

Statistical Analysis and Predictive Models for Expenditures in New York Municipalities

October 22, 2020

Executive Summary: In New York, there are many new housing projects started daily. Given the issue that the town needs to generate more funding through property tax when expenditures increases. However, there is no clear understanding of which factors cause the expenses to increase/decrease. Therefore, a clear goal for this analysis is to find variables that effect expenditures. Then, using the model, we can predict expenditures in New York municipalities Warwick and Monroe in 2005 and 2025. A tool was developed using multiple linear regression models to predict expenditures using a dataset from the New York municipalities from 1992. With an adjusted R-square of 61%, a final model was chosen after stepwise variables selection process with AIC, MSE, and adjusted R-square criteria. The variables in the final model includes population, wealth, income, and percent intergovernmental funds. Diagnostics measurements was also built to check for the quality of the model. Some interesting findings include a 4% increase in expenditures while wealth increases by 10%. On the other hand, there will be a 3% decrease in expenditures while the percentage of intergovernmental funds increases by 10%. Knowing the changes in expenditures caused by wealth, population, intergovernmental funds, and growth rate, construction workers and properties owners would be more caution when starting a new project.

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 expenditures 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 expenditures decrease, property's owners could reimburse the expenses elsewhere. In addition, knowing the expenditures would also help construction managers to order supplies in a more proper manner. If expenditures decrease, then the supplies would have to be less in quantity or cheaper in quality. Numerous questions were proposed in favor of these issues such as 1) What variables cause the fluctuation of expenditures? 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 multiple linear regressions, and diagnostics analysis. With the answered questions, construction workers and property owners would have a better understanding of their expenditures when starting a new project to avoid over or underestimating their budgets. Using the developed model, we will be able to predict expenditures for two municipalities in New York (Warwick and Monrow) in 2005 and 2025.

Methods: A dataset from New York municipalities were provided to access the important measures to predict expenditures. These data contain a total of 916 observations from 1992 with 2

observations contain missing expenditure values. Two observations with missing 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 funding, density, mean income per person, and growth rate. There is a total of 57 distinct county codes implying that there are multiple measurements of expenditures per county in New York. The goal of this data analysis is to predict expenditures of two New York municipalities, Warwick and Monroe. Thus, a projection dataset for Warwick and Monroe town was also provided to generate predictions using the fitted model. To achieve this goal, all analysis will be done using multiple linear regressions for model development process and followed by a diagnostic process to check for the quality of the model. All analysis including coding and writing report was 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 A3, the summary statistics of all independent variables and target variable shows the maximum for expenditures, wealth, population, percent intergovernmental funds, density, income, and growth rate are extremely high compare to their mean and 75 percentiles. This indicates that all the variables mentioned previously are heavily right skewed. Most importantly, expenditures' skewness violates the normality assumption when generating a linear regression. Therefore, a log-transformation was applied to expenditures to normalize the distribution of the target variable. Log-transformation would reduce the values that are very high/low. Figure 1.1 depicts the normality of the outcome variable expenditures after transformation implying the assumption is not violated anymore.

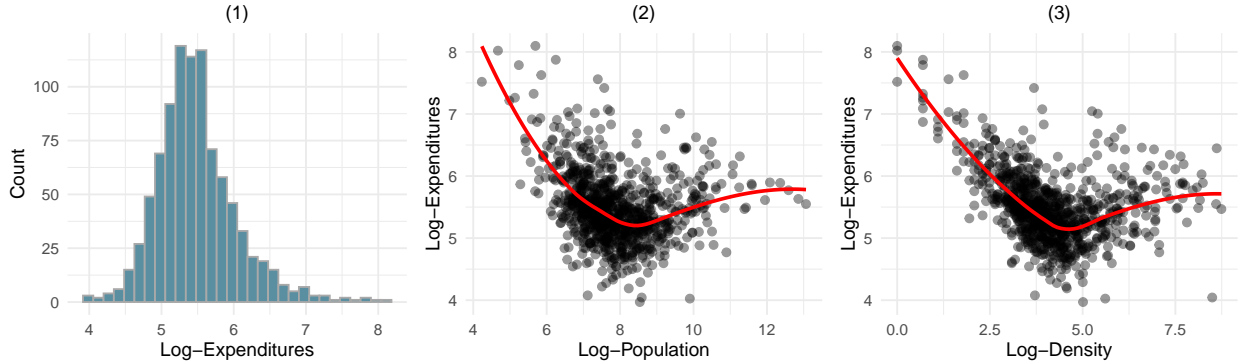


Figure 1: Initial Exploratory plots after log-transformation of 1: Histogram of Log-Expenditures with frequency, 2: Scatter plot of Log-Expenditures vs Log-Population with LOESS smooth line in red, and 3: Scatterplot of Log-Expenditures vs Log-Density with LOESS smooth line in red.

Since all independent variables are right-skewed, a log-transformation also applied to each variable to ensure their linear relationship with expenditures. An amount of 1.01 was added to all growth rate values to account for zero 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, percent intergovernmental funds, income, and grow rate seems to have linear relationship with expenditures. However, population and density have two difference trends 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, represents 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 subsetting 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 using 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 to see if they are significant to predict the outcome variable, expenditures. Expenditures and wealth have a positive relationship indicating the increases of wealth would cause expenditures 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 low income individuals. Other variables like population and density seem to have moderate positive relationship with expenditures, see the left two plots in Figure B4. 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 percent intergovernmental funds and growth rate have moderate negative correlation with expenditures. This indicates that, while percent 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.

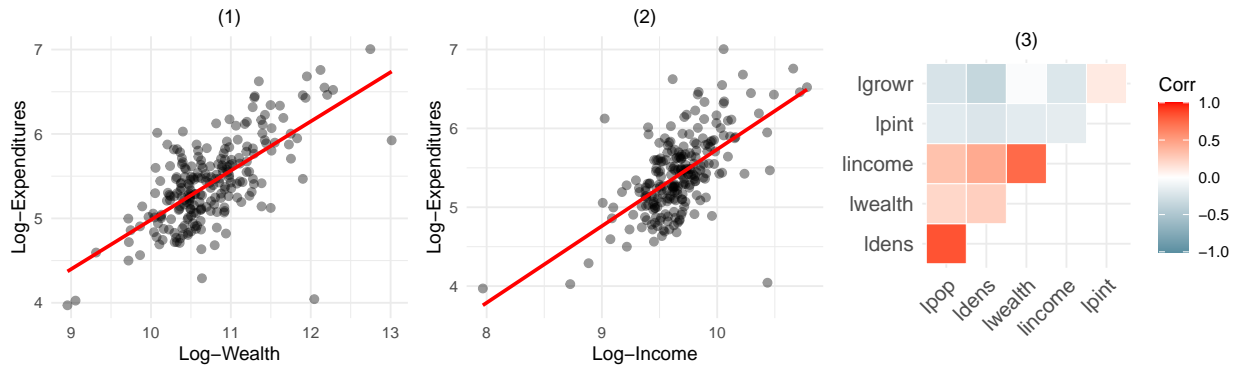


Figure 2: Exploratory plot after log-transformation and subsetting data 1: Scatterplot of Log-Expenditures vs Log-Wealth with Linear Regression line in red, 2: Scatterplot of Log-Expenditures vs Log-Income with linear regression line in red, and 3: Upper Correlation plot of all independent variables.

After detecting the relationship of each independent variable with expenditures, it is important to examine the correlation of each predictor. Higher correlation between two 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, two separated 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 moderate positive correlation of 0.74. Other variables did not mentioned above have little to no relationship with expenditures.

Model Fitting/Inferences: Model selection is a crucial step in a data analysis. Multiple linear regressions will be used to build all the models. After the exploratory data analysis, there are six possible variables that can be included 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, multicollinearity 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 A4. This table clearly states that the second model with predictors wealth, population, percent intergovernmental funds, income, and growth rate is the best model among the two with lower MSE (0.09) and AIC (124.16) values, and higher R-square adjusted (0.61).

After finding the best predictors for the model (wealth, population, percent intergovernmental, income, and growth rate), a stepwise selection method was generated for variables selection using AIC criteria. Unfortunately, stepwise variable selection was not showing any other significant models beside the full model. Furthermore, each combination of the interaction predictors was also taken into consideration to find a better model. Results of MSE, R-square adjusted, and AIC for the model with interaction terms (model 3 in Table A4) are clearly better than model 2. Thus, a stepwise selection method was applied to select the best variables using interaction terms and results are shown in Table A4 as model 4. Model 4 has the best adjusted R-squared due to the higher number of variables contain in the model. However, in terms of the model's quality, model 4 seems to have insignificant predictors including wealth and wealth times population, see p-value column in Table A5. In addition, the coefficient for percent intergovernmental funds was expected to be negative as shown in Figure B4. The coefficient of percent intergovernmental funds contracts the exploratory scatter plot. These two reasons prove that model 4 with interaction terms is not the best model. In addition, growth rate seems to be insignificant in most model with very small slope and high p-value. Therefore, the model with predictors wealth, population, intergovernmental funds, and income will be used as the final model.

After variable selection, diagnostics process is a must to check for the quality of the model. Figure 3.1 shows the studentized residuals. The majority of observations are within the range of -2.5 to 2.5 except one observation, 893. This indicates that is observation is an outlier. In addition, Figure 3.2 shows the cook's distance of all the observations and found that there are a few observations did not make the cutoff point for influential values including index 60, 83, 100, 179, and 225 from the subset data. As seen in Table A6, outliers and influential points have high values of either expenditures or one of the predictors. However, none of the outliers and influential points were removed because there are no good reasons. Last but not least, Figure 3.3 shows that studentized residuals follow a normal distribution implying the error terms are normal.

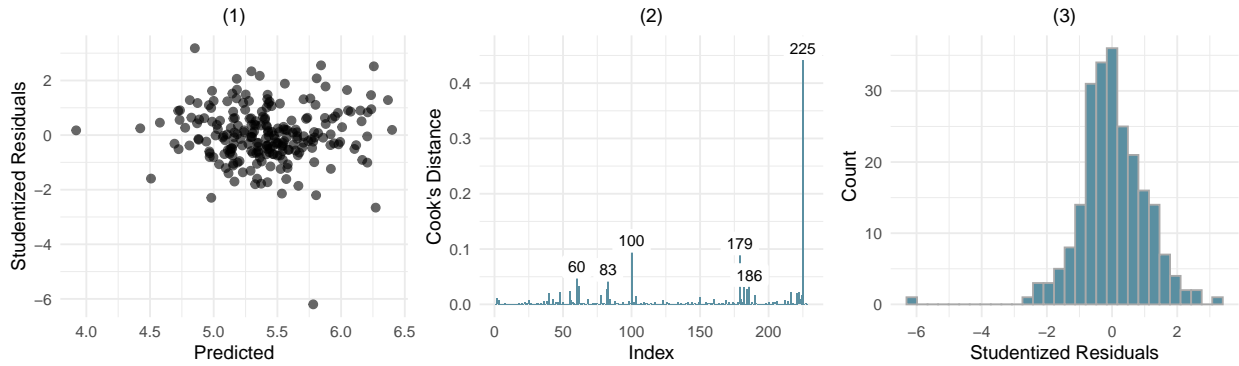


Figure 3: Diagnostic plots of 1: Studentized Residuals vs Predicted values, 2: Cook's distance for each observation, and 3: Distribution of studentized residuals.

After accessing the quality of the model, an interpretation and explanation of the model will be explained. The summary of regression for the final model (wealth, population, intergovernmental, and income) is in Table 1. VIF values are all smaller than 3 indicating there is no multicollinearity problem in the model. Confident intervals for majority of coefficients do not contain zeros and small p-values implying they are statistically significant in the model. For the purpose of interpreting the model, the coefficient will be back transformation. This means that as wealth, population, or income increases by 10%, holding all else constant, expenditures will increase by 4%, 1%, or 2%, on average, respectively. On the other hand, while percent intergovernmental fund increases by 10%, on average, expenditures will decrease by 3%. In addition, this model has an adjusted R-square of 0.6 indicating that 60% of the variation in expenditure can be explained by wealth, population, percent intergovernmental funds, income, and growth rate.

Term	Estimate	SD Error	F-Statistics	P Value	10% Coef	2.5% CI	97.5% CI	VIF
(Intercept)	-1.504	0.687	-2.188	0.030	0.866	-2.858	-0.150	
lwealth	0.398	0.055	7.305	0.000	1.039	0.291	0.506	2.265
lpop	0.086	0.025	3.517	0.001	1.008	0.038	0.135	1.143
lpint	-0.313	0.043	-7.246	0.000	0.971	-0.398	-0.228	1.056
lincome	0.269	0.103	2.618	0.009	1.026	0.067	0.472	2.394

Table 1: Summary regression of final model

Given a good model, the goal of this analysis is to predict expenditures for two New York municipalities Warwick and Monroe. A projection table of values was provided to predict from years 2005 and 2025. Table 2 shows a prediction table of the two towns. The values for expenditures seem to be within the range of expenditures in the training set. Looking closely at the projection table, predicted expenditures are higher, on average, in Monroe than Warwick due to the extremely lower percentage of intergovernmental funds. According to the model, having lower intergovernmental funds leads to higher expenditures. Although, Warwick town has relatively higher population, wealth, and income, on average, Monroe town would still have higher expenditures. The last two columns depict lower and upper 95% prediction interval of predicted values and we can say that expenditures are confident to be inside these ranges. However, the prediction intervals are quite

large making the predictions to have wide ranges. Given the coefficients being fixed, Warwick and Monroe would expect their expenditures to change if their wealth, population, percent intergovernmental fund, or income changes.

Town	Year	Predicted Expen	Lower PI	Upper PI
Warwick	2005.0	264.2	141.4	493.7
Warwick	2025.0	276.4	147.6	517.5
Monroe	2005.0	285.7	153.7	531.0
Monroe	2025.0	288.1	155.1	535.3

Table 2: Prediction table of two municipalities in New York (Warwick and Monroe)

Conclusion: In conclusion, expenditures are one of the most important aspects to know before starting a new construction project. It is difficult to estimate expenditures manually. Therefore, this analysis creates a tool to automate the process of estimating expenditures. The variables that cause the fluctuation of expenditures are population, wealth, percent intergovernmental funds, and income. With this model, construction managers and property owners would have a clear understanding of how much funding they would need for a certain type of project in the city. Using the summary regression results of our model, we are confident to say that an area with higher population would cause expenditures to be higher. Similarly, if the average of wealth and income is higher, then expenditures will also be higher. However, it is clear that if the percentage of intergovernmental funds is lower, then expenditures is obviously higher. As a result, this model does not only provide an estimate of expenditures faster, but construction workers can also have a great understanding of what cause the expenses to shift up and down.

There are a few limitations in this analysis. This model only used a subset of the main data which is 4 times less compared to the full dataset. Therefore, smaller sample size would lead the model to perform not as accurate. Second, the model used a training set from 1992 to project predictions for 2005 and 2025 would also be unreliable. The model was not developed using time as one of the factors. Therefore, it difficult to make accurate forecast using only one year of data. Last but not least, the model is only limited to when population is greater than 4024. Cities with population smaller than 4024 cannot use this model. For future reference, there are many ways to improve the models. We could consider more variables that could lead expenditures to be higher or lower including number of workers, supplies, time of the year, and location of the new project. By doing this, we might be able to build a more accurate model. In addition, we would collect more data from different cities in New York in an expanded time range. With the data from different time range, we will be more confident to predict the projected data in the future years.

Appendix A: Supplemental Tables

Table 3: 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

	Features	MSE	Adj.R.squared	F.statistics	AIC
1	lwealth, lpint, ldens, lincome, lgrowr	0.098	0.592	66.900	132.235
2	lwealth, lpop, lpint, lincome, lgrowr	0.095	0.606	70.910	124.165
3	lwealth, lpop, lpint, lincome, lgrowr, lwealth:lpop, lwealth:lpint, lwealth:lgrowr, lpop:lpint, lpop:lincome, lpop:lgrowr, lpint:lincome, lpint:lgrowr, lincome:lgrowr	0.075	0.676	32.551	89.333
4	lwealth, lpop, lpint, lincome, lgrowr, lwealth:lpop, lwealth:lpint, lwealth:lgrowr, lpop:lincome, lpop:lgrowr	0.075	0.680	49.184	81.874
5	lwealth, lpop, lpint, lincome	0.096	0.603	87.245	125.004

Table 4: Regression validation metrics including MSE, R-squared adjusted, and AIC

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-27.5007	8.5712	-3.21	0.0015
lwealth	0.1411	0.7266	0.19	0.8462
lpop	2.2222	0.8591	2.59	0.0103
lpint	3.1678	0.7536	4.20	0.0000
lincome	3.1988	1.2330	2.59	0.0101
lgrowr	-0.8634	0.2112	-4.09	0.0001
lwealth:lpop	0.1025	0.0728	1.41	0.1604
lwealth:lpint	-0.3227	0.0702	-4.60	0.0000
lwealth:lgrowr	0.0957	0.0171	5.60	0.0000
lpop:lincome	-0.3300	0.1339	-2.47	0.0145
lpop:lgrowr	-0.0203	0.0113	-1.80	0.0727

Table 5: Summary Regression of Model 4

	obs	expen	wealth	pop	pint	income	index
1	342	73	41695	5166	16.70	7219	60
2	480	237	147431	88153	11.40	34775	83
3	519	331	39733	6732	57.00	18375	100
4	778	1102	343364	15247	15.30	23293	179
5	893	57	170195	5371	29.80	34082	225

Table 6: Outlier Obsevatons

Appendix B: Supplemental Figures

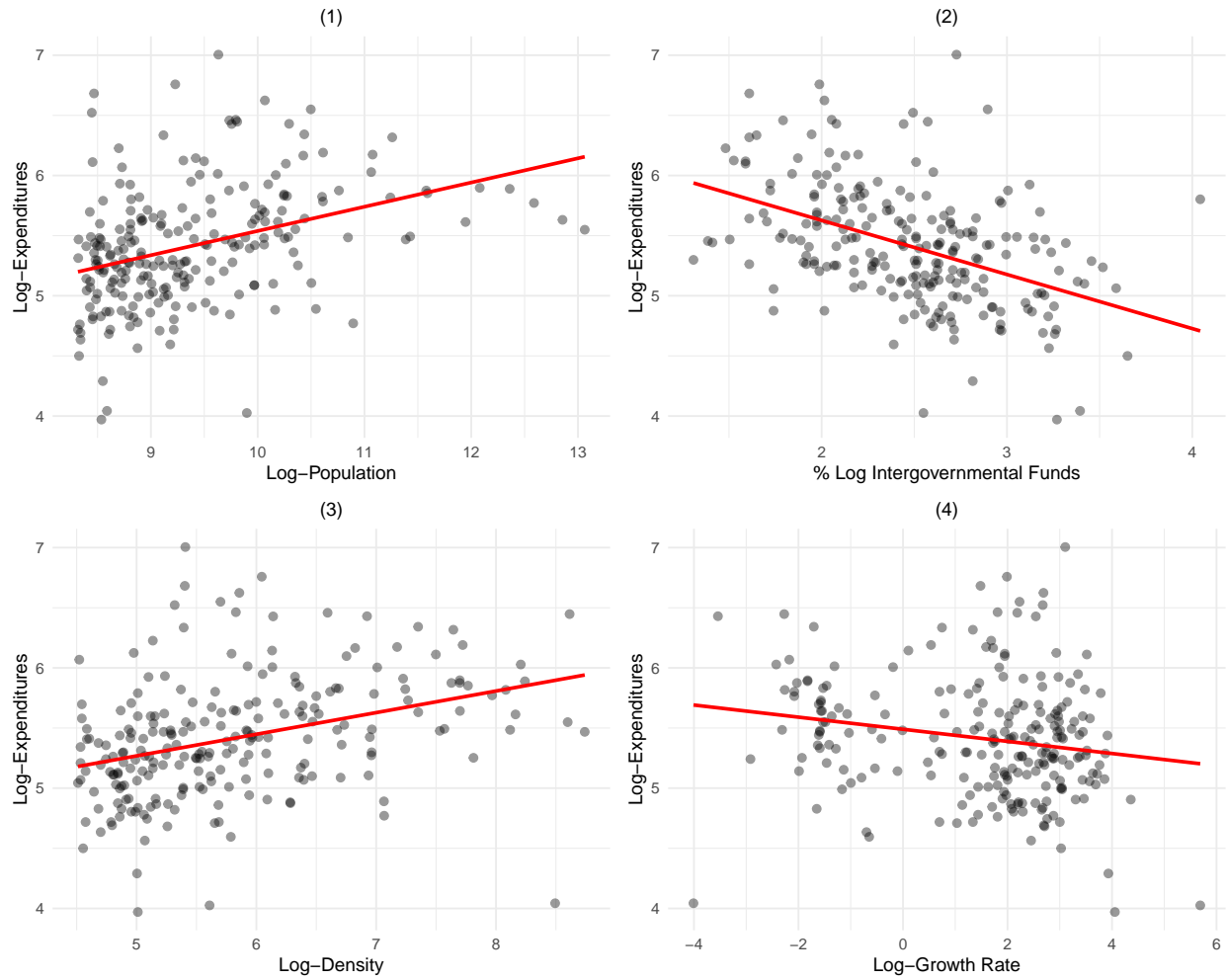
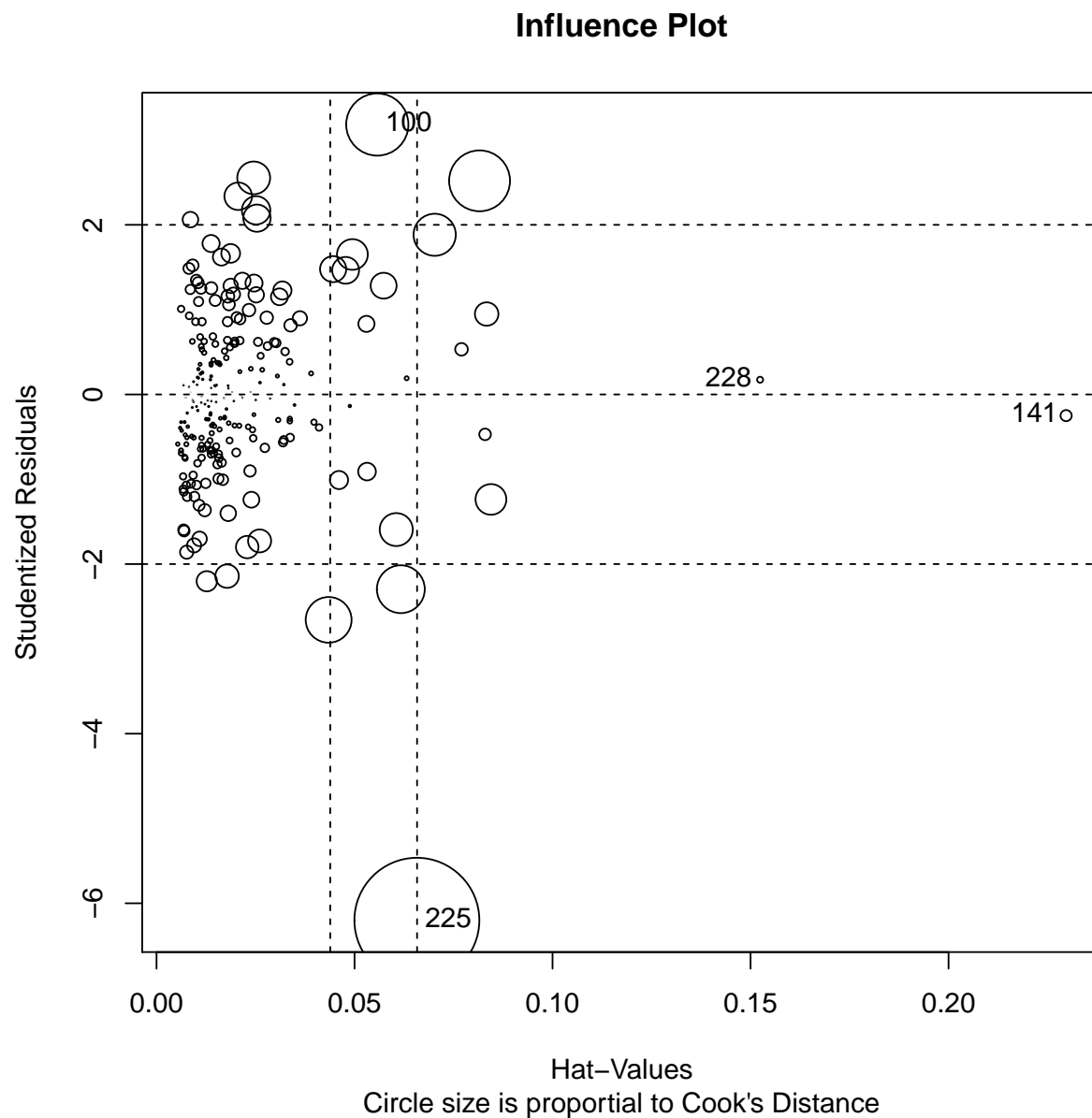


Figure 4: Exploratory plot after log-transformation and subsetting data with linear regression line in red of 1: Log-Expenditures vs Log-Population, 2: Log-Expenditures vs Log-Intergovernmental Funds, 3: Log-Expenditures vs Log-Density, and 4: Log-Expenditures vs Log-Growth Rate.



StudRes Hat CookD 100 3.1826380 0.05574637 0.114896702 141 -0.2493088 0.22969257
 0.003722360 225 -6.1998639 0.06573965 0.463184262 228 0.1731713 0.15240822 0.001083171
 Figure 5: Plot of influential points of studentized residuals vs hat values.

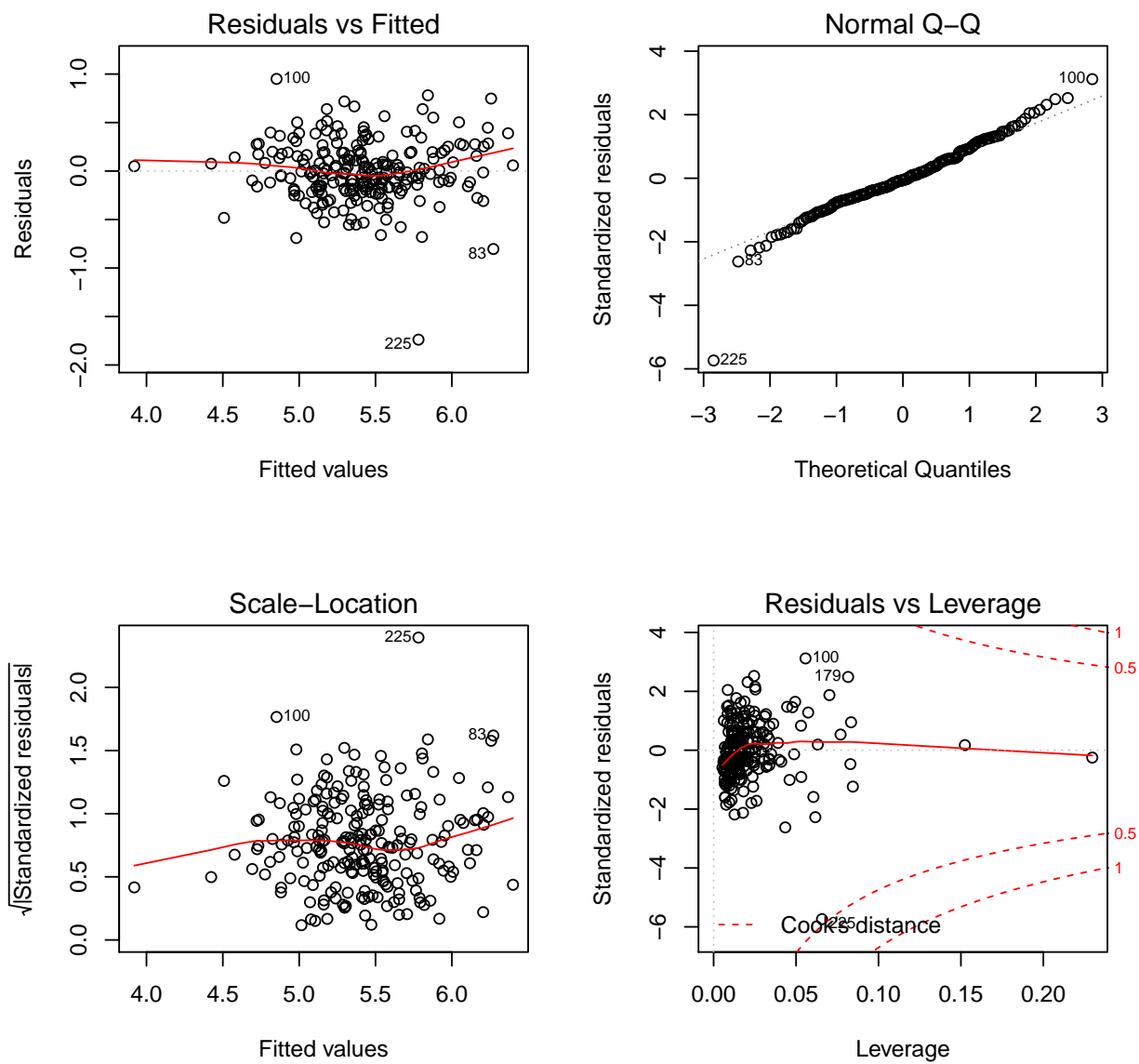


Figure 6: Diagnostics plots

Appendix C: R Code

```

1 ##### Packages #####
2 library(tidyverse)
3 library(magrittr)
4 library(ggcorrplot)
5 library(MASS)
6 library(leaps)
7 library(car)
8 library(stargazer)
9 library(gridExtra)
10 library(broom)
11 library(knitr)
12 library(kableExtra)
13
14
15 ##### Parameters #####
16
17 my_col <- "#5a8fal"
18
19
20 ##### Functions #####
21
22 ScatterPlotFunction <- function(df, xvar, xlab, smooth_method = "loess"){
23   ggplot(df, aes_string(x = xvar, y = "lexpen")) +
24     geom_point(size = 2, alpha = 0.4) +
25     geom_smooth(method = smooth_method, se = FALSE, color = "red", size = 1) +
26     theme_minimal() +
27     labs(y = "Log-Expenditures",
28          x = xlab)
29 }
30
31 ValidationTable <- function(fit){
32   mod <- fit
33   fit_summary <- tibble(Features = paste0((coef(mod) %>% names())[1], collapse = ",
34     ),
35     MSE = mean(mod$residuals^2),
36     Adj.R.squared = summary(mod)$adj.r.squared,
37     F.statistics = summary(mod)$fstastic[[1]],
38     AIC = AIC(mod))
39   return(fit_summary)
40 }
41
42 ##### Load data #####
43
44 ny <- read.table("data/cs73.dat", header = T)
45
46 ny2 <- na.omit(ny)
47
48
49 ##### Summary Statistic Tables #####
50
51 stargazer(ny2 %>% dplyr::select(-obs, -st, -co, -id))
52
53
54 ##### Add transformed variables #####

```

```

55
56 ny2 %<>%
57   mutate(lexpen = log(expen),
58          lwealth = log(wealth),
59          lpop = log(pop),
60          lpint = log(pint),
61          ldens = log(dens),
62          lincome = log(income),
63          lgrowr = case_when(
64            growr > 0 ~ log(growr + 1.01),
65            TRUE ~ -log(-growr + 1.01)))
66
67
68 ##### Explore Target variables #####
69
70 ## generate histogram of expenditure to determine y-var assumption met
71 ## the plot look very right skew indicating we should transform the y variable
72 ggplot(ny2, aes(expen)) +
73   geom_histogram(aes(y = stat(density)), bins = 30, col = "dark gray", fill = my_col
74   ) +
75   geom_density(size = 1, color = "red") +
76   theme_bw() +
77   labs(x = "Expenditures",
78        y = "Density")
79
80 ## after transforming the y variable to using log tranformation
81 ## expen is now seems to be normal distributed
82 ggplot(ny2, aes(lexpen)) +
83   geom_histogram(aes(y = stat(density)), bins = 30, col = "dark gray", fill = my_col
84   ) +
85   geom_density(size = 1, color = "red") +
86   theme_bw() +
87   labs(x = "Log-Expenditures",
88        y = "Density")
89
90 ## perform qq plot
91 ggplot(ny2, aes(sample = lexpen)) +
92   stat_qq(size = 2, alpha = .4) +
93   stat_qq_line(col = "red", size = 1) +
94   theme_bw() +
95   labs(y = "Log-Expenditures",
96        x = "Theoretical")
97
98
99 ##### Explore Independent Variables #####
100
101 ## scatter plots of predictors and target
102 ScatterPlotFunction(ny2, "lwealth", "Log-Wealth")
103 ScatterPlotFunction(ny2, "lpop", "Log-Population")
104 ScatterPlotFunction(ny2, "lpint", "% Log Intergovernmental Funds")
105 ScatterPlotFunction(ny2, "ldens", "Log-Density")
106 ScatterPlotFunction(ny2, "lincome", "Log-Income")
107 ScatterPlotFunction(ny2, "lgrowr", "Log-Grow Rate")
108
109 ##### Subset data #####

```

```

109 ## subsetting the data based on regions of log-population and log-density
110 ## where the relationship with log-expenditure is linear
111 set2 <- ny2 %>%
112   filter(lpop > 8.3 & ldens > 4.5)
113
114 ## use scatterplot to explore the relationship between
115 ## x and y again with new set
116
117 ScatterPlotFunction(set2, "lwealth", "Log-Wealth", "lm")
118 ScatterPlotFunction(set2, "lpop", "Log-Population", "lm")
119 ScatterPlotFunction(set2, "lpint", "% Log Intergovernmental Funds", "lm")
120 ScatterPlotFunction(set2, "ldens", "Log-Density", "lm")
121 ScatterPlotFunction(set2, "lincome", "Log-Income", "lm")
122 ScatterPlotFunction(set2, "lgrowr", "Log-Grow Rate", "lm")
123
124 ## correlation plot ##
125 corr_table <- set2 %>%
126   dplyr::select(lwealth, lpop, lpint, ldens, lincome, lgrowr) %>%
127   cor()
128
129 ## lpop and ldens have high positive correlation = multicollinearity problem
130 ## lincome and lwealth has high positive correlation also
131 ggcorrplot(corr_table, hc.order = TRUE,
132             outline.col = "white",
133             colors = c(my_col, "white", "red"), type = "upper")
134
135
136 ##### Model Development #####
137
138 ## since lpop and ldens are highly correlated, we cannot put them in the same model
139
140 fit1 <- lm(lexpen ~ lwealth + lpop + lpint + ldens + lincome + lgrowr,
141           data = set2)
142
143 ## after doing stepwise selection
144 ## the final model include the same 6 variables
145 fit1_step <- stepAIC(fit1, direction = "both", trace = FALSE)
146
147 ## looking at the summary of regression we can see that
148 ## the coefficient for density is not in the right direction
149 ## scatter plot shows positive relationship between ldens and lexpen
150 ## suspecting the model has high VIF on ldens and lpop
151 summary(fit_step)
152
153 ## try removing lpop
154 fit2 <- lm(lexpen ~ lwealth + lpint + ldens + lincome + lgrowr, data = set2)
155
156 summary(fit2)
157
158 fit2_step <- stepAIC(fit2, direction = "both", trace = FALSE)
159
160 ## try removing ldens
161 fit3 <- lm(lexpen ~ lwealth + lpop + lpint + lincome + lgrowr, data = set2)
162
163 summary(fit3)
164

```

```

165 fit3_step <- stepAIC(fit3, direction = "both", trace = FALSE)
166
167 ## coefficient seems right now
168 ## stepwise for both models also give the same models
169 ## picking model 3 to be the final model since it has
170 ## lower MSE and higher Adjusted R-squared
171
172 ##### Interaction #####
173
174 fit4 <- lm(lexpen ~ .*.,
175           data = set2 %>% dplyr::select(lexpen, lwealth, lpop, lpint, lincome,
176                                         lgrowr))
177
178 fit4_step <- stepAIC(fit4)
179
180 ## coefficient seems quite off for pint
181 summary(fit4_step)
182
183 ##### final model #####
184 ## pick final model to have no interaction terms
185 fit_final <- fit3_step
186
187
188 ##### Model Diagnostic #####
189
190 ## Residuals
191 ## add prediction values and rstudent residuals into set2
192 set2 %>%
193   mutate(predict = predict(fit_final),
194          rstudent = rstudent(fit_final))
195
196 ## plot studentized residuals vs predicted values
197 ggplot(set2, aes(x = predict, y = rstudent)) +
198   geom_point(size = 2, alpha = .6) +
199   theme_bw() +
200   labs(x = "Predicted",
201        y = "Studentized Residuals")
202
203 ## identify the outlier in residual vs predicted plot
204 ## extremely large wealth and dens
205 ## extremely low grow rate
206 set2 %>%
207   filter(rstudent == min(rstudent))
208
209
210 ## Normality of Residuals
211 set2 %>%
212   mutate(dnorm_rstudent = dnorm(rstudent))
213
214 ggplot(set2, aes(x = rstudent)) +
215   geom_histogram(bins = 30, col = "dark gray", fill = my_col) +
216   theme_bw() +
217   labs(title = "Distribution of Studentized Residuals",
218        x = "Studentized Residuals",
219        y = "Count")

```

```

220
221 ## Q-Q plot for studentized resid
222 qqPlot(fit_final, main="QQ Plot", ylab="Studentized Residuals")
223
224 ## Influential Observations
225 ## Cook's D plot
226 ## identify D values > 4/(n-p-1) as a guide;
227 ## Cook and Weisberg recommend 0.5 and 1 (R uses these guides in default diagnostic
    plots below)
228 cutoff <- 4/((nrow(set2) - length(fit_final$coefficients) - 2))
229
230 diag <- broom::augment(fit_final) %>% mutate(Index = 1:nrow(.))
231
232 diag %<>%
233   mutate(high_cooksd = case_when(
234     .cooks_d > cutoff ~ 1, TRUE ~ 0),
235     col_stdresid = case_when(
236       .std.resid > 0 ~ 1,
237       .std.resid < 0 ~ 0),
238     high_hat = case_when(
239       .hat > .1 ~ 1,
240       TRUE ~ 0))
241
242 ## cook's distant ggplot
243 ggplot(diag, aes(x = Index, y = .cooks_d)) +
244   geom_bar(stat = "identity", fill = my_col) +
245   labs(y = "Cook's Distance") +
246   theme_minimal() +
247   geom_label(data = diag %>% filter(.cooks_d > cutoff + .03),
248     aes(label = Index), label.size = NA, size = 3)
249
250 ## Influence plot: studentized residuals vs. hat matrix diagonals (leverage) with
    bubbles a function of Cook's D
251 ## Interactive, so can click to identify high leverage/influential/outlying points
252 influencePlot(fit_final, id.method="identify",
253   main="Influence Plot", sub="Circle size is proportional to Cook's
    Distance" )
254
255 ## VIF
256 ## vif score seem very close to 1
257 vif(fit_final) # closer to 1 the better; 5-10 is moderate
258
259 # All encompassing R default regression model diagnostics
260 par(mfrow=c(2,2))
261 plot(fit_final)
262
263
264 ##### Predictions
265
266 ## Account for sigma^2
267 sd_fit <- sd(fit_final$resid)
268
269 ## Read Warick and monroe data
270 wm <- readxl::read_xlsx("data/warwick-and-monroe.xlsx")
271
272 ## create log transformation variables for wm

```



```

273 wm %>%
274   mutate(lwealth = log(wealth),
275          lpop = log(pop),
276          lpint = log(pint),
277          ldens = log(dens),
278          lincome = log(income),
279          lgrowr = case_when(
280            growr > 0 ~ log(growr + 1.01),
281            TRUE ~ -log(-growr + 1.01)))
282
283
284 ## generate prediction for wm
285 wm$lexpen_pred <- predict(fit_final, newdata = wm)
286
287 wm %>%
288   mutate(lexpen_pred = predict(fit_final, newdata = wm))
289
290 wm %>%
291   cbind(exp(predict(fit_final, wm, interval="prediction") + sd_fit^2/2))

```

Listing 1: Appendix of Code