Expenditures of Warick and Monroe

Kelso Quan

October 9, 2018

1 Executive Summary

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 in its city 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 rasing property tax or looking for other sources of revenue. Looking into six predictor variables, a log linear trend was found between the resulting expenditure of the two cities and its six indictators. 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 intergovermental”, income, growth rate, and wealth), that the expendenitures of both towns were going to increase over time.

2 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 if not moderate expenditure 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 not been log-transformed. For example, every 1 unit (thousands) of wealth increased, the expenditure of Warwick and Monroe goes up by roughly \$1.5k. Pint stands for “Percent Intergovernmental” which is represents the precentage of revenue coming from state and federal grants subsidies. From Table 2, we have the predictions 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 noticeable more pronounce in Warwick’s prediction of expenditures.

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

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

3 Summary Information

Taking a simple look at the predictor varibles showed that they needed to be log transformed at least once before proceeding. A simple histogram of each predictor showed log transformation was much needed. Afterwards, the variable expenditure was plotted against each of its predictor. Each plot looked fairly linear. There didn't seem like an clear trends while plotting the residuals, thus no homoscedasticity. Using the function stepAIC, it was shown that the linear model of the fit

$$\log(expen_i) = \log(wealth_i) + \log(pop_i) + \log(pint_i) + \log(dens_i) + \log(income_i) + \log(growr_i) \tag{1}$$

was the best fit and modeled the data with the least amount of penalties. This fit was pretty good because all, but the log-transformed density variable had less than a 0.5 p-value. Futher 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 extremity 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 model selection. Log linear is the way to go.

4 Satistical Analysis

The analysis was kept to a linear log transformation. Every predictor needed to be log-transformed before further conduting anything else.

5 Conclusion

Appendix B: R Code

```
library(MASS)
library(corrplot)
library(car)
#install.packages(”leaps”)
#install.packages(”stargazer”)
install.packages(”xtable”)
library(xtable)
library(stargazer)
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])
lexpen<-log(expen[set2])
hist(lexpen) #lexpen is ”normal”
hist(wealth[set2])
lwealth<-log(wealth[set2])
hist(lwealth) #lwealth is ”normal”
hist(pop[set2])
lpop<-log(pop[set2])
hist(lpop) #close to ”normal”
hist(dens[set2])
ldens<-log(dens[set2])
hist(ldens) #terrible
hist(income[set2]) #kinda better than lincome
lincome<-log(income[set2])
hist(lincome)
hist(pint[set2])
lpint <- log(pint[set2])
hist(lpint)
hist(growr[set2])
lgrowr<-ifelse(growr[set2]>0, log(growr[set2]+1), -log(-growr[set2]+1))
hist(lgrowr)

#correlation matrix
nydata <- data.frame(lexpen, lwealth, lpop, lpint, ldens, lincome, lgrowr)
cormat <-cor(nydata)
corrplot(cormat)

# plot expense with each covariate
plot(lwealth, lexpen)
lines(lowess(lwealth,lexpen), col=2) #lwealth is linear
plot(lincome, lexpen)
lines(lowess(lincome,lexpen), col=2) #lincome is linear
plot(lpop, lexpen)
lines(lowess(lpop,lexpen), col=2) #lpop is linear
```

```r
plot(lpint, lexpen)
lines(lowess(lpint,lexpen), col=2) #lpint is linear-ish
plot(ldens, lexpen)
lines(lowess(ldens,lexpen), col=2) #ldens is linear-ish
plot(lgrowr, lexpen)
lines(lowess(lgrowr,lexpen), col=2)

# 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)
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
influencePlot(fit1, id.method="identify",
               main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )
vif(fit1) #numbers above 8 is bad
```

Listing 1: Warwick and Monroe