Predicting the Expenditures for two NY towns for 2030

Morgan Buterbaugh

11/29/2022

# 1 Executive Statement

The results of the study suggest that a multiple linear regression model with a transformation using the wealth per person in the town in terms of real estate holdings and the mean income per person in the town as the independent variable successfully predicts the expenditure for the two NY towns Warwick and Tuxedo. Accurate forecasts of the expenditures of the two towns for 2030 were obtained.

# 2 Introduction

The aim of this study is to investigate the estimates of expenditures for Warwick and Tuxedo, two towns in New York, for the year 2030 using a multiple linear regression model.

## 2.1 Study Design

A sample of 914 towns located in New York were randomly selected. The population for each town and population per area of each town was recorded. The mean income of each person and wealth per person in the town in terms of real estate holdings was also recorded. Along with this information, the percentage of revenue from each state and federal grants/subsidies and the expenditure per person in the town was collected.

## 2.2 Variables

The expenditure per person in the town was used as the response variable. In this study, the response variable is named ‘expend’. The predictor variables were the wealth per person in the town in terms of real estate holdings, the town’s population, percentage of revenue from state and federal grants/subsidies, population per area of the town, and mean income per person in the town. In this study, the predictor variables are named ‘wealth’, ‘pop’, ‘pgs’, ‘density’, and ‘income’ in this order respectively. Variables are in Table 2.1.

Table 2.1: Variables used in the study

| Variable | Description |
| --- | --- |
| expend | The expenditure per person in the town. |
| wealth | Wealth per person in the town in terms of real estate holding. |
| pop | The town's population. |
| pgs | Percentage of revenue from state and federal grants/subsidies. |
| density | The population per area of the town. |
| income | Mean income per person in the town. |
| The variable expend is the response variable and the variables wealth, pop, pgs, density, and income are the predictor variables. | |

# 3 Methodology

Statistical analyses are conducted using the R Software package. The expenditure for the year 2030 for the two towns in NY was modeled based on the wealth and income via a multiple linear regression model and making a transformation based off this regression model.

## 3.1 Exploratory Data Analysis

The mean, standard deviation, and interquartile range are calculated for all the variables. Histograms are made of each of variable to assess the need for transformations.

## 3.2 Multiple Linear Regression

Multiple linear regression is a statistical technique that uses several predictor variables and one response variable. The general form of a multiple linear regression is

where is the intercept parameter, are the regression parameters, and are the predictor variables that can either be quantitative or categorical. The Error is a random variable assumed to be normally distributed with a mean of 0 and a variance of .

The form of the model for this study is

where is the intercept parameter, and are the regression parameters, ‘wealth’ and ‘income’ are the quantitative predictor variables, and Error is a random variable.

## 3.3 Stepwise Forward Regression

Stepwise forward regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. The method of fitting the regression model consists of adding one independent variable at a time to the multiple linear regression equation, starting with the independent variable with the most significant probability of the correlation. Lower values indicate a preferred model. Stepwise forward regression is used to figure out which predictor variables play the most significant role in the regression model.

## 3.4 Prediction Intervals

A prediction interval is an estimate of the response variable for the next unit in a study that has a particular value for the predictor variable.

# 4 Results

The first thing I did was make histograms for all of the variables in the data set, which can be found in Figure 6.1. After taking a look at each of the individual histograms, we see that they are all heavily right-skewed. This indicates that we have to make some sort of transformation, either or .

Before completing any transformation, I ran a stepwise forward regression on the data. The results were that the wealth and income predictor variables were the most significant. The results of this model building technique can be found in Figure 6.2. Figure 6.3 shows the Residual vs Predicted and QQ plots for stepwise forward regression results.

With the new regression equation, I made both a or transformation and found that the transformation did a better job of modeling the data. Figure 6.4 shows the Residual vs Predicted and QQ plots for log transformation of the regression equation.

In this study we are looking at the 95% prediction intervals for the two NY towns Warwick and Tuxedo. The projected data for 2030 for Warwick and Tuxedo are in Table 4.1.

Table 4.1: Projected Data for the Year 2030

| Town | Population | Wealth | PGS | Density | Income |
| --- | --- | --- | --- | --- | --- |
| Warwick | 310,333 | 89,000 | 26 | 352 | 20,000 |
| Tuxedo | 292,246 | 115,000 | 7 | 656 | 25,000 |

The 95% prediction interval for Warwick based off the projected data for 2030 is . The 95% prediction interval for Tuxedo based off the projected data for 2030 is .

# 5 Conclusions

We are 95% confident that the expenditure for Warwick will be between $152.74 and $794.29 for the year 2030. Also, we are 95% confident that the expenditure for Tuxedo will be between $161.10 and $840.44 for the year 2030.

The multiple linear regression was successful at prediction the expenditure of the two NY towns Warwick and Tuxedo. It is my belief that this modeling technique can be used to predict the expenditure of towns in NY based on their wealth per person in the town in terms of real estate holdings and the mean income per person in the town

# 6 Appendix

Below are histograms, R output, and plots of the statistical analysis results.

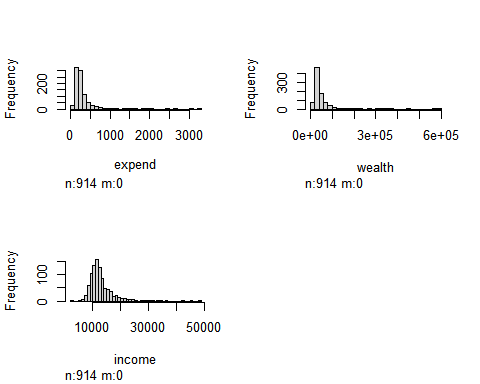


Figure 6.1: Histograms for the variables in the study.

##   
## Call:  
## lm(formula = expend ~ wealth + income, data = ny)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1470.02 -70.46 -17.02 48.96 1447.61   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.225e+02 1.798e+01 12.38 < 2e-16 \*\*\*  
## wealth 3.896e-03 1.078e-04 36.15 < 2e-16 \*\*\*  
## income -1.026e-02 1.420e-03 -7.23 1.02e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 172 on 911 degrees of freedom  
## Multiple R-squared: 0.5939, Adjusted R-squared: 0.593   
## F-statistic: 666.1 on 2 and 911 DF, p-value: < 2.2e-16

Figure 6.2: R output for the results of the stepwise forward regression modeling.

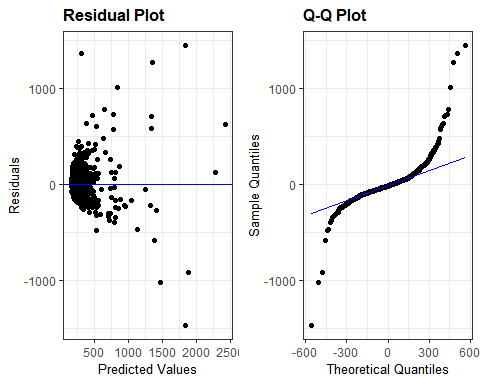


Figure 6.3: Residual vs Predicted and QQ plots for the stepwise forward regression modeling.

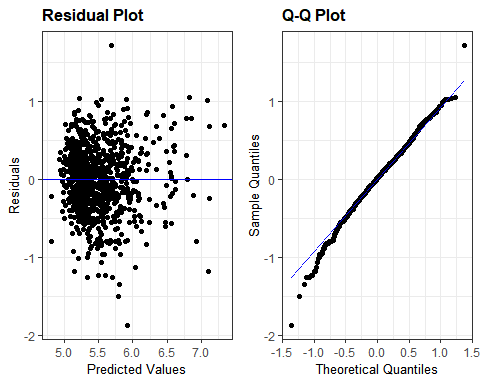


Figure 6.4: Residual vs Predicted and QQ plots for log transformation of the regression equation.