Parul Datta, Amrit Dhillon

ECON124: Machine Learning for Economists

Michael Leung

9 June 2025

Final Paper: Causal Effect of Local Median Income of an Area on the Value of Homes

1. Introduction

This paper investigates the causal effect of median income in an area on the value of a property, using data from the Housing Affordability Data System (HADS), 2002. The goal is to estimate whether and how changes in income levels causally affect housing values, rather than simply establishing a correlation or making predictions.

Understanding the causal effect of the median income of an area on property values has important implications for officials determining housing policy and for urban planning. Take for example when policymakers attempt to improve the affordability of a neighborhood or provide stability for residents with low income using economic development programs. If it is shown that higher median income is linked to increased property values, implementing these programs in these areas could actually have a reverse effect, and increase the property values of the areas, pricing out the local residents, and contributing to gentrification.

Quantifying these relationships allows for analysis that can help make more effective and equitable housing policies. Furthermore, understanding how these relationships differ in different types of cities(urban, suburban, etc.) is important as people in different settings value different things regardless of income.

2. Data

As previously mentioned, the data being used is from the Housing Affordability Data System (HADS), 2002 dataset, and it can be found at this link. The dataset includes a wide range of information on the subjects of the datasets including their local median income, their individual incomes, the value of their homes, the number of people in their household, and much more. Its original purpose pertained to estimating the affordability of a housing unit/home for an individual based on the other metrics available in the dataset.

For this study, we focused specifically on homeowners by excluding observations where the home value was recorded as zero, which corresponded to renters in the dataset. We also refined the dataset by trimming down variables that were less relevant to our research question, such as niche metrics like “Low Income Adjusted for # of Bedrooms” or “Cost08 Relative to Poverty Income (Percent)” while preserving all key indicators necessary for meaningful interpretation.

In this analysis, the treatment variable (D) is local median income(LMED), and the outcome variable (Y) is the value of a property(VALUE). The regressors we are choosing include: BEDRMS(# of bedrooms), BUILT(when property was built), NUNITS(# of units in the property), ROOMS(# of rooms in the property), UTILITY(monthly Utility costs), and AGE1(age of head of household). The reason we chose these variables is because we believe that they are variables that could have an effect on both median household income and property values, so we are assuming these will act as controls in the regression.

While the majority of the research for this paper went well, one limitation did appear early on in our work. Rather than having separate local median incomes for different neighborhoods in an area, the Local Median Incomes were all the same across each subject who fell into that city/area. For instance, one value of the metric “SMSA” which held the different places the subjects were from was “Anaheim-Santa Ana (Orange County), CA”. This can be misleading as it is evident that some areas of Orange County are better than others, with some being crime ridden and others being made up of multi-million dollar homes. In order to combat this, we used another metric in the dataset named “METRO” which assigns the subject a label dependent on the type of area they live in(Primary Central City, Secondary Central City, etc.). Doing so allowed us to see the difference in home values across different types of urban and suburban environments, helping to capture some of the within-city variation that the uniform Local Median Income variable could not account for.

To understand how the METRO category works, take for example the “Anaheim-Santa Ana (Orange County), CA” category of the SMSA variable. METRO 1 in this area would correspond to cities such as Santa Ana which is the county seat(administrative center) of Orange County, has a significant economic and historical impact, has a large population, and holds landmarks such as the Civic Center and Downtown Santa Ana amongst other things. METRO 2 would include cities such as Anaheim, a city which is both populous and economically significant and holds places like Disneyland and the Convention Center, but is not the county seat or center of administration. Other cities such as Fullerton or Garden Grove would fall into METRO 3 as they’re representative of themselves and have their own local governments, populations, downtowns, and other makings of cities but are smaller and less central than the cities listed above. Finally, cities such as Irvine and Tustin would make up METRO 7, and they’re primarily residential areas and suburbs that surround the primary cities of the metro areas.

Figure [1] provides a visualization of the mean values of significant variables: AGE1, ROOMS, UTILITY, and VALUE within each category of METRO. It is evident that the average value of homes is highest in METRO 7 which corresponds to suburban cities and lowest in METRO 3 which corresponds with cities with a decent population and economic value but not within the top 3-4 cities in the area. The utility prices on average are relatively stable from METRO 1 through 3; However, the utilities increase significantly in METRO 7 which may suggest larger homes or larger family sizes are more prominent in METRO 7 as generally larger homes or numbers of people within a home use a larger amount of utilities. Through the barplot it seems as though there is little variation between METRO areas, but this is far from true.

Prior to running our post lasso regressions, we wanted to gain an understanding of the effect of local median income(LMED) on the value of a home, where the mean value of a home in the dataset was $191,079.70. To establish how LMED influenced home values across different regions, we ran logistic regressions on each METRO area on whether a home's value exceeded $80,000 a threshold far below the mean, as well as whether a home exceeded a threshold well above it, $400,000. The results of these regressions reveal important variation across METRO areas. In METRO 1, LMED was significant but was a negative predictor of home value at both thresholds, suggesting that local income levels alone did not drive high-value properties in dense urban areas. In contrast, METRO 7, LMED was both strongly and positively associated with home values at both thresholds, indicating that in suburbs and residential areas, higher local incomes were closely tied to more expensive home values. METRO 2 showed an insignificant relationship between local median income and value at the lower threshold but turned notably negative at the upper bound, which points to non-linear relationships and a sign of possible wealthy homeowners clustering in areas with lower valued properties. Finally, METRO3 showed weak and insignificant relationships, most likely due to the small sample size and less variation in home values.

To further unpack the relationships, we estimated simple OLS regressions using log(VALUE) as our outcome, allowing us to interpret coefficients as percent changes. The results of these regressions mirrored earlier patterns, and METRO 7 once again stood out. LMED had a strong, positive, and highly significant effect(2e-16 > p), and characteristics such as rooms, bedrooms, and utility costs were also very predictive of higher valued homes. Within both METROs 1 and 2, LMED was insignificant, but the size of the home(# of bedrooms and rooms) and the year the home was built had positive effects and were influential in predicting the value of the home. Surprisingly, METRO3 had a negative and statistically significant coefficient for LMED(p = 0.0037), which could indicate unique dynamics within this metro area, omitted variable bias, or some other issue.

3. Methodology

In order to estimate the causal effect of LMED on home value, we used the Post Lasso Selection method we learned in Lecture 3. The method is efficient in estimating the effect of treatments when many control variables are available and it is unclear which should be chosen. Specifically, we ran two separate lasso regressions within the Post Lasso method. One predicted our outcome variable(log(VALUE)), while the other predicted the treatment variable (log(LMED)). Both regressions had a large set of variables to choose from, and after choosing, the union of both sets was taken in order to use only these variables in a final OLS regression of the outcome on the treatment. Doing so helped to ensure that the regression is conditioned on only variables that matter to both the outcome and treatment, thereby reducing omitted variable bias without overfitting. We ran this Post Lasso Selection method upon the datasets four times, once with only our originally selected controls, and three times more, each time utilizing a different interaction term that we assumed would have changed the estimated effect of variables.

As aforementioned, we utilized three interaction terms within our regressions: (LMED * AGE1), (BEDRMS * ROOMS), and (UTILITY * ROOMS). For (LMED * AGE1), we chose to interact these variables together because we believed that older aged heads of household may respond differently to changes in LMED. For example, retired individuals probably would not benefit from increasing LMED in the same way that working households would. The (BEDRMS * ROOMS) interaction was meant to capture how the use of space in a home affects value. For instance, in homes with fewer rooms overall extra bedrooms may add value, but in larger homes with more rooms they might not add as much value. Finally, we included (UTILITY * ROOMS) to account for how monthly utility costs scaled with home size as larger homes typically have higher utilities. Each interaction that was chosen was motivated by the idea that their combined effects could influence both home value and LMED in non-linear ways, and the variables were more complex and dependent on each other.

One major concern for the validity of our causal estimates was omitted variable bias. To be able to make credible claims about LMEDs causal effect on home values, we must assume that all relevant confounding variables which affect both local median income and home value have been controlled for; However, this assumption is likely disregarded in our study. Take for example, neighborhood quality and safety, a key variable in estimating either, which is something we did not account for and could create bias in our results. Areas with access to better public schools, with lower crime rates, or greater access to general amenities most likely attract higher income residents, thereby raising local median income and leading to higher home values. Without controlling for aspects such as neighborhood quality and safety, the estimated effect of LMED on home value will be upwardly biased, and in addition to the true effect of local median income, the influence of unobserved traits will be captured. This directly violates the conditional independence assumption that was discussed in Lecture 3, Section 3.3, which stated that valid causal inference requires no omitted variables that are correlated with both the treatment(LMED) and the outcome(VALUE).

Another important source of potential bias could be local economic growth/investment which leads to booms in employment, infrastructure, development, or population growth within a metro area as these all contribute to increases in both income and home values in an area. These types of booms are likely driven by outside forces such as government investments/policies or industry investments which have not been included within our dataset. By omitting such variables, LMED becomes unfairly associated with increases likely tied to economic growth/investment within an area and attribute their effects to local median income alone. Similarly to neighborhood quality and safety, by omitting these variables, it is plausible that the relationship between local median income and home values will become confounded and violate the conditional independence assumption. Given these limitations, we must be careful in interpreting our estimates causally as the assumptions required for valid causal inference do not appear to be fully met in this context.

4. Main Results

The results of our analysis using the Post Lasso Selection method showed a pattern in the causal relationship between local median income(LMED) and the value of homes. In METRO 7, which represents suburban and residential areas, the LMED consistently had large, positive, and highly significant effects on home values across all of our models. In our baseline Post Lasso model(without any interactions), the coefficient on log(LMED) was approximately 0.0000169, implying that a $1 increase in LMED predicted a .0017% increase in home value or in other words, a $1000 increase in LMED predicted a 1.7% increase in home value. After introducing interaction terms (BEDRMS x ROOMS) and (UTILITY x ROOMS), the coefficients ranged from 0.0000167 to 0.0000169 and the standard error remained below 0.000000931, while the p-values stayed below 0.001 consistently, suggesting that even when accounting for nonlinear housing characteristics and potential confounders LMED is the dominant influence of housing value in suburban areas. It is our belief that although a $1000 increase to LMED might not seem impactful with only a 1.7% increase in value; If it is scaled up, it has a meaningful and economical significant effect. For example, if the average income in a suburb were to rise by $10,000 over a few years, the home value is predicted to rise 17% potentially leading to affordability issues and gentrification in suburban communities.

METRO 1 and METRO 2, which represent more central cities, did not show consistent or significant causal relationship between LMED and the housing value. It should also be noted that other features such as the number of bedrooms and the year the home was built consistently showed both positive and significant effects on log(VALUE), suggesting that in the central cities physical home characteristics have a greater impact for determining value than LMED. For example, in METRO 2 one additional bedroom is associated with a 21.6% increase in the value of a home, and each additional room adds 8.1%. The results of the interaction of AGE1 and LMED (AGE1*LMED) revealed that as the age of the homeowner increases in METRO 2 there is a slight increase in the value of the home, but this effect is small, as for every one-year increase in homeowner age, the marginal effect of LMED on log(VALUE) is only .000124. In other words, with a 20 year age gap and a 10% increase in LMED there is only a 0.024% positive impact on the value of the home.

Surprisingly we found METRO 3 to hold a negative coefficient of approximately -0.0000655 that was statistically significant as p was 0.0053. The findings indicated that with a

$1 increase in LMED there is approximately a 0.00655% decrease in home value which can be translated to be understood as an $1,000 increase in LMED predicts a 6.5% decrease in home value. This result may be a reflection of the small and unrepresentative sample size as METRO 3 only holds 152 observations. METRO 3 also has the highest predictive error rates at over 30%, and it can be concluded that the dataset may not be reliable in understanding the housing dynamic in METRO 3. This can be further understood by the results provided by the interaction of number of bedrooms and rooms(BEDRMS*ROOMS), which showed a small but statistically significant negative coefficient of approximately -0.000626 while the p-value was 0.0105. This suggests that as the number of bedrooms and rooms increase in smaller cities that are less populous than cities in METROs 1 and 2, there is a marginal decrease in the value of homes. This can be explained by the interaction of utilities and rooms (ROOMS*UTILITY). The result of the interaction shows a statistically significant negative correlation as p = 0.0053 and the coefficient is -6.55e-05. As the utility price increases by $50 and the number of rooms increase by 1 the value of the home decreases by .33%. The METRO 3 market might value more cost effective homes, which explains as bedrooms and rooms increase why the value decreases, as extra rooms create additional costs that might be viewed as unnecessary.

To further illustrate the effects of LMED on home value, we conducted counterfactual predictions by increasing LMED by 10% in each METRO area and measuring the resulting change by predicted log(VALUE). In METRO 1 there was an average 1.34% decrease in value, while in METRO 2 there was roughly 0.53% increase in value. METRO 7 again confirmed our main findings, that a 10% increase in LMED was associated with a 10.2% increase in home value, an effect which reflects suburban values. These counterfactuals held almost entirely consistent throughout the different regression models and suggest that while LMED can be a powerful predictor for home value in certain settings, its impact is highly dependent on context and unobserved confounders.

In conclusion, our analysis is indicative that the effect of local median income on home values varies significantly across different types of metropolitan areas. While in suburban areas like METRO 7 a strong, positive, and consistent relationship is shown, in more central and more populous areas, a weaker or even negative association was found. These findings highlight the importance of contextual and geographical factors when forming housing policies. Ultimately, while local median income can be a key driver in home values in certain environments, its influence is molded by complex relationships of household characteristics, unobserved factors, and regional preferences.

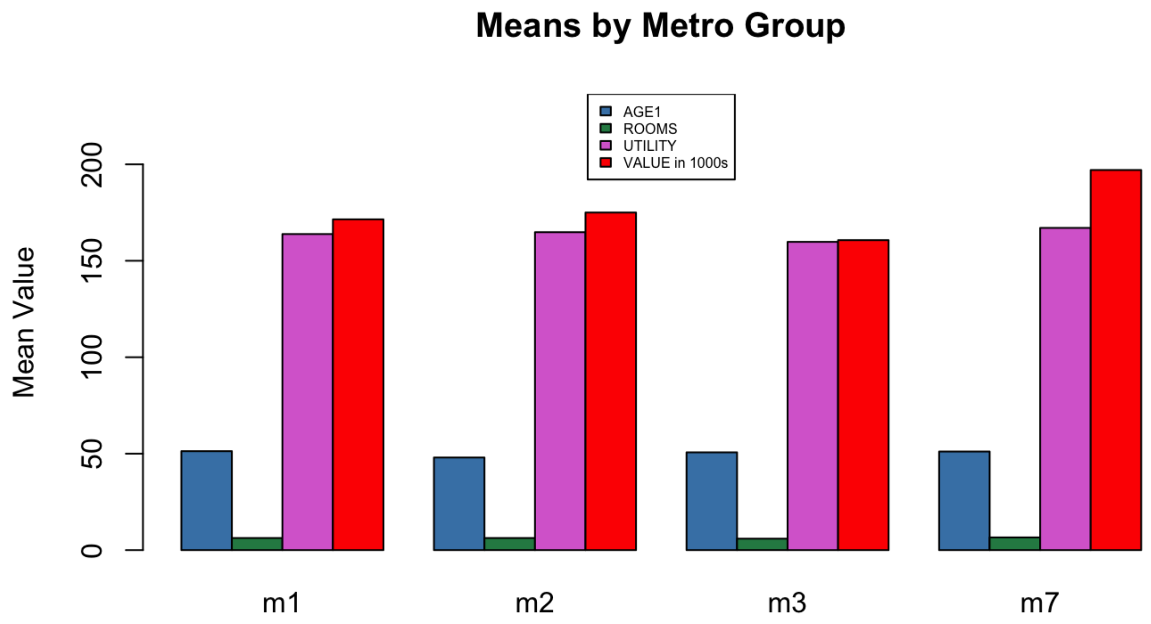
Means by Metro Group
AGE1
ROOMS
UTILITY
VALUE in 1000s
Mean Value
0
50
100
150
200
m1
m2
m3
m7

Figure [1]