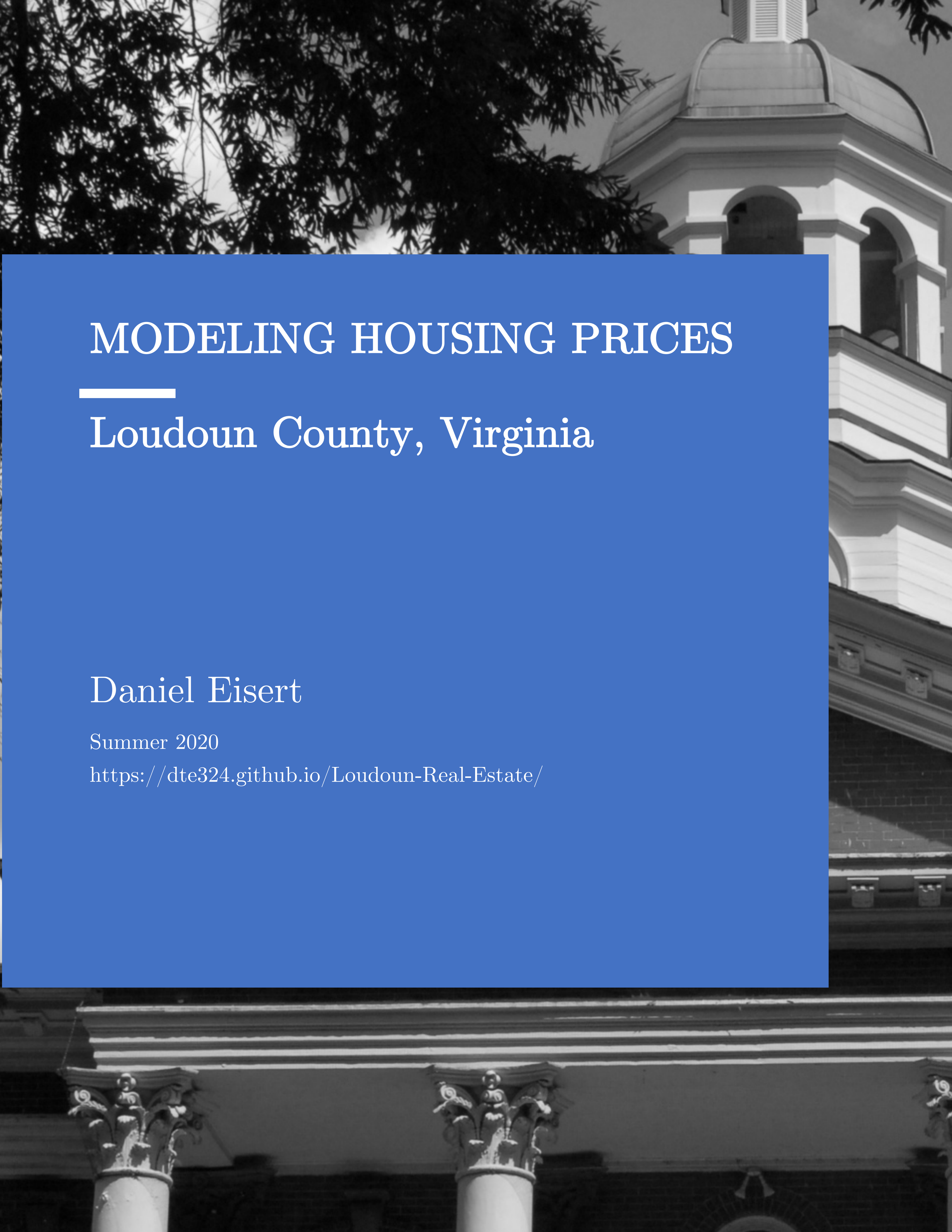

MODELING HOUSING PRICES

Loudoun County, Virginia

Daniel Eisert

Summer 2020

https://dte324.github.io/Loudoun-Real-Estate/

Table of Contents

Executive Summary

Loudoun County, Virginia is a rapidly developing suburb of Washington, D.C. Loudoun County is known for its combination of stunning planned communities along with its charming towns and countryside. Over 500 recently sold homes were web scraped from realtor.com. Only homes sold in Ashburn, Leesburg, and Sterling, Virginia (all six ZIP codes) between June 1, 2020 and July 9, 2020 were included in the model building.

Next, three different models were constructed—one for each town. Contextual knowledge and automatic selection techniques were used to arrive at the following three models. Each model exhibited R^2 values over 90% and relatively low residual standard errors. For Ashburn, home prices can be predicted by the following:

$$\widehat{\text{Price}} = -4309000 + 215200 \log x_1 + 162300 x_2 + 55830 x_{3=2} + 105000 x_{3=3} + 526 \exp\{x_4\} - 64780 x_{5=\text{Townhouse}} + 1576 x_6 + 102000 x_{8=\text{Willowsford}} + 40210 x_{9=\text{OneLoudoun}},$$

and for Sterling, home prices can be predicted by the following:

$$\widehat{\text{Price}} = -4240011 + 182462 \log x_1 + 125498 x_2 + 36089 x_{3=2} + 72733 x_{3=3} + 824 \exp\{x_4\} - 57941 x_{5=\text{Townhouse}} + 1657 x_6 + 33878 x_{10=20165},$$

and for Leesburg, home prices can be predicted by the following:

$$\widehat{\text{Price}} = \exp\{8.877 + 0.512\ \log\ x_1 + 0.100 x_{3=2} + 0.132 x_{3=3} - 0.134 x_{5=\text{Townhouse}} + 0.252 x_{7=3} + 0.251 x_{7=4} + 0.201 x_{7=5,6} + 0.058 x_{11=20176} - 0.221 x_{12=\text{Lucketts}}\}$$

where

- x_1 denotes the square footage,
- x_2 is the lot size in acres,
- x_3 is an indicator variable denoting the amount of garage spaces,
- x_4 is the number of bathrooms,
- x_5 is an indicator variable for home type (i.e., townhouse/condominium),
- x_6 is the year built,
- x_7 is an indicator variable for the number of bedrooms,
- x_8 is an indicator variable for the Willowsford neighborhood,
- x_9 is an indicator variable for the One Loudoun neighborhood,
- x_{10} is an indicator variable for Sterling's 20165 ZIP code,
- x_{11} is an indicator variable for Leesburg's 20176 ZIP code, and
- x_{12} is an indicator variable for the Lucketts area.

Modeling Housing Prices

Data Description

Over 500 homes sold between June 1, 2020 and July 9, 2020 were web scraped from realtor.com using Python's Beautiful Soup library. Observations were collected in all six ZIP codes contained between Ashburn, Sterling, and Leesburg, Virginia. In order to remove sources of bias, all homes sold between June 1 and July 9, 2020 were included in the dataset.

Many potential regressors were included within the dataset. These included number of bedrooms, number of bathrooms, square footage, lot size, year built, number of garage spaces, neighborhood, house type, high school attendance zone, ZIP code, and town.

The dataset included a wide array of homes. The least expensive home sold for $140,000 while the most expensive home sold for $2,400,000. Similarly, a wide array of home sizes were included with the smallest home size being 640 square feet and the largest home boasting 11,006 square feet. Furthermore, single family, townhome, and condominium dwellings were included to maximize model application.

Methods

After data collection was complete, the data had to be cleansed before model building could take place, as many inconsistencies and discontinuities existed within the data. For example, in some listings a home's lot size was recorded in acres, but in other cases, it was listed in square feet, so all the lot sizes were converted to acres for consistency. All condominiums were said to have a lot size of zero since condominiums lack a private yard space.

Many Loudoun County high schools have undergone significant attendance boundary alterations within the last decade. As a result, some realtors put the incorrect high school assignment for a listing. The Loudoun County Public Schools attendance boundary portal was often consulted to resolve these errors. In other cases, high school information was missing, so high school districts were mapped to their respective neighborhoods to ensure that all information was accurate.

Furthermore, some listings included a neighborhood's full name and others included a shortened version (e.g., Loudoun Valley and Loudoun Valley Estates). All these inconsistencies were corrected using Python before model building began.

Subsequently, I transitioned to R to begin the model building phase. Throughout this process, I attempted to capture as much of the variability as possible with the data I had. Adjusted R^2 values, residual standard errors, and other residual diagnostics were monitored throughout the modeling building process. Contextual knowledge along with automatic selection techniques (to minimize the Akaike Information Criterion) were used to arrive at the finalized models.

With each model, residual diagnostics were analyzed and considered. In some cases, these diagnostics along with lack-of-fit statistics drove categorical groupings and mathematical transformations to obtain a better fit. These plots and diagnostics aided in determining whether I could conclude whether the residuals were normally distributed, yielded a constant variance, and were independent from one another. Observations with unusually large residuals were reviewed and investigated to determine whether the house was truly representative of most houses in the area. In order to increase the accuracy of the model, a few observations from each town were removed. Loudoun County has been the county with the highest median household income for more than a decade, and as a result, some extravagant homes have been constructed and are worth millions. Due to their incredibly high leverage and challenging predictability, I decided to remove them from the model building datasets. My goal for these models was to accurately predict a wide array of representative houses of the eastern half of Loudoun County. It was more important for me to accurately predict the several thousand $600,000 houses as opposed to the few $2,000,000 houses.

Results

ASHBURN

Beginning with the Ashburn model, I found that the following terms were significant in predicting real estate prices: logarithm of the square footage, lot size, garage size (discretized with levels 0-1, 2, and 3), the exponential of the number of bathrooms, home type (i.e., single family or townhouse/condominium), and year built. A few residuals were particularly extreme, so I added two neighborhood indicator variables for the Willowsford and One Loudoun developments. This also significantly decreased the residual standard error. All the terms were exceedingly significant with the largest p-value being 0.00588 which is still relatively small. The overall F-statistic was 709.2 on 9 and 237 degrees of freedom with a corresponding p-value of ≈0. The R^2 and adjusted R^2 values were 0.9642 and 0.9628, respectively.

Additionally, I compared differing models with certain variables as continuous and other models with the same variables as discrete. For certain variables including garage size, this was incredibly desirable as it increased the significance of the regressor and improved the overall quality of the model.

For the most part, the residuals vs. fit plot in Figure 1 produced a nice scatter with the mean not far from zero implying that the residuals are independent from one another.

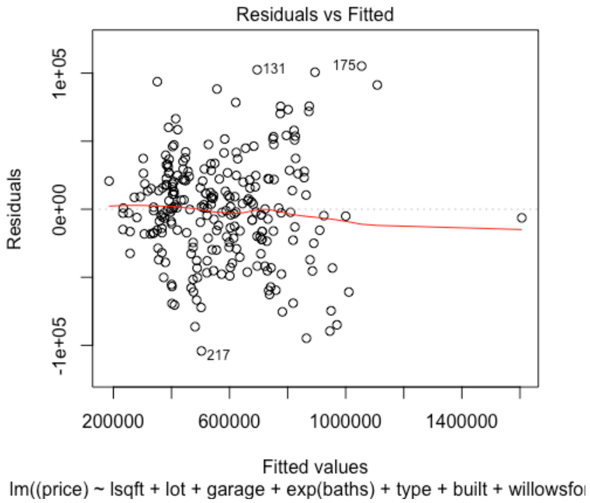

Figure 1. Residuals vs. fitted values plot for the Ashburn model.

Similarly, in Figure 2, the finalized normal probability plot appeared promising. Adding the Willowsford and One Loudoun indicator variables significantly helped bring some of the residuals closer to the plot's straight line. The residuals appeared to be normally distributed.

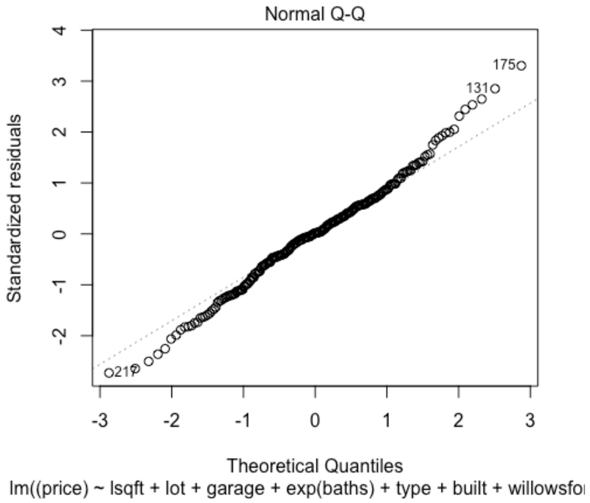

Figure 2. Normal probability plot for the Ashburn model.

STERLING

The Sterling model was quick and easy to construct. Automated selection techniques produced a model with overly significant terms with minimal grouping almost instantaneously. These regressors included the logarithm of the number of square feet, lot size, garage size (discretized with levels 0-1, 2, and 3), the exponential of the number of bathrooms, home type (i.e., single family or townhouse/ condominium), year built, and ZIP code. Though not as overly significant as the terms in the Ashburn model, the regressors in the Sterling model still produced significant p-values. The F-statistic was 243.6 on 8 and 115 degrees of freedom yielding a corresponding p-value ≈0. The R^2 and adjusted R^2 values were 0.9443 and 0.9404, respectively. These values were slightly lower than that of the Ashburn model, but they remained pleasing.

Subsequently, I compared several models with certain variables as continuous and other models with the same variables in a discrete form. Like in the Ashburn model, discretizing the garage term was incredibly helpful. Additionally, including an indicator variable for the ZIP code also added to the predicting power.

Overall, the standardized residuals plot in Figure 3 produced a nice scatter implying that the residuals are independent from one another.

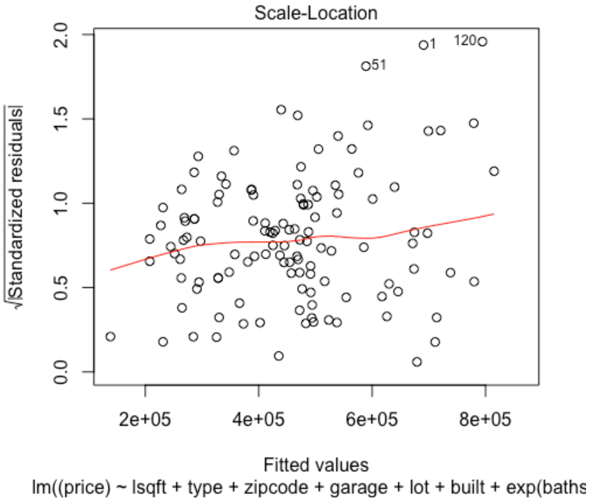

Figure 3. Scale-location plot for the Sterling model.

Like with the Ashburn model, the Sterling normal probability plot shown in Figure 4 produced a relatively straight line with a couple stray points at both ends. These points weren't overly concerning, as they tended to be residences with significantly higher selling prices than observations typical of the Sterling area.

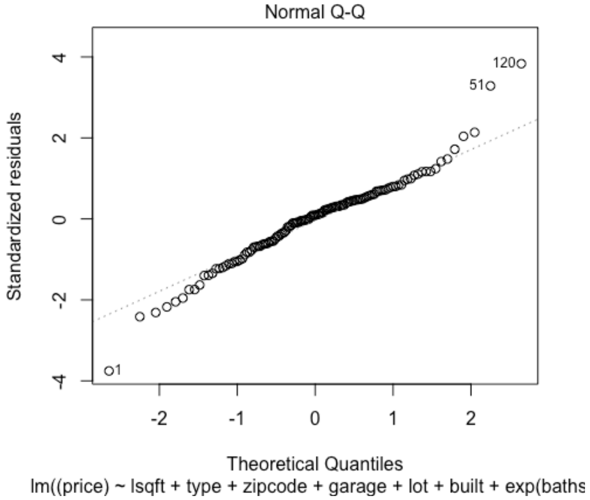

Figure 4. Normal probability plot for the Sterling model.

LEESBURG

Lastly, the Leesburg model was the most complicated to construct. Of the three localities I produced models for, Leesburg's home prices had the widest variability across all levels. Similar homes sell for vastly different amounts for a variety of reasons including neighborhood, style, and location. Leesburg's ZIP codes (particularly 20176) stretch vast areas yielding high amounts of variability.

After trying several different options, I finally arrived at a model with the following significant regressors: logarithm of the square footage, garage size (discretized with levels 0-1, 2, and 3), number of bedrooms (discretized with levels 1-2, 3, 4, and 5-6), home type (i.e., single family or townhouse/condominium), and ZIP code. It is important to note that Leesburg's 20176 ZIP code is incredibly large compared to the other ZIP codes of interest in this project. In order to accommodate the price differences, I decided to include an additional categorical variable indicating whether the listing was in the Lucketts area, a small village within Leesburg's 20176 ZIP code but roughly 7 ½ miles north of Leesburg's downtown area.

Unlike the Ashburn and Sterling models, I decided to take the logarithm of the response variable, price. This was because of the high variability exhibited within the Leesburg model. Though not as overly significant as the regressors in the other two models, this mathematical transformation drastically improved the overall model as well as the residual diagnostics. The terms in the Leesburg model still offered significant p-values. The F-statistic was 180.5 on 9 and 140 degrees of freedom producing a corresponding p-value ≈0. The R^2 and adjusted R^2 values were 0.9206 and 0.9155, respectively. These values were slightly lower than that of the Sterling model but remained pleasing.

Like previously, I compared differing models with certain variables as continuous and other models with the same variables as discrete. In this case, discretizing the garage and bedroom predictors proved to be incredibly helpful. Including an indicator variable for the ZIP code also added to the predicting power.

Looking at Figure 5, the scale-location plot, illustrates a nice scatter of the residuals implying that the residuals are independent from one another with a constant variance. It is important to note the smaller range of standardized residuals for higher fitted values; however, this is not concerning to me, for fewer observations with extremely high price ranges were available at the time of the web scraping.

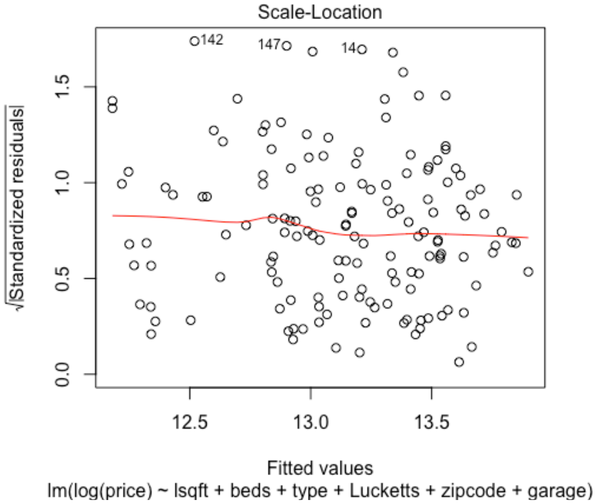

Figure 5. Scale-location plot for the Leesburg model.

Similarly, the normal probability plot satisfies the normality requirement of the residuals. Taking the logarithm of the response variable had an incredibly positive impact on both the normal probability plot and the residual vs. leverage plot. In Figure 6, only three of Leesburg's 151 observations had a leverage over 0.20, and these three points had standardized residuals near zero. This was a good sign, and implied that these points were not having too much influence on the model.

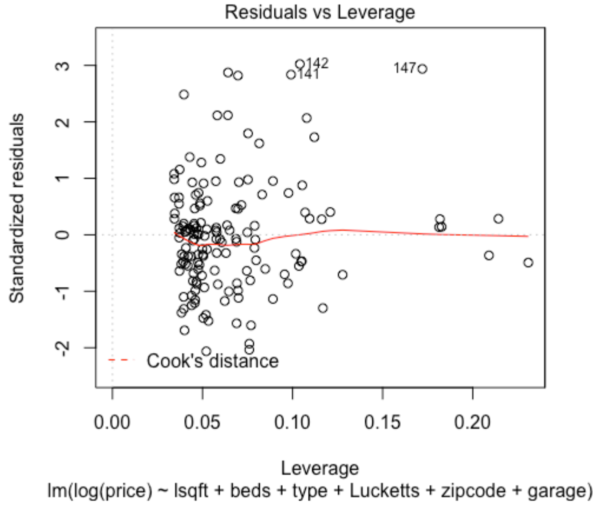

Figure 6. Residuals vs. leverage plot for the Leesburg model.

Conclusion

The three models chosen to predict housing prices in three localities in the eastern portion of Loudoun County (Ashburn, Sterling, and Leesburg) used a variety of continuous regressors—some mathematically transformed—including square footage, number of bathrooms, lot size, discretized regressors including number of bedrooms and number of garage spaces, and strictly categorical regressors including neighborhood and housing type (e.g., single-family or townhouse/condominium). After performing these groupings and transformations along with removing some observations that did not accurately represent the other observations, the three models seemed to fit the data well with nearly all the standardized residuals being around or

less than 3. This model accurately predicts the prices of not only single-family residences, but also townhouses and condominiums. These models could easily undergo slight modifications in order to be applied to other nearby localities in Loudoun and Fairfax Counties since they have similar housing types and neighborhoods.

Even though the regressors are broad and successful in application in most scenarios, the model falls short then predicting extravagant houses that are not representative of other houses in the area. This model would identify such a house as an outlier, and it may not offer an accurate prediction due to extraneous factors.

Regression model building is an extremely powerful tool; however, with great power comes great responsibility. The process is not an exact science, but the models were constructed to do their best to capture most of the variability in the response variable.