Predicting Expenditures for the Municipalities of Monroe and Warwick for the Construction of New Housing Projects

Travis Gubbe

March 26, 2025

Executive Summary

The municipalities of Monroe and Warwick in the state of New York are interested in the construction of new housing projects in their communities. Before starting construction, the villages want a projection of potential expenditures per person after housing projects are completed to determine potential funding increases, including the implementation of various taxations such as property or sales taxes. After transforming the variables to assist in meeting linear assumptions, a linear regression model is fitted to predict the potential expenditures for both municipalities for the years 1992, 2005, and 2025. After performing stepwise regression model selection and running model diagnostics on the log transformed variables, it was determined that expenditures per person has a significant relationship with wealth per person, population, percentage intergovernmental (both linear and quadratic terms), density, mean income per person, and growth rate.

Introduction

The municipalities of Monroe and Warwick are concerned about the potential increase in expenditures per person over time, as they may need to find various ways of meeting the funding requirements for the construction of new housing projects. Using applicable data from the municipalities of New York, Monroe and Warwick are hoping for an accurate model prediction for years to come. The variables provided for the study are the response variable expenditures per person and the predictor variables wealth per person, population, percent intergovernmental (percentage of revenue from state and federal grants and subsidies), density, mean income per person, and growth rate. There are three identifier variables in the data set, which were not needed for the purposes of the linear regression model.

Once a final regression model is chosen, the predictions will be created for Monroe and Warwick. These predictions will assist the two municipalities in forecasting the potential expenditures per person due to new housing projects. A potential challenge in developing a regression model will be the relationships among the variables. It is possible the demographic and income-related variables will exhibit non-linear relationships, particularly with a change in population and income-related characteristics of municipalities.

Methodology

The data set provided contains data from 1992 on various municipalities in New York. After removing two observations with empty values, the data set contains 914 observations from 10 variables, with the seven variables listed in the Introduction above being the primary focus. After performing exploratory data analysis, it was determined to use linear regression to create a prediction model. A subset split of the data was created for the prediction model, with the split occurring where high population and high density occurred. From Figure 1 the subset used for the study will be from the right of both grey lines, as these mark high population and high density areas. The following assumptions were used as part of the regression analysis: the error terms are independent of others and follow a normal distribution, the observations are independent of one another, and the residuals have constant variance. The statistical analysis was performed in R software version 4.4.2.

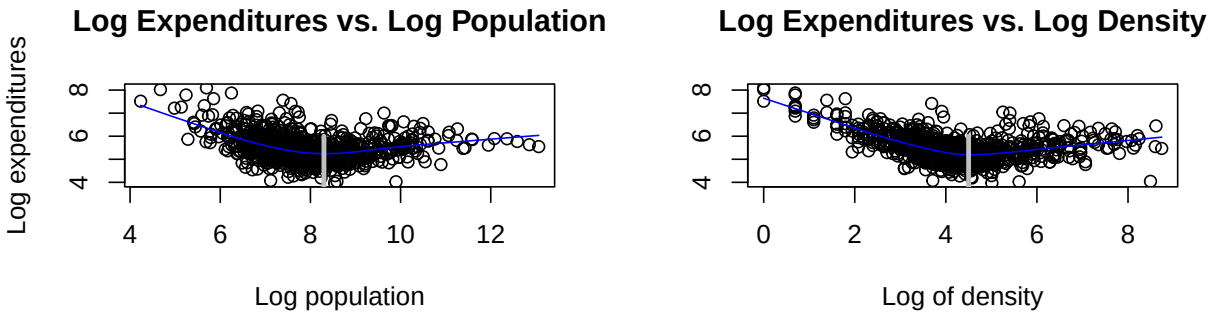

Figure 1: Two scatter plots between log expenditure vs. log population and log expenditure vs. log density. Both scatter plots show two linear segments split by a grey line. The prediction model will focus on the data to the right of the grey line in both scatter plots, which highlights high population and high density areas.

Statistical Analyses

Exploratory Data Analysis

To begin, exploratory data analysis (EDA) was performed on the data set to better understand the relationship between the response and predictor variables as well as the distribution of each variable and any potential impact on the analysis. A variety of EDA visualizations were run, including correlation plot, boxplots, and scatter plots between variables. The correlation plot shows that most variables do not have a strong correlation with one another, which is a sign of independence among the variables. There appears to be some correlation between expenditures and wealth as well as population and density, though this is not completely unexpected since it is not uncommon to have larger expenditures in wealthier communities and the high populations may also have high densities. These correlations are something to note when performing model diagnostics, particularly when running a VIF calculation for potential multicollinearity in the data. From the boxplot of expenditures, the data seems to be heavily right-skewed distribution, which will need to be investigated more with a histogram plot and Q-Q plot.

To determine if the predictor satisfies normality assumptions, a Q-Q plot and a histogram plot for the expenditures variable were created. From the Q-Q plot in Figure 2, the data points stray far from the Q-Q normality line, which means expenditure variable is not normally distributed and transformation may need to be completed. From the histogram plot, the data appears to be heavily right-skewed, meaning the expenditures data is not normally distributed. From these two plots, it was determined to perform a logarithmic transformation on the expenditures variable. Once the transformation was complete, a histogram and Q-Q plot were created to check the normality assumption for the newly transformed expenditures variable **(Appendix A)**.

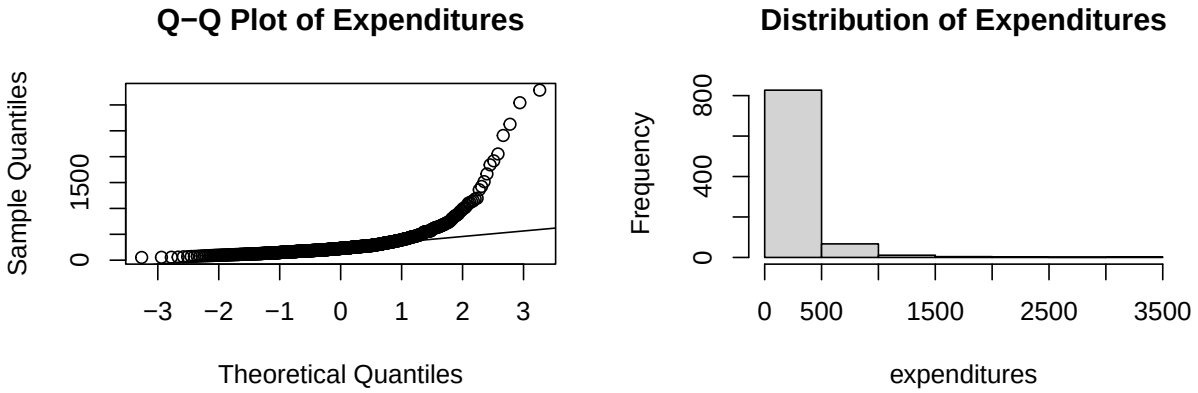

Figure 2: The Q-Q Plot of the expenditure variable on the left helps determine the normality of the explanatory variable. Since the points are generally not close to the Q-Q line (particularly at the tail ends) it can be concluded the expenditure variable is not normally distributed. This is confirmed with the histogram on the right, confirming the expenditure variable is heavily skewed and is not normally distributed.

When recreating the histogram for the log expenditures variable, a normal approximation curve was added to the figure to compare the normality of the log-transformed variable. From Figure 3, the log transformation of the expenditures variable appears to be normally distributed in the histogram, with the density smooth curve very similar to the normal approximation curve. The Q-Q plot for the log-transformed expenditures variable shows the data is much closer to the Q-Q line. While it is not perfect, the tails of the graph are not too extreme. After viewing the log-transformation by these two plots, it can be concluded the log expenditure variable satisfies normal distribution **(Appendix B)**. Next, histograms of each response variable comparing the relationship between log expenditures and each response variable were created to determine the potential distribution of the data. From the histograms, it appeared none of the variables were normally distributed as they were heavily skewed right. It was decided to complete a log transformation of each variable and compare the log transformed histograms with the original histograms created **(Appendix C)**. From the histograms, the log transformations of the variables appear to be normally distributed. In addition, paired scatter plots of each transformed variable with the log expenditure variable were completed to determine the relationships of the log-transformed variables.

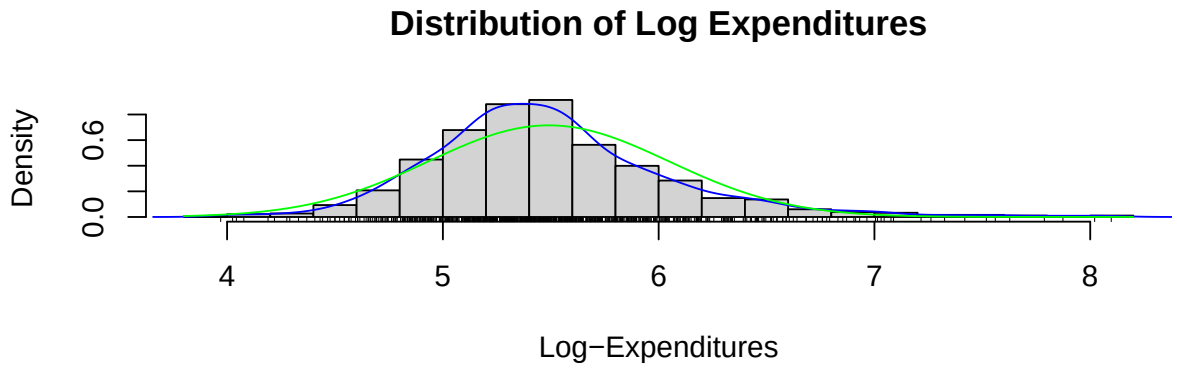

Figure 3: The histogram on the Log Expenditures shows the density smooth in blue and the normal approximation curve in green. The density is similar to the normal curve, which means it can be concluded the log expenditures data is more normally distributed.

When comparing the log transformations of the response variables to log expenditures, it was determined to create a an additional transformation for log income and log percent intergovernmental since they appear to be nonlinear to log expenditures (quadratic transformation for log income and cubic transformation for log percent intergovernmental), while log wealth and log growth rate appear to have linear relationships with our explanatory variable and do not need further transformation **(Appendix D)**. Several new variables are created for the regression model ($logpint^2$, $logpint^3$, and $logincome^2$) to later determine if further transformation of these variables help the prediction model.

Regression Modeling

Now that the variables have been established, the next step in the modeling process is to create an appropriate regression model. Multiple regression models were created and evaluated to determine the best model based on AIC criteria, while also checking BIC and C_p Mallow's regression fit to help determine how many variables to include in the final model. First, a full regression model was run on log expenditures to determine the impact of each individual response variable, including choosing a model based on AIC by stepwise regression. Once run, the AIC report is generated to determine where to minimize the AIC in the model. The stepwise regression both adds and removes the response variables to the regression model, determining the impact of the variables when fit into the model. The final result is the regression model that best minimizes the AIC.

Multiple model attempts were performed by slowly removing non-impactful response variables from the model and performing AIC calculations. The linear regression model chosen removed $logincome^2$ and $logpint^3$, resulting in the selection of the following seven response variables: log wealth, log population, log pint, log pint2, log density, log income, and log growth rate. In addition to the AIC model, a best subsets search was performed, using BIC and C_p Mallow's regression fit to determine the ideal number of variables. To determine the ideal number of variables, the best subsets search determines the BIC and C_p values for each amount of variables in the model, with the lowest possible values determined to be the best model fit. From the BIC and C_p model, the

ideal number of variables was 7 variables for C_p Mallow's and 5 variables for BIC, which is in line with what was found using the AIC model.

Next, there needs to be a check if the correlation of the independent variables may have a potential impact on the regression model via multicollinearity, as noted after creating the correlation plots at the beginning of the EDA cycle. Now that there is a potential linear regression model, a VIF plot is completed to determine if there is multicollinearity among the response variables. A VIF value above 7 would raise concern, while a value above 10 indicates multicollinearity.

Table 1: VIF scores of the response variables.

Variable	VIF Score
lwealth	2.05
lpopulation	6.40
lpint	1.32
ldens	7.81
lincome	2.94
lgrowr	1.03

From Table 1, most of the variables have a low VIF score, indicating low multicollinearity between the response variables and not a concern for our model. The variable of concern is log density with a VIF score of 7.81, which hints that multicollinearity may exist and become a factor in the model. As a test, I removed the log density variable from the model to see if the model improved, which was not the case as the AIC was not minimized. Thus, I chose not to remove log density from the model since the VIF score was not above 10 and the regression model did not improve. As such, the regression model is the following:

Variable	Coefficients	Standard Error	P-Value	95% CI
Intercept	-0.88	0.80	0.27	(-2.47 , 0.70)
LWealth	0.40	0.05	0.01	(0.30 , 0.51)
LPopulation	0.16	0.04	0.01	(0.08 , 0.24)
LPint	-1.26	0.32	0.01	(-1.88 , -0.63)
$LPint^2$	0.19	0.06	0.01	(0.06 , 0.31)
LDensity	-0.09	0.04	0.03	(-0.17 , -0.01)
LIncome	0.30	0.11	0.01	(0.09 , 0.52)
LGrowr	-0.03	0.01	0.03	(-0.05 , -0.01)

$$Adjusted R^2 = 0.63, \quad \text{AIC} = -533.54$$

Table 2: Table of the regression model displaying the Coefficient values, standard error, p-values, and 95% confidence intervals of each response variable. All coefficients are statistically significant based on p-value and confidence intervals.

As seen from Table 2, each response variable is statistically significant with a p-value less than 0.05. In addition, the coefficients describe the relationship between the response variable and log expenditures. For example, as log wealth increases by one unit, this means log expenditures will increase by 0.40 units. Another example is as log density increases by one unit, log expenditures decreases by 0.09 units. From the 95% confidence intervals, each variable does not contain zero, which means the variable is statistically significant in the regression model. Also, the adjusted R^2 of 0.63 means 63% of the variance in the regression model can be explained.

Model Diagnostics

Now that a regression model has been developed, the next step is to perform model diagnostics to determine if the model meets normality assumptions, as well as determine if there are any outliers or influential points of concern. The first diagnostic reports run is a studentized residuals plot and a Q-Q plot for the studentized residuals. The studentized residuals compare the observed and

predicted values in a regression model, which helps identify potential outliers and influential points in the model. As stated earlier, the Q-Q plot helps determine the normality of the data as well as identify potential outliers and influential points.

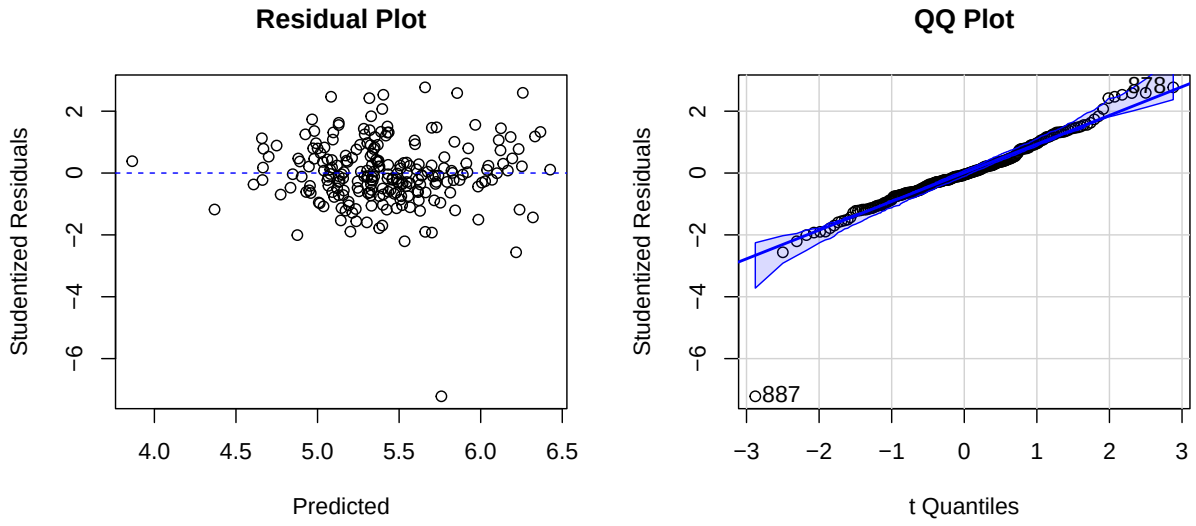

Figure 4: A studentized residual plot and Q-Q plot to identify and determine the potential effect of outliers and influential points in the regression model. It appears observation 887 has a very large influence on the model and may be an outlier.

From the studentized residuals and Q-Q plot, the data appears to be normally distributed but contains a potential influential point in observation 887. In addition to the above plots, a Cook's distance plot is created to investigate potential influential points **(Appendix E)**. At this point, an argument can be made to remove observation 887 from the data and rerun the regression model. However, since it is a real data point, is not an empty value, and does not appear to be a typo or egregious error when viewing its values, I chose to leave the observation in the regression model and continue forward with the prediction process.

Prediction for Monroe and Warwick

Now that a regression model has been developed and model diagnostics have been run, the predictions for expenditures in 1992, 2002, and 2025 can be entered for Monroe and Warwick municipalities. First, the standard deviation was fit on the residuals of the regression model, which will help in the construction of the confidence interval. Next, a data frame was created for each prediction year for each municipality, making it easier to enter the individual prediction values. For each data frame, the prediction values given by the municipalities are entered into the respective variables. Below are the log expenditure predictions for Monroe and Warwick municipalities for the given years:

Municipality	Year	Prediction($)	95% CI
Monroe	1992	237	(129 , 437)
	2005	244	(133 , 450)
	2025	242	(131 , 445)
Warwick	1992	265	(143 , 489)
	2005	287	(155 , 531)
	2025	302	(163 , 560)

Table 3: Predictions of Log Expenditures for Monroe and Warwick for years 1992, 2005, and 2025, as well as a 95% confidence interval for each prediction.

From the predictions, Warwick appears to have higher expenditure predictions than Monroe for all three years, though Monroe has a tighter 95% confidence interval in its prediction compared to Warwick. As such, the model expects Warwick municipality to have higher expenditures and an increase in expenditures over time, but the model appears to have a more precise estimate for the Monroe municipality. With higher predicted expenditures over time, Warwick may need to explore potential funding increases to offset costs. For Monroe, their predictions don't increase dramatically over time, which means they most likely do not need to explore funding increases at this time.

Conclusion

Based on the needs of Monroe and Warwick and exploratory data analysis conducted on the given data, a linear regression model was developed to predict the expenditures for the years 1992, 2005, and 2025. This model determined the log transformations of wealth per person, population, percent intergovernmental (both linear and quadratic terms), density, mean income per person, and growth rate were statistically significant in the prediction of log expenditures per person. Once the log forms are extrapolated, the model believes Warwick municipality will have higher expenditures and an increase over time but is more likely to be precise in predicting the expenditures for Monroe based on the smaller confidence interval. From these results, Warwick may want to explore future funding and offset these costs, while Monroe may not need to explore additional funding since the expenditures remain relatively the same over a long period of time.

While this was the prediction model chosen for this report, different interpretations and analyses can improve the prediction model. For example, the predictor variables can undergo a multitude of transformations based on the author's interpretation. Instead of splitting the data for population and density and creating a prediction model based on the subset, the full data set can be used and a nonlinear regression model or additional transformation on the variables can be performed. These decisions can change and potentially help improve the prediction model. In addition, choosing to remove observations can lead to a different model. It was determined not to remove observation 887 in this report, but an argument can be made to do so, which may influence a change in the coefficients of the regression model and different predictions for the municipalities. One of the challenges of this study is the lack of variables provided. Adding additional demographic and income-related variables may help improve prediction models for future studies, particularly for events decades in the future. Overall, the regression model presented is, at the very least, a good starting point for the municipalities. The regression model can then be adjusted accordingly depending on any additional factors the municipalities want to consider before adjusting their respective budgets.

Design Statement

During the preparation of the lab the author used ChatGPT for code debugging and assistance in the creation and placement of the figures and tables in LaTeX. ChatGPT presented additional packages for LaTeX studio to help with the formation of tables and print out the PDF file. After using this service, the author reviewed and edited the content as needed and takes responsibility for the content of this lab.

Appendix

Appendix A

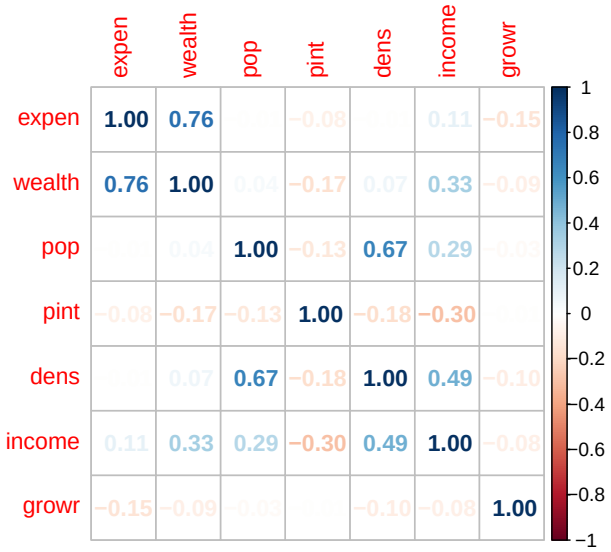

Figure 5: Correlation plot of expenditures, wealth, population, percent intergovernmental, density, income, and growth rate using numerical values. Most variables do not seem to be correlated with one another and thus are independent of one another. Note correlation values of expenditures with wealth and population with density may impact the model.

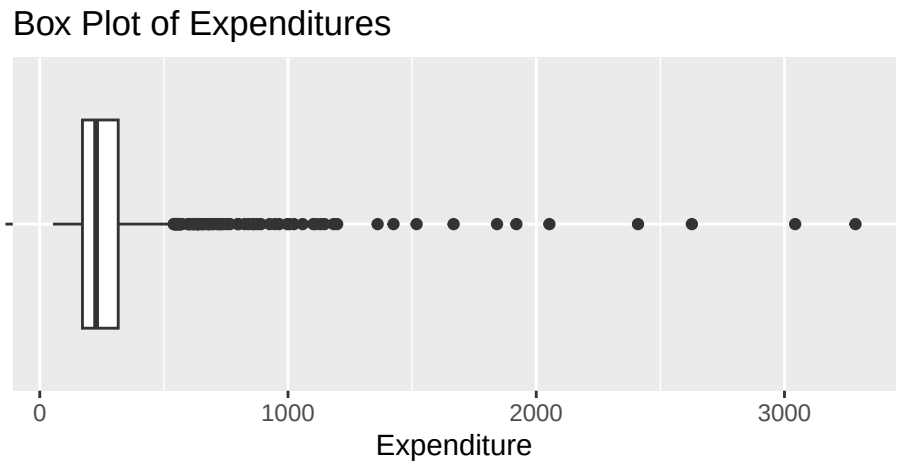

Figure 6: Boxplot of the expenditures shows a heavy right-skew in the data. Further analysis will determine the expenditures variable is not normally distributed and thus needs a transformation.

Appendix B

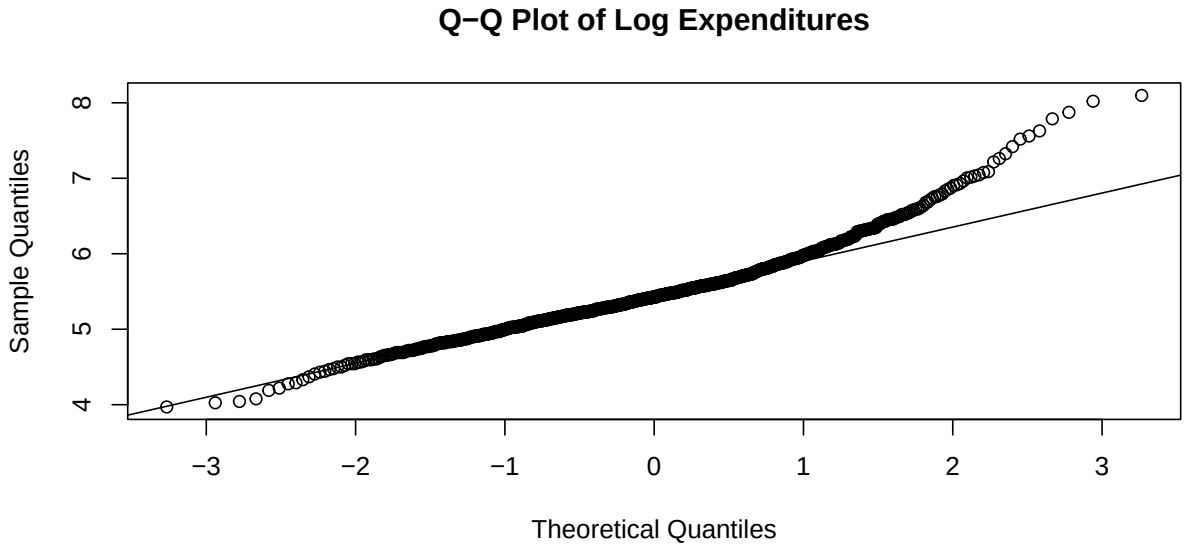

Figure 7: The Q-Q Plot on the right of Log Expenditures follows the Q-Q line much better, appearing to be more normally distributed.

Appendix C

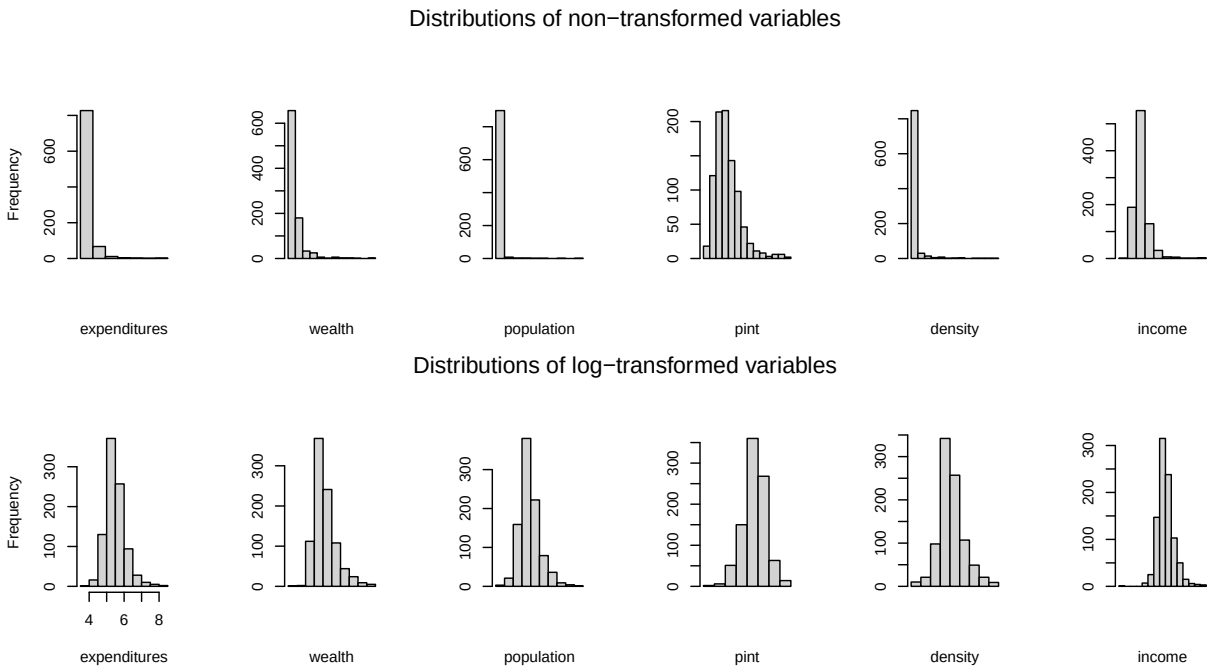

Figure 8: The histograms of non-transformed variables show each variable is not normally distributed. Creating a log transformation of each variable causes each one to follow a normal distribution.

Appendix D

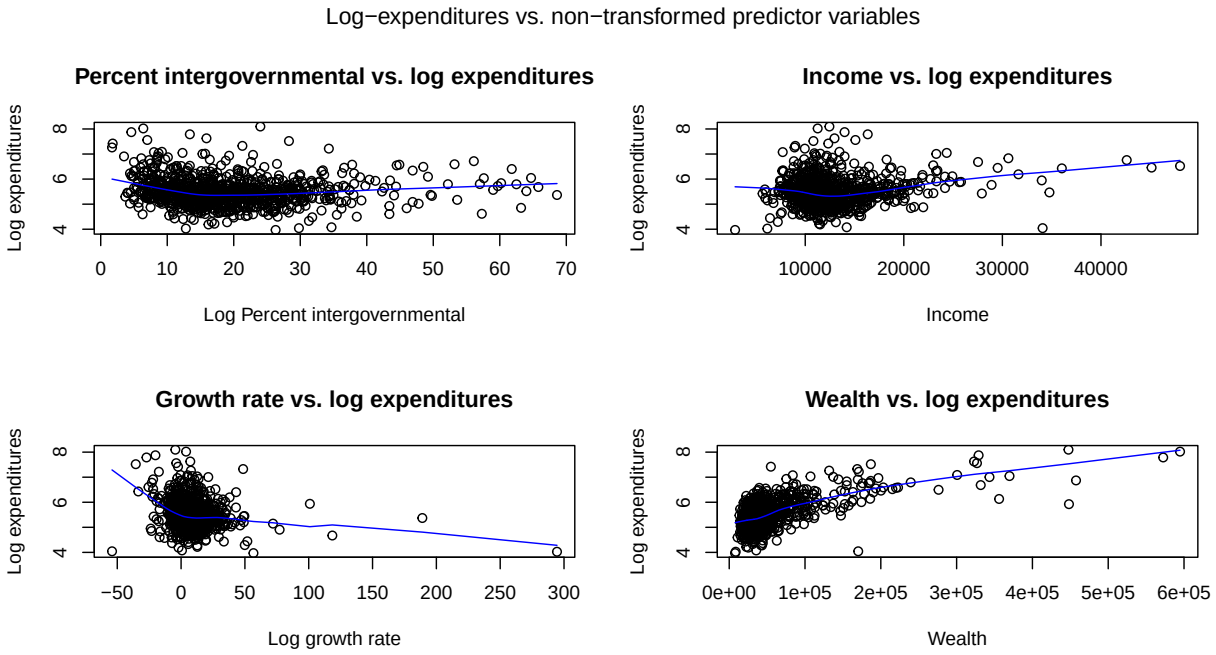

Figure 9: Scatter plots of log-expenditures vs. the remaining predictor variables.From the scatter plots, it is difficult to determine the relationship between the predictors and log expenditures. It is beneficial to log-transform the predictor variables to best see the relationships.

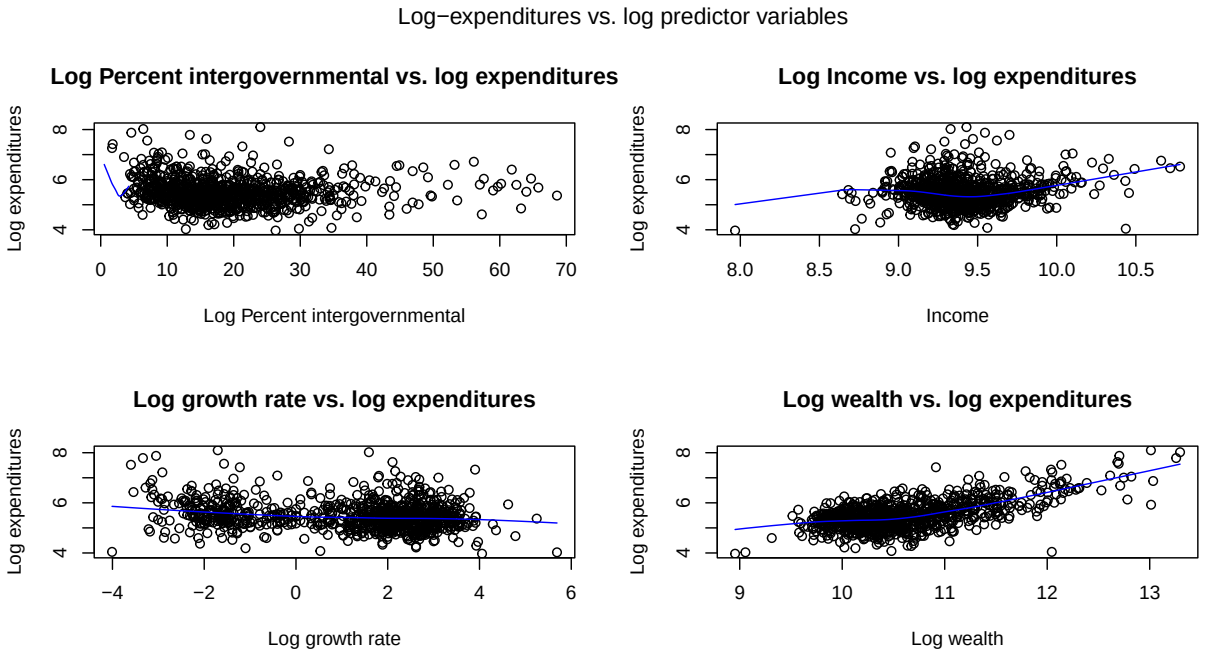

Figure 10: Scatter plots of log-expenditures vs. the remaining log-transformed predictor variables. Log-wealth and Log-growth rate appear linear, while log-percent intergovernmental appears non-linear and would need further transformation.

Appendix E

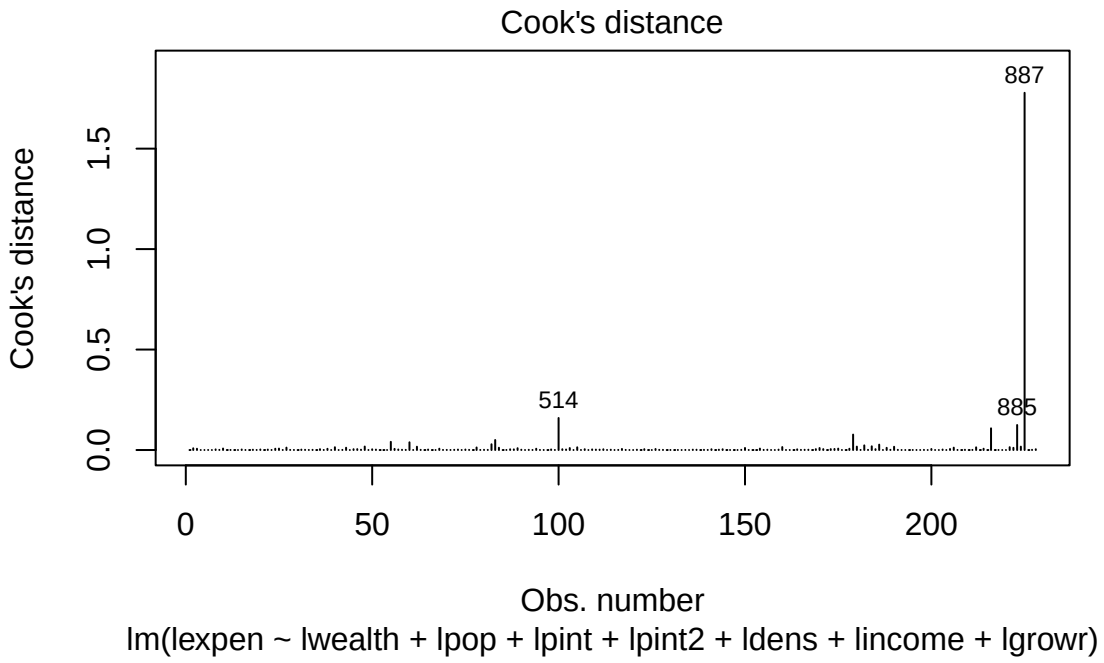

Figure 11: Cook's Distance plot displaying the influence of observation 887.

R Code

```
library(MASS)
library(corrplot)
library(car)
library(leaps)
library(vioplot)
library(ggplot2)
library(xtable)

# Read in data, remove missing data
ny<-read.table("~/Masters Program/Stat 794/Labs/EDA lab/cs73.dat",header=T); dim(ny)
      #  916  11
ny2<-na.omit(ny); dim(ny2) # 914  11
attach(ny2)

## review data set: expen is response
names(ny2)
head(ny2, n=5)

#Log transformations
lexpen<-log(expen)
lwealth<-log(wealth)
lpop<-log(pop)
ldens<-log(dens)
lincome<-log(income)
lpint <- log(pint)
pint2<-pint**2
pint3<-pint**3
lpop2<-lpop**2
lpop3<-lpop**3
ldens2<-ldens**2
ldens3<-ldens**3
i2 = income^2
i3 = income^3
lpint2<-lpint**2
lpint3<-lpint**3
lincome2 = lincome**2
lgrowr<-ifelse(growr>0, log(growr+1), -log(-growr+1))

#Correlation plot
ny2vars = data.frame(expen, wealth, pop, pint, dens, income, growr)
cny2 = cor(ny2vars)
corrplot(cny2, method = 'number')
#Histogram plot of expenditures
hist(expen, breaks = seq(min(expen), max(expen), length.out = 20), xlab = '
    Expenditures', ylab = 'Frequency', main = '')
#Violin plot of expenditures
vioplot(expen)
#Q-Q plot of expenditures
qqnorm(expen, main = "Q-Q Plot of Expenditures")
qqline(expen)

# histogram smoothing
lexpen = log(expen)
# histogram with density plot overlay
```

```
hist(lexpen, prob=T, breaks = 20, xlab="Log-Expenditures", ylab="Density", main="")
# density smooth
lines(density(lexpen), col="blue")
rug(lexpen) # data rug
# normal approximation
curve(dnorm(x, mean=mean(lexpen), sd=sd(lexpen)), add=TRUE, col="green")

par(mfrow=c(2,6), oma = c(0,0,2,0))
# First histogram of non-transformed variables
hist(expen, main = " ", xlab = 'expenditures', xaxt = 'n')
hist(wealth, main = " ", xlab = 'wealth', ylab = "", xaxt = 'n')
hist(pop, main = " ", xlab = 'population', ylab = "", xaxt = 'n')
hist(pint, main = " ", xlab = 'pint', ylab = "", xaxt = 'n')
hist(dens, main = " ", xlab = 'density', ylab = "", xaxt = 'n')
hist(income, main = " ", xlab = 'income', ylab = "", xaxt = 'n')
# Then histogram of log-transformed variables (except Growth Rate)
hist(lexpen, main = " ", xlab = 'expenditures')
hist(lwealth, main = " ", xlab = 'wealth', ylab = "", xaxt = 'n')
hist(lpop, main = " ", xlab = 'population', ylab = "", xaxt = 'n')
hist(log(pint), main = " ", xlab = 'pint', ylab = "", xaxt = 'n')
hist(ldens, main = " ", xlab = 'density', ylab = "", xaxt = 'n')
hist(lincome, main = " ", xlab = 'income', ylab = "", xaxt = 'n')
#Title to separate non-transformed graphs from log-transformed graphs
mtext('Distributions of non-transformed variables', side = 3, line = 0, outer = TRUE
    )
mtext('Distributions of log-transformed variables', side = 3, line = -18, outer =
    TRUE)

#Q-Q plot of log-expenditures
qqnorm(lexpen, main = "Q-Q Plot of Log Expenditures")
qqline(lexpen)

#lexpen vs. pint scatter plot
plot(pint, lexpen, xlab = 'Percent intergovernmental', ylab = 'Log expenditures',
    main = 'Scatter plot of percent intergovernmental vs. log expenditures')
lines(lowess(pint,lexpen), col="blue") # using a LOWESS scatter plot smooth
#lexpen vs. income scatter plot
plot(income, lexpen, xlab = 'Income', ylab = 'Log expenditures', main = 'Scatter
    plot of income vs. log expenditures')
lines(lowess(income,lexpen), col="blue") # using a LOWESS scatter plot smooth

par(mfrow = c(2,2), oma = c(0,0,2,0))
#lexpen vs. lpop scatter plot
lpop = log(pop)
plot(lpop, lexpen, xlab = 'Log population', ylab = 'Log expenditures')
lines(lowess(lpop,lexpen), col="blue") # using a LOWESS scatter plot smooth
lines(c(8.3,8.3),c(0,6), col="grey", lwd=3)
#lexpen vs. lwealth scatter plot
lwealth = log(wealth)
plot(lwealth, lexpen, xlab = 'Log wealth', ylab = "")
lines(lowess(lwealth,lexpen), col="blue") # using a LOWESS scatter plot smooth
lines(c(8.3,8.3),c(0,6), col="grey", lwd=3)
#lexpen vs. lgrowr scatter plot
plot(lgrowr, lexpen, xlab = 'Log growth rate', ylab = 'Log expenditures')
lines(lowess(lgrowr, lexpen), col="blue") # using a LOWESS scatter plot smooth
lines(c(8.3,8.3),c(0,6), col="grey", lwd=3)
```

```
#lexpen vs. ldens scatter plot
plot(ldens, lexpen, xlab = 'Log of density', ylab = "")
lines(lowess(ldens,lexpen), col="blue") # using a LOWESS scatter plot smooth
lines(c(4.5,4.5),c(0,6), col="grey", lwd=3)
mtext('Scatter plots of log-expenditures vs. log predictor variables', side = 3, line = 0, outer = TRUE)

#New data subset with high population & high density areas
set2<-(lpop>8.3 & ldens>4.5)
ny2vars = data.frame(lexpen,lwealth,lpop,lpint,ldens,lincome,lgrowr)
#Final regression model chosen
fit2c = lm(lexpen~lwealth+lpop+lpint+lpint2+ldens+lincome+lgrowr, data = ny2vars, subset=set2)
summary(fit2c)
stepAIC(fit2c, direction="both") #get same final model as fit2c

#VIF plot of the variables - no value above 10
vif_model <- lm(lexpen~lwealth+lpop+lpint+ldens+lincome+lgrowr)
vif(vif_model)

#Removed log density; model performed worse.
fit3 = lm(lexpen~lwealth+lpop+lpint+lpint2+lpint3+lincome+lgrowr, data = ny2vars, subset=set2)
summary(fit3)
stepAIC(fit3, direction="both") #get same final model as fit2c

#BIC and Cp regression
regfit =regsubsets(lexpen~lwealth+lpop+lpint+lpint2+lpint3+ldens+lincome+lincome2+lgrowr, data = ny2vars, subset=set2, nvmax=10)
summary(regfit)
par(mfrow=c(2,1))
plot(summary(regfit)$cp, xlab="Number of Variables", ylab="Cp")
plot(summary(regfit)$bic, xlab="Number of Variables", ylab="BIC")
which.min(summary(regfit)$cp); which.min(summary(regfit)$bic)

plot(predict(fit2c), rstudent(fit2c), ylab="Studentized Residuals", xlab="Predicted")
identify(predict(fit2c), rstudent(fit2c), labels=row.names(ny2)) # 'escape to finish'
predict(fit2c)[rstudent(fit2c)==min(rstudent(fit2c))]


# Q-Q plot for studentized resid
qqPlot(fit2c, main="QQ Plot", ylab="Studentized Residuals", id = list(method = "y", print = FALSE))

cutoff <- 4/((nrow(ny2)-length(fit2b$coefficients)-2))
plot(fit2c, which=4, cook.levels=cutoff) # influence Plot
# Influence plot: studentized residuals vs. hat matrix diagonals (leverage) with bubbles a function of Cook's D
# Interactive, so can click to identify high leverage/influential/outlying points
influencePlot(fit2c, id.method="identify",
              main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )
#XTable with Variable, Coefficients, Standard Error, P-Value, and 95% Confidence Interval
```

```
coefficients <- summary(fit2c)$coefficients
names = rbind("Intercept", "LWealth", "LPopulation", "LPint", "LPint2", "LDensity", "LIncome", "LGrowr")
coef = cbind(coefficients[,1])
standard_error = cbind(coefficients[,2])
p_value = cbind(coefficients[,4])
conf_intervals <- confint(fit2c)
lwr = rbind(conf_intervals[,1])
upr = rbind(conf_intervals[,2])

lm.data <- cbind(names, coef, standard_error, p_value, matrix(paste("(", lwr, ",", upr, ")")))
colnames(lm.data) <- c('Variable', 'Coefficients', 'Standard Error', 'P-Value', "95% CI")
lm.table = xtable(lm.data, digits=3,
                  caption="Table of the regression model displaying the Coefficient values, standard error, p-values, and 95% confidence intervals of each response variable.",
                  label="reginf")
align(lm.table) <- "|l|rrrrr|"
print(lm.table, include.rownames = FALSE)

#Standard deviation of predictions, followed by the setup of Warwick predictions
sdfit=sd(fit2c$resid)

war92 = data.frame(lwealth=log(72908), lpop=log(16225), lpint=log(24.7), lpint2=log(24.7)^2, ldens= log(170),
                   lincome = log(19044),lgrowr=log(30.3+1))
war05 = data.frame(lwealth=log(85000), lpop=log(20442), lpint=log(24.7), lpint2=log(24.7)^2, ldens= log(214),
                   lincome = log(19500),lgrowr=log(35+1))
war25 = data.frame(lwealth=log(89000), lpop=log(31033), lpint=log(26.0), lpint2=log(26.0)^2, ldens= log(325),
                   lincome = log(20000),lgrowr=log(40+1))
#1992 Warwick Prediction
warwick92=predict.lm(fit2c,war92); exp(warwick92+sdfit^2/2)
warwick92_pred = exp(predict(fit2c, war92, interval="prediction")+sdfit^2/2)
#2005 Warwick Prediction
warwick05=predict.lm(fit2c,war05); exp(warwick05+sdfit^2/2)
warwick05_pred = exp(predict(fit2c, war05, interval="prediction")+sdfit^2/2)
#2025 Warwick Prediction
warwick25=predict.lm(fit2c,war25); exp(warwick25+sdfit^2/2)
warwick25_pred = exp(predict(fit2c, war25, interval="prediction")+sdfit^2/2)

#Standard deviation of predictions, followed by the setup of Monroe predictions
sdfit=sd(fit2c$resid)

mon92 = data.frame(lwealth=log(55067), lpop=log(9338), lpint=log(8.8), lpint2=log(8.8)^2, ldens= log(599),
                   lincome = log(16726),lgrowr=log(30.0+1))
mon05 = data.frame(lwealth=log(58000), lpop=log(10496), lpint=log(8.8), lpint2=log(8.8)^2, ldens= log(695),
                   lincome = log(17100),lgrowr=log(35.0+1))
mon25 = data.frame(lwealth=log(60000), lpop=log(13913), lpint=log(10.1), lpint2=log(10.1)^2, ldens= log(959),
                   lincome = log(18000),lgrowr=log(35.0+1))
```

```r
#1992 Monroe Prediction
monroe92=predict.lm(fit2c,mon92); exp(monroe92+sdfit^2/2)
monroe92_pred = exp(predict(fit2c, mon92, interval="prediction")+sdfit^2/2)
#2005 Monroe Prediction
monroe05=predict.lm(fit2c,mon05); exp(monroe05+sdfit^2/2)
monroe05_pred = exp(predict(fit2c, mon05, interval="prediction")+sdfit^2/2)
#2025 Monroe Prediction
monroe25=predict.lm(fit2c,mon25); exp(monroe25+sdfit^2/2)
monroe25_pred = exp(predict(fit2c, mon25, interval="prediction")+sdfit^2/2)
```

Listing 1: R Source Code