Statistical Analysis and Predictive Models for Expenditures in New York Municipalities

October 08, 2020

Executive Summary:

Introduction: Generally, construction companies have numerous aspects in estimating the cost of each new housing project. To estimate the cost of each housing project, expenditures play an important role in increasing or decreasing the cost. For example, higher expenditure would result in an increase in cost of construction. Therefore, the property owners would have to seek for higher funding to fulfill the project. On the other hand, while expenditure decreases, properties owner could spend the reimburse the expenses elsewhere. In addition, knowing the expenditures would also help construction manager to order supplies in a proper manger. If expenditure decreases, then the supplies would also be less in quantity or cheaper in quality. Numerous questions were proposed in favor of these issues such as 1) What variables causes the fluctuation of expenditure? 2) What is the best predictive model that could predict expenditures? 3) How can we validate and implement the model? 3) How accurate is the model? 4) Is there any improvement to the future models? To answer these questions, this analysis will take a deep dive into the data exploratory analysis, model development process using linear regression, and diagnostics analysis. With the answered questions, construction workers and properties owner would have a better understanding of their expenditures when starting a new project to avoid over or underestimating their budgets.

Methods: A dataset from two New York municipalities (Warwick and Monroe) were provided to access the important measures to predict expenditures. These data contain a total of 916 observations from 1992 with 2 observation contains missing expenditure value. Two observation with NA expenditures have been removed from the analysis to improve the assumption of linear regression modeling. In terms of variables, this dataset contains three identifiers including identity number, state code, and county code and six demographic and income-realted variables including wealth per person, population, percent intergovernmental, density, mean income per person, and growth rate.There is a total of 57 distinct county code implying there are multiple measurement of expenditure per county in New York. The goal of this data analysis is to predict the chances in expenditures of two New York municipalities, Warwick and Monroe. A projection dataset for Warwick and Monroe was also provided to generate predictions from using the fitted model. To achieve this goal, all analysis will be done using multiple linear regression models in R Studio with R version 3.6.2.

Exploratory Data Analysis: Table 1 shows the summary statistics of all independent variables and target varable.

Statistical Analysis:

Conclusion:

Appendix A: Supplemental Tables and Figures

Table 1: Summary Statistics for all numerical independent features

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
expen	914	293.818	269.678	53	172	316	3,286
wealth	914	51,837.720	55,994.250	7,744	25,745.2	54,224.8	594,758
pop	914	7,090.270	26,417.210	69	1,258.8	4,816.8	471,283
pint	914	19.231	10.225	1.700	12.400	23.975	68.600
dens	914	189.495	534.188	1	30	111	6,252
income	914	12,724.960	4,250.423	2,884	10,336.8	13,867.5	48,021
growr	914	8.100	17.434	−54.100	−0.300	13.700	294.500
lexpen	914	5.491	0.558	3.970	5.147	5.756	8.097
lwealth	914	10.599	0.627	8.955	10.156	10.901	13.296
lpop	914	7.876	1.143	4.234	7.138	8.480	13.063
lpint	914	2.826	0.522	0.531	2.518	3.177	4.228
ldens	914	4.141	1.296	0.000	3.401	4.710	8.741
lincome	914	9.409	0.278	7.967	9.243	9.537	10.779
lgrowr	914	1.264	1.925	−4.009	−0.270	2.689	5.689