The Effect of New Housing Projects on Expenditures in Two New York Municipalities

Andrew Bates

October 11, 2018

Executive Summary

In this paper we estimate future expenditures for two municipalites in New York based on various projected demographic and income-related factors. These estimates are obtained from a linear regression model chosen via a stepwise model selection procedure. The variables in the model are wealth per person, population, percent intergovernmental funding, and growth rate. The estimated expenditures for Warwick are 1, 2 for the years 2005 and 2025. The estimated expenditures for Monroe are 1, 2 for the years 2005 and 2025.

1 Introduction

Two New York towns, Warwick and Monroe, would like to estimate future expenditures triggered by new housing construction proposals. They are primarily interested in determining whether they need to increase funds to compensate for increased expenditures related to the housing projects. To construct these estimates, Warwick and Monroe obtained data on expenditures along with various demographic and income-related variables from several New York municiplaties.

2 Methods

The data used in this study consisted of 916 observations of seven variables. Each observation corresponds to a New York municipality for which each of the variables were collected. The response variable is expenditure per person. The covariates are as follows: wealth per person, population, revenue from state and federal grants, population density, mean income per person, and growth rate.

For reasons of simplicity and interpretability, a linear regression model was chosen to estimate future expenditures. The correlation between population and population density was high (0.67) so to prevent this from leading to multicolinearity issues, only population was considered in the analysis. The variables in this data set (including the response) were heavily skewed, so log transformations were applied to each. To ensure a linear relationship between the predictors and the response, the data was subsetted to include only those observations for which the population was larger than $4,000$. The projected covariates for Warwick and Monroe fall within the range of the covariates in the subsetted data so we felt this method was appropriate. In addition, this method was favored over a more complicated method like including a quadratic term on log population which would be harder to interpret. The regression model was constructed via stepwise regression using akaike information criterion (AIC) as the selection criterion. This model was then used to estimate future expenditures.

3 Analysis

3.1 Exploratory Data Analysis

In this data set there were two measures of the size of a municipality: population and population density. Unsurprisingly, the correlation between these two measures was relatively high (0.67). In the interest of parsimony and to mitigate possible colinearity issues, we decided to consider only one of these variables to construct our model. Population density had a moderate correlation with mean income per person (0.49) whereas population had a comparably low correlation with income (0.29). To hedge against problems with colinearity, we decided to consider population for the model building process.

Upon an initial examination of the data, it was evident that transformations would be necessary. Each of the covariates, and the response, were skewed (see Appendix A) and some were heavily skewed. As this might pose problems with linearity, we performed transformations on the variables. To remain in line with our goal of having a comprehensible model, we favored log transformations over a more complex procedure. To that end, we performed log transformations on all variables in the data set. However, a straightforward application of the logarithm was not possible for one covariate. Growth rate contains some negative values as well as some zero values. To remedy this, we used the following pseudo-log transformation:

$$\text{p-}\log(\text{growth rate}) = \begin{cases} \log(\text{growth rate} + 0.15) & \text{if growth rate} > 0 \\ -\log(-\text{growth rate} + 0.15) & \text{if growth rate} \leq 0. \end{cases}$$

Note that since this is a one-to-one transformation ...

To ensure the linearity assumption of our model was satisfied, we examined the relationship between log expenditure and the log of each of the covariates. Except for log population, all other covariates had an approximately liner relationship with log expenditure. Figure 1 is a plot of log expenditure vs. log population with the addition of a smoothing line. We see what appears to be a quadratic relationship. However, notice that on either side of the vertical line (corresponding to a log population of 8.3), the relationship is approximately linear. So we were essentially faced with two choices. We could try to include the square of log population as an additional predictor or we could subset the data and have a roughly linear relationship between the response and log population.

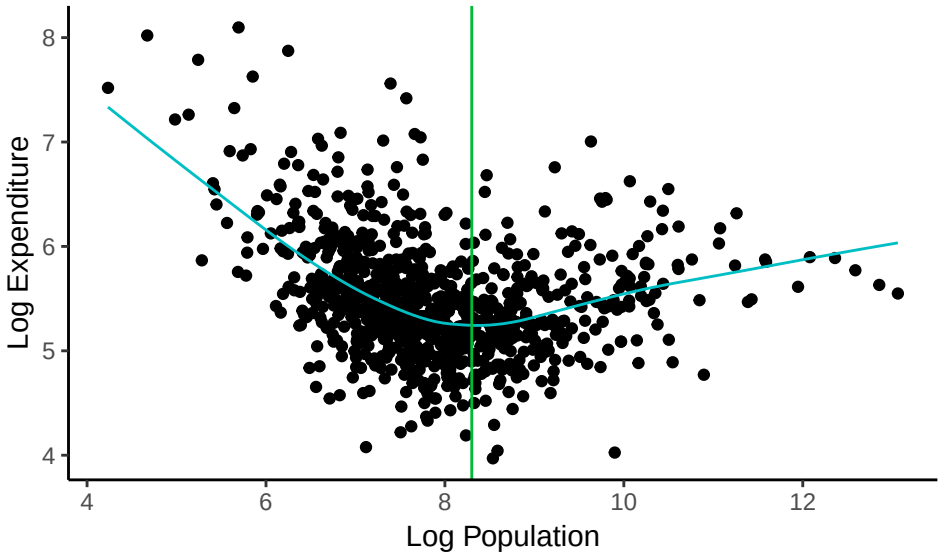

Figure 1: Plot of log expenditure vs. log population with LOWESS smooth line

We chose the latter method. As mentioned previously, we favored a simpler model for the interpretability that comes with it. The log of population squared is not as easy to understand. Also, the projected covariates we would be predicting with fall within the rang of the covariates in the subsetted data. For these reasons, we subsetted the data to include only those observations that had a population greater than $4,000$ (log population above 8.3). This subsetted data set included 266 of the original observations which we felt was large enough to obtain reliable inferences from. Figure 2 is a plot of log expenditure vs. log population for this subsetted data set along with a smoothing line. We can see that the relationship between the two variables is roughly linear. We do see minor deviation from linearity for high values of log population but overall the trend is approximately linear. Additionally, it is much more linear than the relationship we see in Figure 1.

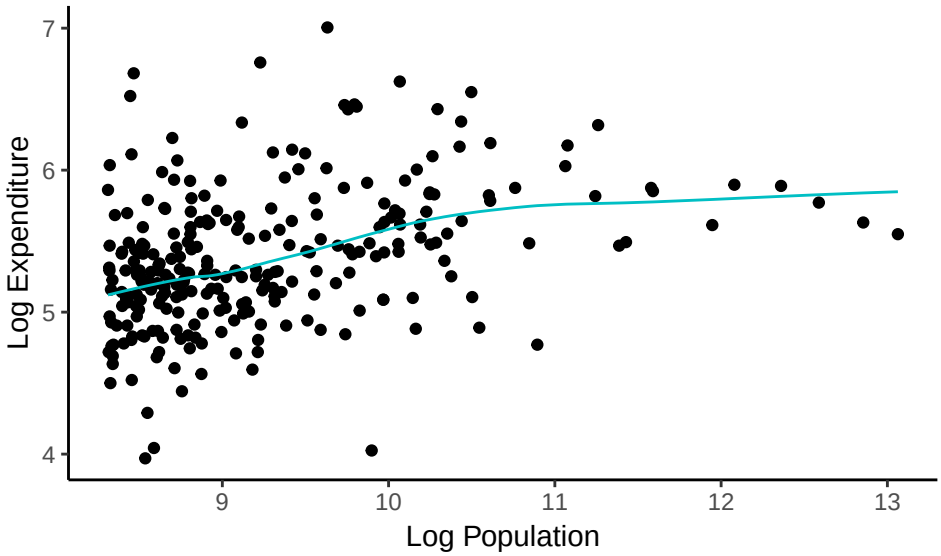

Figure 2: Plot of log expenditure vs. log population with LOWESS smooth line from subsetted data

3.2 Modeling and Diagnostics

After performing variable transformations and data subsetting, our regression model was ready to be constructed. We used both forward and backward stepwise regression to build the model. AIC was used as our model selection criterion. This was partly due to the ubiquity of AIC and partly because it penalizes larger models. This penalization aligns itself with our preference for simpler models. That being said, we did investigate the inclusion of squared log transformations for all the covariates ($\log^2$(income), etc.). However, there was not much of an improvement that could be gained by including these terms. The AIC for the model including these terms was -622 while the AIC for the model without them was -611, not much of a difference. Furthermore, the model with the quadratic variables included contained some quadratic terms without their corresponding linear components. We simply could not justify such a model. The final model used log expenditure as the response with the following covariates: log wealth, log population, log percent intergovernmental funding, and log growth rate.

Diagnostic plots from the fitted model are shown in Figure 3. The left figure is a plot of studentized residuals vs. fitted values. The horizontal line is a reference line at a residual of zero. Notice that there are no apparent patterns and the points are scattered roughly equally about the reference line. Nevertheless, there is a troublesome point with an unusually small value. This point corresponds to a municipality with a rather large value of wealth per person. It's not entirely surprising that such a municipality exists as wealth related data often contains abnormally high observations. Although this observation is a bit outside

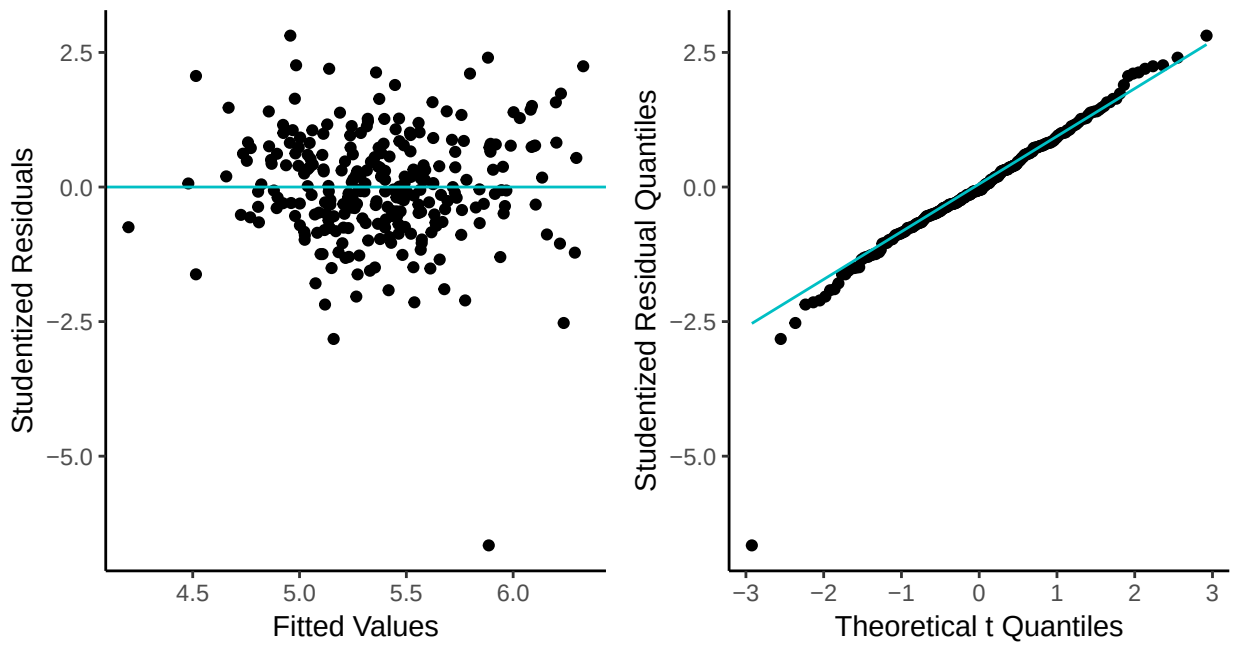

Figure 3: Diagnostic plots for regression model. The left plot shows studentized residuals vs. fitted values. The right plot shows a QQ plot of the studentized residuals vs. t-distribution quantiles.

the range of most of the data, it is nonetheless a valid data point and this municipality may also wish to estimate their future expenditures at some point. For this reason, we decided to not remove this data point. The right-hand side of figure 3 provides an alternative view of the model residuals. Here we have plotted quantiles of the studentized residuals against quantiles of a t distribution which is the distribution that they should follow theoretically. The blue line is a reference line, indicating what the theoretical relationship should be. Aside from the point just discussed, we see that the residuals follow their theoretical distribution quite well. As an additional diagnostic assessment, we computed variance inflation factors which were all less than 1.2. Considering our diagnostic plots and metrics as a whole, we were satisfied with this model.

A summary of the regression model can be found in Table 1. Note that all p-values (except the intercept) significant at the 0.05 level and the 95% confidence intervals are relatively narrow. The largest coefficient is for wealth per person which is 0.49 on the log scale. We can interpret this as follows. For a $\$1,000$ increase in the wealth per person, the municipality can expect expenditures to increase by \$30 per person. This may seem odd in that one might expect wealtier people to require less money from their local government. But perhaps the increase in expenditures comes from providing additional community resources, infrastracture improvements, or increased funding for public schools. Increased expenditures on items such as these tend to be more common in wealthier areas compared to their less wealthy counterparts. The values of the remaining coefficients can be interpreted in a similar manner. But note that the signs of the coefficients for the other covariates are a bit more inline with what we might initially expect. A municipality with a larger population will likely need to spend more money than a smaller one. The more funding a local municipality receives from state and federal governments, the less they will need to spend themselves. A town that sees a quick influx of new residents will probably require increased expenditures eventually but given the typical pace of government entities, this increase will not be immediate.

From here:

make predictions. give table with projected wealth, income, etc. then give table of predictions with PI's. be sure to try to find anything interesting to talk about. this might go in a 'Results' section. reference to prediciton table Table 2

	Estimate	SE	p-value	95% CI
Intercept	0.13	0.44	0.769	(-0.731 , 0.987)
Log Wealth	0.49	0.04	0.000	(0.423 , 0.563)
Log Population	0.08	0.02	0.001	(0.032 , 0.123)
Log % Intergov Funding	-0.28	0.04	0.000	(-0.358 , -0.198)
Log Growth Rate	-0.02	0.01	0.031	(-0.046 , -0.002)

Table 1: Summary table for regressing log expenditure on log wealth, log population, log percent intergovernmental funding, and log growth rate.

	Population	Wealth	% Intergov Funding	Growth Rate	Estimate	95 % PI
Warwick 2005	20442.00	85000.00	24.70	35.00	261.35	(139.64 , 489.16)
Warwick 2025	31033.00	89000.00	26.00	40.00	271.37	(144.64 , 509.12)
Monroe 2005	10496.00	58000.00	8.80	35.00	273.86	(147.02 , 510.13)
Monroe 2025	13913.00	60000.00	10.10	35.00	273.93	(147.08 , 510.18)

Table 2: Projections for Warwick and Monroe along with predicted values and prediction intervals.

4 Conclusion

Appendix A Supplementary Plots