A Statistical Analysis on Recent Housing Data

Matthew Grohotolski, Alexander Waskiewicz
Elizabethtown College
{*GrohotolskiM*, *WaskiewiczA*}@etown.edu

***Abstract*—The purpose of this paper is to identify which variables have the largest influence on house value in the United States and to create a multiple linear regression model to predict house value. Using the 2019 data from IPUMS, we found that age of structure, metropolitan status, number of bedrooms, log value of property tax, and state all have significant effect on the dependent house value variable. State and property tax were two of the most significant predictors within the model.**

These results indicate possible correlation between location and house value as states and counties within those states have varying property taxes based on population size and density in the region. Age of structure was also a good predictor of home value but we believe that it could be skewed in more wealthy areas where older built homes could be renovated and upgraded. This could introduce bias into the reporting of home value as the owner gave an estimate of what they believe their house is worth. Since the housing market is constantly changing with increasing inflation from COVID-19, it is important that more research is conducted to test for other variables that could affect home price to make extrapolation into the future more accurate.

I. Introduction

A. Background

House values in the United States are constantly shifting due to outside political and economic factors that the public does not have control of. Being able to predict house values in this ever changing market is important for home buyers across the country who are looking for the best price on a home. Having a complete understanding of the variables that affect house value will aid home buyers in assessing what their needs and wants are during this process.

B. Data

The data utilized is extracted from the IPUMS USA database from a survey conducted in 2019 by Integrated Public Use Microdata Series. IPUMS USA is a website and database that provides access to samples of the American population from sixteen different federal censuses, the American Community Surveys of 2000 to the present and from the Puerto Rican Community Surveys from 2005 to the present.[1] A random sample of 100,000 units were selected from the original data for the analysis. Of the 17 predictor variables examined in the original data, 5 were used in the regression analysis and model building to predict the response variable of house value; all of which are described below. The other 12 that were omitted were either not of interest in predicting house value or did not have measurable values for the purposes of this study.

C. Variables

State which is described by the name ”STATEFIP”, reports the state in which the house was located in alphabetical order. This geographic variable includes all states and the District of Columbia as well as groupings of states which totals to 62 possible values for this explanatory variable in the original data set.[1] The variables used in the random sample study only included 51 categorical variables which consisted of all the states and the District of Columbia which were represented by the various state abbreviations.

Metropolitan status which is described by the name ”METRO”, indicates whether the household resided within a metropolitan area and, households in metropolitan areas, whether the household resided within or outside of a central city. This geographic variable includes 5 possible values in the original data set.[1] For the study, 4 of those categorical variables were used as there were no values of ”Metropolitan status indeterminable” within the 100,000 units.

Annual property taxes which is described by the name ”PROPTX99”, reports the household’s total real estate tax costs for the previous year. The total tax includes state, local, and any other taxes. Respondents reported the full amount of taxes including mortgage payments, delinquent payments, and if they were paid by another party. This economic variable includes 70 values, all of which are represented by an interval of property tax ranging from 50.00 to 1000.00 intervals in the original data set.[1] With the intention of creating a simpler model, the 70 variables were condensed into 11 categorical variables represented by the median of a $1,000.00$ interval; 500.00, $1,500.00$, $2,500.00$ e.t.c.. We renamed this variable ”LNPROPTX” after taking the natural log of the property tax value to reduce variability.

Age of structure which is described by the name ”BUILTYR2”, reports the decade in which the structure was built. This dwelling characteristic variable has 24 values which range from homes being built in 1939 or before to being built in 2019 in the original data set. For this study, all 100,000 randomly selected homes were built in 2005 or after which creates 14 quantitative variables.[1] It was not necessary to study houses built before 2005 as the population of interest is only in recently built homes. Making the ”BUILTYR” variable quantitative will give a simpler model. The years after 2005 were represented by values of 10 through 24. These values were re-coded into years 2005 to 2019.

Number of bedrooms which is described by the name ”BEDROOMS”, reports the number of bedrooms within the housing unit. This dwelling characteristic variable has 22 possible values with x-1 being the number of bedrooms represented by each value (x) in the original data set. The data in this study did not include any units with greater than 8 bedrooms so there are 9 categorical variables for bedroom.[1] The variable values for bedrooms were all transformed into x-1 values to make this a quantitative representation of the number of bedrooms.

House value which is described by the name ”VALUEH”, reports the value of the housing units in contemporary dollars. This value was collected based on the home owners opinion of the value of their home at the time of the survey. House value is a continuous variable from houses built in 2008 and on. There are 65 possible values in the original data set that are represented by intervals in which a homeowners property tax would be included in.[1] For this study, the categorical intervals were reverted into quantitative values that the owners originally priced their homes at. The variable was renamed ”LNVALUE” after the natural log was taken to reduce variability. This will allow for analysis to become simpler when testing variables that may have a significant effect on house value.

D. *Exploratory Data Analysis*

Fig. 1 shows an overview of the house count in each state from the simple random sample (SRS) of 100,000 data entries.

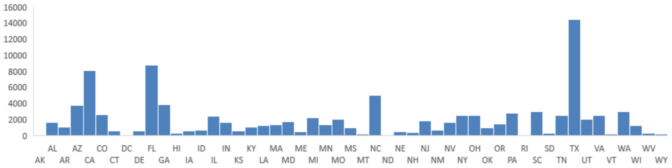

Fig. 1. Count of Houses per State

Fig. 2 shows a plot of the count of houses built by year from the SRS.

Fig. 3 shows a plot comparing the distribution of average property tax in each state with the log distribution of average property tax in each state from the SRS.

Fig 4 shows a plot comparing the distribution of average home values in each state with the log distribution of average home values in each state from the SRS.

II. Methods

A multiple linear regression model was used with house value as the response variable. We tested 5 possible models by using the same 5 explanatory variables examined in the **variables** section but with different selection procedures and criteria. We took the natural log of property tax and predicted value in order to increase the r-squared value while decreasing variance. This means that we will be predicting the log value of a house with the final model. The step wise procedure was used for the first three models where the different selection criteria values of BIC, CP Mallow, and p-value were used. For the other two models, we took the selection criteria that had the best fit from the first three models (p-value) and used it with the forward selection procedure and then the backward selection procedure. We chose this process as a way to standardize

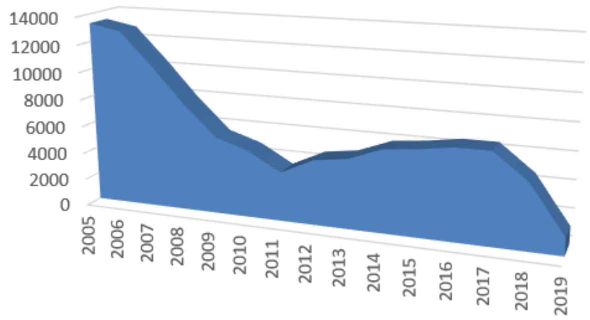

Fig. 2. Count of Houses Built by Year

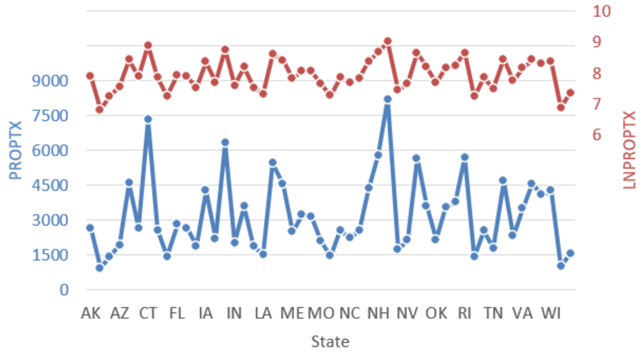

Fig. 3. PROPTX vs. LNPROPTX

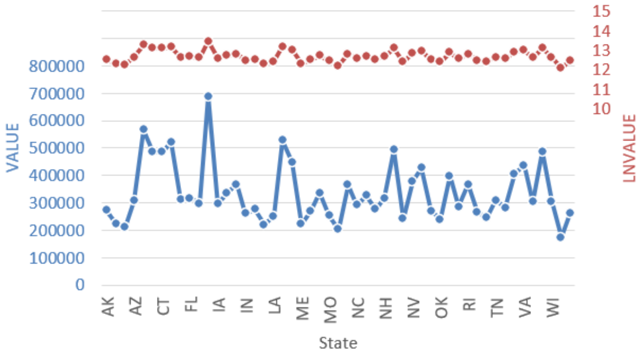

Fig. 4. VALUE vs. LNVALUE

selection procedure and predictor variables which allowed for comparison of the multiple model's ability to predict house price.

III. Results

A. Final Model

The effects in the model are as follows: Intercept, BUILTYR, METRO, NUMBED, LNPROPTX, LNPROPTX*METRO, STATELBL METRO*STATELBL, and LNPROPTX*STATELBL.

TABLE I
Final Model Statistics

Root MSE	0.59124
Dependent Mean	12.71301
R-Square	0.4960
Adj R-Sq	0.4944
AIC	-3836.11727
AICC	-3834.59265
BIC	-83795
C(p)	1164.65786
PRESS	28100
SBC	-81517
ASE (Train)	0.34849
ASE (Test	0.35099

B. Analysis

The final model that was chosen best predicts home value was found using the step wise selection procedure with p-value as the selection criteria. The predictor variables of age of structure, metropolitan status, number of bedrooms, log value of property tax, log value of property tax interacting with metropolitan status, state, metropolitan status interacting with state, and log value of property tax interacting with state all have significant effect on the dependent house value variable. As referenced in Table I, this model had a test MSE value of 0.59124 which was similar to the other selection criteria we tested. However, this model had the least amount of variables with 8 significant predictors compared to step-wise BIC which had 11 variables and step-wise CP which had 12 variables. The simpler model can more easily analyze single variables and their affects on home value.The forward and backward selection processes using the criteria of p-value both had 8 predictor variables but was slightly less effective in predicting home value.

The goodness of fit of the model is 0.4944 which is represented by the adjusted R-squared value. This model will accurately predict home value 49.44% of the time. The model has an f-value of 335.00 with a p-value of less than 0.0001 making at least one of the variables significant in predicting home value.

Fig. 5 shows an overview of the house count in each state from the simple random sample (SRS) of 100,000 data entries. Fig. 6 shows an overview of the house count in each state from the simple random sample (SRS) of 100,000 data entries.

C. Assumptions

While the final model found significant variables that were accurate predictors of house value, it is important to note that this model does not satisfy the necessary assumptions in order to be accurately used for this prediction. The Normal Q-Q plot shown in Fig. 6 suggests that the LNValue variable is left skewed and does not follow the normal distribution of data. The residual line varies greatly from the predicted LNValue line. The residual plot presents a heteroscedastic (funneling shape) pattern which violates both the linearity and equal variance assumptions. This in combination with a presence of outliers towards the negative residual values prevents this model from accurately representing the data and its usage as a predictive method.

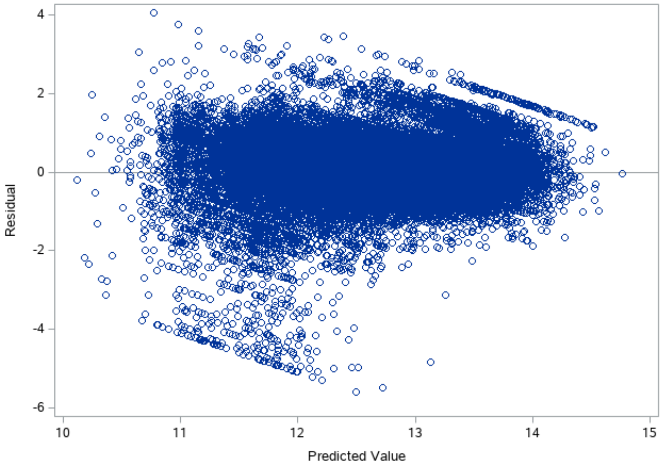

Fig. 5. Residual vs. Predicted Plot of LNVALUE

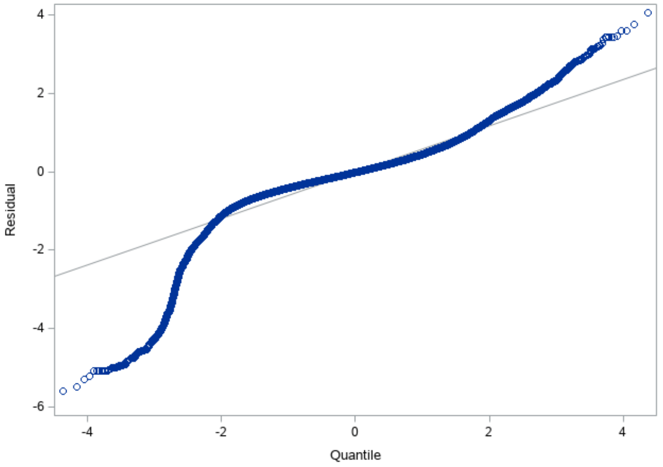

Fig. 6. Normal Q-Q Plot of LNVALUE

IV. Discussion

A. Significant Variables

Of the 8 predictor variables utilized in the model, age of structure, state (Texas), number of bedrooms, and log of property tax were the most significant predictors of home value in the model with each variable mentioned below having a p-value of less than 0.0001. All intercepts are in terms of predicted log house values. To convert these log coefficients to U.S. dollar amounts, use this equation.

$$USD = 10^x$$

where x is a log coefficient.

Age of structure had an intercept value of 0.009527. As the age of structure becomes newer by 1 year, the average predicted log home value increases by 0.009527 holding all other variables constant.

The state of Texas had an intercept value of -0.851144. The average predicted log home value for a house in Texas is 0.851144 less than a house in Pennsylvania holding all other variables constant.

Number of bedrooms had an intercept value of 0.135463. For every 1 bedroom increase, the average predicted log home value increases by 0.135463 holding all other variables constant.

Log of property tax had an intercept of 0.560288. For each 1 unit increase in log of property tax, the average predicted log home value increases by 0.560288 holding all other variables constant.

The coefficients for the significant predictor variables are located are located in a GitHub file due to spacing conflicts with this report. View References section for the link.[2].

B. Limitations

A few of the variables were challenging to represent originally which made the model more complex and confusing to interpret. One of these variables, age of structure had to be made into a quantitative variable in order to obtain the assumption plots within a reasonable amount of processing time in SAS. This limits our ability to interpret the effect on house price that each year has individually and our ability to compare years with our intended year of interest of 2019. With this variable, extrapolation is very dangerous for years outside of the range 2005 to 2019 as there are confounding variables that could affect home value such as economic prosperity, inflation rates, and general value of the U.S. dollar.

It is also important to mention that the years of 2008 and 2009 could include multiple outliers because these observations were collected directly following the Great Recession in late 2007 from the housing market crash. We expect these housing values to be extremely lower compared to 2005, 2006, and 2007 which would make the predicted model hard to use during the years in which the country was recovering from the market crash. More research should be completed to find if variables such as employment status of owners or mortgage value are more significant for predicting home value during this recession.

References

[1] R. Steven, F. Sarah, F. Sophia, G. Ronald, P. Jose, S. Megan, and S. Matthew, "Ipums usa: Version 11.0 [dataset]," 2021. [Online]. Available: https://doi.org/10.18128/D010.V11.0

[2] M. Grohotolski, "Ma252 housing project," 2021. [Online]. Available: https://github.com/MattBcool/MA252-Housing-Project