Expenditures of Warwick and Monroe

Kelso Quan

December 2, 2018

Abstract

The study wanted to predict expenditures of New York cities, Warwick and Monroe. By gaining insight, the city council would like to know if its city will be spending more money due to proposed construction of new housing projects and a possible growing population within city limits. While the cities are small to begin with, it is still a good idea to prepare for the future whether that be raising property taxes or looking for other sources of revenue. Looking into six predictor variables, a log linear trend was found between the resulting log transformed expenditure of the two cities and its six indicators. Furthermore, the pattern continued into the 2005 and 2025 predicted expenditures. It is noted that the predictions were made with a subset that required ($\log(pop) > 8.3$ & $\log(dens) > 4.5$). Using the all six log transformed predictors, it was possible to come up with expenditure predictions. It was evident that there was a log positive trend with the six predictors (population, density, "percent intergovernmental", income, growth rate, and wealth), and the expenditures of both towns were going to increase over time.

1 Introduction

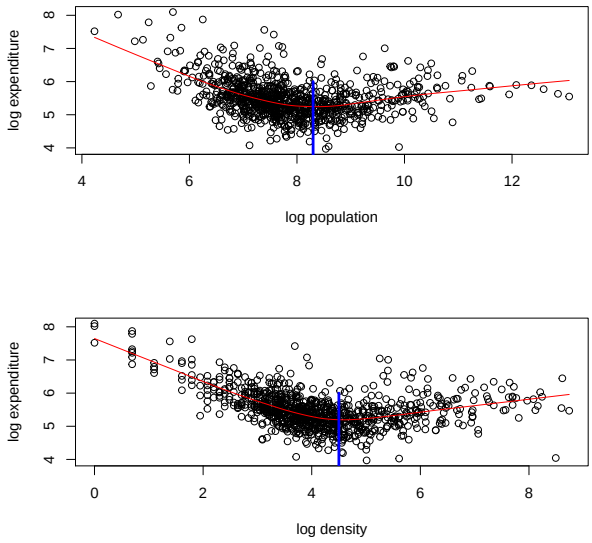
In 1992, data was taken over a number of municipalities in New York state. By using the 914 observations taken, Warwick and Monroe city council should indeed plan for slightly higher expenditures by the city to keep up with the growing population possibly due to the planned housing projects. The 99% confident intervals of each coefficient can be seen in Table 1. These figures have been rounded. Please see the appendix for more exact numbers. The numbers in Table 1 have been transformed back. For example, every 1 unit (thousands) of wealth increased, the expenditure of Warwick and Monroe goes up by roughly \$1.50.

Pint stands for "Percent Intergovernmental" which represents the percentage of revenue coming from state and federal grants subsidies. From Table 2, we have the prediction of expenditures in thousands. For exact figures, refer to the appendix. There is a slight positive linear trend while looking at the prediction intervals for Monroe, but it is noticeably more pronounce in Warwick's prediction of expenditures.

2 Methods

The data was drawn from 916 municipalities in the state of New York in 1992. Two observations (i.e. Warwick and Monroe) were missing dataset. The purpose of the analysis is to predict expenditures of Warwick and Monroe from the data of 914 other municipalities. There was a municipality of interest which was included in the analysis. The observation 225 is influential point shown in figure 3. While it is possible to remove this influential point, observation 225 was still a municipality of New York state. Removing it didn't change the model selection and thus was kept. The model selection and all analysis was done in R Studio to predict future expenditures.

Figure 1: Log Dens and Log Pop before subsetting



3 Results

3.1 Exploratory Data Analysis

A boxcox transformation confirms that the response should be log transformed since λ is near zero. Creating a histogram of the response variable expenditure shows a need for a log transformation. After the log transformation, the histogram of log-expenditure in figure 2 shows lexpen is approximately normal. Taking a simple look at the predictor variables showed that they needed to be log transformed at least once before proceeding further. Histograms can be seen in figure 6 showed that there was heavy right skewness to all predictors before log transformations. After log transformation, seen in figure 7 showed nicer normality of each predictor.

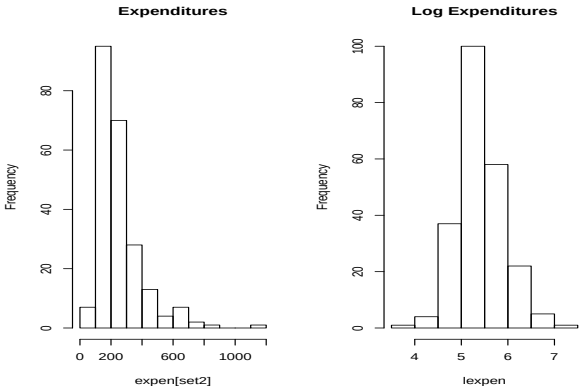
Afterwards, the variable log expenditure was plotted against each of its predictor illustrated in figure 8. Each plot looked fairly linear. There didn't seem like an clear trends while plotting the residuals, thus no violation of homoscedasticity which can be seen in figure 5. There's no clear trends in the residuals and appeared to be random which is good news for linear regression. The QQ plot appears linear which suggest that the data is approximately normal after log transformations.

In figure 1, shows that there's a possible quadratic function, but the data could also be subsetted. This analysis focused on the subsetted data where $\log(\text{pop}) > 8.3$ and $\log(\text{dens}) > 4.5$ before more exploratory analysis was done. It is clear that the data must be subsetted. The blue lines shown in figure 1, where the data should be subsetted.

3.2 Model Fitting/Inferences

After log transforming all variables, a correlation matrix should be examined. The correlation matrix showed that there were some highly correlated covariates shown in the appendix. Log pop and log density .8 correlation with log wealth and log income had .75 correlation. Such highly correlated variables could be dropped from the analysis due to parsimony, but a quickly glance at the Variance Inflation Factor table 3 shows no serious violations of numbers being greater than 7. With that said, the analysis kept all of its

Figure 2: Log Transforming Response Variable: Expenditures



predictors.

Using the function stepAIC, it was shown that the linear log log model was the best fit and modeled the data with the least amount of penalties. stepAIC is a function that fits possible models from the data given. Every possible model is given a penalty marker. The least amount of penalties, the better the model comes out to be. Looking at the stepAIC results were preferable compared to looking at coefficients from summary of the model and then manually dropping predictors one at a time before achieving the best fitted model. This fit was reasonable because all, but the log-transformed density variable had less than a 0.05 p-value. Since the p-value of log-density wasn't that much higher than .05 and .05 was an arbitrary benchmark, the predictor log-density was left in. Further investigation into this model showed an influential point (obs 225) that was left in. While it is feasible to remove an outlier, it was kept in the analysis to account for extremities of the New York state. It is quite possible that observation 225 was Manhattan. With 914 observations, removing a single outlier probably will not change the model selection. Looking at figure 3, the plot shows the predicted model with observation 225. The removal of observation 225 only slightly increases R^2 coefficient. R^2 is a statistical measure of how close the data are to the fitted regression line. I.e. how well the model explains the data set. Thus the outlier was retained in this analysis.

Table 1: 95% Confidence Intervals of model's coefficients

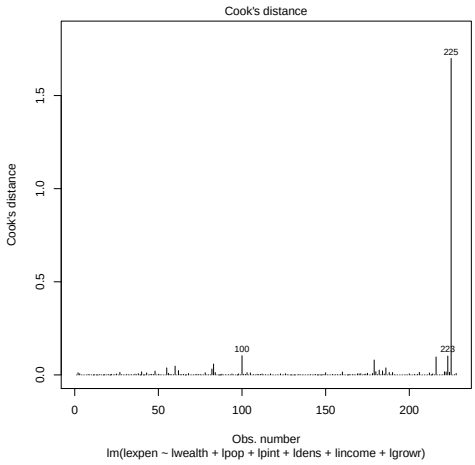
Coefficient	**Lower Bound**	**Upper Bound**
Wealth	1.3	1.7
Pop	1.1	1.3
Pint	0.67	0.79
Dens	0.85	1.0
Income	1.1	1.7
Growr	0.95	0.99

From the model, expenditures by Warwick and Monroe can be predicted for the years 1992, 2005, and 2025. Those predicted expenditure values can be seen in table 2. These predictions should be kept in the minds of the city councils of Monroe and Warwick when they create annual budgets. In addition, the coefficients of each predictor can be found in table 1. While it may look like the prediction intervals and estimated expenditures for Monroe remain fairly constant, the unrounded figures in the table within the appendix shows that there are slight increases over time. This may be due to Monroe being a small town.

Table 2: 95% Predictions Intervals of Warwick and Monroe Expenditures

Town	Year	Expen Est.	Lower Bound	Upper Bound
Warwick	1992	250	130	460
	2005	270	140	500
	2025	280	150	520
Monroe	1992	250	130	460
	2005	250	140	470
	2025	250	140	470

Figure 3: Cook's Distance



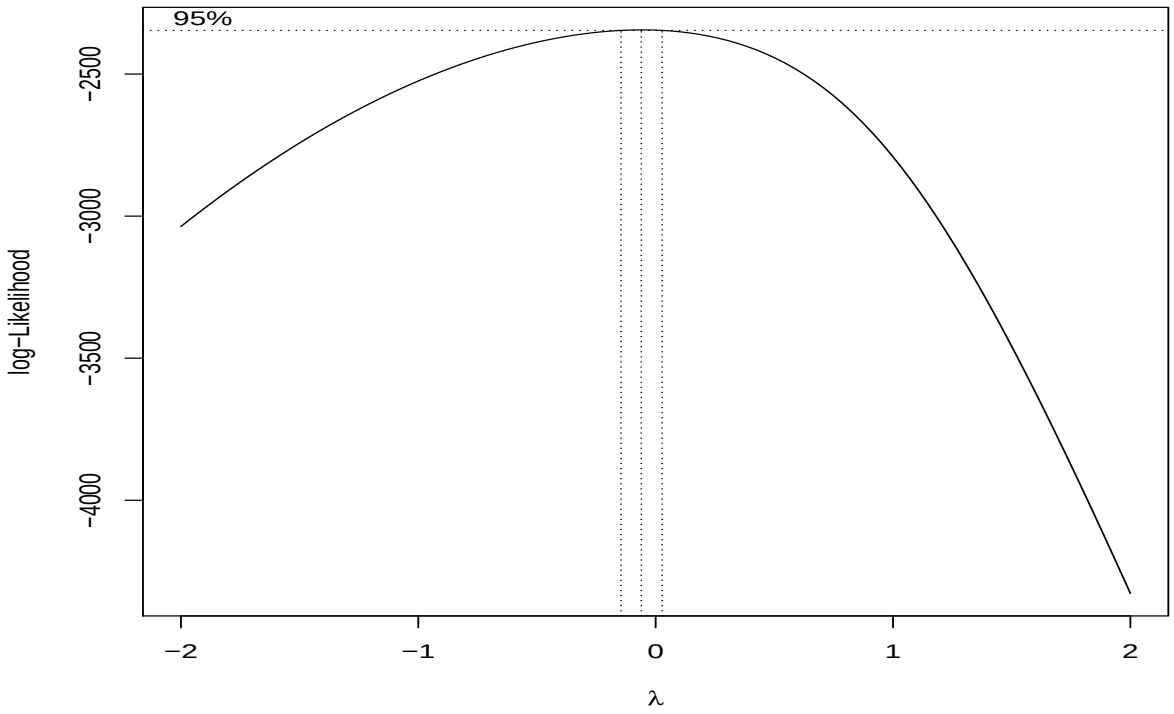
4 Conclusion

In this analysis, it has shown that the expenditures of the New York municipalities Warwick and Monroe will most likely increase due to proposed construction of new housing projects. The two towns should look for methods to increase funds whether it be increasing property tax, increasing sales tax, or another method. Table 2 has shown positive linear trend in expenditures for both towns. Warwick should find sources of revenue quicker than Monroe due to its ever increasing expenditures.

This analysis was limited in three ways. The first way was that the outlier was not taken out of the dataset even though it was looked at. R^2 may have increased by .05 without the outlier, but the model would have stay the same either way. Second, the analysis stuck to a linear log log model. There were a couple of predictors that were worth examining at a higher power. But not having variables log transformed would have prove disastrous. The last limitation was that the data was subsetted ($\log(\text{pop}) > 8.3$ & $\log(\text{dens}) > 4.5$), thus data set with $\log(\text{pop}) < 8.3$ and $\log(\text{dens}) < 4.5$ was not considered. Presently, it is not clear why a subset of that nature should be looked at, but that data set may hold valuable information for Warwick and Monroe. For the future, these three limitations could be explored, but for now, this analysis should suffice.

Since there is data on 1992 and 2005, it is possible to compare those actual values to the analysis' predicted values. It would be very interesting to see if those values lined up. If the values are no where close, then another model must be considered. In addition, would this model hold up to states similar to New York? With a little further investigation, surely, that question can be answered.

Figure 4: BoxCox Transformation of Response Variable

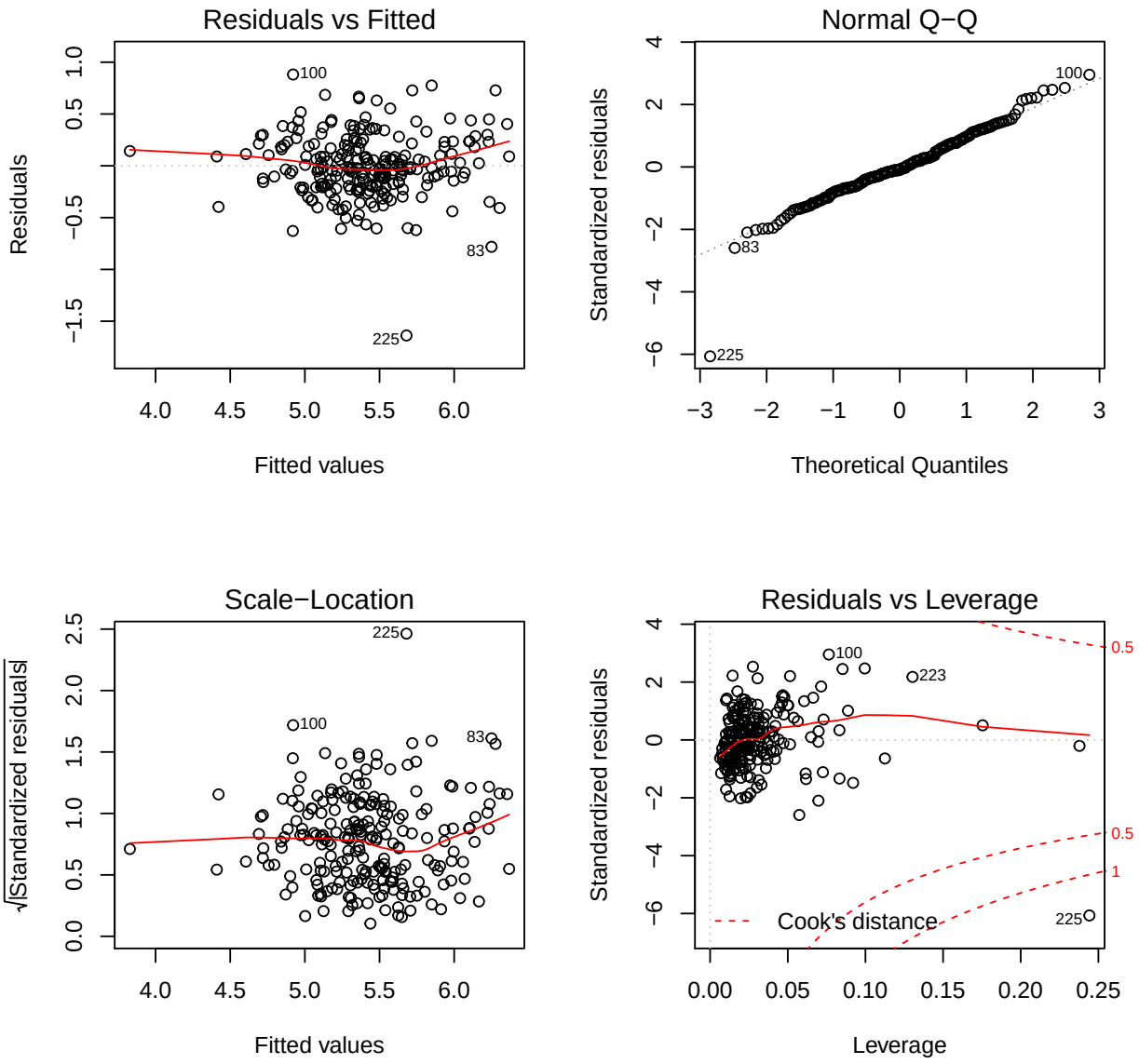



```
## [1] -0.06060606
```

Appendix A: Auxiliary Graphics and Tables

```
## corrplot 0.84 loaded
```

Figure 5: Homoscedasticity



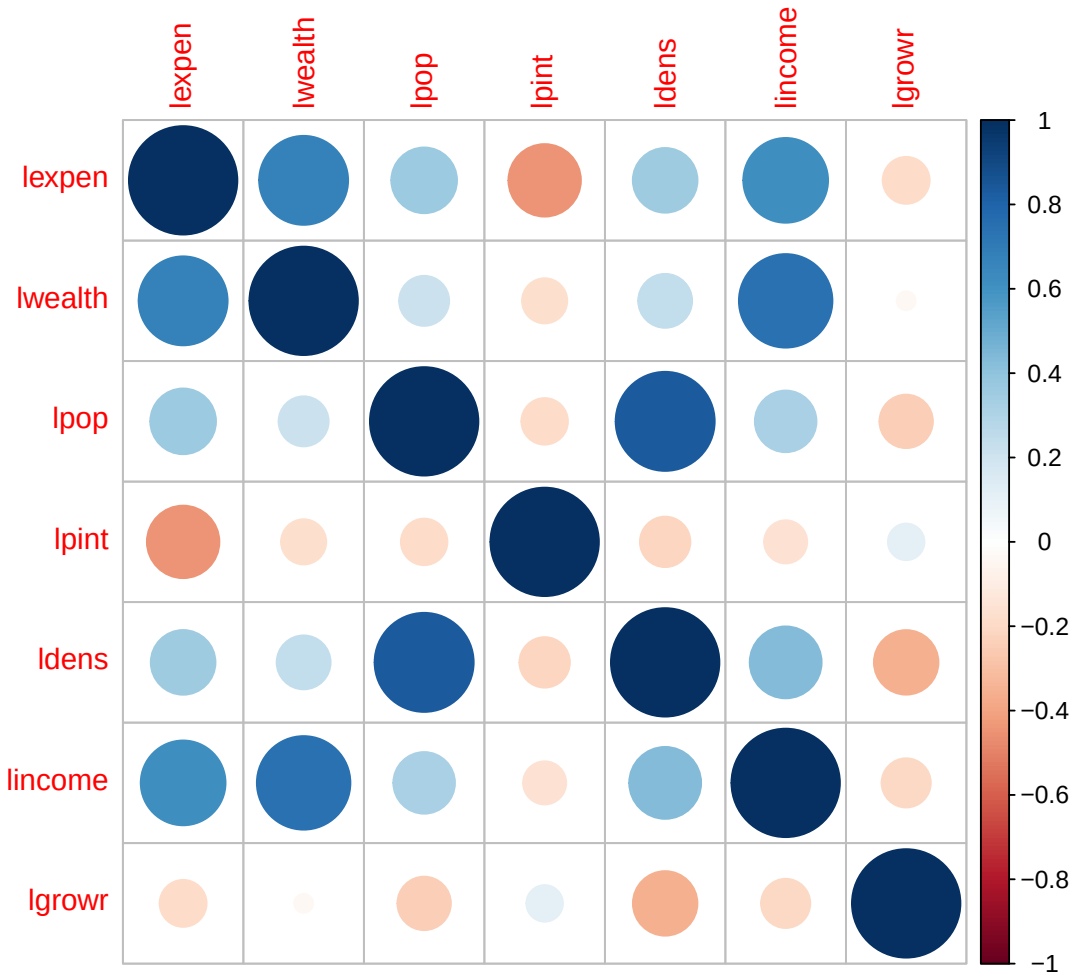


Table 3: Variance Inflation Factor

Lwealth	**Lpop**	**Lpint**	**Ldens**	**Lincome**	**Lgrowr**
2.4	3.4	1.1	4.0	2.8	1.2

Figure 6: Histograms of Not Transformed Predictors

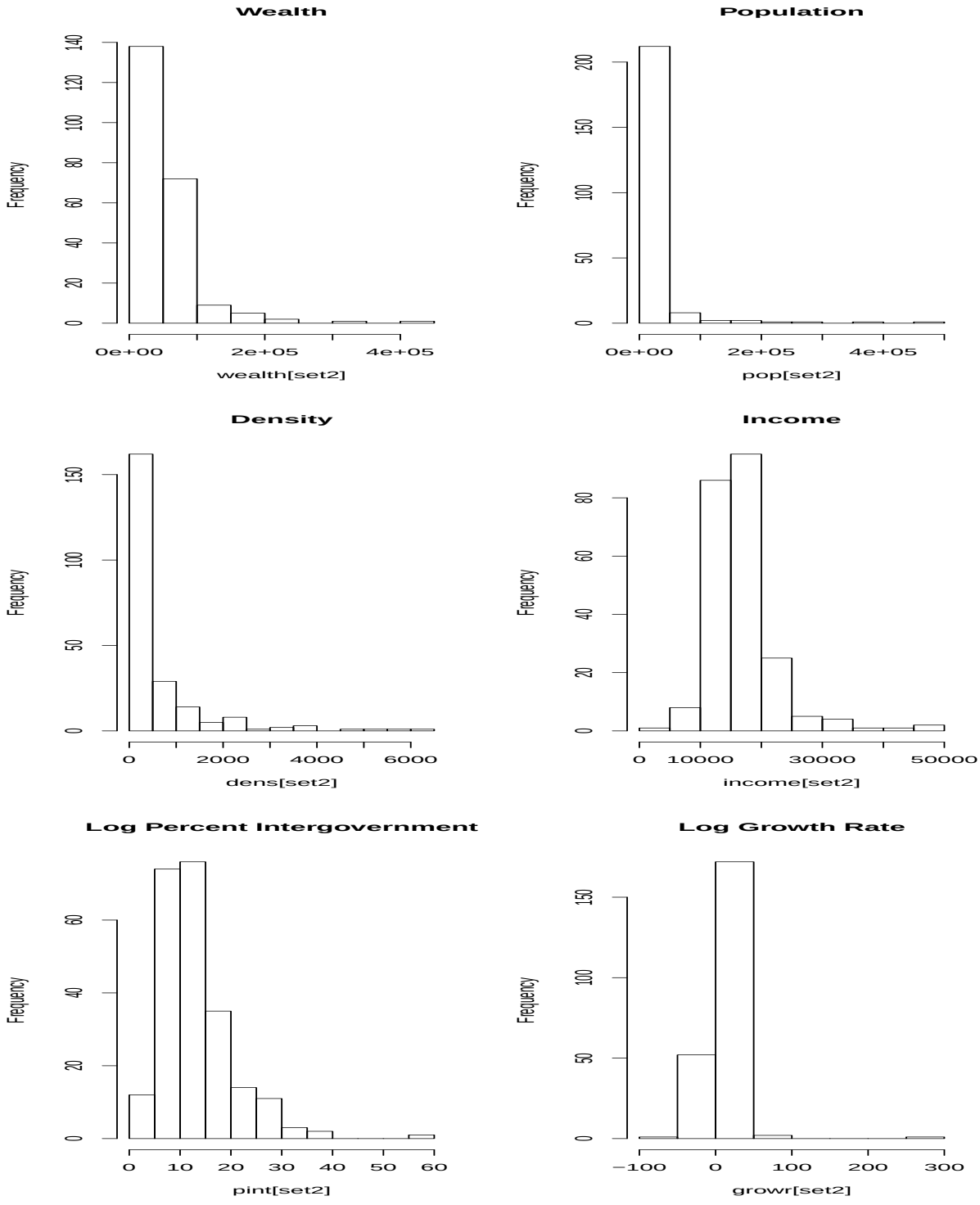


Figure 7: Histograms of Log Transformed Predictors

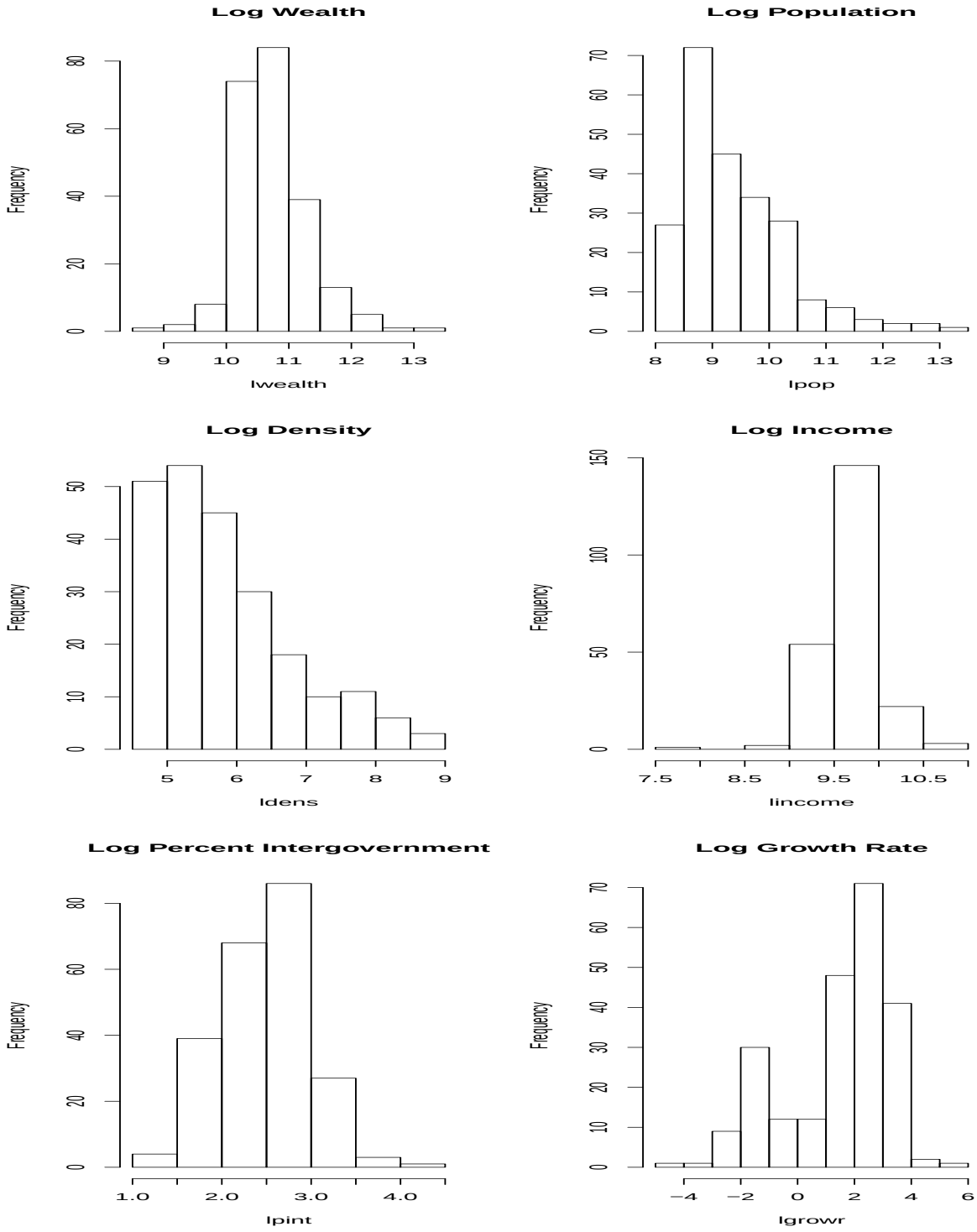


Figure 8: Log Expenditures vs Covariates

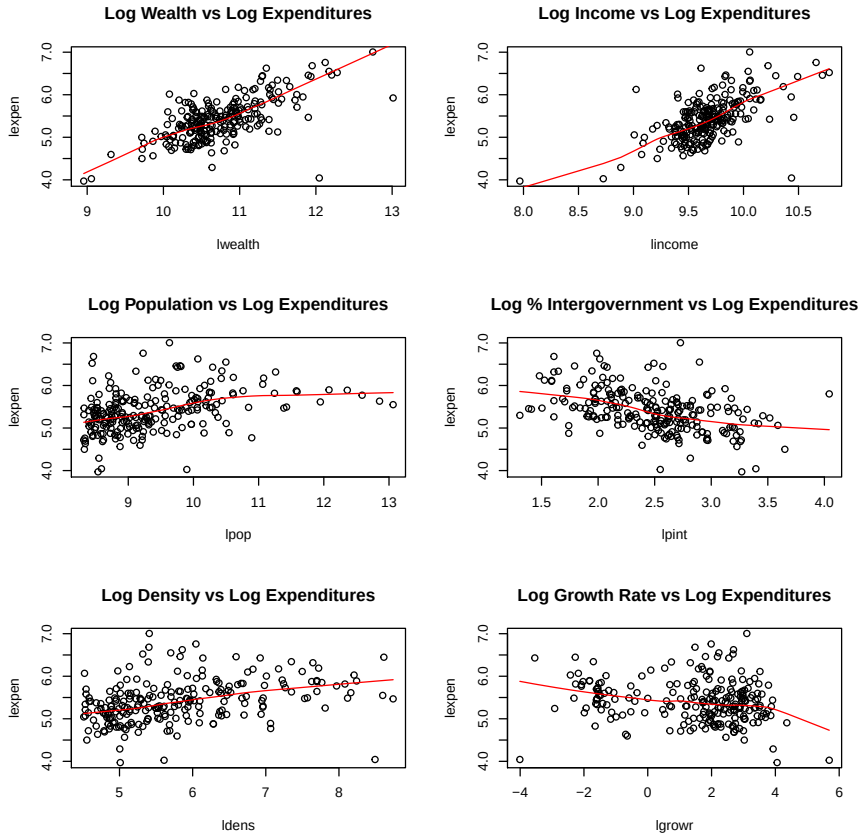


Table 4: 95% Confidence Intervals of model's coefficients

Coefficient	Lower Bound	Upper Bound
(Intercept)	0.03791628	0.7578978
Lwealth	1.33904808	1.6666506
Lpop	1.06202032	1.2521650
Lpint	0.67044390	0.7946076
Ldens	0.85051014	1.0037347
Lincome	1.07781031	1.6605132
Lgrowr	0.95239659	0.9989585

Table 5: 95% Predictions Intervals of Warwick and Monroe Expenditures

Town	Year	Expen Est.	Lower Bound	Upper Bound
Warwick	1992	248.2336	133.1536	462.773
	2005	268.8325	144.0641	501.658
	2025	277.8198	148.6064	519.3843
Monroe	1992	248.0169	133.4198	461.0441
	2005	253.7929	136.4834	471.9316
	2025	254.4445	136.7075	473.5805

Appendix B: R Code

```
library(MASS)
library(corrplot)
library(car)
#install.packages("le  aps")
library(leaps)
par(mfrow = c(1,1))
options(warn=-1)  # forces R to ignore all warning messages
ny<-read.table("C:/Users/Kelso Quan/Documents/SchoolWork/Stat696/cs73.dat",header=T); dim(
    ny)
ny2<-na.omit(ny); dim(ny2) # 914  11
attach(ny2)
names(ny2)

# look at the density of each variable especially response
# because response looks log normal argue that it needs a log transform
lpop=log(pop)
ldens = log(dens)
lexpen = log(expen)
plot(x = lpop, y = lexpen)
lines(lowess(lpop,lexpen), col=2)
lines(c(8,8), c(0,6), col = "blue", lwd = 3)
plot(x = ldens, y = lexpen)
lines(lowess(ldens, lexpen), col = 2)
lines(c(4.3,4.3), c(0,6), col = "blue", lwd = 3)
set2 = (lpop > 8.3 & ldens > 4.5) #specification of the analysis
hist(expen[set2], main = "Expenditures")
lexpen<-log(expen[set2])
hist(lexpen, main = "Log Expenditures") #lexpen is "normal"
hist(wealth[set2], main = "Wealth")
lwealth<-log(wealth[set2])
hist(lwealth, main = "Log Wealth") #lwealth is "normal"
hist(pop[set2], main = "Population")
lpop<-log(pop[set2])
hist(lpop, main = "Log Population") #close to "normal"
hist(dens[set2], main = "Density")
```

```
ldens<-log(dens[set2])
hist(ldens, main = "Log Density") #terrible
hist(income[set2], main = "Income") #kinda better than lincome
lincome<-log(income[set2])
hist(lincome, main = "Log Income")
hist(pint[set2], main = "Percent Intergovernment")
lpint <- log(pint[set2])
hist(lpint, main = "Log Percent Intergovernment")
hist(growr[set2], main = "Growth Rate")
lgrowr<-ifelse(growr[set2]>0, log(growr[set2]+1), -log(-growr[set2]+1))
hist(lgrowr, main = "Log Growth Rate")

#correlation matrix
nydata <- data.frame(lexpen, lwealth, lpop, lpint, ldens, lincome, lgrowr)
cormat <-cor(nydata)
corrplot(cormat, main = "Correlation between Log Variables")

# plot expense with each covariate
plot(lwealth, lexpen, main = "Log Wealth vs Log Expenditures")
lines(lowess(lwealth,lexpen), col=2) #lwealth is linear
plot(lincome, lexpen, main = "Log Income vs Log Expenditures")
lines(lowess(lincome,lexpen), col=2) #lincome is linear
plot(lpop, lexpen, main = "Log Population vs Log Expenditures")
lines(lowess(lpop,lexpen), col=2) #lpop is linear
plot(lpint, lexpen , main = "Log % Intergovernment vs Log Expenditures")
lines(lowess(lpint,lexpen), col=2) #lpint is linear-ish
plot(ldens, lexpen , main = "Log Density vs Log Expenditures")
lines(lowess(ldens,lexpen), col=2) #ldens is linear-ish
plot(lgrowr, lexpen, main = "Log Growth Rate vs Log Expenditures")
lines(lowess(lgrowr,lexpen), col=2) #lgrowr is linear-ish

# finding a fit
fit1<-lm(lexpen~lwealth+lpop+lpint+ldens+lincome+lgrowr)
par(mfrow=c(2,3))
plot(lexpen~lwealth+lpop+lpint+ldens+lincome+lgrowr)
plot(fit1)
par(mfrow=c(1,1))
stepAIC(fit1, direction = "both") #this one
summary(fit1)
exp(confint(fit1)) #confident interval for coefficients
#predictions for 1992, 2005, and 2025
sdfit <- sd(fit1$resid)
war92 <- data.frame(lwealth=log(72908), lpop=log(16225),  lpint=log(24.7),
                    ldens= log(170),  lincome=log(19044),
                    lgrowr=log(30.3 + 1))
war05 <- data.frame(lwealth=log(85000), lpop=log(20442),  lpint=log(24.7),
                   ldens= log(214),  lincome=log(19500),
                   lgrowr=log(35+1))
war25 = data.frame(lwealth=log(89000), lpop=log(31033), lpint=log(26.0),
                   ldens = log(325), lincome=log(20000),
                   lgrowr=log(40+1))
warick92=predict.lm(fit1,war92); exp(warick92+sdfit^2/2)
exp(predict(fit1, war92, interval="prediction")+sdfit^2/2)
warick05=predict.lm(fit1,war05); exp(warick05+sdfit^2/2)
exp(predict(fit1, war05, interval="prediction")+sdfit^2/2)
warick25=predict.lm(fit1,war25); exp(warick25+sdfit^2/2)
exp(predict(fit1, war25, interval="prediction")+sdfit^2/2)


mon92 <- data.frame(lwealth=log(55067), lpop=log(9338), lpint=log(8.8),
                    ldens = log(599), lincome=log(17100),
                    lgrowr=log(35+1))
mon05 <- data.frame(lwealth=log(58000), lpop=log(10496), lpint=log(8.8),
                    ldens = log(695), lincome=log(16726),
                    lgrowr=log(30+1))
```

```
mon25 <- data.frame(lwealth=log(60000), lpop=log(13913), lpint=log(10.1),
                    ldens = log(959), lincome=log(18000),
                    lgrowr=log(35+1))
monroe05 <- predict.lm(fit1 ,mon92); exp(monroe05+sdfit^2/2)
exp(predict(fit1 , mon92, interval = "prediction")+ sdfit^2/2)
monroe05 <- predict.lm(fit1 ,mon05); exp(monroe05+sdfit^2/2)
exp(predict(fit1 , mon05, interval = "prediction")+ sdfit^2/2)
monroe25 <- predict.lm(fit1 ,mon25); exp(monroe25+sdfit^2/2)
exp(predict(fit1 ,mon25, interval = "prediction")+ sdfit^2/2)

#outliers
plot(predict(fit1), rstudent(fit1), ylab="Studentized Residuals", xlab="Predicted")
identify(predict(fit1), rstudent(fit1), labels=row.names(ny2)) # 'escape to finish'

predict(fit1)[rstudent(fit1)==min(rstudent(fit1))]
sresid <- studres(fit1)
hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit, col = 2)

qqPlot(fit1, main="QQ Plot", ylab="Studentized Residuals")
cutoff <- 4/((nrow(set2)-length(fit1$coefficients)-2))
plot(fit1, which=4, cook.levels=cutoff) # influence Plot via cook's distance
influencePlot(fit1, id.method="identify",
              main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )
vif(fit1) #numbers above 8 is bad

bc = boxcox ( expen~ . , data=ny2 )
bc $x [ bc $y==max( bc $y ) ]
```

Listing 1: Warwick and Monroe