Holly Josephs
GTECH 733 -Geocomputation II
Literature Review for Final Project
5/21/2020

Evictions and Multiple Linear Regression Modeling: A Literature Review

Introduction

In today's day and age, huge amounts of data are available on almost any subject imaginable. If this data is found, cleaned, and used correctly, it can explain the past and present and it can predict future trends. Many different statistical and modeling methods have been developed to allow analysts to make sense of almost any kind of data. In the past, the data had to be processed and the modeling equations had to be calculated and tested by humans. Now, there are thousands of tools available through different softwares, packages, and libraries to aid in statistical analysis and modeling. The tools are available and can run at the click of a button, but the user still must have an excellent understanding of the data being analyzed, the different modeling techniques available, and ways to validate their chosen models.

Given the scope of GTECH 733 and my interests in applying geospatial data science to the field of housing, I decided to research and build a project related to predicting evictions. With the goal in mind of building a model to predict evictions based on new unemployment rates, I selected several papers to provide background in the following areas: eviction and its causes and effects, previous housing studies using regression modeling, the process of multiple linear regression, and Python tools related to regression modeling. This literature review is structured along those categories to serve as a roadmap for preparing for my final project.

Evictions - Causes and Effects

Eviction is the legal process of stripping a tenant of their right to tenancy in a property. It comes in two forms: holdover or non-payment. Holdover evictions are evictions incited due to the tenant breaking some aspect of the lease or the owner's desire to have the unit vacant. There are rules regarding what may constitute grounds for a holdover eviction.[1] For example, a homeowner cannot evict a tenant who has a lease and is paying their rent on time and has not broken the lease in another way. In cases of non-payment, the tenant can pay their overdue rent to stop the eviction process. In the case of a holdover, depending on the case, the tenant may correct their behavior to comply with the lease to stop the eviction proceedings.

In these two formal types of eviction, there is a strict process that varies state by state. It begins with a formal notice of termination of tenancy. If the renter decides to continue to stay beyond the termination of tenancy, the homeowner then has to go to the court to register their initial letter and begin the formal eviction. The filing then brings a small period of time where the tenant may provide a formal answer to the eviction. At that point, the homeowner and tenant can come to an agreement or a trial will be scheduled. After the trial, the tenant may have won the case outright, may be offered the opportunity to comply with a settlement in exchange for continued tenancy, or may be given some time to vacate. If the tenant goes beyond the time given to vacate, the homeowner can schedule a marshall to come and remove the tenant's belongings and change the locks.[2]

[1] Ready,1998

[2] NYC Housing Court

Beyond these formal types of evictions, there are circumstances under which a tenant is effectively evicted or feels that they have no other option but to leave. This can be due to hostility with the landlord or other tenants, poor conditions, inability to pay rent, etc. Of course, a tenant has the right to remain in the property and await a formal process. However, many do not possess the knowledge of their rights or find the circumstances so unbearable that leaving seems to be the only option. The case of a tenant leaving due to unaffordable rent is called economic eviction.[3]

In terms of collecting data on evictions, only the formal ones may be counted. There are records of evictions filed and evictions completed. Many jurisdictions do not publish their eviction data. Princeton Eviction Lab is one of the only sources of aggregated eviction data. They provide data form the range of block groups, census tracts, zip codes, counties, and states. However, many areas have no data at the higher resolutions.[4]

When designing a study regarding evictions, it is important to consider the causes and effects to avoid overlooking potential covariants. In many ways, eviction can be part of a cycle where the causes and effects are the same. Causes of eviction can be considered on an individual level and on a societal level. On an individual level, income loss, rent increase, illegal behaviors, lease-breaking behaviors, a change of ownership of the property, a landlord/tenant conflict, etc. can all be causes of an eviction. On a societal/aggregated level, gentrification leading to rent increases, lack of control on rent increases, an economic collapse leading to major job losses, and mental health issues can all be causes of eviction. Even more generally, eviction rates have been found to be highly correlated with single-parent households, urban density, and households of color.[5] Households that face eviction often have resulting traumas. It can a be trigger for mental health issues (including suicide) and lead to further financial insecurity.[6] The effects of eviction can be part

[3] Chum, 2015
[4] Princeton Eviction Lab
[5] Chum 2015, Rojas and Sternberg 2015, and Ready 1998
[6] Rojas and Sternberg 2015.

of the cause of a future eviction, and people who have been evicted once are more likely to be evicted in the future than people who have never experienced eviction.

Previous Housing Studies Using Regression Analysis

Chum's 2015 study, *The Impact of Gentrification on Residential Evictions*, used multiple linear regression to model evictions due to gentrification in Toronto using data from 1999 to 2001. Gentrification itself is difficult to define or identify. It is related to changes in price of living, displacement of a lower income community by a higher income community, etc. It is also difficult to measure displacement because it is hard to know if people move by choice or due to "economic eviction." Because it is such a broad phenomenon and Chum was attempting to complete the study using gentrification to predict evictions, Chum had to correct for factors that are correlated with both gentrification and evictions, in order to get the isolated relationship between gentrification and eviction. Using previous studies, Chum decided to control for the factors of single-person households, unemployment, single-parent households, income, and type of employment. He controlled for these variables by finding the bivariate relationships between evictions and each of these variables through correlation analysis. His final step was to perform a multivariable linear regression. One of his main findings was that older waves of gentrification actually have a negative relationship with evictions in a neighborhood while newer waves of gentrification have a highly positive relationship with evictions.

Rojas and Sternberg's 2015 study, Evictions and Suicide, sought to model the relationship of eviction as a cause of suicide. Like Chum, Rojas and Sternberg had to control for factors that may be both a cause of eviction and suicide such as mental health and behavioral issues, high urban density, unemployment, social welfare recipiency, possession of a criminal record, being foreign born, single-parent households, age, gender, educational attainment, and substance abuse history. Schizophrenia diagnosis was also a controlled factor separate from the general mental health/behavioral disorder factor. The study then set the exposed group as people who had experienced eviction and suicide and the copmarison group as people who had experienced suicide but not eviction. Then, Rojas and Sternberg used maximum likelihood regression to find the isolated relationship. Maximum likelihood regression is a technique used in situations where the dependent event (suicide after eviction) is extremely rare in comparison to the non-event (eviction without suicide). This regression technique is available in STATA. The conclusion of this study was that people who experience eviction are nine times more likely to commit suicide than people who do not experience eviction. When adjusted for the basic demographic control variables, the result is practically unchanged. However, when adjusted for mental health and schizophrenia diagnoses, people who experience eviction are about 5-6 times more likely to commit suicide than people who do not experience eviction.

Ready’s 1998 study, An Analysis of the Factors Influencing Eviction Decisions in Six Public Housing Sites, used similar control variables as the previous studies. This study identified the age of the leaseholder, number of minors in the household, single-parent household, number of years living in public housing, monthly rent, and number of non-payment-of-rent infractions issued to the household as the control variables. However, Ready mentions that some of these factors, such as single-parent households may be difficult to define due to the prevalence of off-the-record members of households. This study used multiple logistic regression to model the relationships comparing the results between evicted and non-evicted households. The results of the study were that a significant number of non-evicted households reported at least part-time employment in comparison to evicted households. Additionally, a significant number of non-evicted households were of Social Security receiving age in comparison to the generally younger evicted group. Finally, the non-evicted group had significantly lower amounts of non-payment of rent infractions and criminal charges than the evicted group. Overall, evicted households had more children, had lived in public housing for a shorter period of time, and pay *less* rent than non-evicted households. The study found that lease/behavioral infractions were a better predictor of eviction than criminal issues or non-payment of rent. The study calls for a standardized eviction process for public housing authorities to prevent biased evictions.

The Process of Multiple Linear Regression

The studies listed above all use some sort of regression technique for describing the relationship between dependent and independent variable(s). The process of building a regression model is to first find data points that represent the independent and dependent variables of interest, then use some of the data points to allow the model to find the pattern (train the model), and then validate the model using the remainder of the data points (test the model). In the case of multiple linear regression, the model is in the form of $y=b1x1+b2x2+b3x3.....+b0$. Other regression models will try to fit the data to a different type of equation. Additionally, when building the model, one can transform a variable to increase the efficacy of the model. For example, one could use log(x1) instead of just x1.

Regression models had to be calculated by hand at one point. Now, computers do the work. Even further, there are many packages and libraries that have the formulas ready. The only real decisions the person building the model must make are what type of model to fit the data to, how much of the data should be used for testing and training, which variables to include in the model, if any variables should be transformed, and how well the model must perform to be usable for the task at hand. The process of deciding which model and variables to use is called exploratory data analysis (EDA). SOme parts of EDA may include building charts and graphs of the data, looking for covariances, etc.[7]

Python Tools for Regression Modeling

Igual’s chapter on Regression Analysis and Wu’s study on Housing Price Prediction Using Support Regression Vectors both provide some practical tips for performing regression analysis using Python. Both papers show the results and display the data using Python.

Igual uses Pandas and SKLearn. He recommends Pandas for building a dataframe of the data to be modeled. He also uses Seaborn to build charts of the data in the Exploratory Data Analysis phase. Seaborn has a plot that shows the spread of the dependent variable at each point of the independent variable. Seaborn also has this capability and can even change to a polynomial regression by changing the order of the model. He also recommends the use of SKLearn. SKLearn has the code for actually running the regression models on data. His last two recommendations are the use of Seaborn’s heatmap and Panda’s scatter matrix. Both are for the exploratory data analysis phase for seeking relationships between different variables. The heat map builds a matrix of the different variables and turns red at intersections where the variables are highly positively correlated, blue at intersections where the variables are highly negatively correlated, and white where there is minimal relationship. Scatter matrices give the same information plus they show the spread of the data at each cell in the matrix. [8]

Wu’s study on using regression analysis to predict housing prices also uses SKLearn and Seaborn. Wu uses the Seaborn library for visualizing the relevant independent variables, at one point employing the correlation heatmap mentioned above. He also uses principal component analysis to determine how much variance

[7] Igual, 2017
[8] Igual, 2017

is explained by the variables. SKLearn has this functionality called ‘Standard Scalar.’ In Wu’s study there were tens of variables at play. To find the ones to include the model, Wu used another SKLearn method called recursive feature elimination which uses an estimator model to rank the features. This method selects the usable features by recursively calling the smaller set of features. Wu also uses Random Forest which is an algorithm built with many decision trees. Sklearn has RandomForestRegressor(), and after fitting the data with RandomForestRegressor(), we can call feature_importances to get the importance score for every variable. The higher the score, the more important the variable is. Finally, Wu also uses SKLearn in the model evaluation stage by finding the R square score, MAE, MSE, and RMSE. [9]

Conclusion

As can be seen by the various studies, regression modeling is a solid choice for data that has multiple independent variables that relate to a dependent variable. Not only is the method itself a good choice for the data, but Python and its available packages makes the modeling process extremely straightforward. Just by learning the syntax, one can build an entire model, test it, and evaluate it.

When building my study regarding evictions caused by coronavirus unemployment, I will have to be careful to control for covarying variables. I will spend some time learning the functions of Pandas, SKLearn, and Seaborn as it seems these packages will provide some excellent functions that would make the study more robust and make my data more presentable in a Python Notebook.

[9] Wu, 2017

Bibliography

Chum, Antony (2015) The impact of gentrification on residential evictions, Urban Geography, 36:7, 1083-1098, DOI: 10.1080/02723638.2015.1049480

Desmond M, Gromis A, Edmonds L, Hendrickson J, Krywokulski K, Leung L, & Porton A (2018). Eviction lab national database. Princeton University, 2018, <www.evictionlab.org>

NYcourts.gov/courts/nyc/housing

Igual L., Seguí S. (2017) Regression Analysis. In: Introduction to Data Science. Undergraduate Topics in Computer Science. Springer, Cham

Ready, J., Mazerolle, L. G., & Revere, E. (1998). Getting evicted: Social factors influencing eviction decisions in six public housing sites.

Rojas Y, Stenberg S-Å. Evictions and suicide: a follow-up study of almost 22 000 Swedish households in the wake of the global financial crisis. J Epidemiol Community Health 2016;70:409–413

Wu, Jiao Yang, "Housing Price prediction Using Support Vector Regression" (2017). Master's Projects. 540. DOI: https://doi.org/10.31979/etd.vpub-6bgs