Preliminary engineering (PE) for a highway project encompasses two efforts: planning to minimize the physical, social, and human environmental impacts of projects and engineering design to deliver the best solution. PE efforts begin years in advance of the project’s actual construction operations, often five years or more. An efficient and accurate method to estimate PE costs would benefit transportation departments by facilitating funding allocation projections. Lacking an effective tool to estimate PE cost based on project characteristics, departments of transportation typically estimate PE costs as a fixed percentage of estimated construction costs disregarding other project-specific parameters.

By analyzing 505 North Carolina Department of Transportation (NCDOT) bridge projects awarded for construction from 1999 through 2008, a multiple linear regression model was developed to link variation in PE costs to a set of distinctive project data. Published bid summaries, bridge inventory and assessment reports, NCDOT’s project management system, and published environmental reports were used as sources of project data. The model explained approximately 60 percent of the PE cost variation between projects. Results indicate that right-of-way costs, regional location, and scope delineators are among the project-specific parameters that most influence PE costs of bridge projects. By considering numerous parameters (expressed as independent variables) a more accurate prediction for future projects’ PE costs can be developed.

INTRODUCTION

For this study, preliminary engineering (PE) is defined as the planning and design of a highway project for construction. PE begins when a specific highway project first receives funding authorization for planning and/or design activities. The delivery of the construction documents used for solicitation of construction contract bids (known as project letting) marks the end of PE. Consistent with other investigators’ definitions, PE in this study does not include right of way (ROW) acquisition or construction activities [Turochy et al. 2001; WSDOT 2002].

In stark contrast to the amount of research aimed at improving construction estimates, minimal research directed at PE estimation, especially in the transportation field, exists. Investigators Knight and Fayek (2002) noted the lack of predictive tools to estimate design costs when studying preconstruction project management.

The need for higher quality and reliable PE estimates for transportation infrastructure projects is critical. With better PE estimates, funding allocations can be proactive and more closely matched to the specific needs of each project. Such procedural enhancements promote fiscal
responsibility through improved budgeting and accounting accuracy. PE costs comprise a significant portion (on the order of 10 percent) of total project costs, but current budgeting processes often do not consider the unique characteristics of each project.

PE estimates are frequently based on estimated project construction costs. As part of a comparative analysis of construction costs, Washington State DOT [WSDOT 2002] collected information from twenty-five DOTs whose members served on the AASHTO Subcommittee on Design. Survey participants were asked to identify their typical project PE cost as a percentage of construction cost. PE was defined as “the work that goes into preparing a project for construction.” The average PE cost among respondents was 10.3 percent of construction costs and the range of costs reported was between 4 and 20 percent. This study indicates that continuing to estimate PE costs using a fixed percentage method is inefficient over the project cycle. On a project by project basis, this results in under-allocation or over-allocation of PE funding, which, in turn, necessitates management actions to redistribute PE funds. Avoiding such redistributions improves total project cost control and aids in reducing financial risk for current and future projects.

The Virginia Transportation Research Council (VTRC) assisted Virginia DOT (VDOT) during 2004 to find and implement a construction estimating tool. The estimating tool selected for statewide implementation was based on an existing spreadsheet application developed by the Fredericksburg District of VDOT. With this tool, PE costs can be estimated separately for roadways and bridges. If necessary, they can then be combined to provide a total PE estimate. For roadways, a cost curve, relating PE costs to construction costs, was derived using data from 30 completed VDOT roadway projects. The resulting ratio of PE costs to construction costs ranged from 8 to 20 percent. To verify that the tool’s PE cost curve was applicable for statewide use, an additional 135 completed VDOT roadway projects were analyzed and a modified PE cost curve was derived. For bridges, a similar PE cost curve was derived and confirmed using data from 23 completed bridge projects [Kyte et al. 2004a, 2004b].

Regression techniques have been used to predict construction-related costs. When more than one independent variable is used to predict the response variable, the technique is termed multiple regression. If the relationship between the independent variables and the response variable is assumed to be linear, the technique is multiple linear regression (MLR). Nonlinear regression techniques do not rely on a linear relationship between variables.

Lowe et al. (2006) utilized a multiple linear regression technique to predict the building construction costs using data from 286 buildings constructed in the United Kingdom. A predictive tool was desired that could be used during the early stages of construction cost estimation before the detailed design has been completed. Lowe et al. identified 41 input variables for use in the regression model. The input variables were categorized as either strategic, site related, or design. Lowe et al. concluded that the linear drivers of cost were predominantly design specific. The coefficient of determination ($R^2$) and mean absolute percentage error (MAPE) was used to judge model performance. Lowe et al. reported that their model yielded a $R^2$ of 0.928 and a MAPE of 19.3 percent for predicting the cost of building construction. However, a closer review of prediction error highlighted underestimation of very expensive projects and overestimation of very inexpensive projects [Lowe et al. 2006].
Odeck (2003) used nonlinear multiple regression to identify project factors associated with construction cost overruns for 620 Norwegian road projects. Odeck sought to determine if the impact on cost overrun depended on the magnitude of project cost, project delay, and project duration. Odeck’s regression model only explained about 20 percent of the variation in cost overruns (adjusted $R^2$ value of 0.21). Other project factors, not identified in the regression model, influenced the variation in cost overruns. From his model’s partial regression coefficients, Odeck concluded that estimated cost overruns decreased with increased project costs, increased with increased project duration up to a point and then decreased, and varied with geographic region. Odeck concluded that cost overruns were more predominate among smaller road projects in Norway [Odeck 2003].

**MODEL DEVELOPMENT PLAN**

This paper presents the development of a multiple linear regression model, consisting of four stages. A brief description of each stage follows.

1. Seek descriptive project data to populate predictive variables.
2. Apply statistical analyses to filter predictive variables.
3. Develop a multiple linear regression model using significant predictor variables.
4. Test and validate the model for predictive purposes.

To obtain data, information was requested from NCDOT for bridge projects let for construction. Requested data included project descriptive data, cost estimates, and actual cost expenditures. The initial acquisition strategy sought out as many electronic data sources as possible. The project identification number established in the State Transportation Improvement Plan (STIP) served as the key field linking all data sources and identifying all projects. Preconstruction project data is housed in several independent databases maintained by NCDOT units. Complete preconstruction data does not fully migrate to the construction database maintained after construction contracts are awarded. The construction database proved more accessible electronically, due to online posting of construction data on NCDOT’s public webpage.

Collecting project data from multiple sources produced redundant data. For example, one data source recorded a project’s length in feet while another source tabulated the same length in meters. Redundant fields were filtered from the compiled database. Some data were found to be highly correlated; the value of one data field changes in concert with changes to the value of another data field. As an example, the construction contract for bridge projects is commonly divided into a structure portion (bridge or culvert) and a roadway portion. The cost percentages reported for each contract portion were found to be highly correlated. As the percentage of the construction contract for structures increased, the percentage allocated for roadway decreased. A statistical assessment of candidate variables was performed using correlation and sensitivity analyses. For highly correlated variables, only one variable was selected for inclusion in model development. After reviewing all candidate variables, 28 independent variables describing project-specific parameters were used in model development. These 28 variables are listed in Table 1.
Data from 505 bridge projects were used to build a multiple linear regression (MLR) model. Scatter and box plots were used to visualize the relationship between the 28 predictor variables and PE percentage (the response variable). Through iterative variable selection techniques (using SAS statistical software), a MLR model was fit using the adjusted $R^2$ value as fit criteria. The adjusted $R^2$ value expresses how much of the response variation is captured by the model. It is preferred over $R^2$ alone because it is adjusted for the number of variables included. Whereas $R^2$ continues to increase with each additional variable added to the model, adjusted $R^2$ identifies the fewest variables needed to optimize fit, yielding a parsimonious model. A higher adjusted $R^2$ value indicates a better model fit. Of the 28 predictor variables considered, two variations of the MLR model was developed, the first containing 8 predictor variables, and the second containing 14 variables. Selection of the final MLR model will be dependent on the model’s prediction performance as described below.

The two variations of the MLR model will be applied to 121 bridge projects not used for model building. Each project’s predicted PE percentage will be compared to its recorded PE percentage. This comparison will yield the prediction error. The mean prediction error indicates bias in the model’s prediction capability. Bias is a model’s tendency to systematically under or over predict PE percentage. The prediction error may be positive (overestimated PE percentage) or negative (underestimated PE percentage). Squaring the prediction errors provides a measure of absolute error. The mean of the squared prediction errors quantifies prediction precision. Confidence intervals for both prediction bias and precision will be computed. Predictive capability will be assessed by reviewing the bias and precision. A model with lower values in both bias and precision is preferred over another candidate model [Sheiner and Beal 1981].

DATA COLLECTION AND ANALYSIS

This section describes the strategy employed to develop a multiple linear regression model to predict the PE percentage for bridge projects. Figure 1 summarizes the strategy graphically. The tasks labeled A, B, and C are complete and are reported on herein. The initial performance of Task D is also complete, but model revisions are anticipated after the validation results of Task E have been completed.

Data Sources Utilized for Data Acquisition (Figure 1, Task A)

Ten data sources were investigated and project data were successfully extracted from eight of the ten sources. The sources consisted of the following:

1. NCDOT Online Bid Tabulations & Annual Bid Averages Summary
2. NCDOT Pre-2002 Project Management Data System (obsolete mainframe system)
3. NCDOT Post-2002 Project Management Data System (SAP based)
4. NCDOT 12-Month Projected Letting List
5. NCDOT National Bridge Inventory System Data (NBIS)
6. NCDOT State Transportation Improvement Plan (STIP)
7. North Carolina State Publications Clearinghouse
8. NCDOT Transpot Program Modification - Project Type Coding
Project data were not obtained from online construction plans (source 9) because plans were not available for all 505 bridge projects. Online plans became available starting with projects let in August 2004. Projects dating back to January 1999 were used in model building. Online Board of Transportation meeting agendas and minutes (source 10) were available in a PDF format. However, the PDF document format did not support efficient searching (utilizing multiple criteria) to extract specific project data.

Project data from the remaining eight sources were grouped by data function: classification, cost, date, design, dimensional, environmental, and geographical. Table 1 lists the 28 independent variables (grouped by category) identified for each bridge project.

Values for the 28 independent variables shown in Table 1 were acquired from the first eight sources mentioned above for 505 North Carolina bridge projects let for construction between January 1, 1999 and June 30, 2008.

![Figure 1. Regression Modeling Flowchart](image-url)
Table 1. Independent Variables (28) for Bridge Data Analysis

<table>
<thead>
<tr>
<th>Category</th>
<th>Independent Variable</th>
<th>Variable Levels or Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>New Location, Replacement, Culvert</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Lanes on Bridge</td>
<td>Numerical Count</td>
</tr>
<tr>
<td></td>
<td>Type of Service on Bridge</td>
<td>Highway, Railroad, Pedestrian</td>
</tr>
<tr>
<td></td>
<td>Route Signing Prefix</td>
<td>Interstate, US Hwy, State Hwy</td>
</tr>
<tr>
<td></td>
<td>Capacity Rating of Live Load</td>
<td>Metric Tons</td>
</tr>
<tr>
<td></td>
<td>Road System</td>
<td>Arterial, Collector, Local</td>
</tr>
<tr>
<td></td>
<td>Structure Type</td>
<td>Bridge or Culvert</td>
</tr>
<tr>
<td>Cost</td>
<td>ROW Cost to STIP Estimated Construction Cost</td>
<td>Cost Ratio</td>
</tr>
<tr>
<td></td>
<td>Roadway Percentage of Construction Cost</td>
<td>Cost Ratio</td>
</tr>
<tr>
<td></td>
<td>STIP Estimated Construction Cost</td>
<td>Cost in Dollars ($)</td>
</tr>
<tr>
<td>Date</td>
<td>Year of Letting</td>
<td>Calendar Year</td>
</tr>
<tr>
<td></td>
<td>Year of Environmental Document Approval</td>
<td>Calendar Year</td>
</tr>
<tr>
<td></td>
<td>PE Duration After Environmental Doc</td>
<td>Days</td>
</tr>
<tr>
<td>Design</td>
<td>Deck Structure Type</td>
<td>Concrete, Steel, Aluminum, Wood</td>
</tr>
<tr>
<td></td>
<td>Design Live Load</td>
<td>M9, M13.5, MS13.5, M18</td>
</tr>
<tr>
<td></td>
<td>Main Span Structure Type</td>
<td>Concrete, Steel, Wood, Masonry</td>
</tr>
<tr>
<td></td>
<td>Design Type</td>
<td>Slab, Girder, Box Beam, Truss</td>
</tr>
<tr>
<td>Dimensional</td>
<td>Project Length</td>
<td>Feet</td>
</tr>
<tr>
<td></td>
<td>Bypass Detour Length</td>
<td>Kilometers</td>
</tr>
<tr>
<td></td>
<td>Number of Spans in Main Unit</td>
<td>Numerical Count</td>
</tr>
<tr>
<td></td>
<td>Horizontal Clearance for Loads</td>
<td>Meters</td>
</tr>
<tr>
<td></td>
<td>Length of Structure</td>
<td>Meters</td>
</tr>
<tr>
<td></td>
<td>Water Depth</td>
<td>Feet</td>
</tr>
<tr>
<td>Environmental</td>
<td>NEPA Document Classification</td>
<td>EIS, EA, CE, PCE, Min Criteria</td>
</tr>
<tr>
<td></td>
<td>Planning Document Responsible Party</td>
<td>NCDOT or PEF</td>
</tr>
<tr>
<td>Geographical</td>
<td>NCDOT Division</td>
<td>DIV 01 through DIV 14</td>
</tr>
<tr>
<td></td>
<td>Geographical Area of State</td>
<td>Coast, Piedmont, Mountains</td>
</tr>
<tr>
<td></td>
<td>Classification of Route</td>
<td>Rural or Urban</td>
</tr>
</tbody>
</table>

Database Adjustments for the Preferred Response Variable (Figure 1, Task B)

For the 505 bridge projects database, a ratio of PE costs to construction costs was desired as the response (dependent) variable. Using a cost ratio rather than actual cost values allows comparisons to be made among projects having differing construction costs. The ratio of actual PE cost to the estimated STIP construction cost was tabulated for all 505 projects. This ratio is referred to as the project’s PE Ratio.

For the 505 bridge projects, identified above, the distribution of PE Ratio values is shown in Figure 2. The PE Ratio distribution exhibits a left-skewed, nonnormal shape. A normal
distribution would have a center interval as the peak with decreasing intervals symmetrically located both left and right of the peak.

Figure 2. Distribution of Response Variable “PE Ratio”

Power transformations were applied to the response variable, PE Ratio, to improve normality. The response variable was raised to an exponential power resulting in a transformed variable. Applying the Box-Cox statistical procedure to the nonnormal response distribution identified the optimal normalized distribution as the cube root of PE Ratio, $(PE \text{ Ratio})^{1/3}$. Figure 3 shows the distribution for the transformed response variable, Cube Root of PE Ratio; the distribution is normal. Subsequent regression analyses use Cube Root of PE Ratio as the response (dependent) variable.

Normality of the response variable is sought to satisfy multiple linear regression assumptions. Those assumptions are that regression errors are independent, exhibit a constant variance, and be normally distributed. The requirement that error is normally distributed is interpreted to mean that the response variable is normally distributed at all values of the predictor variables. Thus, a normal distribution of the response variable is desired.

Sampling for Validation (Figure 1, Task C)

Models created for prediction purposes require a validation set of data to assess the accuracy of their prediction. The validation set is not used for model building. A set of 75 projects were randomly selected from the 505 bridge database projects to serve as a within-sample validation
set. This set size is 15 percent of the bridge database. The remaining 430 bridge projects were used for model building.

A prediction for a within-sample project is statistically stronger than an out-of-sample project prediction. When using within-sample projects, the full range of values for independent variables have been included in model building. The same is not true with out-of-sample projects. Using past projects (a historically based model) to predict future project performance is considered an out-of-sample prediction. There is no way to know if a future project’s actual independent variable values will fall within the range of historical values used in model building. The intended use of the model is primarily for future project predictions. Out-of-sample validation projects were included to strengthen the validation process.

An additional out-of-sample validation set was created from 46 bridge projects not included in the original 505 bridge projects database. These 46 projects satisfy three criteria:

- Projects let for construction since July 1, 2008
- Projects included in NCDOT’s Structure Inventory and Assessment (NBIS) database
- Projects with identified PE costs

The within-sample projects (75) and out-of-sample projects (46) were combined to form a complete validation data set containing 121 bridge projects. Thus, the model was built using 430 bridge projects and will be validated using 121 projects.

Figure 3. Distribution of Transformed Response Variable “Cube Root of PE Ratio”
Model Selection Techniques (Figure 1, Task D)

Simple linear regression (SLR) offers a mechanism to determine the linear relationship between one independent variable and the response (dependent) variable. However, SLR has only limited use because all of the other independent variables and the interrelationships between them are ignored. Multiple linear regression (MLR) overcomes the limitations of SLR. However, selecting the “best” model using MLR can be difficult if there are a large number of independent variables. Common model selection techniques involve forward, backward, and stepwise variable selection methods. To assist in model selection, the GLMSLECT procedure within SAS statistical software was utilized. In addition to forward, backward, and stepwise, GLMSELECT provides two additional variable selection methods: least angle regression (LAR) and least absolute shrinkage and selector operator (LASSO). The GLMSELECT procedure provides an efficient starting point for model selection. Model refinement can then follow using intuitive insights gained from data familiarity. [Cohen 2006]

FINDINGS

Using the GLMSELECT procedure with SAS statistical software, two candidate models containing eight and fourteen independent variables were selected. Table 2 identifies the independent variables in each model.

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Model 1 8 Variables Selected</th>
<th>Model 2 14 Variables Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route Signing Prefix</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>ROW Cost to STIP Estimated Construction Cost</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Roadway Percentage of Construction Cost</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>STIP Estimated Construction Cost</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Year of Letting</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Year of Environmental Document Approval</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Project Length</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Bypass Detour Length</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Number of Spans in Main Unit</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Horizontal Clearance for Loads</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Planning Document Responsible Party</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>NCDOT Division</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Geographical Area of State</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Classification of Route</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Coefficient of Determination (R²)</td>
<td>0.65</td>
<td>0.72</td>
</tr>
</tbody>
</table>

For each candidate model, R² expresses the effectiveness of the model in explaining the variation in the response variable, Cube Root of PE Ratio. The R² values are shown in the bottom row of Table 2.
CONCLUSIONS
Although the validation process has yet to be completed, the resulting $R^2$ values obtained indicate that regression modeling has potential as a prediction tool for PE costs. By considering numerous individual project factors (expressed as independent variables) a more accurate prediction for project PE costs can be developed. Multiple linear regression modeling shows promise as a tool to support improvement in PE estimate preparation and cost budgeting. With more accurate PE costs, combined with ROW and construction costs, state transportation agencies can better allocate funding resources to capital projects.

CONTINUING WORK
The regression model described has yet to be validated as illustrated in Figure 1, Task E. Once validation is completed, using the 121 bridge projects comprising the validation set, the model can be finalized (Task F). The optimal model will maximize the coefficient of determination values, $R^2$ and Adjusted $R^2$, while minimizing prediction error expressed by mean square error (MSE). After finalization, a user-friendly interface could be developed to provide a software tool to readily determine PE costs. In the meantime, additional work should be done with the model. Key remaining questions are what are the fewest number of parameters that are needed to obtain a good $R^2$ value, and thus, a good prediction and what is the optimum number of parameters?

Advanced statistical techniques such artificial neural networks and fuzzy logic can be used to overcome the assumptions that are necessary when using multiple linear regression, most notably that the predictors and response be linearly related. Other potential advantages include the ability to address complex interactions between predictors, and modeling the subjective nature of predictors when deterministic data are unavailable. We are continuing to study whether or not use of such alternate techniques yields an improved prediction model for PE costs of bridge projects.

ACKNOWLEDGEMENTS
The authors acknowledge the project support provided by the NCDOT Research and Development Unit. Key NCDOT units providing guidance include the Program Development Branch (Calvin Leggett and Majed Al-Ghandour), the Structure Inventory and Appraisal Unit (Cary Clemmons), and the Schedule Management Office (Kim So, Rose Simson, and Anna Twohig). Many additional NCDOT personnel have provided suggestions and insights. Their contributions continue to influence this research effort in a positive manner. The authors also thank the Southeastern Transportation Center whose early support was instrumental in transitioning this research topic into a funded project.

The contents of this paper reflect the views of the authors and not necessarily the views of NCDOT, North Carolina State University, or any other institution. The authors are responsible for the accuracy of the data and findings.
REFERENCES


AUTHOR CONTACT INFORMATION

Donna A. Hollar
Teaching Instructor
East Carolina University
Department of Construction Management
Rawl Building, Room 329
Greenville, NC  27858-4353
Phone 252-328-6968
Fax 252-328-1165
Email hollard@ecu.edu

Ingrid Arocho
PhD Student
North Carolina State University
Department of Civil Engineering
Campus Box 7908
Raleigh, NC 27695-7908
Fax 919-515-7908
Email idarocho@ncsu.edu

Joseph Hummer
Professor
North Carolina State University
Department of Civil Engineering
Campus Box 7908
Raleigh, NC 27695-7908
Phone 919-515-7733
Fax 919-515-7908
Email hummer@ncsu.edu

Min Liu
Assistant Professor
North Carolina State University
Department of Civil Engineering
Campus Box 7908
Raleigh, NC 27695-7908
Phone 919-513-7920
Fax 919-515-7908
Email min_liu@ncsu.edu

William Rasdorf
Professor
North Carolina State University
Department of Civil Engineering
Campus Box 7908
Raleigh, NC 27695-7908
Phone 919-515-7637
Fax 919-515-7908
Email rasdorf@ncsu.edu