**An Econometric Analysis of Housing Demand**

**Motivation**

A common concept within the field of economics is elasticity (i.e., the idea that price for goods/services will generally change in context of the quantity of an item available within a market) (Hayes, 2021). Housing is one of many commonly mentioned market goods in both macro and microeconomics, and housing is often subject to price fluctuations within the market. However, housing units are also vastly variable in quality (from studio apartments to mansions), so analysis of housing markets generally requires *ceteris paribus* conditions to analyze. Thankfully, the U.S. Department of Housing and Urban Development (HUD) has an Office of Policy Development & Research that provides publicly available data on the 50th percentile/median market rate for housing (specifically for studio and one-four-bedroom housing units) in every area within the United States, including territories such as Puerto Rico and Guam. (US Housing and Urban Development - Office of Policy Development & Research, 2021) As medians are centers of mean generally less prone to skews and bias, my intent is to see whether the basic concept of elasticity holds well for housing across the United States, with a secondary goal of using linear regression to establish a general idea of exactly *how* elastic (or inelastic) the demand for housing is, assuming median housing quality is present for the median market rate present in the 2020 housing data from HUD.

**Data Description**

As referenced in the above motivation statement, HUD has publicly available data for 50th percentile (median) rents for the entire United States as far back as 2001. The number of areas within the dataset is defined as n= 4766, accounting for several (if not all) counties, parishes, and municipalities within all states and territories in the United States. For the sake of using the most recent full calendar year, I am using the dataset for 2020 — as 2021 is not entirely over at the time of this project.

The dataset is an MS Excel spreadsheet with the following

* fips2010
* rent50\_0
* rent50\_1
* rent50\_2
* rent50\_3
* rent50\_4
* state
* cbsasub21
* areaname21
* county
* cousub cntyname
* name
* pop2017
* hu2017
* state\_alpha

For the purposes of SAS coding for this simple regression project, I have used “rent50\_0”, “rent50\_1”, “rent50\_2”, “rent50\_3”, “rent50\_4” as potential predictor variables, as well as pop2017, hu2017, and transformations of pop2017 and hu2017 as potential predictor variables. I will specifically refer to “rent50\_2” as my response variable in this written report from here on out, as two-bedroom housing units are often used as a benchmark for what constitutes affordable housing (Adamczyk, 2021), (Simone, 2021). Additionally, “rent50\_2” serves as the median number of bedrooms within the range of rent prices, giving another good reason to use it as a metric.

The other variables outside of the assorted rent variables, population variable, and housing unit variable are all categorical variables that entail things such as metropolitan statistical areas, counties, states/territories, and other assorted government codes that further detail each region.

**Data Exploration**

I used the below SAS code snippet to import a slightly cleaned up version\* of the HUD data referenced earlier for this paper:

*FILENAME CSV "/home/u49665201/sasuser.v94/STA3064/Copy 2 of FY2021\_50\_County.csv" TERMSTR=CRLF;*

*/\*\* Import the CSV file. \*\*/*

*PROC IMPORT DATAFILE=CSV*

*OUT=HUDdata*

*DBMS=CSV*

*REPLACE;*

*RUN;*

*/\*\* Print the results. \*\*/*

*PROC PRINT DATA=HUDdata;*

*RUN;*

From this point, I created a secondary dataset I called “Housing” with a few additional changes.

*data Housing;*

*set HUDdata;*

*PtoH = pop2017/hu2017;*

*HtoP = hu2017/pop2017;*

*transPop = pop2017\*\*-1;*

*transHu = hu2017\*\*-1;*

*run;*

In this instance, the most classic example of elasticity would be present in the model “rent50\_2=hu2017” to mimic a classical “price=quantity” scenario. However, I made some additions for additional experimentation and exploration of this data.

The first addition I made was creating population: housing (defined as ‘PtoH’) and housing: population (defined as ‘HtoP’) ratios from pop2017 and hu2017. While these ratios would not necessarily create the scenario for classic elasticity, my initial assumption was that they could serve as proxy variables for housing scarcity in lieu of being able to use multiple regression for this project.

The second addition I made was adding transformations to the variables pop2017 and hu2017, based on the recommended Box-Cox transformation recommended by the output generated by the below code:

*proc transreg data=Housing;*

*model boxcox(rent50\_2)=identity(HtoP);*

*run;*

*proc transreg data=Housing;*

*model boxcox(rent50\_2)=identity(PtoH);*

*run;*

*proc transreg data=HUDdata;*

*model boxcox(rent50\_2)=identity(pop2017);*

*run;*

*proc transreg data=HUDdata;*

*model boxcox(rent50\_2)=identity(hu2017);*

*run;*

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

As shown above, both the ratio variables as well as pop2017 and hu2017 are recommended to be transformed to the -1st power.

In addition to producing scatterplots for the classical “rent50\_2=hu2017” model, I also generated five additional scatterplots using the pop2017 variable, as well as the two transformed versions of those variables and the ratios of those variables using the following code:

*proc sgplot data=Housing;*

*scatter X=rent50\_2 Y=pop2017;*

*run;*

*proc sgplot data=Housing;*

*scatter X=rent50\_2 Y=hu2017;*

*run;*

*proc sgplot data=Housing;*

*scatter X=rent50\_2 Y=HtoP;*

*run;*

*proc sgplot data=Housing;*

*scatter X=rent50\_2 Y=PtoH;*

*run;*

*proc sgplot data=Housing;*

*scatter X=rent50\_2 Y=transHu;*

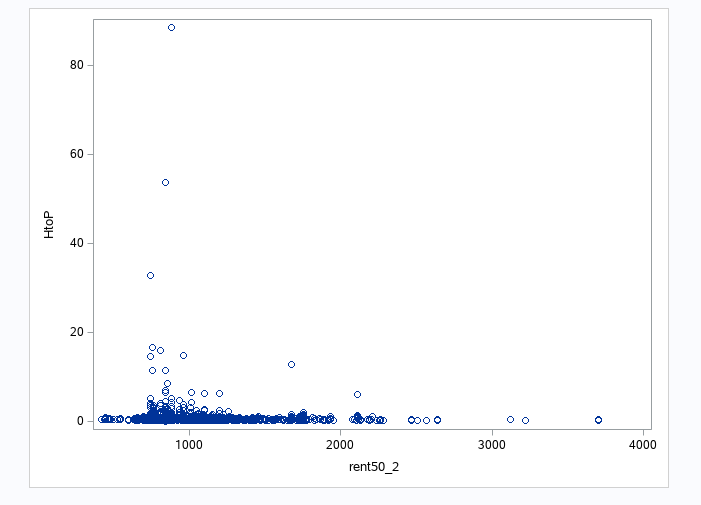
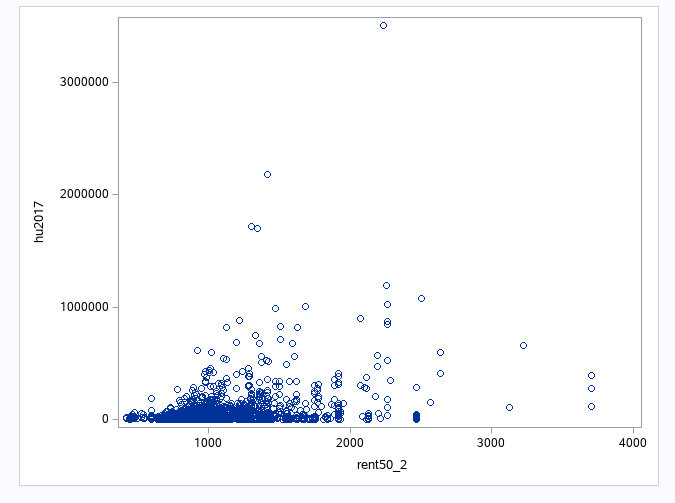
*run;*

*proc sgplot data=Housing;*

*scatter X=rent50\_2 Y=transPop;*

*run;*

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generated

In addition to the previous code and output for proc sgplot, I also used the following code to generate LOESS curves:

*proc loess data=Housing;*

*model pop2017=rent50\_2;*

*run;*

*proc loess data=Housing;*

*model pop2017=rent50\_2;*

*run;*

*proc loess data=Housing;*

*model pop2017=HtoP;*

*run;*

*proc loess data=Housing;*

*model pop2017=PtoH;*

*run;*

*proc loess data=Housing;*

*model transPop=rent50\_2;*

*run;*

*proc loess data=Housing;*

*model transHu=rent50\_2;*

*run;*

Graphical user interface, chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface, diagram, application

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface, diagram

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface, application, Word

Description automatically generatedChart

Description automatically generated with low confidenceChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface, application, Word

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface, chart

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface

Description automatically generatedChart, scatter chart

Description automatically generatedGraphical user interface, chart

Description automatically generatedChart

Description automatically generatedGraphical user interface

Description automatically generated

**Model Fitting and Analysis**

While I generated ***several*** potential models, I would like to start with the analysis of the initially assumed relationship of “rent50\_2” being the response variable, and “hu2017” being the predictor variable. Using the below code with the original HUD dataset;

*proc reg data=HUDdata;*

*model rent50\_2=hu2017;*

*run;*

I generated the following output: Table

Description automatically generated

Diagram, engineering drawing

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Using the above information, we can note that a model estimate of

rent50\_2 = 992.40167 + 0.00077344X +ε

is not a very practical regression model. With an R-Square value of 0.0422, we can also note that the relationship between these two single variables is close to non-existent, despite what would make economic sense. It is, however, interesting to note that the 95% confidence intervals get wider with a larger number of housing units available.

Looking at the residual graphs, this model violates *several* assumptions; predominantly on constant variance and normality, as the observations curve away from a normal distribution, and seem to clump together on top of each other in a line. I would hesitate to say that the relationship is not linear…although it specifically lines up in a vertical pattern that is almost close to having an undefined slope. In economics, this data would still be useful, as it shows housing has extremely inelastic demand that veers close to perfectly inelastic demand (Minnesota State University, 2017), although this model is effectively useless for statistical prediction, and does not meet the statistical definition for linearity, either.

As noted from the previous section with Box-Cox transformations, I used the modified housing data with a transformation of -1 and coded this proc reg statement

*proc reg data=Housing;*

*model rent50\_2=transHu;*

*run;*

to generate the following output: Table, Word

Description automatically generated

Graphical user interface, diagram, engineering drawing

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generated

Surprisingly, the transformation recommended by the Box-Cox output from proc transreg made this model even less useful. The above output leads to the following equation of

rent50\_2 = 1016.11525 -821.67499X +ε

that provides practically no use, as it has an R-Square value of 0.0003. I would almost say that there is almost *no relationship* between these two variables when it is transformed like this.

There is *one* model that produced slightly better results than the original “rent50\_2=hu2017” model. Namely, using the ratio of population to housing units as a proxy for housing scarcity produced marginally better results, although not by much. By using this code;

*proc reg data=Housing;*

*model rent50\_2=PtoH;*

*run;*

I was able to produce the below output:Table

Description automatically generatedDiagram

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

The above output shows a regression model that follows the below line:

rent50\_2 = 645.80518 + 179.79938X +ε

The above output also presents with an ever so slightly more promising R-square value of 0.0581. Additionally, it can be noted that the residuals still show a linear, clumped pattern, but they *are* somewhat more spread out than the other models I was able to generate and fit somewhat closer to a normal distribution than evaluating the quantity of housing units without the context of population sizes…which would model situations where multiple people might be applying for one unit simultaneously, but only one person can get that housing unit.

Lastly, I ran a proc ANOVA data step for both the classic “Price=Quantity” demand curve that would be reflected with “rent50\_2=hu2017”, as well as my modified “rent50\_2=PtoH” model with the following code:

*proc ANOVA data=Housing;*

*class hu2017;*

*model rent50\_2=hu2017;*

*run;*

*proc ANOVA data=Housing;*

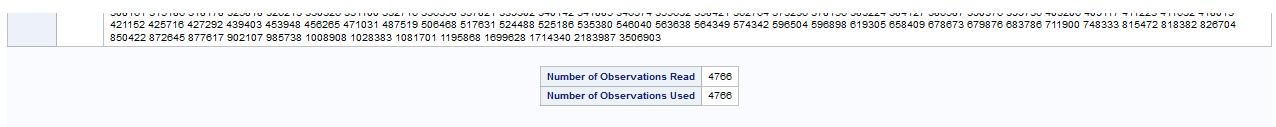
*class PtoH;*

*model rent50\_2=PtoH;*

*run;*

The above code generated the following output:

*A picture containing text

Description automatically generatedTable

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, application, table, Word

Description automatically generatedGraphical user interface, text, application

Description automatically generated*

Due to the *contradictory* ANOVAoutput, I am honestly unsure as to whether I should reject the null hypothesis or not. The F-Test statistics are 1.30 and 2.64 for the first and second model, which would imply to keep the null hypothesis that there is no relationship for now (which would make sense, given the other metrics). However, the p-values for the ANOVA tests are both <.0001, which would imply to reject the null hypothesis in favor of the alternative hypothesis that there is a statistically meaningful relationship of rent to housing quantity.

**Conclusion**

While the data applied to a simple regression model does *not* support the traditionally understood economic idea that prices are caused by quantity alone, it does yield other interesting insights. Namely, the distribution of the observations and residuals seems to imply that the demand for housing is close to perfectly inelastic.Graphical user interface, chart

Description automatically generated with medium confidence

In layman’s terms, without other factors to consider such as house quality (which is wildly variable, considering that everything from studio apartments to mansions exist), region, proximity to work, etc; people will pay nearly anything for shelter. This would still make economic sense, considering that necessities such as food and healthcare are typically inelastic, while goods such as video games, luxury goods, and other non-essentials typically have more elastic demand. As shelter is one of the core things necessary for human survival and a basic standard of living, it would make sense that it also has an extremely inelastic demand akin to food and healthcare. While this observation is neutral, it can be interpreted in a vast number of ways on the political spectrum, which can and does impact economic policy regarding affordable housing.

**Additional Notes**

\*The original dataset from HUD was slightly modified to account for SAS being unable to read ‘ñ’, ‘ó’, ‘í ‘, ‘á’, or ‘ü’ in context of several observations of Puerto Rican regions without additional coding— the following characters were changed to ‘n’, ‘o’, ‘I’, ‘a’, and ‘u’ in the dataset for the sake of SAS being able to read the observations. (Although I would like to fix this for the sake of language integrity, the documentation that detailed fixing this would probably take more time to learn than one week — especially given how complex the dataset is for analysis purposes as is. Thus, I deemed it more time efficient to change the special characters in MS Excel.)

**References**

Adamczyk, A. (2021, July 14). *CNBC*. Retrieved from cnbc.com: https://www.cnbc.com/2021/07/14/full-time-minimum-wage-workers-cant-afford-rent-anywhere-in-the-us.html

Hayes, A. (2021, August 01). *Investopedia*. Retrieved from investopedia.com: https://www.investopedia.com/terms/e/elasticity.asp

Minnesota State University. (2017). *Minnesota State University*. Retrieved from http://mnstate.edu: http://web.mnstate.edu/stutes/Econ202/Econ202/Fall%202017/study3.htm

Simone, J. (2021, July 20). *Coalition for the Homeless*. Retrieved from https://www.coalitionforthehomeless.org/: https://www.coalitionforthehomeless.org/todays-read-out-of-reach-2021/

US Housing and Urban Development - Office of Policy Development & Research. (2021, January 01). *HUDUser*. Retrieved from HUDUser.gov: https://www.huduser.gov/portal/datasets/50per.html#2022