Forecasting Completed Cost of Highway Construction Projects Using LASSO Regularized Regression

Yuanxin Zhang[1]; R. Edward Minchin Jr., PE, A.M. ASCE[2]; Duzgun Agdas, PE, M. ASCE[3]

[1]Graduate Assistant, Rinker School of Construction Management, Univ. of Florida, 324 Rinker Hall, Gainesville, FL 32611-5703, Email: yuanxinzhang@ufl.edu

[2]Rinker Professor, Rinker School of Construction Management, Univ. of Florida, 304 Rinker Hall, Gainesville, FL 32611-5703, (corresponding author) Email: minch@ufl.edu

[3]Lecturer, School of Civil Engineering and Built Environment, Queensland University of Technology, Gardens Point, Brisbane, QLD 4001, Australia. Email: duzgun.agdas@qut.edu.au

ABSTRACT

Finishing highway projects within budget is critical for state highway agencies (SHAs) because budget overruns can result in severe damage to their reputation and credibility. Cost overruns in highway projects have plagued public agencies globally. Hence, this research aims to develop a parametric cost estimation model for SHAs to forecast the completed project cost prior to project execution to take necessary measures to prevent cost escalation. Ordinary least square (OLS) regression has been a commonly used parametric estimation method in the literature. However, OLS regression has certain limitations. It, for instance, requires strict statistical assumptions. This paper proposes an alternative approach—least absolute shrinkage and selection operator (LASSO)—that has proved in other fields of research to be significantly better than the OLS method in many respects, including automatic feature selection, the ability to handle highly correlated data, ease of interpretability, and numerical stability of the model predictions. Another contribution to the body of knowledge is that this study simultaneously explores project-related variables with some economic factors that have not been used in previous research, but economic

conditions are widely considered to be influential on highway construction costs. The data were separated into two groups: one for training the model and the other for validation purposes. Using the same dataset, both LASSO and OLS were used to build models, and then their performance was evaluated based on the mean absolute error, mean absolute percentage error, and root mean square error. The results showed that the LASSO regression model outperformed the OLS regression model based on the criteria.

Keywords: Highway Construction Cost; LASSO; Completed Cost; Ordinary Least Square; Parametric Cost Estimation.

INTRODUCTION

Cost management is one of the most important responsibilities for the decision makers that lead state highway agencies (SHAs). Completing highway projects within budget is critical to SHAs because such performance is tied to their public image and credibility. Cost overruns on highly visible highway projects are likely to attract tremendous attention from the public, press, and legislators (Wilmot and Cheng 2003; Wilmot and Mei 2005). Unfortunately, cost overruns on highway projects have been a ubiquitous phenomenon that plagues public agencies globally (Anderson et al. 2007; Cantarelli et al. 2010; Flyvbjerg et al. 2002). Flyvbjerg et al. (2002) discovered that nine out of ten infrastructure transportation projects exceed their original contract prices and that cost overruns occurred across 20 countries on five continents. Cost management is a challenging task for SHAs (Anderson et al. 2007) because numerous factors can contribute to cost overruns in highway projects, including project complexity, duration, contractor experience and capability, weather conditions, site conditions, economic situations, local political and social climates, and so forth.

Theoretically speaking, the probability of cost overrun and underrun should be equal (Emhjellen et al. 2001). However, many studies have shown that cost overrun is more prevalent than underrun in the highway construction industry, and its magnitude is large (Odeck 2004). Because SHAs have a fixed amount of financial resources to build highway projects (Flyvbjerg et al. 2002), both cost overrun and underrun are bad for SHAs and the traveling public. Cost underruns result in a suboptimal number of projects being selected and completed (Minchin et al. 2004; Minchin et al. 2005), which may leave insufficient funds for other critical projects (Asmar et al. 2011). Cost overrun, on the other hand, leads highway agencies to overspend a given year's budget and then "steal" from next year's to cover the shortfall. Consistent occurrences of this phenomenon can cause "fiscal and political complications" (Minchin et al. 2005) as well as impaired credibility and lost public trust (Wilmot and Cheng 2003).

Most projects are let through a competitive setting in the United States. In this competitive setting, contracts are usually awarded to the lowest responsive bidders under the unit price arrangement. However, the lowest bid occasionally turned out not the least expensive in many cases because the completed cost is subject to change due to a variety of reasons (Williams 2002). More so than in building construction, highway projects entail high risks because of the more complicated unforeseen site conditions, which can enormously increase the quantities of work and in turn raise the completed cost. Unexpected cost increases and decreases can also be tied to the different bidding strategies that contractors use to win the bid. In some cases, contractors make huge mistakes in their estimates by missing some important pay items. Under this circumstance, projects may be delayed or disserted because of the serious mistakes by contractors. The parametric cost estimation model is capable of discovering the pattern between the completed cost and significant factors using a large number of historical data. The model can

69 forecast the completed cost during the preconstruction phase to provide information help SHAs
70 take appropriate actions during the construction phase (Williams 2002). For instance, if the
71 model predicts that the completed cost of a project is likely to be much higher than the low-bid
72 price, the project likely requires careful supervision during construction (Williams 2002).
73 An effective cost-estimating system should be able to estimate project costs quickly and
74 accurately (Anderson et al. 2007; Herbsman and Mitrani 1984). The conventional quantity
75 takeoff and adjusted unit price approach adopted by most Departments of Transportation (DOTs)
76 (Chou et al. 2006) heavily depends on the estimators' experience and sound judgment to create
77 highly accurate estimates. The subjective judgments made by estimators are not always
78 consistent and reliable, which leads to errors in estimations (Chou et al. 2015). Therefore, this
79 study introduced an objective method that does not require significant experience in estimating
80 and can avoid human errors, but it can recognize the pattern between the completed cost and its
81 significant contributors. In addition, the parametric estimation method using LASSO regression
82 yields reliable estimates in a much faster fashion compared to the conventional estimating
83 methods and OLS regression.

84 **PROBLEM STATEMENT**

85 The reliability of cost estimates for highway projects is of great importance; thus, a great deal of
86 effort has been devoted to the development of cost prediction models for highway construction
87 projects. Since its first use in the 1970s (Kim et al. 2004), ordinary least square (OLS) regression
88 has been the most commonly used approach to building cost estimation models (Wilmot and Mei
89 2005). There are a number of factors contributing to its popularity. It is easy to grasp and master
90 the concept from which OLS regression models are derived, and it is straightforward to interpret
91 the models developed compared to artificial neural network (ANN) models. In addition, multiple

92 software programs (e.g., SAS, SPSS, R, Minitab, Matlab, and Stata) are readily available for
93 OLS regression model development. Its popularity is also ascribed to the fact that it is a good
94 tool for examining and explaining the relationships between predictor and response variables.
95 The least square fitting theory is still held, and it is unnecessary to check the assumptions behind
96 OLS regression if the primary purpose of developing an OLS model is for the discovery and
97 explanation of relationships. Finally, in application to practical problems, OLS linear models
98 generally have distinct advantages in terms of inference and extrapolation compared to nonlinear
99 models (James 2013).

100 Despite of all these merits, the OLS method also has many serious drawbacks. First, the
101 OLS method requires external efforts to select critical predictors (e.g., best subset selection,
102 forward and backward stepwise method, and hybrid stepwise method). Second, to be reliable for
103 prediction, the OLS model is bound by strict assumptions that are highly technical and tedious.
104 Specifically, residuals of linear regression models must meet the following assumptions:
105 (i) residuals must be independent; (ii) residuals must be normal or approximately normal; (iii) the
106 mean of the residuals must be zero, and the variance must be a constant. Frequently, the
107 empirical data gathered from field do not meet these assumptions. Third, OLS regression is
108 susceptible to multicollinearity, which can be defined as strong correlation among independent
109 variables in the regression models. Multicollearirity can cause instability in model predictions
110 (Fu 1998; Tibshirani 1996). Last but not least, handily available statistical analysis programs can
111 cause overreliance on these solutions and overlooking the conditions to be met before OLS
112 regression can be appropriately used.

113 This study, therefore, proposes an alternative process—the least absolute shrinkage and
114 selection operator (LASSO). The LASSO approach shares the advantages of the OLS method

115 and offsets the aforementioned shortcomings associated with the OLS approach when building a
116 regression model. Compared to OLS regression, the LASSO regression coefficients are not
117 unbiased estimators because of the biased term added to the diagonal of the design matrix
118 (Tibshirani 1996), which results in numerous other advantages. Not only does the LASSO model
119 perform better when handling multicollinearity but it also is better in terms of numerical
120 stability. Computationally, LASSO is also superior to the stepwise and best-subset methods in
121 terms of model development (James 2013).

122 Additionally, it has been commonly believed by professionals in the construction sector
123 that the economy has a substantial influence on highway construction costs (Anderson et al.
124 2007; Flyvbjerg et al. 2002; Herbsman 1983). Specifically, economic fluctuation causes changes
125 in highway construction cost. During economic booms, for instance, when there are more
126 projects available in the market, contractors might accumulate an extensive backlog and become
127 more selective when bidding for jobs, which increases completed construction costs. In contrast,
128 during economic recessions, available jobs in the market are scarce, but contractors do not want
129 to lay off their productive crews and efficient management teams, because they want to survive
130 these difficult times with their good people and begin making profits again when the economy
131 recovers. During these times, contractors hungry for new jobs are willing to take jobs with small
132 profits or even no profit to keep the crews and management team working.

133 Previous research, surprisingly, has not simultaneously considered economic factors with
134 project-related variables when building a regression model so that significant economic factors
135 that have an effect on construction cost estimates can be identified through the model. As a
136 result, this research investigates the effect of leading economic indicators on completed costs
137 when building the OLS and LASSO regression models.

PREVIOUS RELATED RESEARCH

Public Agencies' Efforts

The construction industry has focused continuous interest on improving the quality of cost estimates. In the early 1980s, the Florida Department of Transportation (FDOT) initiated a research project to create a model that could forecast highway project costs in their budgeting process (Herbsman 1983). Using statistical methods to analyze the main types of highway construction projects, Herbsman (1983) considered the general features of the highway construction industry and specific site conditions. In the late 1980s, a group of researchers were hired to improve the accuracy of preliminary cost estimates for federally sponsored projects. As part of that research, Morrs (1990) proposed a system with standard terms and uniform formats, training the estimating personnel to ensure their capability in generating accurate early estimates and encouraging more use of parametric and probabilistic methods. Circa 2006, the Texas Department of Transportation (TxDOT) funded research to improve its preliminary cost estimates (Chou et al. 2006). The researchers developed a statistical model to first predict quantities of the major work items and then use the predicted quantities to estimate the total cost of a project during the early phases.

Investigation of State of the Practice in Cost Estimation

Turochy et al. (2001) investigated the practices of DOTs in estimating highway project costs during early phases and categorized them into three groups: (1) some DOTs use rough estimates of quantities of the major pay items as well as unit prices produced by a collection of prices culled from previous or present construction contracts; (2) many DOTs rely on cost-per-mile and cost-per-item tables consisted of generalized costs for several project designs, which sometimes requires engineering experience and judgement to adjust the prices from those tables.

Adjustment of the prices improves the accuracy of cost estimates as the prices are tailored to project specific conditions; (3) a statewide standard procedure in cost estimating is absent in few DOTs. Those DOTs use whatever method they prefer, including the above two methods or methods purely dependent on engineering judgement and experience.

Byrnes (2002) also supported the findings of Turochy et al. (2001) and concluded that the most widely used technique in practice by DOTs was the lane-mile cost average for highway projects. However, using such single-value estimation has numerous downsides (Chou et al. 2005; Chou et al. 2006). Thus, researchers and engineers adopted many other superior approaches. Most of them were based on OLS regression analysis.

Application of Ordinary Least Square Regression

Herbsman (1983) used a multiple regression analysis to establish long-term projections on highway construction costs. The model was based on four aspects. First, the input costs included direct and indirect costs. Second, a series of indices allowed the prediction of the "input" cost elements. Third, the one major factor influencing highway project costs was bidding volume, the total volume of work in a specific area during a certain time period. In addition, the estimator's influence was also regarded as an important factor. In predicting final project costs, Wright and William (2001) developed a regression model using logarithmically transformed data, in which the median bid and normalized median absolute deviation were selected as the best independent variables to enter into the final model. William (2002) used bid data, contract duration, and the number of bidders as independent variables in both regression models to predict completed project costs. William (2003) developed a multiple regression model via the natural log transformation of the low-bid and final project cost, which revealed a near-perfect linear relation between the predictor variables and the response variable and excellent predictive performance

184 using an independent dataset. Sonmez (2004) used both linear regression and neural networks to
185 discover that simultaneous use of both methods can lead to a satisfactory model with acceptable
186 prediction performance. In addition, range estimates proved to help evaluate uncertainties
187 associated with conceptual cost estimation. Minchin et al. (2004) developed a regression model
188 using empirical data from several state DOTs, and the number of bidders was identified as the
189 most important factor increasing the deviation between the engineer's estimate and the low bid.
190 In addition, the economic influence was also analyzed, but separately.

191 To avoid the problems caused by multicollinearity among the independent variables,
192 Chan and Park (2005a) utilized principal component analysis for the original data and then
193 generated a regression model to forecast the construction project cost. Petroutsatou et al. (2006)
194 employed regression to forecast road tunnel costs during the preliminary phases of project
195 development. The predictor variables for the final model were determined by a literature review
196 and the input of experts from both academia and the construction industry. Later, several linear
197 regression models were produced by Mahamid and Bruland (2010) to predict the cost of three
198 major road construction activities using road length, pavement width, base course width, terrain
199 condition, soil drillability, and soil suitability as independent variables. Mahamid (2011) used a
200 multivariable regression model to project road construction early costs. They found that the
201 model using bid quantity as a predictor variable performed better than those which utilized road
202 length and width. Wang and Liu (2012) built a multiple regression model to predict the mixture
203 price of asphalt. The independent variables used for model development included economic
204 factors (Kentucky diesel price index [KDPI] and Kentucky asphalt price index [KAPI]), the
205 number of bidders, and the occurrence of economic recession. Hollar et al. (2012) formed a
206 multiple regression model to forecast preliminary engineering costs using bridge projects data

from North Carolina DOT. The final model contained seven predictors comprised of four numerical variables and three categorical variables. Using multiple regression analysis and ANNs, Cirilovic et al. (2014) forecasted the cost of road rehabilitation and reconstruction projects on the basis of the projects completed in European and Central Asian countries. A total of 19 variables under three major categories (oil price-related, country-related, and project-related variables) were analyzed in the model development.

In previous research in cost estimation, no one used the LASSO approach to build cost estimation models for highway construction projects. With regard to the commonly used OLS method, no research has displayed and elaborated on the rigorous process concerning residual diagnosis, which is the guarantee of model reliability. Furthermore, economic factors, to the authors' best knowledge, have not simultaneously been used as predictor variables with project-related variables when developing a parametric cost estimation model. Nevertheless, the economy has been widely regarded as having a significant influence on highway construction costs (Anderson et al. 2007; Flyvbjerg et al. 2002; Williams 2003). This economic perspective has been verified through an online survey of all DOTs in the United States (nine out of ten respondents think that the economy has an influence on highway cost).

RESEARCH METHODOGY

Ordinary Least Square Method

The principle behind the OLS method is to minimize the sum of the "distances" (or squared difference) between the predicted and observed values. The optimal regression line ensures that the sum of squared residuals (SSR) or "distances" is the smallest, which depends on the coefficients ($\hat{\beta}$) or slope. Thus, the key is to find the regression coefficients. The mathematical formulas to find the best regression coefficients are shown as Eq. 1 and 2.

$$\text{Min } \{\text{SSR}\} = \text{Min } \{ (\boldsymbol{Y} - \boldsymbol{X}b)^T (\boldsymbol{Y} - \boldsymbol{X}b)\} \quad (1)$$

$$\hat{\beta}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2)$$

where X is the design matrix that holds all independent variables (e.g., number of bidders, contractor past performance ratings, project duration, total lengths of the projects, number of lanes, weather days, contract price, Consumer Price Index, construction spending, prime loan rate, Producer Price Index); *b* represents the coefficient vector. Y is the matrix containing the dependent variable (completed cost); SSR represents the sum of the squared residuals; and $\hat{\beta}^{OLS}$ represents the OLS regression coefficients.

When building the OLS regression model, this study chose the best-subsets selection procedure because it is this procedure that is least likely to miss the optimal model (James 2013), compared to forward, backward, and stepwise techniques. The best subset procedure entails more computation work; nonetheless, it is no longer a significant issue because of the tremendous advancement in computer science. The best subset model selection procedure includes the following steps (see Fig. 1):

- Step 1. All possible regression models consisting of all possible combinations of the candidate predictors were identified. The total number of possible models identified was 2^{11}.
- Step 2. From the possible models identified in the prior step, the "best" models with one predictor, two predictors, three predictors, and so on, were determined according to some well-defined criteria—R^2, adjusted R^2, Mallows' Cp, and Schwarz's Bayesian Information Criterion (BIC). Using multiple criteria for model fitting provides a balance in selecting the best model because each of them has its pros and cons. R^2, most used in practice, tends to overfit a model, meaning including more than necessary number of variables in the model. This is because the R^2 value increases as the number of variables increases (see Eq. 3). The

253 higher R^2 is, the better the model fits the data. However, goodness of fit does not necessarily
254 lead to good predictive power (Rawlings et al. 2001). Subsequently, other criteria were
255 proposed to offset this shortcoming with R^2. $R^2_{adjusted}$ does not increase as dramatically as
256 R^2 (see Eq. 4) (Rawlings et al. 2001), which reduces overfitting. Malow's Cp and BIC
257 decrease when adding a large number of variables (see Eq. 5 and 6), which prevents
258 overfitting as well (Rawlings et al. 2001).

259 $$R^2 = \frac{SSR}{SST} \quad (3)$$

260 $$R^2_{adjusted} = \frac{SSR/_{n-k}}{SST/_{n-1}} \quad (4)$$

261 $$C_p = \frac{SSR_p}{MSE_{full}} - (n - 2p) \quad (5)$$

262 $$BIC = n\,Ln\,(SSR) - n\,Ln\,(n) + p\,Ln\,(n) \quad (6)$$

263 Where SST denotes total sum of squared residuals; SSR represents sum of squared residuals,
264 n is sample size; k is equal to the number of regression coefficients (including the intercept);
265 SSR_p stands for sum of squared residuals based on the model containing p number of
266 predictors. MSE_{full} denotes mean square error of the full model consisted of all predictors.

267 The "best" models with the same number of variables are whose with the highest R^2,
268 adjusted R^2, and the lowest BIC value. The "best" model should have a Mallows' C_p value
269 less than or equal to the number of independent variables in the model.

270 ▪ Step 3. Based on a group of models selected in the last step, some further evaluations must be
271 conducted to determine the final model from the "best" models. The values of the criteria for
272 the models with different numbers of variables were compared.

273 ▪ Step 4. After the final model was decided, residual diagnosis was carried out to check the
274 satisfaction of the assumptions, which is a critical step to warrant reliable performance of the

model. The tests for those purposes include the Shapiro–Wilk normality test, Durbin–Watson test for autocorrelated residuals, Breusch–Pagan test for heteroscedasticity (constant variance), and Variance Inflation Factor test for multicollinearity [for details, (Rawlings et al. 2001)].

- Step 5. The influential data points also must be identified. Cook's distance values were checked to identify the highly influential data points. The outliers Bonferroni test was performed to discover and eliminate outliers. This study also applied the Box–Cox test to decide how to transform data to improve the linearity and fitness of the model [for details, (Rawlings et al. 2001)].

LASSO Method

In the OLS procedure, estimation of the regression coefficients is based on **X'X** being susceptible to collinearity, which can cause the regression model to behave poorly (e.g., increasing variance of the estimated regression coefficients). The LASSO approach, conversely, increases a small amount of bias by adding a term in the diagonal of design matrix **X** by a magnitude of λ but increases robustness with correlated predictor variables when making predictions (Tibshirani 1996). Estimation of the coefficients by the LASSO approach is shown as Eq. 3 and 4.

Additionally, unlike the OLS method, the LASSO approach adds a constraint to the sum of the regression coefficients, which is a penalty for adding too many variables to the model (see Eq.7). In the OLS method, the more independent variables a model contains, the smaller its SSR becomes. However, this has several negative consequences including undermined prediction accuracy and reliability. LASSO places a penalty on having a large number of variables in a

297 model by making the sum of the absolute values of the coefficients less than a constant (see Eq.
298 8), which can constrain overfitting.

299 $$\text{Min } \{\text{SSR}\} = \text{Min } \{ \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{n}\left|\beta_j\right| \} \quad (7)$$

300 $$\hat{\beta}^{LASSO} = (\mathbf{X'X} + \lambda \mathrm{I})^{-1}\mathbf{X'Y} \quad (8)$$

301 where SSR represents the sum of the squared residuals, and β_j represents the regression
302 coefficients; λ is a nonnegative tuning parameter; y_i is the value of the completed cost; x_i is the
303 value of each predictor variable; X is the design matrix that contains all the same independent
304 variables evaluated in the OLS method; Y is the matrix containing the dependent variable
305 (completed cost); and $\hat{\beta}^{LASSO}$ represents the matrix of the estimated LASSO regression
306 coefficients.

307 In the LASSO procedure, the value of the tuning parameter λ was divided into small,
308 equal segments. Each segment of λ yielded a set of regression coefficients with a minimum SSR.
309 The λ value ranged from a model that yielded a nonzero coefficient to one that resulted in all
310 zero coefficients. As a result, there were many models corresponding to the λ values. There is a
311 tradeoff between λ and the final $\hat{\beta}^{LASSO}$. Large λ causes small $\hat{\beta}^{LASSO}$—namely, fewer
312 independent variables. Without external effort in the model selection process, the LASSO
313 procedure automatically selects significant variables based on the λ that generated the minimal
314 SSR.

315 In summary, the OLS and LASSO methods have a few things in common, but there are
316 stark differences as well. They both aim to fit a regression model to minimize the SSR. OLS
317 models can still be used to display the relationship between the predictor variables and the
318 response variable even if high collinearity among the independent variables exists; it is reliable
319 for prediction only when the residuals satisfy the assumptions; otherwise, there will be numerous

problems in prediction. The LASSO model, on the other hand, is robust with the multicollinearity issue. The following subsections first show the model building processes using these two methods, and then the performance of the models is evaluated based on the identical criteria.

Model Performance Assessment

Using a new dataset, the final models developed through the OLS and LASSO methods were evaluated based on three commonly used criteria: the mean absolute error (MAE) and the mean absolute percentage error (MAPE), and the root mean square error (RMSE). The MAE and MAPE values were calculated to compare the predictive performance of the OLS and LASSO models (see Eq. 9, 10, and 11).

$$\text{MAE} = \frac{\sum_{i=1}^{i=n}|y_i - p_i|}{n} \tag{9}$$

$$\text{MAPE} = \frac{\sum_{i=1}^{n}\left(\left({}^{|y_i - p_i|}/_{y_i}\right) \times 100\right)}{n} \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - p_i)^2}{n}} \tag{11}$$

where y_i represents the actual final contract amount; p_i represents the completed cost; and n is the number of data points (projects).

DATA COLLECTION

Data used in this research were gathered from a variety of sources. The project-related data were retrieved from a single source, but the economic data (leading economic indicators) were obtained from several different sources because they were published and stored by different entities.

340 **Project-Related Data**

341 According to the abundance of data availability, this research team decided to use resurfacing
342 project data to evaluate the proposed approach because an abundance of data is critical to a
343 reliable model. Data from resurfacing projects completed between 2006 and 2013 by the Florida
344 Department of Transportation (FDOT) were used in this study. The variables available in the
345 FDOT database include type of work (e.g., interchange, widening, new construction, bridge
346 rehabilitation, maintenance, traffic control projects), letting date, contract type, project
347 engineer's estimate, original contract price, completed cost, change orders, bituminous
348 adjustments, contract days, completed days, project location (county), number of bidders,
349 weather days, contractors' past performance ratings (CPPR), total project length, and number of
350 lanes. This team chose to use the following variables: number of bidders, CPPR, contract days,
351 length of the project, number of lanes, and contract price/low bid. These variables were selected
352 to initiate model training because they are not categorical variables, which undermine model
353 performance and should be avoided in empirical modeling (Flood and Issa 2010).

354 A total of 1249 projects contained resurfacing work, but 1199 of them were resurfacing
355 projects only, among which 925 projects contained information about weather days, total length,
356 contractor's past performance rating, and number of lanes. There were a few projects that did not
357 have data concerning the number of lanes. Eventually, the data mining process resulted in 741
358 projects for further analysis, of which 641 were picked randomly for model selection, and the
359 remaining 100 were allocated for validation. The same datasets were used for the OLS and
360 LASSO model development and evaluation.

361 According to the completed cost, the smallest project amounted to $114,514, and the
362 largest was approximately $21.39 million. Project duration ranged from 40 days to 855 days.

With respect to the number of bidders, the minimum and maximum were 1 and 14, respectively. Project length fell into the range between 0.16 and 36.19 km. The maximum number of lanes and weather days were eight and 280. CPPR scores were between 38 and 110.

Leading Economic Indicators

The majority of highway construction costs comprises labor, material, equipment, overhead, and profit, the prices of which are closely related to the economy. The following leading indicators have a close link to the price of labor, material, equipment, and overhead. Some of the leading indicators can also reflect the state of the macro-economy and construction market competition, which can affect the contractor's profit margin.

The Consumer Price Index (CPI) is calculated using a weighted average on a wide range of representative goods and services and is an indicator of the price fluctuation of consumption commodities. CPI is a good indicator of the price fluctuation of construction material, labor, and equipment. The data were retrieved from the Department of Labor's website.

Construction spending (CS) is an indicator of the total volume of construction work that has been put in place. It is a good indication of the competition in the construction market. The data are published and stored on the website of the Census Bureau.

The prime loan rate (PLR) is the interest rate that banks charge for short-term loans between commercial banks or to companies. The data are stored in the Board of Governors of the Federal Reserve System's website.

The Producer Price Index (PPI) reflects the wholesale prices of manufacturers or producers that make up the U.S. economy. PPI is a complement of CPI because it does not include customer level prices but leaves those to CPI. The data publisher is the Bureau of Labor Statistics.

386 The CPI ranged from 157.60 to 212.37, and highway construction spending spanned from
387 $788,332 million to $1,161,282 million. The PLR was between 0.1% and 5.02%. Regarding PPI,
388 the minimum was 133.7, and the maximum was 222.40.

389 **MODEL DEVELOPMENT AND PERFORMANCE EVALUATION**

390 Two regression models through the OLS and LASSO methods were developed using the 641
391 resurfacing projects. The respective regression coefficients of the models were presented and
392 incorporated into the model equations. The models' predictive performance was then evaluated
393 with the same criteria (MAE, MAPE, and RMSE) using the held-out dataset.

394 **Pearson's Correlation Analysis**

395 Pearson's correlation analysis measures the linear association between two random variables.
396 The coefficient values are between -1 and 1, of which the numerical value represents the strength
397 of the relationship, and the sign indicates the direction of the relation.

398 Checking correlation between every two variables is a standard procedure in developing
399 an OLS regression model (Rawlings et al. 2001). Pearson's correlation analysis was conducted to
400 detect multicollinearity between the independent variables. With highly correlated independent
401 variables, fluctuation of one independent variable affects other correlated independent variables,
402 consequently, the variance of predictions is inflated, which reduces accuracy of predictions. If
403 strong multicollinearity exists, countermeasures (e.g., use LASSO) should be taken to prevent
404 jeopardy of OLS model performance. The correlation coefficients between the selected predictor
405 variables in this research revealed strong correlation among several predictors, which could pose
406 a risk to OLS model performance. The particular highly correlated predictors in pairs were
407 contract duration and contract price (0.66), PLR and CS (0.92), CPI and CS (-0.51), PPI and CPI
408 (0.68), PPI and CS (-0.82), weather days and contract price (0.43), and PLR and PPI (-0.74).

409 The results of Pearson's correlation test indicated strong collinearity between the predictors in
410 this research. This necessitated a close examination of the Variance Inflation Factor (VIF) of
411 each regression coefficient.

412 **OLS Regression Model**

413 The coefficient of determination R^2 is a commonly used measure to evaluate the goodness-of-fit
414 of regression models. Nonetheless, simply relying on R^2 is prone to overfit a model, which
415 undermines model predictive performance (Rawlings et al. 1998). This study initially obtained a
416 series of "best" models according to R^2, BIC, Mallows's C_p, and adjusted R^2. Table 1 provides a
417 list of top four "best" models containing the same number of variables based on these four
418 criteria. However, three best models were then determined only based on the BIC, Mallows' C_p,
419 and adjusted R^2 because these criteria can alleviate the overfitting issue. Fig. 2 provides the plots
420 to show fluctuation of the adjusted R^2, Mallows' Cp, and BIC along with the change of number
421 of predictor variables in "best" models.

422 It is shown that among the three best models, the one with four predictors (contract price,
423 CPI, PLR, and CS) has the lowest BIC (-2722.10) and is considered the optimal of the best
424 models (see Table 1 and Fig. 2). Coincidentally, Mallows' Cp (2.80) and adjusted R^2 (0.9864)
425 determined two best models with the same five predictors (contract days, contract price, CPI,
426 PLR, and CS) (see Table 1 and Fig. 2). This team finally chose to use the former model with four
427 variables determined by the BIC because it is more parsimonious than the other two.
428 Parsimonious model means a model with fewer variables can explain the variation of the
429 dependent variable as good as other models with more variables without compromising
430 goodness-of-fit of the model (Seasholtz and Kowalski 1993). Furthermore, there is no significant

difference in the Mallows' C_p and adjusted R^2 among the three best models. This is a common practice in statistical (Seasholtz and Kowalski 1993).

The final OLS regression model (see Eq.12) included contract price, CPI, CS, and PLR as the significant predictors, and their coefficients were analyzed by t-tests. Table 2 contains the results of the tests. All coefficients of the variables entered into the final model have a p-value less than 5%. Moreover, the reported multiple R^2 was 0.9875, and the adjusted R^2 was 0.9874. The residual standard error was approximately 3.65×10^6. According to the ANOVA test, the F statistic was 1.25×10^4 with a p-value of 2.2×10^{-16}.

$$\text{Completed Cost} = 2807098 + 1.04 \times \text{Contract Price} - 8194.11 \times \text{CPI} - 1.56 \times \text{CS} + 8520.22 \times \text{PLR} \quad (12)$$

To ensure reliable and consistent performance of the selected OLS regression model, its residuals were diagnosed concerning the normality assumption, independence assumption, and the assumption of constant error variance. With regard to the normality assumption, the Shapiro–Wilk test reported test statistic W = 0.8121 and p-value = 2.2×10^{-16}, suggesting that the assumption was violated, but a histogram of the residuals (see Fig. 3) showed an approximately normal distribution. According to the Durbin–Watson test for detection of autocorrelation of the residuals, the test statistic DW was 1.9740, and the p-value was 0.76, indicating that the residuals were independent. The non-constant variance score test (testing for heteroscedasticity) reported that the chi-square test statistic was 789.7906 with a corresponding p-value of 8.94×10^{-174}, suggesting violation of the constant variance assumption. Violation of this assumption can increase the possibility of gaining an erratic estimation of the regression coefficients and reduce the consistency of the model performance. In addition, the square roots of the VIF concerning the predictors were 1.08 for contract price, 1.54 for CPI, 8.64 for CS, and 7.33 for PLR, which

indicates no violation. However, the VIFs for CS and PLR were close to ten, which pose potential risks to model performance. The influential data points were then checked through the test of Cook's distance, and six outliers were identified and eliminated via the Bonferroni test. This study also attempted the Box-Cox method to see whether the data needed transformation. The best lambda was 0.91, chi-square was 110.55, and the corresponding p-value was approximately zero, meaning that no transformation was necessary.

LASSO Regression Model

In Fig. 4, each line represents the trajectory of a coefficient in the LASSO model. It shows the profiles of coefficients of all predictor variables decreasing toward zero as the tuning parameter λ increases from log (λ) = -5 to log (λ) = 14. This indicates that more regularization reduces the number of nonzero regression coefficients—namely, predictors. The numerals on top are degrees of freedom—the number of non-zero coefficients. The coefficients start to dramatically change after increasing log (λ) to five and even greater.

The vertical axis in Fig. 5 represents mean squared errors (MSEs) from the regression's cross-validation procedure, plotted as a function of log(λ), shown along lower horizontal axes. The numbers above the upper horizontal axis are the numbers of predictor variables in the LASSO regression models based on different log(λ) values. The vertical dash lines indicate the number of nonzero coefficients determined by the minimum MSE, where the model provides its best fits to the data. The dots in the center indicate average MSE values for all models resulting from the corresponding λ value, and the vertical bars through the dots show upper to lower MSE values. It is shown that the MSE value increases tremendously after log(λ) becomes greater than 14, and there is only one predictor.

476 Through the cross-validation process, the best tuning parameter λ for the best model was
477 identified as 9378 or log (λ) = 9.15 because it has the lowest MSE (approximately equal to zero)
478 and smallest variance (see Fig. 5). The LASSO procedure selected seven predictors out of the
479 eleven: number of bidders, CPPR, contract days, weather days, contract price, CPI, and CS. The
480 LASSO model is shown as Eq. 13:

481 $$\text{Completed Cost} = 850076.08 - 580.77 \times \text{No. of Bidders} - 126.49 \times \text{CPPR} + 404.49 \times$$
482 $$\text{Contract Days} + 302.17 \times \text{Weather Days} + 1.03 \times \text{Contract price} -$$
483 $$0.15 \times \text{CS} \quad (13)$$

484 **Discussion of the Selected Variables**

485 The OLS and LASSO methods picked different numbers of predictors to enter into the final
486 model. It is noteworthy that both models included certain leading economic indicators, which
487 supports that economic factors affect the completed costs of highway construction projects.

488 The nominal values of the regression coefficients can be interpreted as weights of the
489 predictors, but not for ranking the importance of the predictors because the scale of the predictors
490 varies considerably. Another meaningful aspect of the model equations would be the signs
491 (positive or negative) of the selected predictors, which reflect the relationship between
492 completed cost and each predictor variable. The number of bidders has a negative coefficient,
493 suggesting that the completed cost decreases when the number of bidder increases in the bidding
494 process. In practice, more bidders mean more competition; thus, contractors are forced to lower
495 their tender prices to win the jobs. The negative sign for CPI in the model indicates a negative
496 relation between CPI and the completed cost. The reason is that CPI is a price index of
497 commodities; according to basic economic theory, when product price grows, demand generally
498 declines (Baye and Beil 2006). In construction, when low bid prices go up, fewer projects can be

initiated by the SHAs; as a result, the competition becomes fierce, which induces contractors to reduce the tender price to win the bids. Following the same theory, lower CS means fewer jobs available and higher competition; hence, the low bid price drops, as does completed cost, and vice versa. PLR can leverage construction product supply. When PLR is low, a contractor can loan more money with lower interest rates, which may drive down the product price (low bid/completed cost). Contract days has a positive coefficient, indicating that the completed cost would rise if the contract duration increases. This is probably because the longer the project takes to finish, the more risks could arise, and the greater the cost for labor and pricy equipment sitting on the site. The same principle applies to weather days. With regard to CPPR, higher scores suggest high qualification and capabilities of the contractors. With high qualification and capabilities, they are more likely to finish projects more quickly and at low cost.

Assessment of the Models' Predictive Performance

After the models were selected via the OLS and LASSO approaches, they were tested using a new dataset (the 100 projects that were not used for model training). The observed completed cost, predicted completed cost, and their difference in percentage are exhibited in Table 3.

With regard to the OLS regression model performance, the evaluation criteria were MAE = 176,263, MAPE = 7.60%, and RMSE=259,987. With respect to the LASSO regression model performance, the corresponding values of the criteria were MAE = 174,864, MAPE = 7.10%, and RMSE=269,667. Based on the first two criteria, the LASSO regression model performed better than the OLS regression model. The third criterion triggers a controversy that LASSO underperformed because LASSO model has greater RMSE than OLS model. Nevertheless, this does not disapprove that LASSO is better than OLS as the model developed via OLS violated the statistical assumptions. This mean the OLS model is only valid to show the relationship between

predictor and response variables, and is not reliable to make predictions. There are several factors contributed to this controversy. One explanation is the formula itself, which quadratically exaggerates the deviation between the observed and predicted values. Another cause probably stems from high VIFs of some regression coefficients, which are close to 10. High VIFs possibly have caused inconsistent predictions even though they are below the conventionally accepted threshold. Finally, violation of the assumptions is the most important reason because it results in unstable performance.

CONCLUSIONS

One contribution of this study is that it introduced an alternative tool—the LASSO method—for building a parametric cost estimation model to forecast the completed costs of resurfacing projects in the preconstruction phase. LASSO is developed to deal with multicollinearity, which is common to the empirical data (Chan and Park 2005). In contrast, multicollinearity renders the OLS method ineffective. The second contribution to the construction engineering and management community is the developed model because the model for forecasting the completed costs in this study can serve two purposes (Williams 2003): first, it would be useful to have good knowledge about the completed cost of highway projects for planning purposes. SHAs can use such predictions to better budget their financial resources; second, if abnormally high cost escalation during the construction phase is predicted by the model, necessary measures (e.g., careful supervision) can be taken when executing highway projects.

To compare the LASSO method, the OLS regression was used as a benchmark, and this study demonstrated the rigorous residual diagnostics process because the procedure is a precondition for reliable performance. This is another contribution to the body of knowledge for the construction engineering and management community. In this study the statistical

545 assumptions concerning a constant variance was violated, which is detrimental to consistent
546 predictive performance. Compared to the LASSO method, the OLS procedure is more tedious
547 and requires extensive effort to determine the important predictor variables. In addition, the
548 residuals of the OLS regression model must be carefully examined to ensure that the assumptions
549 are satisfied so that the model can be used to serve long-term purposes. In contrast, the LASSO
550 process is relatively concise and automatic in feature selection. In determining the best model,
551 the OLS approach requires human judgment to choose the predictors for the model. In contrast,
552 the tuning parameter λ automatically determines the best model in the LASSO procedure.
553 Moreover, the LASSO regression model outperformed the OLS regression model in terms of the
554 MAE and MAPE values calculated through the identical validation dataset used for the OLS
555 method. Although the RMSE caused a controversy, this is mainly due to the OLS model's
556 violation of the statistical assumption, high VIFs of some regression coefficients, and the
557 formula itself. Violation of the statistical assumption makes the OLS model unreliable for
558 prediction.

559 Although most professionals in the construction industry acknowledge that the economy
560 has substantial influence on highway project costs, no previous research has analyzed the leading
561 economic indicators simultaneously with project-related factors while developing a parametric
562 cost estimation model. This research explored several leading economic indicators in developing
563 the prediction models using both the conventionally used OLS method and the alternative
564 approach—LASSO regression. Both methods exhibited that some economic indicators are
565 closely related with the completed cost of highway projects.

566 **LIMITATIONS**

The model developed in this study works best for FDOT because the data used for model development were retrieved from the FDOT database. The pattern recognized by the model better reflects the situation in Florida. However, to maximize reliability, other DOTs can copy the procedure to develop another model using their own data. To be useful for types of projects other than resurfacing projects, the modeler must train the model using data for a specific type of project. Other variables may potentially have a significant influence on the completed cost of highway projects as well, but this study was limited by the available data. Regarding the leading economic indicators, more can be explored.

ACKNOWLEDGEMENT

The authors would like to acknowledge FDOT for furnishing the data used for this study and providing clarifications during the process of data analysis. The authors also thank Dr. Hanguk Ryu and Dr. Zezhou Wu for the insightful inputs regarding the structure of the manuscript. It should be noted that the opinions represented here are the authors' and do not reflect opinions of FDOT.

DATA AVAILABILITY STATEMENT

The interested readers of this research can request for project-related data from FDOT. Leading economic indicators are publicly available and the authors can also share the gathered data when requested.

REFERENCES

Anderson, S. D., Molenaar, K. R., Schexnayder, C. J. (2007). *Guidance for Cost Estimation and Management for Highway Projects during Planning, Programming, and Preconstruction,* Transportation Research Board, Washington DC.

589 Asmar, M., Hanna, A., Whited, G. (2011). "New Approach to Developing Conceptual Cost
590 Estimates for Highway Projects." *J. Constr. Eng. Manage.,* 137(11), 942-949.

591 Baye, M. R., and Beil, R. O. (2006). *Managerial Economics and Business Strategy,* McGraw-
592 Hill New York, NY.

593 Byrnes, J. E. (2002). "Best Practices for Highway Project Cost Estimating." *MS Thesis, Arizona*
594 *State University*.

595 Cantarelli, C. C., Flyvbjerg, B., Molin, E. J., Van Wee, B. (2010). "Cost Overruns in Large-Scale
596 Transportation Infrastructure Projects: Explanations and Their Theoretical Embeddedness."
597 *European Journal of Transport Infrastructure Research,* 10(1), 5-18.

598 Chan, S. L., and Park, M. (2005). "Project Cost Estimation Using Principal Component
599 Regression." *Constr. Manage. Econ.,* 23(3), 295-304.

600 Chou, J., Lin, C., Pham, A., Shao, J. (2015). "Optimized Artificial Intelligence Models for
601 Predicting Project Award Price." *Autom. Constr.,* 54, 106-115.

602 Chou, J., Peng, M., Persad, K. R., O'Connor, J. T. (2006). "Quantity-Based Approach to
603 Preliminary Cost Estimates for Highway Projects." *Transportation Research Record:*
604 *Journal of the Transportation Research Board,* 1946(1), 22-30.

605 Chou, J., Wang, L., Chong, W. K., O'Connor, J. T. (2005). "Preliminary Cost Estimates Using
606 Probabilistic Simulation for Highway Bridge Replacement Projects." *Proc., American*
607 *Society of Civil Engineers, Construction Research Congress,* 939-948.

608 Cirilovic, J., Vajdic, N., Mladenovic, G., Queiroz, C. (2014). "Developing Cost Estimation
609 Models for Road Rehabilitation and Reconstruction: Case Study of Projects in Europe and
610 Central Asia." *J. Constr. Eng. Manage.,* 140(3), 04013065.

611 Emhjellen, M., Emhjellen, K., Osmundsen, P. (2001). "Cost Overruns and Cost Estimation in the
612 North Sea." <

https://brage.bibsys.no/xmlui/bitstream/handle/11250/165854/A52_01.pdf?sequence=1&isAllowed=y> (March 11, 2017).

Flood, I., and Issa, R. R. (2010). "Empirical Modeling Methodologies for Construction." *Journal of Computing in Civil Engineering,* 8(2), 131-148.

Flyvbjerg, B., Holm, M. S., Buhl, S. (2002). "Underestimating Costs in Public Works Projects: Error or Lie?" *Journal of the American Planning Association,* 68(3), 279-295.

Fu, W. J. (1998). "Penalized Regressions: The Bridge Versus the Lasso." *Journal of Computational and Graphical Statistics,* 7(3), 397-416.

Herbsman, Z., and Mitrani, J. (1984). "INES-an Interactive Estimating System." *J. Constr. Eng. Manage.,* 110(1), 19-33.

Herbsman, Z. (1983). "Long-Range Forecasting Highway Construction Costs." *J. Constr. Eng. Manage.,* 109(4), 423-434.

Hollar, D. A., Rasdorf, W., Liu, M., Hummer, J. E., Arocho, I., Hsiang, S. M. (2012). "Preliminary Engineering Cost Estimation Model for Bridge Projects." *J. Constr. Eng. Manage.,* 139(9), 1259-1267.

James, G. (2013). *An Introduction to Statistical Learning with Applications in R,* Springer, New York, NY.

Kim, G., An, S., Kang, K. (2004). "Comparison of Construction Cost Estimating Models Based on Regression Analysis, Neural Networks, and Case-Based Reasoning." *Build. Environ.,* 39(10), 1235-1242.

Mahamid, I., and Bruland, A. (2010). "Preliminary Cost Estimating Models for Road Construction Activities." *Proc., the FIG Congress, Sydney, Austrialia*.

635 Mahamid, I. (2011). "Early Cost Estimating for Road Construction Projects Using Multiple
636 Regression Techniques." *Australasian Journal of Construction Economics and Building,*
637 11(4), 87-101.

638 Minchin, R. E., Glagola, C. R., Thakkar, K. V., Santoso, A. (2004). "Maintaining preliminary
639 estimate accuracy in a changing economy." *Proc., American Society of Civil Engineers*
640 *Specialty Conference on Leadership and Management in Construction,* 120-128.

641 Minchin, R. E., Campo, M., Glagola, C.,R., Thakkar, K. (2005). "Managing preliminary
642 estimates in a changing economy." *Proc., The Counseil International Du Batiment,* 70-71.

643 Morris, M. (1990). "Improving the Accuracy of Early Cost-Estimates for Federal Construction
644 Projects." *BRB, National Research Council, Washington, DC, USA*.

645 Odeck, J. (2004). "Cost Overruns in Road Construction—What Are Their Sizes and
646 Determinants?" *Transp. Policy,* 11(1), 43-53.

647 Petroutsatou, C., Lambropoulos, S., Pantouvakis, J. (2006). "Road Tunnel Early Cost Estimates
648 using Multiple Regression Analysis." *Operational Research,* 6(3), 311-322.

649 Rawlings, J. O., Pantula, S. G., Dickey, D. A. (2001). *Applied Regression Analysis: A Research*
650 *Tool,* Second Ed., Springer, Verlag, New York.

651 Rawlings, J. O., Pantula, S. G., Dickey, D. A. (1998). *Applied Regression Analysis: A Research*
652 *Tool,* Springer Science & Business Media, New York.

653 Seasholtz, M. B., and Kowalski, B. (1993). "The Parsimony Principle Applied to Multivariate
654 Calibration." *Analytica Chimica Acta,* 277(2), 165-177.

655 Sonmez, R. (2004). "Conceptual Cost Estimation of Building Projects with Regression Analysis
656 and Neural Networks." *Canadian Journal of Civil Engineering,* 31(4), 677-683.

657 Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal*
658 *Statistical Society. Series B (Methodological),* 267-288.

Turochy, R. E., Hoel, L. A., Doty, R. S. (2001). *Highway Project Cost Estimating Methods used in the Planning Stage of Project Development,* Virginia Transportation Research Council Charlottesville, Virginia.

Wang, Y., and Liu, M. (2012). "Prices of Highway Resurfacing Projects in Economic Downturn: Lessons Learned and Strategies Forward." *J. Manage. Eng.,* 28(4), 391-397.

Williams, T. P. (2003). "Predicting Final Cost for Competitively Bid Construction Projects using Regression Models." *Int. J. Project Manage.,* 21(8), 593-599.

Williams, T. P. (2002). "Predicting Completed Project Cost using Bidding Data." *Constr. Manage. Econ.,* 20(3), 225-235.

Wilmot, C., and Cheng, G. (2003). "Estimating Future Highway Construction Costs." *J. Constr. Eng. Manage.,* 129(3), 272-279.

Wilmot, C. G., and Mei, B. (2005). "Neural Network Modeling of Highway Construction Costs." *J. Constr. Eng. Manage.,* 131(7), 765-771.

Wright, M. G., and Williams, T. P. (2001). "Using Bidding Statistics to Predict Completed Construction Cost." *The Engineering Economist,* 46(2), 114-128.

683 **Table 1**. The Top 4 OLS Models in Terms of R^2, Adjusted R^2, C_p, and BIC

No. of Predictors	(Intercept)	No. Biders	Weather Days	Length	Lane	Contract Days	Contract Price	CPPR	CPI	CS	PLR	PPI	Rsq	AdjR2	Cp	BIC
1	x						x						0.9857	0.9857	32.25	-2709.09
1	x					x							0.4294	0.4285	26043.66	-346.66
1	x		x										0.1828	0.1816	37569.02	-116.51
1	x											x	0.0891	0.0877	41952.81	-46.88
2	x						x		x				0.9860	0.9859	20.82	-2715.62
2	x					x	x						0.9858	0.9858	26.97	-2709.64
2	x		x				x						0.9858	0.9857	30.83	-2705.91
2	x						x	x					0.9857	0.9857	33.30	-2703.54
3	x						x		x	x			0.9862	0.9861	13.82	-2718.02
3	x					x	x		x				0.9861	0.9861	15.11	-2716.74
3	x						x		x			x	0.9861	0.9860	17.97	-2713.91
3	x		x				x		x				0.9860	0.9860	19.93	-2711.99
4	**x**						**x**		**x**	**x**	**x**		0.9864	0.9863	5.27	**-2722.1**
4	x					x	x		x	x			0.9863	0.9862	10.85	-2716.5
4	x					x	x		x			x	0.9862	0.9861	13.26	-2714.1
4	x		x				x		x	x			0.9862	0.9861	14.94	-2712.43
5	**x**					**x**	**x**		**x**	**x**	**x**		0.9865	**0.9864**	**2.80**	-2720.16
5	x		x				x		x	x	x		0.9864	0.9863	5.83	-2717.08
5	x						x		x	x	x	x	0.9864	0.9863	7.01	-2715.89
5	x			x			x		x	x	x		0.9864	0.9863	7.05	-2715.85
6	x		x			x	x		x	x	x		0.9865	0.9864	4.06	-2714.44
6	x					x	x		x	x	x	x	0.9865	0.9864	4.34	-2714.15
6	x				x	x	x		x	x	x		0.9865	0.9864	4.45	-2714.04
6	x			x		x	x		x	x	x		0.9865	0.9864	4.58	-2713.91
7	x		x			x	x		x	x	x	x	0.9865	0.9864	5.30	-2708.75
7	x		x		x	x	x		x	x	x		0.9865	0.9864	5.67	-2708.37
7	x		x	x		x	x		x	x	x		0.9865	0.9864	5.82	-2708.22

7	x	x	x			x	x		x	x	x		0.9865	0.9864	5.84	-2708.2
8	x		x		x	x	x		x	x	x	x	0.9865	0.9864	6.80	-2702.8
8	x	x	x			x	x		x	x	x	x	0.9865	0.9864	6.81	-2702.78
8	x		x	x		x	x		x	x	x	x	0.9865	0.9864	6.99	-2702.6
8	x		x			x	x	x	x	x	x	x	0.9865	0.9864	7.19	-2702.4

684

685 **Table 2.** Summary of Statistical Analysis of The Regression Coefficients

Predictors	Coefficients	Std.Error	t value	Pr(>|t|)
(Intercept)	2.81E+06	4.45E+05	6.30	5.48E-10
Contract price	1.04E+00	4.85E-03	215.29	< 2.00E-16
CPI	-8.19E+03	1.36E+03	-6.04	2.70E-09
CS	-1.56E+00	2.94E-01	-5.29	1.70E-07
PLR	8.52E+04	1.98E+04	4.31	1.94E-05

686

687 **Table3.** OLS and LASSO Regression Model Test Case Predictions

	OLS			LASSO				OLS			LASSO		
Index No.	Observed	Predictions	Deviation (%)	Observed	Predictions	Deviation (%)	Index No.	Observed	Predictions	Deviation (%)	Observed	Predictions	Deviation (%)
16	1397161	1517314	8.60	1397161	1515142	8.44	349	144815	115774	-20.05	144815	108584	-25.02
29	1933489	2038165	5.41	1933489	1997367	3.30	353	6276857	5960793	-5.04	6276857	5987706	-4.61
46	2987231	2910715	-2.56	2987231	2851286	-4.55	354	332920	304915	-8.41	332920	307251	-7.71
49	835052	873720	4.63	835052	828863	-0.74	398	984593	926317	-5.92	984593	991426	0.69
53	894870	690685	-22.82	894870	816081	-8.80	399	5423581	5344113	-1.47	5423581	5396936	-0.49
54	3280622	3577893	9.06	3280622	3679722	12.17	430	8388367	8034070	-4.22	8388367	7994480	-4.70
60	2686469	2909822	8.31	2686469	2994895	11.48	432	10156994	10008640	-1.46	10156994	9985175	-1.69
65	1002302	685208	-31.64	1002302	807661	-19.42	435	4277435	4368784	2.14	4277435	4383953	2.49
81	2090914	2146527	2.66	2090914	2140801	2.39	440	3279016	3319865	1.25	3279016	3312512	1.02

104	508663	406590	-20.07	508663	388276	-23.67	458	2881682	3077193	6.78	2881682	3214980	11.57
109	1202015	1115978	-7.16	1202015	1102714	-8.26	459	3224059	3411969	5.83	3224059	3583206	11.14
119	1268538	1304636	2.85	1268538	1266822	-0.14	461	8965798	9592862	6.99	8965798	9695291	8.14
124	3324313	3238962	-2.57	3324313	3158878	-4.98	474	10521895	10101424	-4.00	10521895	10213532	-2.93
128	1589316	1536596	-3.32	1589316	1472182	-7.37	476	1140865	1146225	0.47	1140865	1157148	1.43
132	1468991	1352147	-7.95	1468991	1313495	-10.59	490	893033	923810	3.45	893033	946194	5.95
133	3906882	4098989	4.92	3906882	4069189	4.15	502	3289177	3348593	1.81	3289177	3445619	4.76
154	4997942	5774770	15.54	4997942	5773734	15.52	504	903090	887686	-1.71	903090	881921	-2.34
155	6014938	6625667	10.15	6014938	6543651	8.79	512	2767472	3006086	8.62	2767472	3025214	9.31
167	2042399	2119787	3.79	2042399	2058607	0.79	531	1732633	1908306	10.14	1732633	1880269	8.52
173	1475380	1558384	5.63	1475380	1510313	2.37	533	252095	254870	1.10	252095	187070	-25.79
179	2569687	2627736	2.26	2569687	2595249	0.99	548	1796843	1989659	10.73	1796843	1972380	9.77
185	6445523	6181331	-4.10	6445523	6102194	-5.33	555	1915700	1872548	-2.25	1915700	1904927	-0.56
191	3096772	3189602	3.00	3096772	3324119	7.34	556	2666319	2578784	-3.28	2666319	2597458	-2.58
197	2399967	2305087	-3.95	2399967	2447067	1.96	561	1843976	1995439	8.21	1843976	1990711	7.96
201	3927052	3401036	-13.39	3927052	3528873	-10.14	562	2505391	2724048	8.73	2505391	2733679	9.11
202	1231709	977827	-20.61	1231709	1086425	-11.80	575	4907937	5228261	6.53	4907937	5160059	5.14
205	2749183	2796608	1.73	2749183	2803935	1.99	582	1905546	2071261	8.70	1905546	2044585	7.30
210	5524054	5462336	-1.12	5524054	5488386	-0.65	592	2482615	1862035	-25.00	2482615	1987589	-19.94
212	389209	347222	-10.79	389209	335813	-13.72	600	373683	203710	-45.49	373683	328636	-12.05
213	1073402	1070064	-0.31	1073402	1064603	-0.82	606	1600971	1834511	14.59	1600971	1945436	21.52
219	5131124	5227855	1.89	5131124	5207543	1.49	607	1736517	1642708	-5.40	1736517	1765591	1.67
229	851352	819388	-3.75	851352	802729	-5.71	629	4211601	4198652	-0.31	4211601	4194828	-0.40
238	2301763	2083216	-9.49	2301763	2094816	-8.99	632	3666142	3601719	-1.76	3666142	3586052	-2.18
243	1759100	1642811	-6.61	1759100	1623227	-7.72	651	2356564	2187832	-7.16	2356564	2118292	-10.11
245	9265416	9532442	2.88	9265416	9449526	1.99	656	2967618	3010613	1.45	2967618	3036947	2.34
247	3034822	3020123	-0.48	3034822	3017808	-0.56	658	1742875	1807648	3.72	1742875	1815668	4.18
250	3902631	3865810	-0.94	3902631	3862563	-1.03	664	2377151	2144795	-9.77	2377151	2141490	-9.91
256	13555829	13161556	-2.91	13555829	13078216	-3.52	672	1916203	2044765	6.71	1916203	2201199	14.87
262	1466138	1484660	1.26	1466138	1405862	-4.11	673	4759691	4502230	-5.41	4759691	4667986	-1.93
271	1773283	1668748	-5.89	1773283	1619667	-8.66	674	972432	807670	-16.94	972432	959692	-1.31

277	1591464	1683754	5.80	1591464	1633802	2.66	686	2462796	2516661	2.19	2462796	2521648	2.39
283	765641	926936	21.07	765641	888850	16.09	691	810344	756416	-6.65	810344	780947	-3.63
293	974513	947205	-2.80	974513	959757	-1.51	695	7528736	7663384	1.79	7528736	7680025	2.01
300	4539364	4169743	-8.14	4539364	4193117	-7.63	702	2792481	2866456	2.65	2792481	2897548	3.76
302	5610608	5330879	-4.99	5610608	5222405	-6.92	706	1184190	1178927	-0.44	1184190	1176244	-0.67
303	758839	977506	28.82	758839	824831	8.70	709	324263	202295	-37.61	324263	182903	-43.59
322	15971631	16419215	2.80	15971631	16308476	2.11	713	2984278	2918396	-2.21	2984278	2932284	-1.74
324	7739275	6733936	-12.99	7739275	6658517	-13.96	714	2353145	2263126	-3.83	2353145	2279435	-3.13
330	2626291	2501797	-4.74	2626291	2466693	-6.08	730	3592439	3744742	4.24	3592439	3696364	2.89
333	3974463	5000418	25.81	3974463	5186649	30.50	735	1707340	1889858	10.69	1707340	1852497	8.50