Michael Frajman (mhf8621) and Lu Zhang (lz3148)
Multiple Regression and Econometrics - Spring 2024
May 9, 2024

Examining the Relationship Between Average Commute Time and Demographic Characteristics in New York City

Abstract

This paper examines the determinants of commute times across New York City's census tracts from 2010 to 2019. The variables used encompass modal choice, racial, age and disability demographics, employment, and income. By performing regression analysis on spatial and demographic data, this research seeks to explore how modal choice influences commute time and how demographic data from these census tracts may in turn influence modal choice. While most coefficients were found to be significant at high levels of confidence, the differences between commute time as predicted by modal choice were not large. Further, the effects of the examined demographics variables were significant but not large. Future analyses should take distances to employment and other magnet sites or origin-destination pairs of commuters into consideration as these were omitted from the available data and stand to be the strongest predictors of commute time.

Introduction

New York City (NYC) is renowned for its dense population, mixed use neighborhoods, and diverse transportation options. These factors make it an ideal location for studies on urban commute patterns and their influencing factors. The decade from 2010 to 2019 was marked by substantial developments in NYC's urban infrastructure, including the opening of phase 1 of the Second Avenue Subway, massive expansion of the City’s bike infrastructure, and a refocusing of urban design toward the mixed use. These changes could stand to impact commuter behaviors significantly.

Commute time is a critical aspect of urban living, influencing everything from personal health and productivity to environmental sustainability. Longer commutes not only increase stress and decrease life satisfaction but also contribute to more traffic congestion and vehicle emissions, which will lead to negative environmental externalities. As urban areas continue to grow, understanding the factors that influence commute times becomes essential for city planners and policymakers aiming to improve urban livability and equity. Recent studies have highlighted the impact of transportation infrastructure on urban efficiency, emphasizing the role of public transit systems in shaping commuting behaviors (Zhai et al., 2018). Furthermore, the adoption of bike-sharing systems and their integration into the urban transportation matrix have also been shown to significantly affect commuting patterns (Blanford et al., 2020). While these studies provide insights into specific aspects of urban mobility, there is a gap in comprehensive research examining the interplay of multiple factors over extended periods, particularly at the granularity of census tracts.

This study aims to fill this gap by analyzing a decade of data on NYC's commute times, focusing on the influence of modal choice, racial, age and disability demographics, and other employment and income-related factors. Using panel data regression techniques, we explore how these factors collectively influenced commute times in NYC. By providing a detailed examination of the factors influencing NYC's commuting patterns, this study contributes to the broader discourse on urban development and transportation planning. The insights gained are intended to inform future infrastructural and policy interventions aimed at reducing commute times and enhancing the quality of urban life, aligning with global efforts towards more sustainable and livable cities.

Description of Data

The primary data for this study came from an aggregated dataset exploring crime and demographics across NYC neighborhoods at a census tract level from 2010 through 2019. The data was aggregated from the American Community Survey (ACS), NYPD reports, and NYC Open Data. This dataset covers 2,165 census tracts within New York City, as organized by the Department of City Planning. While the dataset was oriented towards crime, it provided several data points related to commuting and transit. The dependent variable selected for this study was average commute time, while the independent variables include factors related to transportation modes, demographic characteristics, and socio-economic conditions. See Table 1 in the appendix for a complete list of variables and their explanations.

Table 2 presents the descriptive statistics for all variables, other than dummy variables created for year and borough. This table forms the empirical foundation for analyzing how different factors influence average commute times across the City's census tracts.

Racial composition variations show that on average, 13.58% of the population is Asian, 22.40% is Black, and 26.40% is Hispanic, with the percentages varying widely across tracts. These figures could suggest significant community-specific transit needs and preferences due to diverse demographic makeups. In terms of commuting patterns, the average commute time for NYC residents over this time was approximately 40.59 minutes, with a standard deviation of about 7.2 minutes. Average commute times by tract over time range widely, from a minimum of 11 minutes to a maximum of 66.2 minutes, highlighting significant disparities in transportation access and efficiency across different neighborhoods. Public transit over this period was on average the largest modal share commuting, with 56.8% of the population relying on it, underscoring the city's heavy dependence on its public transportation system. Biking to work is

less common, accounting for only 0.96% on average. Walking and driving are more prevalent, with on average 9.75% of residents walking to work and on average 30.9% commuting by car by tract.

An average of 7.89% of the population reports having a disability per tract. An average of 13.2% of the population is elderly, 65 years and older, per tract over time. The median household income was $61,776.69, with substantial variation, ranging from $9,001 to $250,001, indicating diverse economic conditions across different tracts. About 4.18% of households received public assistance, reflecting the socio-economic challenges some residents face. Additionally, 24.19% of renters experience a rent burden greater than 30% of their income, impacting commuting choices and highlighting the need for affordable transportation options. Analyzing these diverse and comprehensive points allows us to gain insights into how different population segments experience and navigate the city's transportation infrastructure.

Before including the data in a model it was cleaned. Tracts with no population were removed. Tracts with missing commute time entries were excluded as well. Missing entries for independent variables were filled in using the average of that variable's other values for that year over all tracts. For the pop_disabled_pct variable no data was collected for the years 2010 and 2011 so these values were instead averaged by the tract over all years. The rent burden variable prentburden was created by calculating the percent of all people with a rent burden of 30% or more of their income by applying the law of total probability to several separate variables representing different levels of rent burden.

Model Design

Several regression models were run using STATA in order to come up with equations for predicting commute time. Five models were created in total. A general equation listing all variables that were used in these models is listed below:

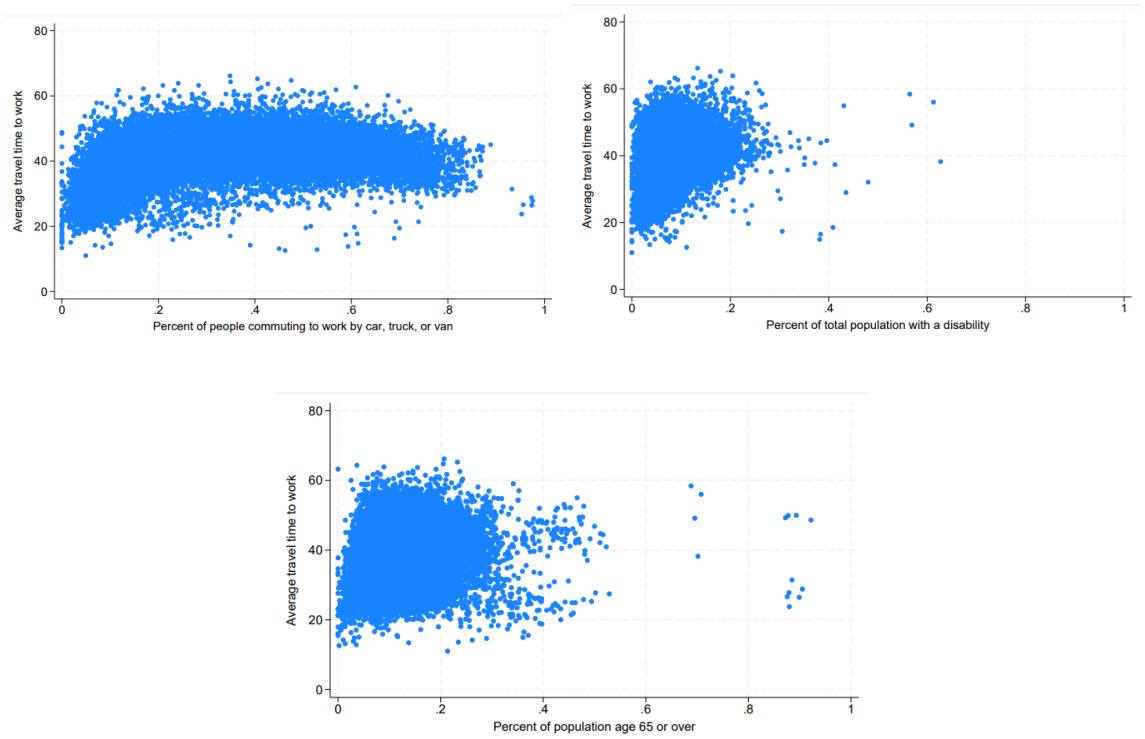
$$
\begin{aligned}
\text{pop\_commute\_time\_avg}_i = \\
&\beta_0 + \beta_1 \text{pop\_commute\_pubtrans\_percent}_i + \beta_2 \text{pop\_commute\_walk\_percent}_i + \\
&\beta_3 \text{pop\_commute\_car\_percent}_i + \beta_4 \text{pop\_commute\_car\_percent}_i^2 + \\
&\beta_5 \ln(\text{pop\_commute\_car\_percent}_i) + \beta_6 \text{pop\_race\_asian\_pct}_i + \\
&\beta_7 (\text{pop\_commute\_pubtrans\_percent}_i)(\text{pop\_race\_asian\_pct}_i) + \beta_8 \text{pop\_race\_black\_pct}_i + \\
&\beta_9 (\text{pop\_commute\_pubtrans\_percent}_i)(\text{pop\_race\_black\_pct}_i) + \beta_{10} \text{pop\_race\_hisp\_pct}_i + \\
&\beta_{11} (\text{pop\_commute\_pubtrans\_percent}_i)(\text{pop\_race\_hisp\_pct}_i) + \beta_{12} \text{pop\_disabled\_pct}_i + \\
&\beta_{13} \text{pop\_disabled\_pct}_i^2 + \beta_{14} \ln(\text{pop\_disabled\_pct}_i) + \\
&\beta_{15} (\text{pop\_commute\_pubtrans\_percent}_i)(\text{pop\_disabled\_pct}_i) + \beta_{16} \text{pop16\_64\_emp\_pct}_i + \\
&\beta_{17} \text{pop\_65p\_pct}_i + \beta_0 \text{pop\_65p\_pct}_i^2 + \beta_{18} \ln(\text{pop\_65p\_pct}_i) + \\
&\beta_{19} (\text{pop\_commute\_pubtrans\_percent}_i)(\text{pop\_65p\_pct}_i) + \beta_{20} \text{prentburden}_i + \beta_{21} \text{hh\_inc\_med}_i + \\
&\beta_{22} (\text{pop\_commute\_pubtrans\_percent}_i)(\text{hh\_inc\_med}_i) + \beta_{23} \text{hh\_inc\_pubasst\_pct}_i + \\
&\beta_{24} (\text{pop\_commute\_pubtrans\_percent}_i)(\text{hh\_inc\_pubasst\_pct}_i) + \beta_{25} \text{year10}_i + \beta_{26} \text{year11}_i + \beta_{27} \text{year12}_i \\
&+ \beta_{28} \text{year13}_i + \beta_{29} \text{year14}_i + \beta_{30} \text{year15}_i + \beta_{31} \text{year16}_i + \beta_{32} \text{year17}_i + \beta_{33} \text{year18}_i + \beta_{34} \text{boromn} + \\
&\beta_{35} \text{borobk} + \beta_{36} \text{boroqn} + \beta_{37} \text{borobx} + a_i + e
\end{aligned}
$$

Model 1 was a linear regression only using three percent modal choice variables to establish a baseline. Dummy variables for year were included and carried forward into all other models. Some modal choice variables were excluded in order to avoid multicollinearity as they add up to one hundred percent.

Model 2 was another linear regression which introduced all social variables of interest namely percent race by tract (with exclusions to avoid multicollinearity), percent of population with disability, percent elderly, percent employed of working age, percent rent burdened, median household income and percent of households on public assistance. These variables were chosen as these demographics may be more inclined to be regular commuters and to make use of modes like public transit, as well as to examine relationships between wealth inequality, transit, and

commute time. Dummy variables for NYC's boroughs were also created as a way to begin disentangling the social and modal variables from time and space.

Models 3 and 4 made use of interactions and non-linear variables. Racial variables, disability, old age, median income and public assistance were all multiplied against public transit under the assumption that these demographics may be more inclined to make more use of public transit due to difficulty using other modes or income constraints. The variables pop_commute_car_pct, pop_disability_pct, and pop_65p_pct were also found to have non-linear shapes to their scatterplots shown in figures 1,2, and 3 below. Model 3 added a quadratic form to these variables while model 4 used a logarithmic form. In model 4, any value that was equal to 0 had its natural log value defaulted to zero to maintain the integrity of the observation. For both models the year and borough dummies were again included.



Figures 1,2, and 3: pop_commute_car_pct, pop_disability_pct, and pop_65p_pct plotted against average commute time respectively. These variables exhibited a curved/"ramping-up" shape compared to all other variables.

Model 5 built on the quadratic model, model 3, and included spatial fixed effects by tract. The choice of the quadratic model as a base is further explained in the results section below. The dummy variables for boroughs were removed while year dummies continued to be included.

Results

Major deviations in commute time were not immediately obvious from these models primarily relying on modal choice as a dependent variable. All results are listed in Table 3 in the appendix. Models 1 and 2 demonstrated very similar effects induced by both the percent of people commuting by public transit or car. In model 1, a one percent increase in public transit users by tract and a simultaneous one percent decrease in all other modes led to a faster commute of only 0.62 minutes on average with all other variables held equal. In model 2 with demographic variables introduced, that effect decreased to 0.27 minutes on average ceteris paribus.

Introducing interactions widened the gap between the public transit and car coefficients while generally maintaining significance among all variables. It is worth noting that the contributions of all social variables were incredibly small even if significant. Most interactions had fractional effects lost with only two decimal places of precision. This was despite all interactions except that between public transit and public assistance having significance to a high level of confidence again as seen in join F tests at the bottom of table 3. Public assistance shows the overall signs of being an unnecessary variable. Introducing interactions and non-linear forms did increase R-squared and therefore better represented the variability of the independent variable. Using the model with quadratics produced an R-squared of 0.68, higher than that for logarithms at 0.67 and the linear model 2. For this reason it was decided to carry the quadratic

model onward into the final fixed effect model. Accounting for spatial effects in model 5 rendered most social variables insignificant or with small effect but had an incredibly high R-squared.

Of additional note, both models 3 and 4 demonstrated heteroskedasticity via a White test. All models were rerun using robust standard errors which are those reported in Table 3. Two observations of note were that household income never held as significant in any model and that the coefficient for the percent of employed working age people was constantly negative. To the first point, income can likely be declared a poor predictor of travel time even if theory suggests income may determine the neighborhoods and locations available for someone to live. Most other demographic variables had small positive effects on commute time which more aligned with theories that people with disabilities, older people, or racialized communities may have longer commutes due to having less commuting options. To the latter point, it is likely that workers are optimizing their living location to minimize their commute and that this phenomenon could be getting disentangled once spatial effects are accounted for. Overall most demographic variables had little effect in predicting commute times. Once the social variables were introduced, the coefficients of modal choice did not vary greatly.

Conclusion and Future Considerations

As mentioned, the social variables introduced were unable to add much weight toward predicting commute time. Modal choice was always the strongest predictor for commute time. Once spatial effects were considered, the social variables lost much significance. The small difference between public transit and car coefficients may be due again to people optimizing their location. This may be further evidenced by walking having a negative coefficient as people are living very close to their job or preferred amenitiesGiven the nature of this dataset was for

predicting crime, it stands to reason that if more spatially related variables were introduced that a more robust regression model could be created. Introducing origin-destination data for employment or distances to major job centers would be one step. Finding a way to measure transit accessibility and choice would be another major variable as some New York City neighborhoods have abundant transit, like Downtown Brooklyn, asn others are transit deserts like outer-borough Queens. These types of variables would complement the notion that people commuting to jobs optimize their location. All of this shows a clear direction forward even if the regression models explored were limited by the available data.

Reference:

Blanford, J. I., & MGIS Geog 586 Students. (2020). Pedal power: explorers and commuters of New York citi bikesharing scheme. *PLoS one*, *15*(6), e0232957.

Zhai, W., Bai, X., Peng, Z. R., & Gu, C. (2019). A bottom-up transportation network efficiency measuring approach: A case study of taxi efficiency in New York City. *Journal of Transport Geography*, *80*, 102502.

Appendix - Data Tables

Table 1

Variable Descriptions for Commute Time Regression Models

name	varlab
pop_18_64_pct	Percent of population between age 18 and 64
pop_65p_pct	Percent of population age 65 or over
pop_race_asian_pct	Percent of total population that is Asian
pop_race_black_pct	Percent of total population that is Black
pop_race_hisp_pct	Percent of total population that is Hispanic/Latino
pop_disabled_pct	Percent of total population with a disability
pop16_64_emp_pct	Percent of population age 18 to 64 in the labor force that is employed
pop16_64_nonlaborforce_pct	Percent of population age 18 to 64 not in the labor force
pop_commute_time_avg	Average travel time to work
pop_commute_pubtrans_pct	Percent of people commuting to work by public transit
pop_commute_bike_pct	Percent of people commuting to work by bike
pop_commute_walk_pct	Percent of people commuting to work by walking
pop_commute_car_pct	Percent of people commuting to work by car, truck, or van
hh_inc_med	Median household income
hh_inc_pubasst_pct	Percent of households receiving income from public assistance
prentburden	percent of people paying 30%+ of income toward rent
year10	=1 for 2010
year11	=1 for 2011
year12	=1 for 2012
year13	=1 for 2013

year14	=1 for 2014
year15	=1 for 2015
year16	=1 for 2016
year17	=1 for 2017
year18	=1 for 2018
year19	=1 for 2019
boromn	=1 for tract in Manhattan
borobx	=1 for tract in Bronx
borobk	=1 for tract in Brooklyn
boroqn	=1 for tract in Queens
borosi	=1 for tract in Staten Island
pubtrans_pasian	interaction of percent transit and percent asian
pubtrans_pblack	interaction of percent transit and percent black
pubtrans_phispanic	interaction of percent transit and percent hispanic
pubtrans_pdisabled	interaction of percent transit and percent disabled
pubtrans_pelderly	interaction of percent transit and percent elderly
pubtrans_prent	interaction of percent transit and percent rent burdened
pubtrans_income	interaction of percent transit and median household income
pubtrans_pubasst	interaction of percent transit and percent on public assistance
pcar2	percent car squared
p65p2	percent people 65 years+ squared
pdis2	percent people disabled squared
lnpcar	logarithm of percent car
lnp65p	logarithm of percent 65 years+
lnpdis	logarithm of percent disabled

Table 2

Descriptive Statistics of Commuting Patterns and Socio-economic Indicators in New York City (2010-2019)

	(1)	(2)	(3)	(4)
VARIABLES	mean	sd	min	max
pop_65p_pct	13.10	6.43	0.00	92.21
pop_race_asian_pct	13.58	16.23	0.00	91.61
pop_race_black_pct	22.40	28.58	0.00	100.00
pop_race_hisp_pct	26.40	22.54	0.00	100.00
pop_disabled_pct	0.08	0.05	0.00	0.63
pop16_64_emp_pct	91.07	5.18	39.57	100.00
pop_commute_time_avg	40.59	7.17	11.02	66.16
pop_commute_pubtrans_pct	56.84	16.87	0.00	94.04
pop_commute_bike_pct	0.97	1.73	0.00	23.01
pop_commute_walk_pct	9.75	10.20	0.00	85.19
pop_commute_car_pct	30.90	19.66	0.00	97.56
hh_inc_med	61,776.69	30,519.87	9,001.00	250,001.00
hh_inc_pubasst_pct	4.18	4.03	0.00	31.10
prentburden	24.19	9.13	0.00	100.00

Table 3

Regression Models of Commuting Patterns and Socio-economic Indicators in New York City (2010-2019)

VARIABLES	(1) Commuting Variables	(2) Demographic Variables	(3) Interactions and Quadratics	(4) Interactions and Logarithms	(5) Fixed Effects AREG
pop_commute_pubtrans_pct	0.62***	0.27***	0.28***	0.28***	0.22***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
pop_commute_walk_pct	0.19***	-0.02	-0.05***	-0.03**	-0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
pop_commute_car_pct	0.60***	0.26***	0.49***	0.13***	0.17***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
pcar2			-0.00***		-0.00***
			(0.00)		(0.00)
pop_race_asian_pct		0.08***	0.07***	0.07***	0.03*
		(0.00)	(0.01)	(0.01)	(0.01)
pubtrans_pasian			0.00	0.00	0.00
			(0.00)	(0.00)	(0.00)
pop_race_black_pct		0.08***	0.12***	0.14***	-0.01
		(0.00)	(0.01)	(0.01)	(0.01)
pubtrans_pblack			-0.00***	-0.00***	0.00***
			(0.00)	(0.00)	(0.00)
pop_race_hisp_pct		0.05***	-0.04***	-0.01	0.02
		(0.00)	(0.01)	(0.01)	(0.01)
pubtrans_phispanic			0.00***	0.00***	0.00
			(0.00)	(0.00)	(0.00)
pop_disabled_pct		8.37***	19.89***	-2.48	7.19*
		(1.15)	(4.01)	(4.41)	(4.23)
pdis2			-48.91***		1.35
			(7.82)		(10.27)
pubtrans_pdisabled			0.04	0.10	-0.07
			(0.06)	(0.06)	(0.06)
pop16_64_emp_pct		-0.07***	-0.06***	-0.06***	-0.03***
		(0.01)	(0.01)	(0.01)	(0.01)
pop_65p_pct		0.11***	-0.09***	-0.08***	-0.04
		(0.01)	(0.03)	(0.01)	(0.04)
p65p2			0.00***		0.00*
			(0.00)		(0.00)
pubtrans_pelderly			0.00***	0.00***	0.00
			(0.00)	(0.00)	(0.00)
prentburden		-0.00	-0.02**	-0.02**	-0.02*
		(0.00)	(0.01)	(0.01)	(0.01)
pubtrans_prent			0.00**	0.00**	0.00**
			(0.00)	(0.00)	(0.00)
hh_inc_med		-0.00***	0.00***	0.00***	0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
pubtrans_income			-0.00***	-0.00***	-0.00***
			(0.00)	(0.00)	(0.00)
hh_inc_pubasst_pct		-0.00	-0.02	0.04	0.12***
		(0.01)	(0.05)	(0.05)	(0.04)
pubtrans_pubasst			0.00	-0.00	-0.00***
			(0.00)	(0.00)	(0.00)
year10	-2.78***	-2.55***	-2.59***	-2.64***	-2.27***
	(0.17)	(0.15)	(0.14)	(0.15)	(0.12)
year11	-2.68***	-2.52***	-2.52***	-2.60***	-2.27***
	(0.17)	(0.15)	(0.14)	(0.14)	(0.11)
year12	-2.67***	-2.58***	-2.55***	-2.63***	-2.35***
	(0.16)	(0.15)	(0.14)	(0.14)	(0.11)
year13	-2.51***	-2.50***	-2.46***	-2.53***	-2.32***
	(0.16)	(0.15)	(0.14)	(0.14)	(0.10)
year14	-2.40***	-2.39***	-2.32***	-2.39***	-2.21***
	(0.16)	(0.15)	(0.14)	(0.14)	(0.10)
year15	-1.93***	-1.90***	-1.79***	-1.85***	-1.77***
	(0.16)	(0.14)	(0.13)	(0.14)	(0.09)
year16	-1.51***	-1.48***	-1.44***	-1.47***	-1.40***
	(0.15)	(0.14)	(0.13)	(0.13)	(0.09)
year17	-1.02***	-1.00***	-0.98***	-1.00***	-0.92***
	(0.15)	(0.14)	(0.13)	(0.13)	(0.09)
year18	-0.51***	-0.51***	-0.51***	-0.53***	-0.44***
	(0.15)	(0.14)	(0.13)	(0.13)	(0.09)
boromn		-5.91***	-5.97***	-5.50***	
		(0.24)	(0.23)	(0.24)	
borobk		-0.80***	-1.96***	-1.35***	
		(0.20)	(0.20)	(0.20)	

borobx		-1.77***	-3.02***	-2.25***	
		(0.21)	(0.21)	(0.21)	
boroqn		-1.33***	-2.61***	-1.99***	
		(0.19)	(0.19)	(0.19)	
lnpcar				3.23***	
				(0.17)	
lnpdis				0.58***	
				(0.14)	
lnp65p				0.20	
				(0.21)	
Constant	-13.35***	21.56***	18.54***	16.58***	28.72***
	(1.29)	(1.57)	(1.74)	(1.92)	(1.71)
Observations	18,867	18,867	18,867	18,867	18,867
R-squared	0.52	0.64	0.68	0.67	0.89
Tract FE	NO	NO	NO	NO	YES
Pasian=pubtrans_pasian			0	0	
Pblack=pubtrans_pblack			0	0	
Phispanic=pubtrans_phispanic			0	0	
Prentburden=pubtrans_Prentburden			0.0514	0.0612	
Household income=pubtrans_hhincome			0	0	
Ppublicasst=pubtrans_Ppublicasst			0.921	0.726	
Pdisabled and lndis=pubtrans_pdisabled				0.0463	
Pelderly and lnelderly=pubtrans_pelderly				0	
Pdisabled and dis2=pubtrans_pdisabled			5.68e-05		
Pelderly and elderly2=pubtrans_pelderly			0		

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1