$50, then is increased by \$50, and ends up with \$2600 as the highest cost. For the fifth and sixth

Table 4.3: Descriptions of Real-world Data Sets

Data set	Predictors	Observations	Description of response variable
1	52	1994	Total number of violent crimes per 100K population
2	52	372	Actual sales prices
3	52	372	Actual construction costs
4	99	1994	Total number of violent crimes per 100K population
5	99	5000	The number of comments in the next 24 hours

data sets, the cost of selected variables starts with \$200, then is increased by \$100, and ends up with \$10000 as the highest cost. Budget limits are set at 15%, 25%, 50%, and 75% of total costs of all variables. These different budget limits cover the cost of few variables to the most variables which can give us a broader picture of the performance evaluation.

4.4.2 Computational Experiments on Cutting Planes

In this section, we compare the time spent for solving the MINLP models with vs. without minimal cover inequalities (knapsack cuts) on the five real-world data sets of Table 4.3 given budgets to cover 15%, 25%, 50%, and 75% of total costs. The method proposed in Section 4.2.2 is used to generate all possible knapsack cuts for five data sets under different budget costs. The computational results are reported in Table 4.4.

From Table 4.4, we have the following observations.

- For most experiments on the five data sets, adding knapsack cuts helped reduce the time for solving the MINLP models by 20% to 40%.

- For most experiments on the five data sets, the time for solving the MINLP models is not as dramatically reduced by adding the knapsack cutting planes as might be expected. One reason is that there are many cutting planes already included in GUROBI solver. The solver adds all those cutting planes in the solution process to tighten the formulation by removing undesirable fractional solutions. And there is no option to turn off all those existing cutting planes implemented by the solver.

Table 4.4: Solution Time with Knapsack Cutting Planes

Budget	Knapsack Cuts	CPU time (sec)				
$k\%$		Dataset1	Dataset2	Dataset3	Dataset4	Dataset5
15%	No	22	4	8	28113	53
	Yes	18	3	4	25107	32
	% improve	18%	32%	45%	11%	39%
25%	No	102	52	119	99313	3563
	Yes	85	52	91	93527	2365
	% improve	16%	0%	24%	6%	34%
50%	No	430	210	41	47974	3977
	Yes	425	154	19	43684	2427
	% improve	1%	26%	54%	9%	39%
75%	No	82	40	8	17373	2373
	Yes	47	16	6	15401	1234
	% improve	43%	61%	24%	11%	48%

4.4.3 Computational Experiments on Warm Start Methods

In this section, we solve warm start knapsack problems by replacing v_i with continuous relaxation value and two statistical solutions proposed in Sections 4.3.2 and 4.3.3. The computational results in terms of percentage by which the warm start solution exceeded the optimal the objective value are reported on five real-world data sets under different budget limits in Table 4.5.

From Table 4.5, we have the following observations.

- For most experiments on five real data sets, the warm starts generated based on partial R square value and the ratio between partial R square and the corresponding cost yields better results than other methods in terms of minimizing the objective starting solution value .

- For most experiments on five real data sets, the continuous relaxation value of the binary variables q_i does not provide a good warm start as compared to the other methods.

- For most experiments on five real data sets, the warm starts generated based on the

Table 4.5: Test Results of Objective Value Improvement on Warm Starts

Budget	Data set	Percent Deviation from Optimal Value (%)					
$k\%$		$\bar{q}_i$	$\frac{\bar{q}_i}{c_i}$	$\lvert t_i\rvert$	$\frac{\lvert t_i\rvert}{c_i}$	R_p^2	$\frac{R_p^2}{c_i}$
15%	1	13.12	31.45	12.54	54.35	1.30	1.07
	2	28.64	17.94	77.80	48.32	34.24	27.09
	3	116.46	111.46	25.40	20.30	9.63	10.40
	4	35.80	28.65	14.65	22.00	4.03	4.03
	5	22.64	31.55	0.55	0.54	1.10	1.10
25%	1	37.68	31.09	11.79	40.36	1.27	1.80
	2	44.46	39.43	112.15	28.43	12.98	12.81
	3	130.85	128.86	38.93	23.29	8.98	11.18
	4	11.90	12.18	10.99	23.40	2.35	2.32
	5	23.85	32.82	0.64	0.79	1.18	1.18
50%	1	6.19	9.74	6.22	6.76	1.98	2.68
	2	35.61	35.61	7.57	16.01	6.71	6.18
	3	134.09	134.47	1.57	4.97	4.89	4.59
	4	13.68	13.68	4.81	5.41	2.01	2.10
	5	17.36	17.30	0.13	0.47	0.04	0.08
75%	1	5.47	7.22	2.52	4.05	2.71	2.31
	2	36.00	36.52	3.31	3.68	4.37	3.04
	3	134.85	136.73	0.08	0.46	0.55	0.69
	4	8.83	8.83	0.51	1.27	1.54	1.52
	5	5.33	17.13	0.02	0.17	0.02	0.02

$|t_i|$ are better than continuous relaxation value of binary variable but worse than partial R square value.

4.4.4 Solution Time Results with and without Cutting Planes and Warm Starts

In this section, instead of letting GUROBI solver automatically picks the starting points while solving MINLP models using Branch and Bound (B&B) method, we choose the feasible solutions produced by both partial R^2 value and the ratio between partial R^2 and the corresponding cost as warm starts. Other than this, we also combine both knapsack cutting planes and warm starts together to see whether or not we can further speed up the solving process.

The time for solving the MINLP models on five real data sets using different enhancements is reported in Table 4.6. For completeness, some computational results in Table 4.4 are also included in Table 4.6. Since knapsack cuts are generated manually using the proposed method in Section 4.2.2, the time for generating those cuts is not included in Table 4.6. Also, the time for solving the unconstrained statistical models and corresponding warm start knapsack problems to find a good feasible solution is less than a second which can be ignored compared to B&B solution time. Therefore, those times are also not included in Table 4.6.

In Table 4.6, "NN" means no cutting planes no warm starts, "YN" means with cutting planes no warm starts, "NY_{PR2}" means no cutting planes with warm starts based on partial R^2, "NY_{PR2_Ratio}" means no cutting planes with warm starts based on the ratio between partial R^2 and cost, "YY_{PR2}"and "YY_{PR2_Ratio}" means with cutting planes and two different types of warm starts.

From Table 4.6, we have the following observations.

- For all experiments on five real data sets, the time for solving MINLPs with knapsack cutting planes added is smaller than the time without cutting planes added. Still, the difference see in terms of percentage of time saved across most of experiments on five data sets is only modest .

- For most experiments on five real data sets, using warm starts generated by either partial R^2 or partial R^2 ratio help more dramatically to speed the MINLPs solving process of the B&B algorithm. The larger the data set, the more time can be saved.

- For most of time, there is some gain by combining both knapsack cutting planes and warm starts together especially for larger data sets.

4.5 Conclusions and Extensions

In this chapter, we conduct extensive computational experiments to validate the performances of the proposed knapsack cutting planes and warm starts for solving the

Table 4.6: Solution Time with and without Cutting Planes and Warm Starts

Budget	Solution Type	CPU time (sec)				
k%		Dataset1	Dataset2	Dataset3	Dataset4	Dataset5
15%	NN	22	4	8	28113	53
	YN	18	3	4	25107	32
	NY$_{PR2}$	18	3	8	9406	40
	NY$_{RP2\_Ratio}$	18	6	7	9406	40
	YY$_{PR2}$	15	2	6	8025	27
	YY$_{PR2\_Ratio}$	17	3	6	8025	27
25%	NN	102	52	119	99313	3563
	YN	85	52	91	93527	2365
	NY$_{PR2}$	79	55	101	19168	1682
	NY$_{RP2\_Ratio}$	85	50	111	17314	1682
	YY$_{PR2}$	73	56	93	18074	907
	YY$_{PR2\_Ratio}$	75	56	92	19803	907
50%	NN	430	210	41	47974	3977
	YN	425	154	19	43684	2427
	NY$_{PR2}$	346	208	22	7054	1578
	NY$_{RP2\_Ratio}$	348	200	29	4463	1628
	YY$_{PR2}$	353	178	36	6112	1250
	YY$_{PR2\_Ratio}$	375	148	21	4890	1248
75%	NN	82	40	8	17373	2373
	YN	47	16	6	15401	1234
	NY$_{PR2}$	53	41	7	5166	561
	NY$_{RP2\_Ratio}$	52	30	6	4487	473
	YY$_{PR2}$	70	20	7	2636	132
	YY$_{PR2\_Ratio}$	37	16	7	3093	202

budget constrained multiple linear regression model using B&B algorithm. Our major findings are summarized below.

4.5.1 Conclusions

The most important results of the above research can be summarized as follows:

- Developed knapsack cutting planes techniques that can sharpen relaxations and materially reduce B&B solution times.
- Developed warm start methods based on unconstrained statistical computation

produced that can significantly greater B&B time improvements.

- Combined both methods together, to further improve the solving process and make larger instances more possible to be solved to the optimality.

4.5.2 Extensions

Several directions for further extending this research are also suggested:

- Further refining the methods presented above to deal with even bigger instances. One example is instead of adding all those knapsack cutting planes at the root node, the opportunity of adding them during the solution process could be explored. This way, would make sure constraints would be added only if they will help.

- In this chapter, our proposed enhancements were tested only on the constrained multiple linear regression model. Another extension would adapt those enhancements to be tested on budget constrained categorical regression and logistic techniques as discussed in (Zhang et al., 2018).

Reference

Egon Balas. Facets of the knapsack polytope. Mathematical programming, 8(1):146–164, 1975.

Egon Balas and Eitan Zemel. Facets of the knapsack polytope from minimal covers. SIAM Journal on Applied Mathematics, 34(1):119–148, 1978.

Egon Balas and Eitan Zemel. Lifting and complementing yields all the facets of positive zero-one programming polytopes. Mathematical Programming, 9:13–24, 1984.

Egon Balas, Sebastián Ceria, and Gérard Cornuéjols. A lift-and-project cutting plane algorithm for mixed 0–1 programs. Mathematical programming, 58(1-3):295–324, 1993.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Ralph Gomory. An algorithm for the mixed integer problem. Technical report, RAND CORP SANTA MONICA CA, 1960a.

Ralph E Gomory. Solving linear programming problems in integers. Combinatorial Analysis, 10:211–215, 1960b.

Ralph E Gomory. An algorithm for integer solutions to linear programs. Recent advances in mathematical programming, 64:260–302, 1963.

Ralph E Gomory et al. Outline of an algorithm for integer solutions to linear programs. Bulletin of the American Mathematical society, 64(5):275–278, 1958.

Peter L Hammer, Ellis L Johnson, and Uri N Peled. Facet of regular 0–1 polytopes. Mathematical Programming, 8(1):179–206, 1975.

László Lovász and Alexander Schrijver. Cones of matrices and set-functions and 0–1 optimization. SIAM journal on optimization, 1(2):166–190, 1991.

Hugues Marchand. A polyhedral study of the mixed knapsack set and its use to solve mixed integer programs. Faculté des Sciences Appliquées, Université catholique de Louvain, 1998.

Hugues Marchand and Laurence A Wolsey. Aggregation and mixed integer rounding to solve mips. Operations research, 49(3):363–371, 2001.

George L Nemhauser and Laurence A Wolsey. Integer programming and combinatorial optimization. Wiley, Chichester. GL Nemhauser, MWP Savelsbergh, GS Sigismondi (1992). Constraint Classification for Mixed Integer Programming Formulations. COAL Bulletin, 20:8–12, 1988.

Manfred W Padberg. Covering, packing and knapsack problems. In Annals of Discrete Mathematics, volume 4, pages 265–287. Elsevier, 1979.

Hanif D Sherali and Warren P Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. SIAM Journal on Discrete Mathematics, 3(3):411–430, 1990.

Laurence A Wolsey. Faces for a linear inequality in 0–1 variables. Mathematical Programming, 8(1):165–178, 1975.

Jingying Zhang, Ronald L Rardin, and Justin R Chimka. Budget constrained model selection for logistic regression, 2018.

5. Conclusion and Future Research Directions

This dissertation has proposed and studied two budget-constrained regression models for continuous and categorical variables respectively using Mixed Integer Nonlinear Programming (MINLP) to select the best explanatory variables to be included in solutions. Section 5.1 summarizes its contributions. Section 5.2 suggests some directions for future research.

5.1 Summary of Contributions

As a variable or feature selection method, forward, backward and stepwise methods are commonly used. The basic concept of those three methods is adding one variable at a time to the model to minimize the sum of squared of errors or maximize the likelihood function and dropping variables from the model if they are redundant. Therefore, there is no guarantee that a truly best subset of features will be selected.

In recent years, along with hardware and algorithm improvements, different optimization models have been implemented for solving variable selection problems. However, none of those existing optimization models has considered budget constrained variable selection.

In this dissertation, two budget or count-constrained MINLP models for continuous and categorical response variables respectively are proposed to choose an optimal subset of variables to be included in the model. Two enhancements of the MINLP model for the continuous response variable are also studied to speed up the optimization model solving process using B&B algorithm. Specific contributions of this dissertation are summarized as follows.

- We have proposed a budget or count-constrained regression model for a continuous response variable using MINLP. One of the most commonly used data standardizing methods has been implemented to reduce the value of big-M coefficients to 1 in the formulation. Properties of constructed MINLP model such as solvability and global optimality have been studied.

- Computational experiments on the realistic retail store data sets have been conducted to investigate the performance of the proposed MINLP model for continuous response variable. The computational results indicate that, (i) our proposed MINLP model outperforms the statistical software outputs in optimizing the objective function under a limit on the number of explanatory variables selected, and (ii) our proposed MINLP is shown to be capable of selecting the optimal combination of explanatory variables under a budget limit covering the cost of acquiring data sets.

- We have also proposed a budget or count-constrained logistic regression model for categorical response variables limited to the binary case. Different data standardizing methods have been studied. Variance-based fully standardized coefficients method has been implemented to reduce needed big-Ms in the MINLP formulation in order to speed up the solving process.

- Computational experiments on nine realistic data sets indicate that our proposed optimization model outperforms the standard heuristic methods in terms of minimizing the negative log-likelihood function, especially for bigger data sets. Studies varying prices of variables and budget limits demonstrate that our proposed model can be used for deciding what data sources to consider and how large a budget is needed to obtain satisfactory results.

- We have proposed to adjust the objective function of the logistic case to either **AIC** or **BIC** value to overcome the over-fitting issue. The adjusted model is able to reduce the risk of over-fitting by introducing a penalty term to the objective function which grows along with the number of parameters.

- We have proposed and developed tools for cutting plane and warm start solutions as, two types of enhancements to speed up the solving process of the MINLP model for a continuous response variable. Extensive computational experiments results indicate that our two proposed enhancements significantly reduce the computational time, especially for bigger data sets.

5.2 Future Research

In chapter 2, linear regression, also known as ordinary least squares (OLS) is the method we used to build the objective function of our proposed budget constrained MINLP model. There are some known weaknesses related to OLS algorithm such as sensitivity to outliers and multicollinearity and prone to overfitting. To address these problems, several advanced methods have been proposed by researchers, such as ridge regression, lasso regression, and partial least squares regression (PLS). Therefore, the idea of incorporating our budget or count limit constraint to lasso, ridge or PLS to overcome the weaknesses of OLS model and at the same time select the best subset of variables within budget limit could be an interesting area to explore.

In chapter 3, our proposed budget or count-constrained regression model for categorical response variable is limited to the binary case. Categorical variable can be ordinal, nominal or even count data. Different logistic regression models such as ordinal, nominal logistic regression can be used to model the categorical responsible variable with more than two discrete values. Poisson regression or alternatives to Poisson for example negative binomial or zero-inflated models can be used to model count data. Combining the budget constraint with all those different regression models, or the generalized linear model (GLM) could be another interesting area for consideration.

In chapter 4, our two proposed enhancements are only developed and tested on the constrained multiple linear regression model. One extension would adapt those enhancements to be tested on budget constrained categorical regression models as well.

Another potential extension is to further refine the two proposed enhancements to deal with even bigger instances.