Durham County Property Tax Analysis

By Lane Childers, Xavier Lee, and Conrad Lin

Introduction

This data set is focused on the property tax information collected in Durham North Carolina in 2017. This information was collected with 2017 tax filings, and recorded and stored by Durham County. We want to know what types of properties pay more tax, and what aspects of a house/property are more correlated with the amount of tax you pay on it.

Property taxes are an important part of local government revenue; they are a majority of what goes into funding public schools, parks, and other local public goods. Local governments have revenue goals which are used to set property tax rates.

Property taxes are typically decided on by local governments and each county in a state will have their own tax rate. How much you pay in taxes is based on the assessed value of your property. Durham County is right next to the capital and is a mostly urban and centralized county. There are over 3000 counties in the US with North Carolina having 100 of those.

We are not as interested in prediction as you can just calculate your tax bill, rather we are interested in determining the factors that contribute the most to tax bills (inference).

Our main research question is: “What factors affect property taxes paid for a single-family in Durham County, North Carolina?”.

A simple null hypothesis for this research question is : “The Beta values for value, size, age, etc. in an OLS regression model on property tax bill are all 0”. More specifically, we are interested in testing the null hypothesis H0: Bi = 0 for all Bi against the alternative H1: Bi != 0 where Bi is just a beta on one of the parameters of interest such as value, size, etc.

Data are from: https://data.world/durhamnc
(City and County of Durham Data)mnc/durham-county-property-2017

Note: the original durham county website is not available anymore:
https://opendurham.nc.gov/page/home/

Data Description

Our data consists of 114,283 observations with 64 variables. Most of the variables are centered around the address and general information surrounding the property, such as its assessed value or acreage. Some information also describes tax codes or building use. For our model, we are only using data for single family homes. There also are a few binary variables surrounding identifying features such as garages or fireplaces.

There are a few missing values, but we do see some zero values in the tax bill variable, likely due to exemptions to property taxes from low income or some other status.

Table 1: Summarizes the 8 Variables we Chose to Focus on for this Analysis

Variable Name	Variable Type	Description
Tax Bill (response)	Quantitative Continuous	The total tax owed on the property in USD
total_ass_value	Quantitative Continuous	The assessed property value in USD
heated_sqft	Quantitative Continuous	The amount of square feet on the property that are heated (indoors)
map_acre	Quantitative Continuous	Acreage of the property
actual _year_built	Quantitative Discrete	Year that the property was constructed
Number of bathrooms	Quantitative Discrete	Number of bathrooms on the property
Number of bedrooms	Quantitative Discrete	Number of bedrooms on the property
Attached Garage	Categorical (T/F)	Whether the property has a garage attached to it

Exploratory Data Analysis

Before we did anything with the data, we looked at every one of the 64 variables and decided on the ones we thought were most important based on our knowledge of economics and some other resources.

We started EDA with our response variable, tax bill. Figure 1 shows a boxplot for tax bill for all the data.

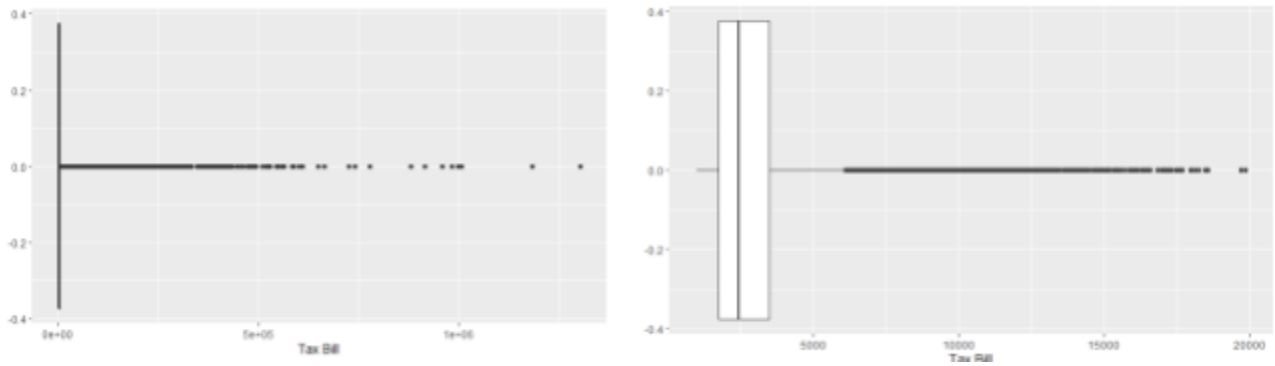

Figure 1: Boxplot showing the tax bills for the entire data set (Left) and Cleaned Boxplot (Right)

From the left plot, we noticed that the median value for our response value is around 3,000. However, the max is around 1.3 million, which affected both the exploratory plots we made as well as our future models. Due to this, we removed all tax bill values that were greater than $20,000 as we wanted to keep our focus more on properties that would typically be owned by the average single-family household.

There were very few NA values overall, but we still removed them. We also noticed from our boxplot that there were many values for our response that were 0. We removed every property with a tax bill of 0 as we just assumed these were tax exempt. This left us with 63,097 observations from our original 114,283. While removing all 0 values may not be the best idea, we agreed that doing this would allow for a more meaningful interpretation of our results when we got to that point due to our research question. The right plot shows how this cleaned up our data and made the visualizations much more interpretable.

We also looked at the overall shape and distribution of our predictors and response which is now tax bills less than $20,000 and noticed a few issues.

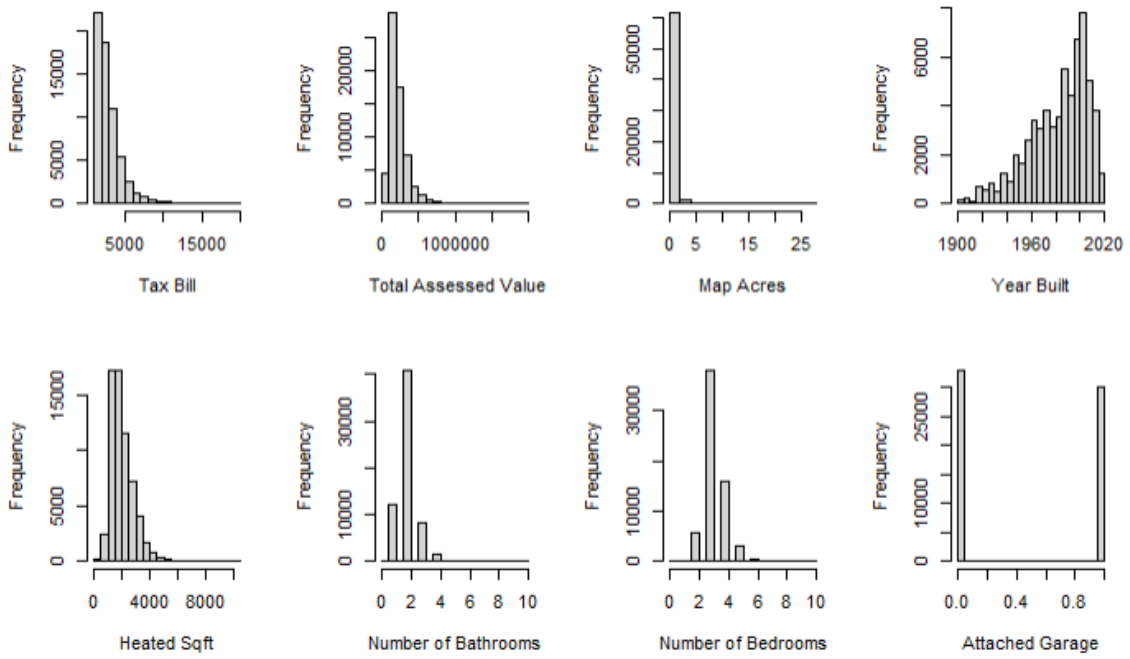

Figure 3: Distributions of our 1 response variable and 7 predictors.

Figure 3 shows that the variables that deal with price and size are right skewed and the variable dealing with time is left skewed, indicating some transformation may be needed to smooth out these distributions.

We know that economic data tends to violate OLS assumptions, specifically the normality and homoskedasticity assumptions for the residuals. So, we took a model with our response regressed on all 7 of our predictors in order to just take a look at some residual plots

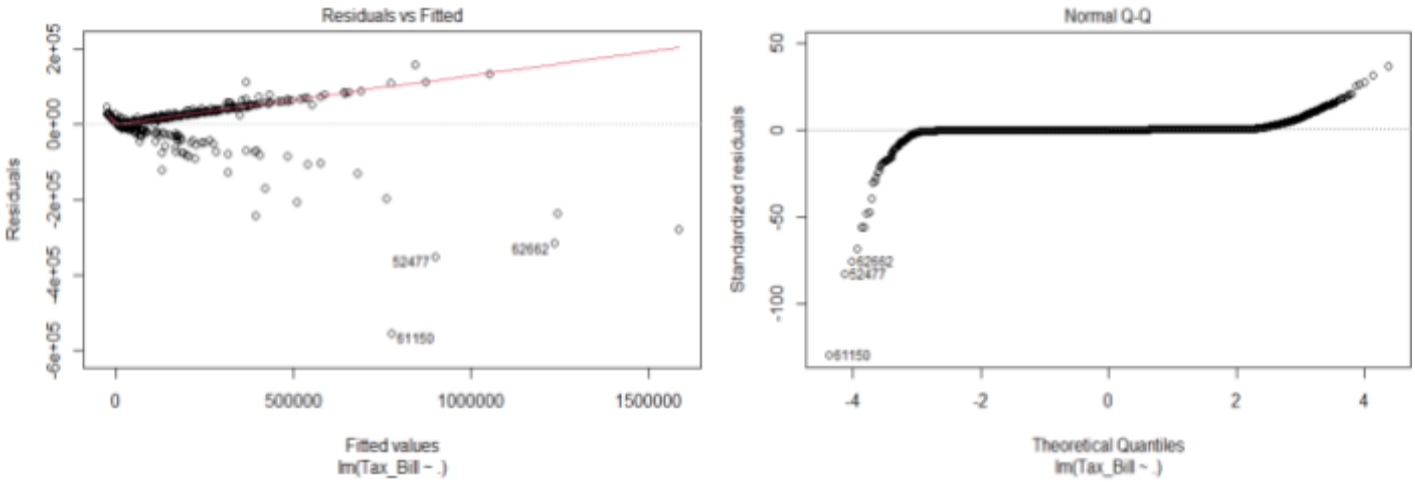

Figure 4: Residual plot (Left) and Normal Q-Q plot (Right).

Figure 4 shows that the residuals are initially not homoskedastic nor do they appear to be normally distributed.

Methods

Assumptions:

We are assuming the relationship between tax bill and predictors can be modeled using OLS. That our model residuals are conditional mean zero, with little multicollinearity as none of the predictors were really correlated, and is linear in its parameters.

From our EDA we can't really assume that our data are homoskedastic, although we do help account for this by using robust standard errors (see Results).

As our data are from a full population, we do not really need to assume that the data are a random sample either.

Log Transformations:

Due to our heteroskedastic residuals we decided we needed to transform some of our data to meet the constant variance assumption. We chose our monetary variables (total assessed value and tax bill) to log transform because they were both heavily right skewed. The log

transformation helps to symmetrize the distribution of these variables and make them appear more normally distributed. Taking the log transformation is quite common in dealing with economic data because monetary values tend to be right skewed. This will also change our interpretation of our model slightly. Now instead of interpreting our variables in terms of a 1 unit increase in total assessed value or tax bill, rather we interpret them in terms of a 1% increase instead.

Polynomial Regression Error:

To see if any of our predictors have a non-linear relationship with our response variable, we ran polynomial regression on each variable and tax bill. This was done with the poly() function in R, for degrees 1 through 5 using lm.

The data for the polynomial regression were split into a 50-50 test and training set, randomly selected 5 separate times for 5 runs. Having a test set increases error with overfitting, which is common with higher degree polynomials. Mean squared error was calculated for each degree of polynomial for each run, for 25 error values per predictor.

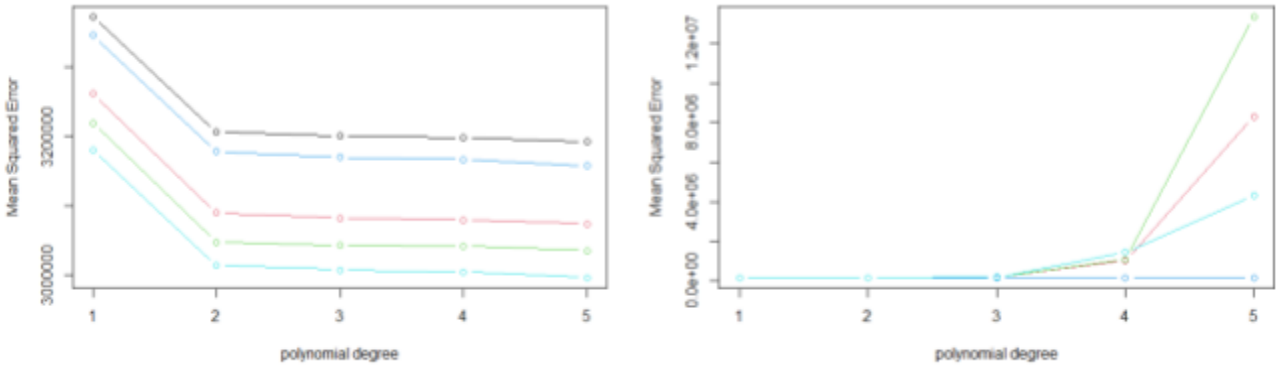

Figure 5: Year Built Polynomial Regression (Left), Assessed Value Polynomial Regression (Right)

In Figure 5 we have graphs of mean squared error values (on the y-axis) for each degree (on the x-axis). Each color represents a different run.

We can see that for the actual year built, (Figure 5), there is a noticeable decrease when going from degree one to degree two, but there is not much of a decrease as the degree increases after that. For this predictor we decided to go with a polynomial relationship of degree two as higher degrees of polynomial don't decrease error as significantly and may lead to overfitting.

For total assessed value, (Figure 5), mean squared error increases as polynomial degree goes up, so we decided that a linear relationship would be best in this instance.

Overall we decided that heated square feet and the actual year built have a polynomial relationship with degree two with tax bill, with the other variables (besides attached garage) retaining their linear relationship.

We then added these variables back into the model with the proper degree of polynomial, under the assumption that they retain their relationship with tax bill even with the other predictors.

Model Stepwise Selection:

Using the previously mentioned modified parameters, including the original linear parameters, we then ran stepwise selection on the full model. The stepwise model selection was done using BIC, with both forward and backward selection. The model given by the stepwise selection results in removing the attached garage predictor from our final model. Table 2 shows all of the estimated coefficients as well as standard errors and p-values for the model.

Table 2: Stepwise Selection Full Model

Coefficients	Estimate	Standard Error	T value	P value
Intercept	-4.63	0.025	-185.2	<2e-16
log(total_ass_value)	1.02	0.002	500.5	<2e-16
map_acres	-0.08	0.001	-96.6	<2e-16
actual_year_built, degree 1	0.95	0.192	5.0	7.30e-07
actual_year_built, degree 2	5.54	0.148	37.5	<2e-16
heated_sqft, degree 1	-2.56	0.277	-9.2	<2e-16
heated_sqft, degree 2	2.95	0.146	20.2	<2e-16
of_bathrooms	-0.01	0.001	-11.3	<2e-16
of_bedrooms	0.01	0.001	7.5	7.06e-14
attached_garage	0.004	0.001	2.6	0.00965

This is inline with the full model p-values and standard errors. We can see from Table 2 that attached garage has a higher standard error relative to its estimate, as well as a higher p-value compared to the other p-values for our model.

Our AIC resulted in no predictors being removed, however as the goal of our project is inference, we decided to use the model obtained from BIC stepwise selection for model simplicity.

The model we are interested in takes the form:

$$log(tax\_bill) = \beta_0 + \beta_1 log(total\_ass\_value) - \beta_2(map\_acres) + \beta_3(actual\_year\_built)$$
$$+\beta_4(actual\_year\_built)^2 + \beta_5(heated\_sqft) + \beta_6(heated\_sqft)^2 + \beta_7(of\_bedrooms) + \beta_8(of\_bathrooms)$$

All analyses were conducted in R version 4.2.0 using RStudio version 2022.12.0+353.

Results

The results of our final model are shown below in Table 3. Due to our heteroskedastic residuals we moved forward using robust standard errors. This is quite typical in dealing with economic data since dealing with large sums of monetary values tends to lead to skewed data and thus heteroskedasticity. We have a score of 92.5% for both R squared and adjusted R squared.

Table 3: Final Model Summary Output

Coefficients	Estimate	Robust Standard Error	T value	P value
Intercept	-4.64	0.0238	-194.8	<2.2e-16
log(total_ass_value)	1.03	0.0019	532.9	<2.2e-16
map_acres	-0.08	0.0035	-22.5	<2.2e-16
poly(actual_year_built, 2)1	1.22	0.1601	7.6	2.570e-14
poly(actual_year_built, 2)2	5.61	0.1752	32.0	<2.2e-16
poly(heated_sqft, 2)1	-2.51	0.3218	-7.8	6.174e-15
poly(heated_sqft, 2)2	2.91	0.2145	13.6	<2.2e-16
of_bathrooms	-0.01	0.0013	-10.9	<2.2e-16
of_bedrooms	0.01	0.0010	7.8	6.235e-15

Our final model is:

$$\widehat{log(tax\_bill)} = \underset{(0.024635)\ [0.0238195]}{-4.64} + \underset{(0.002003)\ [0.0019242]}{1.03log(total\_ass\_value)} - \underset{(0.000835)\ [0.0035814]}{0.08(map\_acres)} + \underset{(0.161007)\ [0.1600961]}{1.22(actual\_year\_built)}$$

$$+\underset{(0.144882)\ [0.1752230]}{5.61(actual\_year\_built)^2} - \underset{(0.276419)\ [0.3218120]}{2.51(heated\_sqft)} + \underset{(0.145228)\ [0.2144670]}{2.91(heated\_sqft)^2} - \underset{(0.0013249)\ [0.001248]}{0.01(of\_bathrooms)} + \underset{(0.001039)\ [0.0010145]}{0.01(of\_bedrooms)}$$

Our estimate for the intercept (-4.64) is the percent change in tax bill when holding all variables at 0. This doesn't really make sense nor is it important since we aren't interested in a tax bill for an observation where all other variables are 0 because that's a house. The estimate (1.03) for total assessed value is the percent change in the tax bill for a one percent change in total assessed value. The estimate (-0.08) for map acres is the percent change in the tax bill for a one unit increase in map acres. The estimate (1.22) for actual year built is the percent change in tax bill for a one unit increase in actual year built. The estimate (5.61) for actual year built squared is the percent change in tax bill for a one unit change in actual year built squared. The estimate (-2.51) for heated square feet is the percent change in tax bill for a one unit change in heated square feet. The estimate (2.91) for heated square feet squared is the percent change in tax bill for a one unit change in heated square feet squared. The estimate (-0.01) for the number of bathrooms is the percent change in tax bill for a one unit change in the number of bathrooms. The estimate (0.01) for the number of bedrooms is the percent change in tax bill for a one unit change in the number of bedrooms.

The normal standard errors are in the parentheses and the robust standard errors are in the brackets.

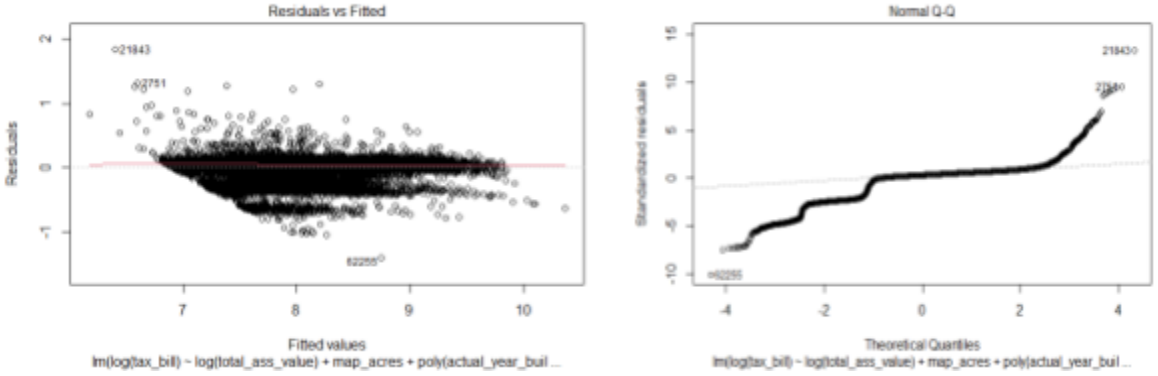

Figure 8: Final Residual plot (Left) and Normal Q-Q plot (Right).

Here in Figure 8, on the left we can see our residuals begin to look slightly more homoskedastic than before although still not perfect. On the right our Q-Q plot also begins to look slightly more normal than before but still has issues with its shape and slope.

We can't trust inference regarding our estimates fully due to the normality issue, but passing over those issues we see that on average a 1% increase in total asset value leads to a 1.03% increase in the property tax paid. However, we want to continue working to try to better fulfill the OLS assumptions before truly concluding which factors seem to have an effect on the tax bill.

Next Steps

We are planning to include 0 value tax bill properties that we initially excluded while we were cleaning up our data. Hopefully we can try to further address the normality issue as well. We also will look at a direct comparison of total asset value and tax bill. Lastly, we plan on reversing our model to see how our current predictors affect total assessed value rather than tax bill. In this last case we would shift our research question towards prediction rather than inference since we might be interested in predicting asset value.

References

City and County of Durham Data. "2017 Property Tax List." *Data.World*, 2017, https://data.world/durhamnc/durham-county-property-2017. Accessed 3 February 2023.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Appendix

Lane led our EDA, doing a lot of the initial data cleanup and early graphs. Conrad focused on model selection, doing a lot of the polynomial regression analysis and stepwise selection steps. Xavier worked mostly on results, such as robust standard errors and model interpretation. We all collaborated on the other sections (introduction and data description), also helping in our respective focuses to ensure accuracy and quality.