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, property's 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-related 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 for model development process and accompany by diagnostics process to check for the quality of the model. All analysis including coding and writing report is done in R Studio with R version 3.6.2.

Exploratory Data Analysis

During the exploratory analysis process, it is important to access all the significant relationship of each variable with the target variable. Initially, looking at Table 2, the summary statistics

of all independent variables and target variable shows the maximum for expenditures, wealth, population, pint, density, income, and growth rate are extremely high compare to their mean and 75 percentiles. This indicates that all the variables mention previous are heavily right skewed. Most importantly, expenditure's skewness violates the normality assumption when generating a linear regression. Therefore, a log transformation was applied to expenditure to normalize the distribution of the target variable. Log-transformation would reduce the values which would account for outliers. Figure 1.1. depicts the normality of the outcome variable expenditure after transformation implying the assumption is not violated.

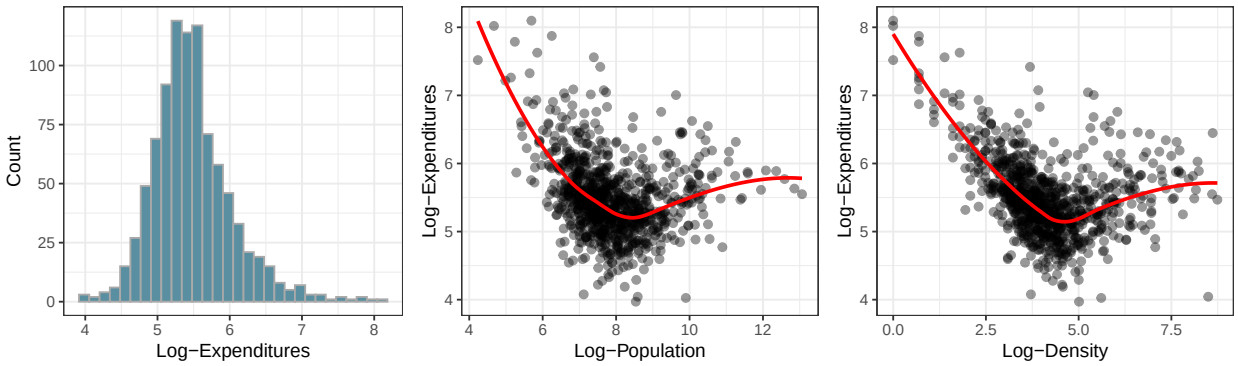

Figure 1: Initial Exploratory Plots after Log-Transformation 1: Histogram of Log-Expenditures with Frequency, 2: Scatter plot of Log-Expenditures vs Log-Population with LOESS smooth line in red, and 3: Scatterplot of Log-Expenditures vs Log-Density with LOESS smooth line in red.

Since all independent variable are right-skewed, a log-transformation also applied to each variable to ensure linear relationship with expenditures. An amount of 1.01 was added to all growth rate values to account for zeros values while taking logarithm. All variables mentioned in the rest of this article are log-transformed variables unless otherwise specified. With that being said, wealth, intergovernmental funds, income, and grow rate seems to have linear relationship with expenditures. However, population and density have different two different trend of expenditures within their plots. Figure 1.2. shows a scatterplot of expenditures and population with a dip at approximately 8.3 to change direction of correlation. Population is self-explanatory variable which represent the number of people living in the county during the year. While population is less than 8.3, as population increases, expenditures decrease on average. When population is greater than 8.3, expenditures increase as population increases. Similar issue happens to density at 4.5, see Figure 1.3. Density, here, represent the population of other substances like animals, environment, or other objects. To account for this problem, the data set of New York city will be subsetted into different groups according to each trend. Subsetting data will help the relationship between density and population and expenditures be linear. This analysis will only model data when population is greater than 8.3 and density is greater than 4.5 since the projection data is within these ranges. After selecting a subset of data, only 228 observations are left in the data. A second round of data exploratory was conducted to ensure the relationship of each measure if significant to the outcome variable expenditures. Expenditures and wealth have a positive relationship indicating the increases of wealth would cause expenditure to be higher, see Figure 2.1. Intuitively, this makes sense since wealthier individuals would spend more resulting in higher expenses. Similarly, Figure 2.2. shows a strong positive correlation between expenditure and income. This relationship is expected since the mean income per person is higher, their expenses would also be higher compared to lower

income individuals. Other variables like population and density seem to have moderate positive relationship with expenditure, see the first two plots in Figure 3. Clearly, population and density are important measurements to predict expenditures. As population and density increases, the amount of expenses also increases, on average. On the other hand, predictors including intergovernmental funds and growth rate have moderate negative correlation with expenditures. This indicate that, while intergovernmental funds and growth rate increases, the amount of expenses should decrease. This makes perfect sense since if the growth rate in economic is slow, then there would be lower expenses.

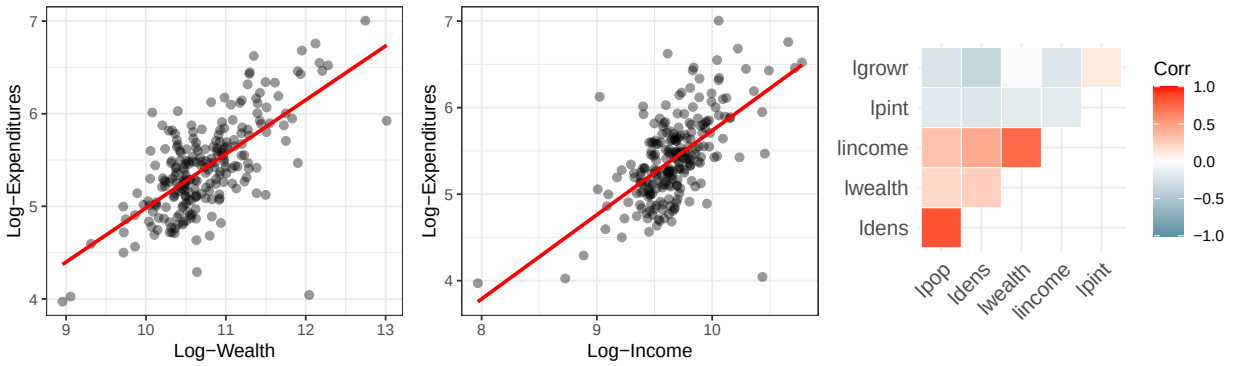

Figure 2: Exploratory Plot after Log-Transformation and Subsetted Data 1: Scatterplot of Log-Expenditures vs Log-Wealth with Linear Regression line in red, 2: Scatterplot of Log-Expenditures vs Log-Income with linear regression line in red, and 3: Upper Correlation plot of all independent variables.

After detecting the relationship of each independent variables with expenditures, it is important to examine the correlation of each predictors. Higher correlation between predictors mean there might be multicollinearity issues when including both variables in the model. This leads the model to have unstable and unreliable coefficients. Figure 2.3. shows the upper diagonal of the correlation matrix plot. As one can see, density and population have a high positive correlation of 0.83. To solve the multicollinearity issues, separate models with density in one and population in other, along with other variables, were generated. The results of the models will be compared and selected as the best model. Furthermore, wealth and income also have high positive correlation of 0.74. However, wealth and income are not strongly correlated. Other variables not mentioned above have little to no relationship with expenditures.

Statistical Analysis

```
fit1 <- lm(lexpen ~ lwealth + lpint + ldens + lincome + lgrowr, data = set2)

fit2 <- lm(lexpen ~ lwealth + lpop + lpint + lincome + lgrowr, data = set2)

fit1_summary <- tibble(Features = "lwealth, lpint, ldens, lincome, lgrowr",
                       MSE = mean(fit1$residuals^2),
                       Adj.R.squared = summary(fit1)$adj.r.squared,
                       F.statistics = summary(fit1)$fstatistic[[1]],
                       AIC = AIC(fit1))
```

```
fit2_summary <- tibble(Features = "lwealth, lpop, lpint, lincome, lgrowr",
                       MSE = mean(fit2$residuals^2),
                       Adj.R.squared = summary(fit2)$adj.r.squared,
                       F.statistics = summary(fit2)$fstatistic[[1]],
                       AIC = AIC(fit2))
```

Model selection is a crutial step in a data analysis. A multiple linear regression will be used to build the model. After the exploratory data analysis, there are six possible variables that can be include in the model including wealth, population, percent intergovernmental funds, density, income, and growth rate. As mentioned before, if population and density are in the same models, multicolinearity issues will occur. Therefore, two initial models will be built to compared using three validation metrics such as Akaike information criterion (AIC), R-squared adjusted, and mean square error (MSE). In this case, the ideal best model would have smaller AIC score, higher R-square adjusted, and lower MSE. After building both models, a table of model comparison was generated to select the best model among the two, see Table 1. This table clearly state that the second model with predictors wealth, population, percent intergovernment funds, income, and grow is the best model with lower MSE and AIC values, and higher R-square adjusted proportion of 0.09, 124.16, and 0.61, respectively. An R-square of 0.61 states that 60.63 of variation in expenditures can be explained by all of the predictors in this model.

Features	MSE	Adj.R.squared	F.statistics	AIC
lwealth, lpint, ldens, lincome, lgrowr	0.0983	0.5921	66.8997	132.2354
lwealth, lpop, lpint, lincome, lgrowr	0.0949	0.6063	70.9101	124.1646

Table 1: Regression validation metrics including MSE, R-squared adjusted, and AIC

After finding the best predictors, a stepwise selection was generated for variables selection using AIC values. Unfortunately, the full model appears to be the final model with the following predictors: wealth, population, percent intergovernment, income, and growth rate.

Conclusion

Appendix A: Supplemental Tables and Figures

Table 2: Summary Statistics for all numerical independent features

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
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

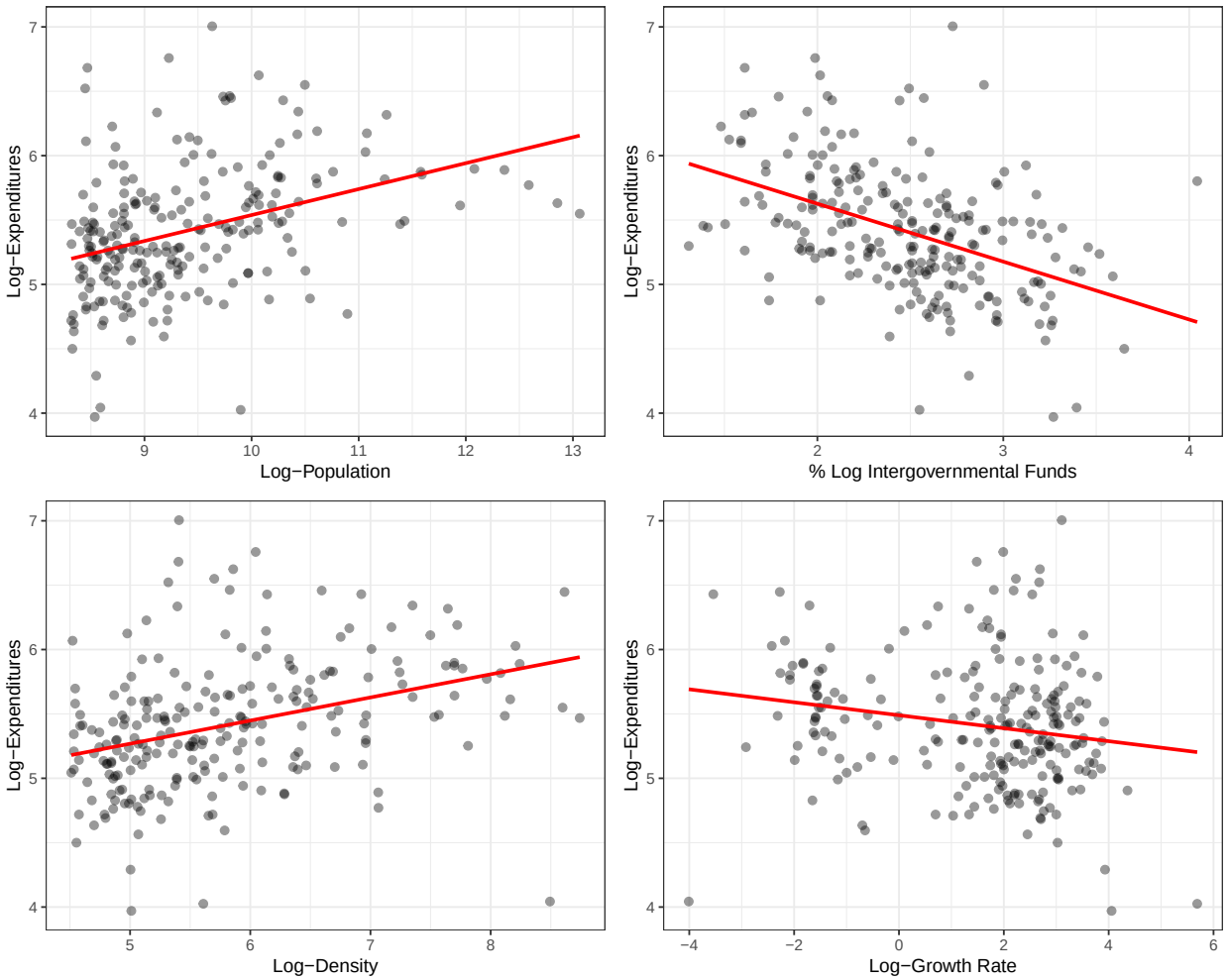

Figure 3: Exploratory Plot after Log-Transformation and Subsetted Data 1 with linear regression line in red: 1: Log-Expenditures vs Log-Population, 2: Log-Expenditures vs Log-Intergovernmental FUnds, 3: Log-Expenditures vs Log-Density, and 4: Log-Expenditures vs Log- Growth Rate.