Jingzhi Chang, Perry Schaffner, Tingting Huang, Xintian Li (Team Autos)
CPLN 501
Guerra
12/16/2019

Demographic and Spatial Factors Related to Commuting by Public Transportation

Introduction

What factors are related to commuting by public transportation? We were interested in finding out which factors were most closely related to who were riding public transportation to work. We found this to be interesting because we knew that many factors can correlate with why people choose to use public transit for their commutes instead of other modes. Additionally, the four of us had some preconceived ideas about which groups of people would be higher users of public transportation, but through trying to answer our question objectively, some of our assumptions were proven wrong. We were interested in examining a question related to public transportation because we all are in the transportation concentration and were interested in using transit data to create a story about who is using this mode for their commute to work.

To do this accurately we felt it best to focus on Metropolitan areas of the United States because these are the places where public transportation infrastructure exists. Based on this we felt that we would get the most accurate picture of the factors of people who were using public transportation. By learning about who are riding public transportation, we would also be able to learn about which groups are not riding public transportation. This could have huge implications for the ways in which transit companies market themselves to users. Transit companies could work to provide better service to those underserved groups which could result in higher ridership and more revenues. Moreover, it would also benefit the whole society to have more public transit ridership since there might be less congestion and emission.

We first did some country level analysis to describe the big picture and steer us in the right direction so that we would could estimate the factors relevant to commuting by public transit. We chose our variables of our census tract level regression mainly based on what we found from our country level analysis. The demographic factors that we chose were gender, race, median income, age, poverty, education, and disability. Then we also considered some spatial characteristics of population density, what the county type relative to its metropolitan area was, as well as the number of public transit stations.

Originally, we hypothesized that the data would tell us a few different things. We thought that fewer women would ride transit than men because women care more about safety and comfort. We thought that higher median income would result in lower ridership since wealthier people were likely to own cars and drive. We thought areas

with higher population density would have higher ridership, and that more stations would lead to greater ridership. The data supported some of our original hypotheses, but we also found factors that we hadn't hypothesized which were correlated with higher percentages of people using public transportation. In summary, population density was an important factor, and the number of stations was also closely correlated with ridership levels. What we found was that the spatial factors have a stronger relationship with commuting by public transit than the demographic factors.

Methodology

Data Source

We collected most of our data from the 2017 ACS 5-year estimates. These data are the demographic and commuting data for each census tract within metropolitan areas. (For which county is in certain metropolitan area, we referred to the list of definition given by the U.S. Census Bureau) Some of these data we viewed at a national or county level, but we used census tract level ACS data for regression. Additionally, we used the Bureau of Transportation Statistics for our transit stops data. We believe that there are missing data for some metropolitan areas, however, this was the best we could found. We used ArcGIS to visualize this data clearly. Once we downloaded and cleaned our data, we were ready to begin our research and analysis.

Methods

As a rough overview of our methods, the first step in our process was to download all the data that we thought necessary. This included all the ACS data above, and the Bureau of Transportation Statistics data with the transit stops. We then analyzed the data both geographically and through some charts in excel. This gave us a basis of understanding how our public transit data was geographically distributed. This also showed us how some of the demographic data was distributed across the country.

Next, we built our regression models. First, we selected our variables based on our original theory and the country level analysis. Then we built the regression with the selected variables and left out those statistically insignificant ones that the model showed. Once we had come up with the best model for our regression, which included the highest adjusted R squared and had statistically significant variables but was also concise, we settled on our final model.

After that, we discuss the relationship shown in the regression and our coefficients. We also did some more analysis to talk about the performance of the regression, including error terms. Finally, we compared the findings in the regression with our initial theory. And speculated as to why there were differences or contradictions between our hypothesis and our findings. In the sections below we will explain in detail the steps of our analysis for each method.

Findings

The Spatial Characteristics of American Commuting by Public Transportation

We used ArcGIS first to show the spatial distribution of the percentage of workers commuting by public transportation. It is important to note that we grouped the various modes of public transportation together here to include bus, streetcar, subway or elevated, railroad, and ferry boat. The first two maps below show the percentage of people who are commuting by public transportation, and population density. We saw that there was a relationship between population density and people commuting by public transportation. From these maps we can see that this relationship is higher percent of workers commuting by public transit in area with greater population density. Figure 3 shows the spatial distribution of public transportation stations in the metropolitan areas of the United States that have public transportation station data available. We can also see that the east coast of the United States has more public transit infrastructure than the west coast. This is the case for both the number of systems, as well as the density of stations. There is a potential limitation of our data, as we do not know if we are missing transit station data from certain areas of the United States. This first part of our exploratory analysis provided us with a spatial understanding of the geographic areas where people are mainly using public transportation to commute to work.

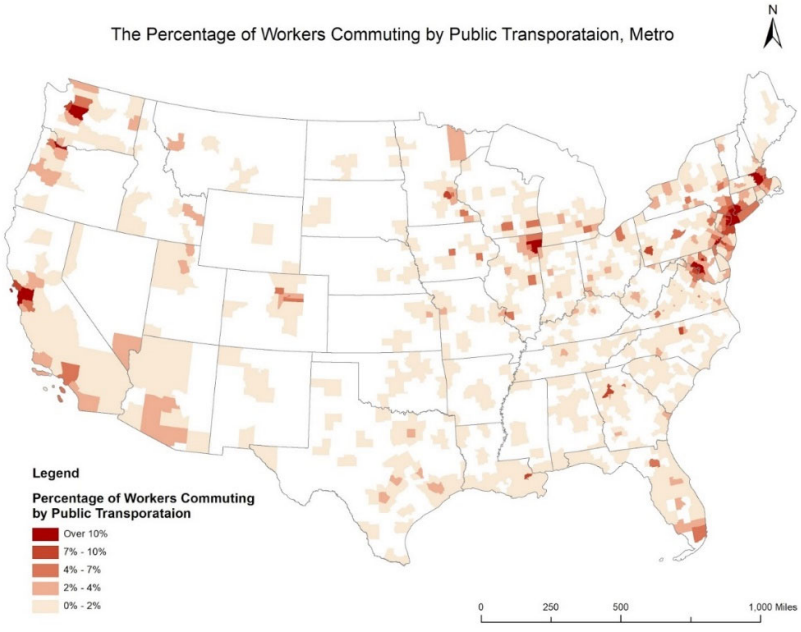

Figure 1 The percentage of Workers Commuting by Public Transportation, USA Metro
(Data Source: 2017 ACS 5-year estimates)

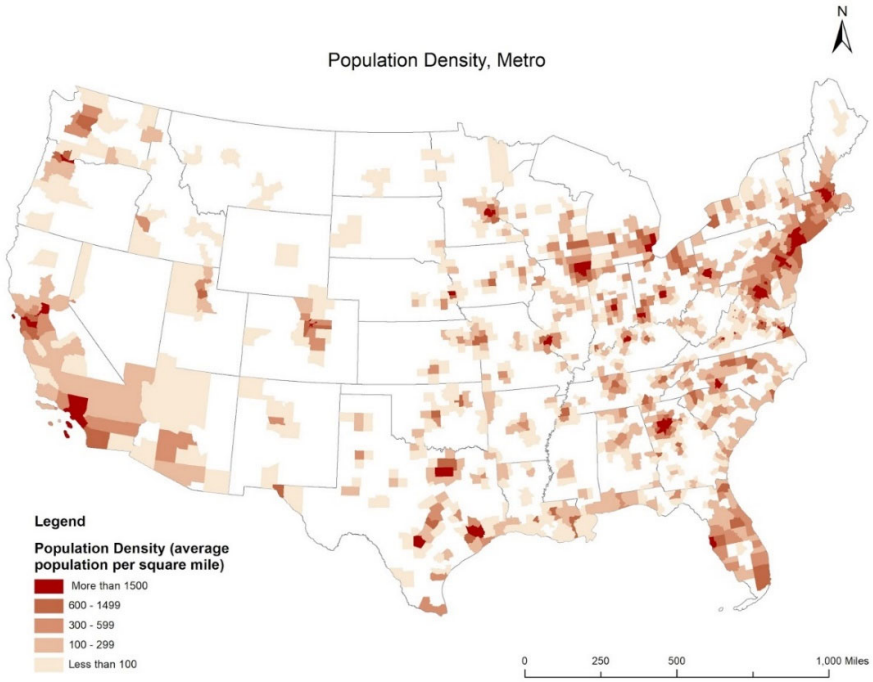

Figure 2 Population Density, USA Metro
(Data Source: 2017 ACS 5-year estimates)

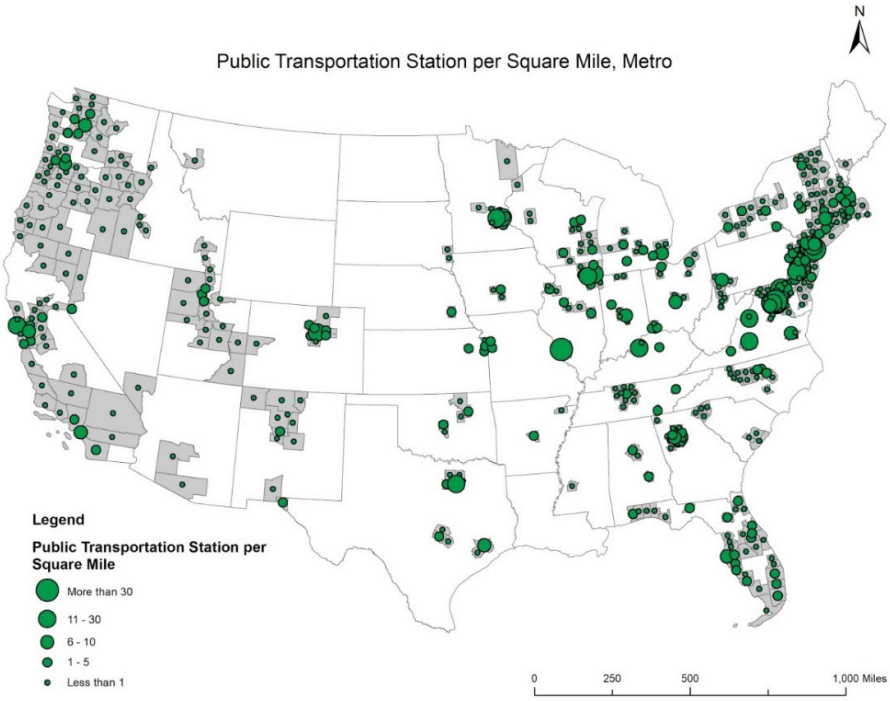

Figure 3 The Distribution of Public Transportation Stations
(Data Source: Bureau of Transportation Statistics)
Notes: this graph only plot the metro areas with public transportation station data available.

The Demographic Characteristics of American Commuting by Public Transportation

Following our analysis of our spatial factors, we needed to look at our demographic factors at the country level. Instead of seeking out any specific characteristics that we thought would have causal relationships, we looked at large group factors to see if there was any significant characteristics at the country level. Below we looked at gender distribution, racial distribution, both the median age of commute mode as well as which age groups were commuting the most, and finally median income. We predicted that these factors would indicate some guided focus for our regression analysis later. We would like to point out that we chose to eliminate driving alone as a unit of analysis for race and gender because it greatly skewed our results, since commute by other modes represented a smaller proportion in relation to driving.

Gender:

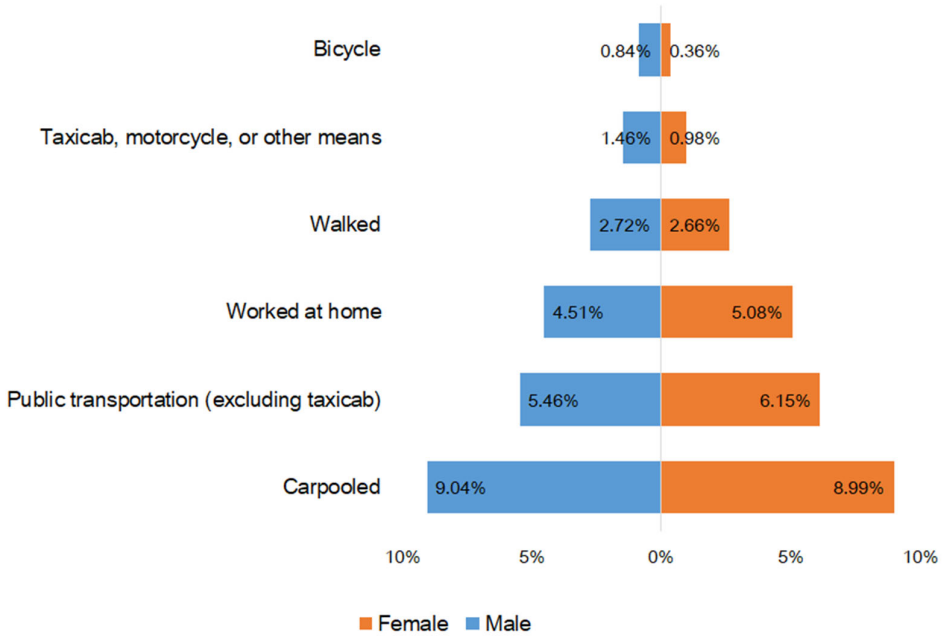

Figure 4 Commuting Mode Choice by Gender
(Data Source: 2017 ACS 5-year estimates)

We found that when we compared the mode choices by gender there were slightly more females who were riding public transportation and working at home than males.

This led us to predict that there is a relationship between gender and commuting by public transportation. We decided that because the female percentage was higher than males, that we would use the percent of females as one of our independent variables in our regression model.

Race：

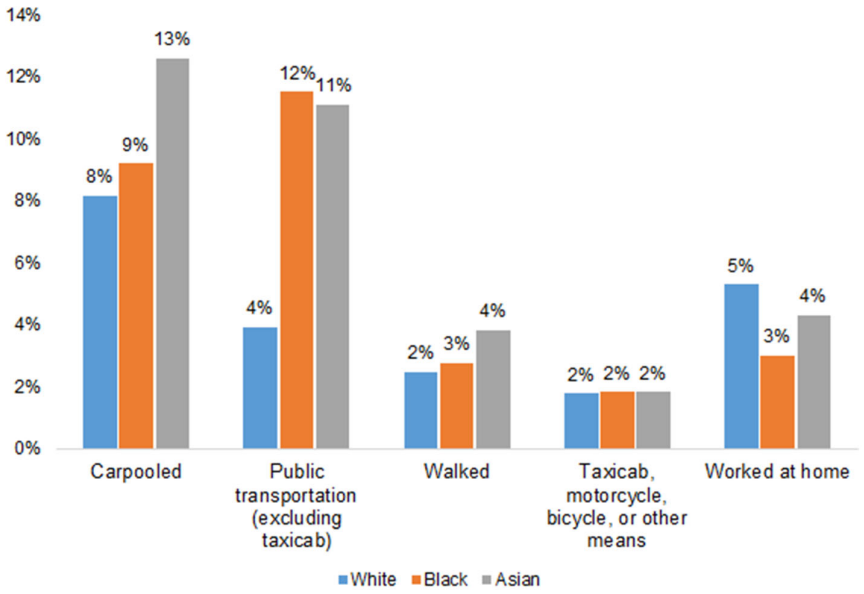

Figure 5 Commuting Mode Choice by Race
(Data Source: 2017 ACS 5-year estimates)

Race seemed like another demographic factor worth considering at a country level scale because we wanted to see what the racial distribution was like between modes. Here we saw more substantial findings than we did with gender. We can see that Black people were three times as likely to ride public transportation as White people. And Asian people were nearly three times as likely to ride public transportation as White people. This racial distribution can also be seen in the other modes, except for working from home where Whites make up the largest percentage. Because the percentage of white people commuting by public transportation was so low, we decided to use this as one of our independent variables because we assumed there could be a significant negative relationship between commuting by public transportation and the percentage of White people.

Age:

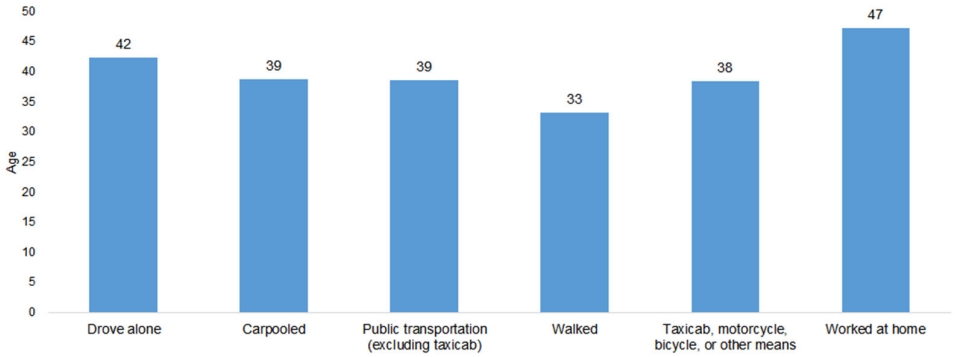

Figure 6 Median Age by Commuting Mode
(Data Source: 2017 ACS 5-year estimates)

We also felt that age was a demographic factor that could have a relationship with commuting to work. To understand this relationship a little bit better on a country level scale we looked at the median age distribution of commute mode. We found that younger people are likelier to walk, bike, or motorcycle to work, with a median age of 33 and 38 respectively. Then those carpooling and riding public transportation for their commute had a slightly higher median age of 39 years old. Then an even higher median age for those driving alone or working at home, with median ages of 42 and 47 respectively. This shows that there is a correlation between commuting mode and age, because the modes that tend to require more physical agility have a lower median age of riders. We were not sure whether this relationship will be relevant enough as a demographic factor so we would decide that in our regression model.

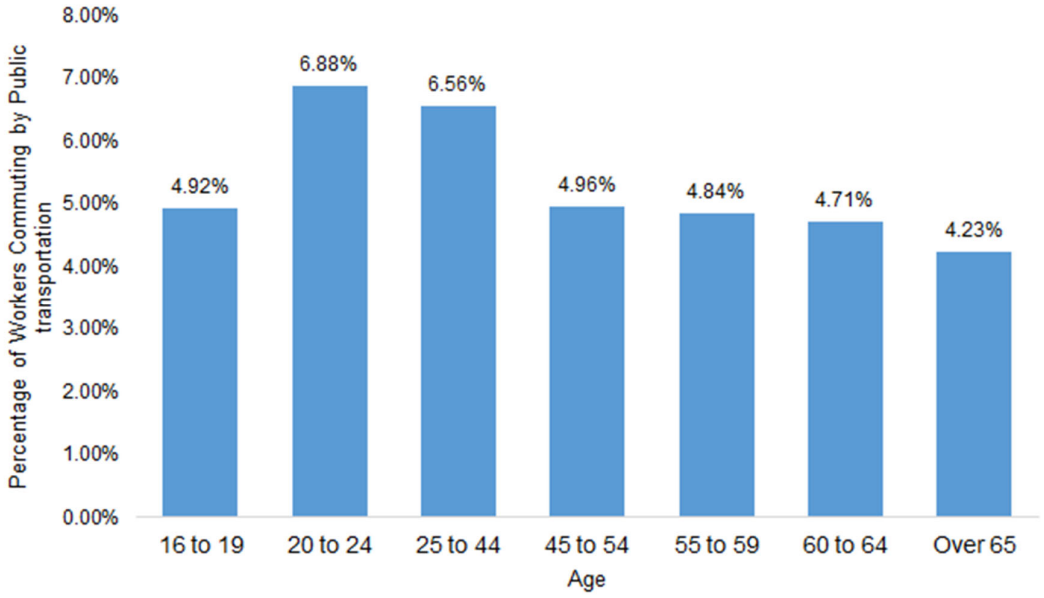

Figure 7 Proportion of workers commuting by public transportation by age
(Data Source: 2017 ACS 5-year estimates)

We also wanted to look at the public transportation mode share distribution

between age groups. Figure 7 above shows the percentage of people in each age group who are commuting by public transportation. As we can see from the graphic, a larger percentage of younger people are commuting by public transportation than older people. People whose age are between 20 and 44 tend to have a higher possibility to commute by public transit than other age groups. We would try including this in our regression.

Income:

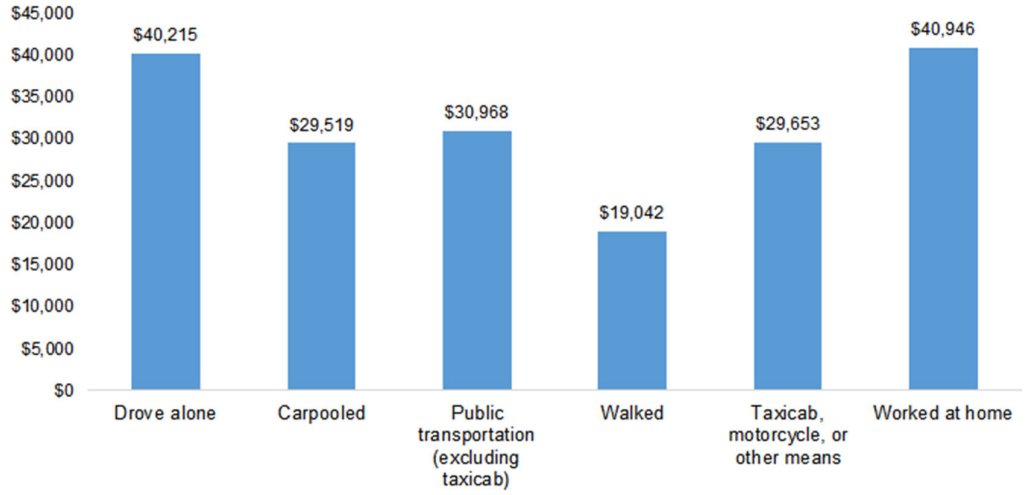

Figure 8 Median Income (Earnings in the past 12 months) by Commuting Mode (Data Source: 2017 ACS 5-year estimates)

Finally, we wanted to look at the relationship between median incomes (earnings in the past 12 months) of people and their commute modes. The relationship between these two factors isn't anything particularly surprising, we can see that a mode like walking has the lowest median income. And more costly modes like driving have much higher median incomes. The median income for public transit falls right in the middle. To detect an accurate relationship between income and taking public transportation we would put the factor into our regression model.

Modeling a Linear Regression from Spatial and Demographic Characteristics

For the regression, we used census tract level data to get the most accurate

results. All these census tracts are within the metropolitan counties. We took our factors from the exploratory analysis above and some other factors that we thought to be relevant and used them as independent variables for our linear regression model. Our dependent variable is the percent of workers commuting by public transportation. Below we will explain our regression in more detail, as well as discussing some of the limitations that we had through the process.

	Dependent variable:
	Percent of commutting by public transit
Population density	673.546***
	(7.944)
Percent of white	-7.632***
	(0.122)
Percent of females	0.036***
	(0.006)
Percent of poverty	-0.116***
	(0.004)
Whether lies in central areas	-0.338***
	(0.102)
Median income of the workers	0.0001***
	(0.00000)
Percent of disabled workers	-3.589***
	(0.811)
Percent of workers with higher education	2.648***
	(0.200)
The number of Transit stations	0.011***
	(0.003)
Whether transit station data is missing	-1.496***
	(0.068)
No Car Ownership	0.687***
	(0.004)
Constant	3.220***
	(0.379)
Observations	61,050
R^2	0.739
Adjusted R^2	0.739
Residual Std. Error	6.391 (df = 61038)
F Statistic	15,723.350*** (df = 11; 61038)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 9 Regression Result

For our independent variables, we incorporated both spatial and demographic factors. Our spatial factors were population density, whether the tract was centrally located within the metropolitan areas, the number of transit stations, and whether the station data in the census tract was missing. Our demographic factors were the percentage of white, percentage of females, percentage of people in poverty, median income of workers, percentage of workers with higher education, and percentage of workers having no cars. All these independent variables were statistically significant at α=0.01 level. The adjusted R squared was 0.739, indicating this was a good fit.

We found that population density showed a strong positive relationship among all our independent variables. The regression shows that for 1 person/m^2 increase in population density, there will be a 673% increase in the percentage of people commuting by public transit. Population density has a statistically strong relationship

with higher percentage of workers taking public transit.

Another spatial factor was whether the tract is in a central county of the metropolitan area or not. Large metropolitan areas that encompass many counties could have some central counties and some that are outlying. We found this variable to have a negative relationship with the percentage of people commuting to work by public transportation. If the tract is in the central counties of the metropolitan area, there will be a 0.338% decrease in the percentage of people who commute with public transit.

We assumed that we would find a strong relationship between ridership and the number of transit stations, and this is proved by our regression. For every one more transit stations in the census tract is associated with a 0.011% increase in percentage of people riding public transit to work. Considering the low average level of public transit ridership in the U.S., this is quite significant.

An associated factor we had to consider was whether transit station data were missing. This could very easily be a limitation of our project, because we had to assume complete transit data for the metropolitan areas, but we in fact did not have all the transit data that exist. This resulted in a negative relationship between missing station data and the percentage of people riding public transportation to work. If the tract doesn't have transportation station data, we saw a –1.496% decrease in the percent of people commute to work by public transit.

The regression shows a strong negative relationship with the percentage of commuters taking public transit and the percentage of White people. For each one-unit increase in the percentage of white people, there will be a 7.632-unit decrease in the percent of people commuting by public transportation. This confirms our findings from the national level above that tells us that a much lower percentage of White people are commuting to work by public transportation.

The percentage of female workers has a slightly positive relationship with the percentage of people commuting to work by public transportation. We saw in the country level that only a slightly higher percentage of females are commuting to work by public transit, so seeing a small positive relationship here affirms that observation. For each one-unit increase in the percent of female workers, there will be a 0.036-unit increase in the percentage of people commuting to work by public transportation.

The relationship between median income of workers and the percent of people commute by public transit is positive. Technically for each one-dollar increase in the median income, there is a 0.0001% increase in the percentage of people commuting to work by public transit. The coefficient is statistically significant at $\alpha=0$ level. This is a surprising finding, for it indicates that people with higher income may be likelier to commute by public transportation.

The percentage of people in poverty has a small negative relationship with the percentage of people commuting to work. For every one-unit increase in percent of people living under the poverty line, we will see a 0.116-unit decrease in percent of people commuting to work by public transit. This finding is consistent with what we find in with median income of workers.

We determined that disability could be a factor in choosing whether to ride public transportation to work. We used this as another independent variable in our regression

model. The model showed a negative relationship. For every 1-unit increase in percentage of disabled workers, there would be a 3.589-unit decrease in percentage of people riding public transit to work. This relationship likely shows that public transportation isn't always accessible to workers with disabilities.

We also were curious to see if education played a role in people's decision to commute to work by public transit. We found a very strong positive relationship between the percentage of workers with higher education (bachelor and higher degree) and the percentage of people commuting to work by public transportation. For every 1 unit increase in the percent of workers with higher education, percent of people commuting to work by public transit would increase by 2.648 units. This tells us that many educated people are choosing to commute using public transit.

Near the end of our regression modeling we decided to add car ownership as an independent variable in our regression because we thought that could influence people's choice to commute by public transportation. We did find a positive relationship between commuting to work by public transportation and not owning a car. For every 1-unit increase in percent of people who do not own a car, there will be a 0.687 increase in percentage of people commuting to work by public transportation. However, we are aware that car ownership and commuting by public transportation mutually affect each other.

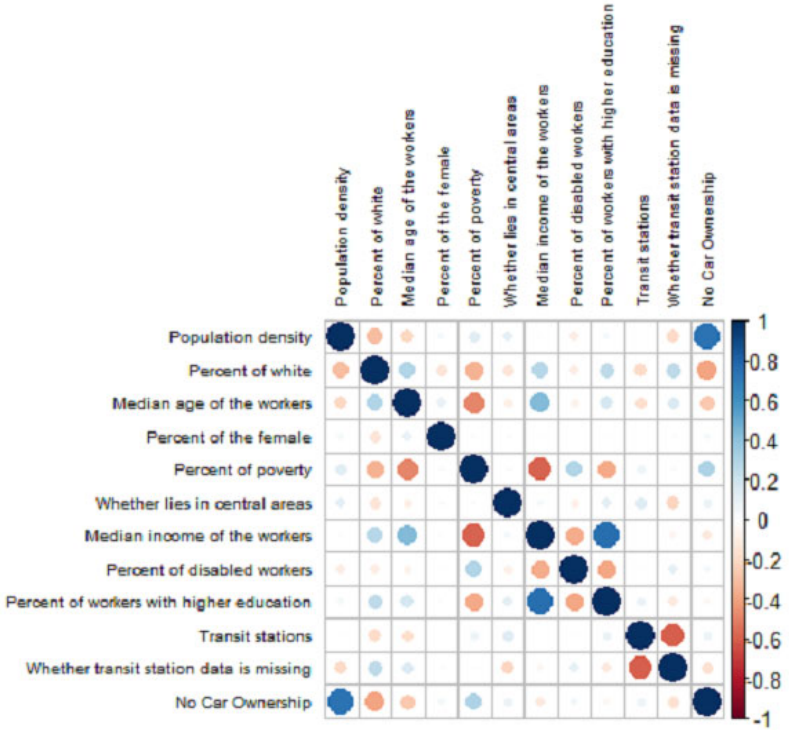

Figure 10 Correlation Table of Independent Variables

From the correlation table of the independent variables, we can see that for most of our variables, there is no obvious collinearity. There is a collinear relationship between the population density and the percent of non-vehicle households, as is the

case between education attainment and income level. Therefore, we should be careful when discussing the coefficients of these variables in the regression model. Generally speaking, our model is concise and accurate without much collinearity between independent variables.

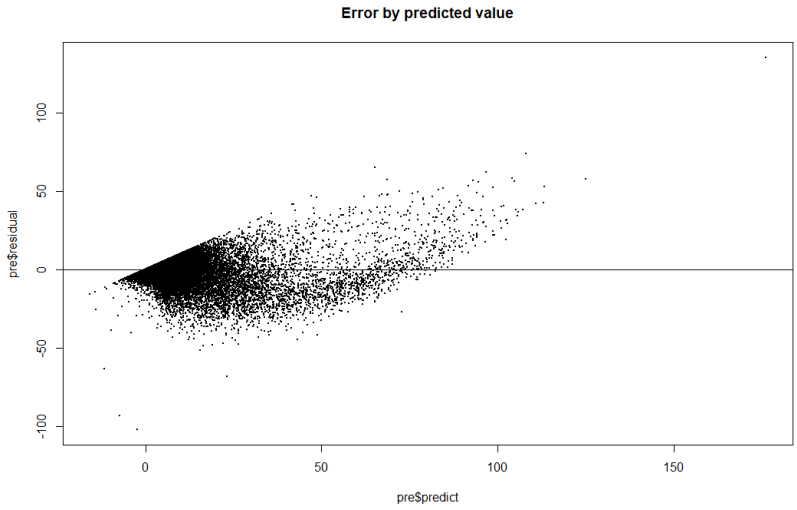

Figure 11 Error term

For the error term, we found that the errors were not random. The residuals were larger when predicted value was higher. This means that the linear model fits less satisfying with high-public-transit-ridership census tracts. Although we tried using log transformation and excluding outliers, this problem has not been fixed and the model above was the best we got. In further exploration, we may try other types of models to get a better fit.

Discussion

Because our methodology involved several different steps, we are going to try to bring all these steps together here and discuss what our findings could mean for the future of commuting by public transportation. We focused only on metropolitan areas in our analysis. As we saw above in our maps, the places with a strong spatial relationship between high percentages of commuting by public transportation and high population density are also the places where transit infrastructure exists. We knew that if we included the entire United States, where many people don't have access to public transportation, our results would be skewed. However, we do think this could be an area for future research if we wanted to continue to pursue the influencers of public transportation commuting.

Our country level analysis showed that people who are commuting by public transportation the most are women, Black and Asians, and young people. We think the only finding about this that was surprising to our group was that women were commuting by public transportation more than men. We assumed that women would be less likely to ride public transportation because they wouldn't feel as safe or comfortable. Our results showed that this is in fact not the case, as women were more likely to ride public transportation than men. Our findings about race were maybe slightly surprising as we assumed that Black people would be the highest users of public transportation when in fact the Asian demographic group has ridership that is almost as high. Getting this general understanding of who is riding public transit was essential for our understanding of the demographic groups involved with public transit.

When we finally ran our regression, we determined several things. Higher population density is one main factor associated with high percent of commuting to work by public transportation. Additionally, having more stations is likely to relate with greater ridership, based on accessibility. Race, poverty, central county location and disability all had negative relationships with ridership. We tried putting Blacks and Asians in our regression model, but they were less significant than the Whites and caused colinearity. Poverty could have a negative relationship because maybe those people can't afford to use public transportation, and instead rely on other modes for their commute like walking. Disability could have a negative relationship with commuting by public transportation because it might be a mode that those with disabilities cannot physically access, as not all transit is ADA accessible.

The positive relationships in our regression model were with population density, females, higher education, and not owning a car. We speculate that the strong correlation with population density means that with more people the demand would be higher, so that explains this relationship. Females is harder to explain, as are many of our demographic factors, because a lot of ridership has to do with choice, and if people have the choice to use other modes, it is likely that they will for any number of reasons. This is certainly a limitation of our entire project, because the data cannot account for human choice. Having higher education could mean that people are making more money, living closer in city cores, and choosing to commute by public transportation. We also saw this correlation at the country level with income. We used to think that people with lower incomes would ride public transportation more but found that wasn't the case. We thought the cause of this could be the same phenomenon of wealthier people living in city cores in recent years. Not owning a car is very clearly related with riding public transportation to work, because if you do not own a vehicle, you are certain to rely on other modes to get around. That other mode does not have to be public transportation, but there is a strong likelihood that it could be.

Overall, we found that any number of factors can be related to people commuting by public transportation. Each person might have a different reason to choose to ride public transit. Our demographic data just so happens to show the combination of factors of people who have been commuting by this mode, however it doesn't mean they represent everyone. Also, since the spatial variables are more significant in ridership than the demographic ones, it can be argued that demographics are not as important

when it comes to commuting by public transportation.

Moving forward this could have some policy implications for planning. If transportation agencies want to increase their ridership levels, specifically for commuting to work, then they should target the densest areas of the cities for routing. Additionally, building more stations could help improve accessibility and increase ridership. Spatial factors will be the largest factors in ridership, so these agencies do not necessarily have to target specific demographic groups for new riders. However, from a more equitable prospective, they should provide more facilities to help the disable workers and the poor to get easy access to public transit. Besides, we do think that they could try to market themselves to the demographic groups with lower levels of current usage. This could result in an increase in commute ridership.

Conclusion

We found that spatial factors such as population density and number of stations are strongly associated with higher percentage of commuters by public transit. Some demographic factors, such as race and income also have a strong relationship with percentage of workers taking public transit.

Our original hypothesis was that fewer females would ride transit than men, higher median income would result in lower ridership, higher population density would have higher ridership, and that more stations would lead to greater ridership. And while some of these hypotheses were true, for many of them the data told a different story. I think we did have some assumptions about who was riding transit, that could have played into our original hypothesis. Some of these assumptions were causal in nature, which is problematic. We assumed that a higher median income would result in lower ridership due to higher car ownership or living in the suburbs. We had also assumed that less females would use public transportation than males due to it being inconvenient and unsafe. However, I think because we expanded our independent variable to include a broader spectrum of considerations, we were able to see how many factors could play into choosing to commute by public transportation.

In the end we have found that the spatial factors are the biggest contributors to riding public transportation. However, some demographic factors are still important too. More women are riding public transportation than men, which shows a surprising gender split. Additionally, we found that census tracts with higher median incomes were associated with higher levels of ridership. Analyzing both spatial and demographic features was essential in drawing these conclusions. For next steps we could examine how changing spatial factors might affect ridership levels, or how transit companies could attract a more demographically diverse group of riders to commute through this mode.