UNDERGRADUATE RESEARCH

Public Transportation Ridership Levels

Christopher R. Swimmer and Christopher C. Klein[1]

Abstract

This article uses linear regression analysis to examine the determinants of public transportation ridership in over 100 U. S. cities in 2007. The primary determinant of ridership appears to be availability of public transportation service. In fact, the relationship is nearly one to one: a 1% increase in availability is associated with a 1% increase in ridership. The relative unimportance of price may be an indicator of the heavy subsidization of fares in most cities, leaving availability as the more effective policy tool to encourage use of public transport.

Key Words: identification, public transportation, ridership.

JEL Classifications: A22, C81, H42

Introduction

What makes one city more apt to use public transportation relative to another? This is an important issue that has been studied by others in various ways. Glaeser et al. (2008), find that the availability of public transportation is a major explanatory factor in urban poverty. Glaeser and Shapiro find evidence that car cities, where a large percentage of people drive themselves to work, grew at the expense of public transportation cities as the percentage of cities' population taking public transportation declined between 1980 and 2000. Murray et al. (1998), conclude that the performance of a public transport system is determined largely by the proximity of public transport stops to the regional population.

Initially, the data were gathered for the top 136 metropolitan statistical areas in the U.S. using the raw number of unlinked trips on public transportation as the measure of ridership. Due to the wide variation in population and ridership across cities, the per-capita unlinked trips were calculated for use as the dependent variable. Missing values reduced the number of observations to 105.

The regression analysis utilizes the process of backwards selection by eliminating a single variable per regression based on the highest P-value that is attained. This process will continue until each coefficient's p-value is less than .10. First, however, Park's test for heteroscedasticity is performed. After the backwards selection, an F-test is performed to confirm the appropriateness of the resulting equation.

---

[1] Mr. Swimmer, as an undergraduate Economics major at Middle Tennessee State University, prepared the initial draft of this paper for an Econometrics class taught by Dr.Klein in the Economics and Finance Department at Middle Tennessee State University during the Spring Semester 2009. Christopher C. Klein is Associate Professor, Economics and Finance Department, Middle Tennessee State University, Murfreesboro, TN. cklein@mtsu.edu.

Data and Variables

Eleven independent explanatory variables were chosen as described below.

- *Metropolitan Density*: Density in this analysis will be defined as the population divided by the metropolitan area. The reasoning for selecting this is that public transportation is more efficient in areas of higher density. The coefficient should be positive in value.
- *Metropolitan Area*: The Metropolitan Statistical Area is generally characterized as the official area of an incorporated city along with its immediate sphere of economic influence. There are two opposing theories for this variable. The first is that a larger area makes public transportation less efficient for those riding. The second is that a larger area simply implies greater economic activity.
- *Average distance*: Average distance is the typical number of miles traveled using public transportation. The theory is that a person who expects a longer trip will be less likely to endure the apparent inconvenience of using public transit. The coefficient should be negative.
- *Service availability*: Service availability is the maximum availability of transit cars in each metropolitan area. As in Glaeser (2008), this variable measures the convenience of public transportation. Its coefficient is expected to be positive.
- *Gas price index*: The gas price index is indexed relative to the national average to account for the volatility of gas prices. The gas price index will serve as the price of an input to a substitute mode of transportation, the private automobile. The coefficient is expected to be positive.
- *Commute time:* Commute time is defined as the average time in minutes one spends traveling from home to work in a particular metropolitan area. The theory is that a higher commute time implies a higher density of traffic, meaning people will be inclined to drive less. This would give the coefficient a positive sign.
- *Poverty rate*: The poverty rate is the percentage of inhabitants living below the poverty line calculated by the US Census Bureau. If automobiles are considered luxury goods, then impoverished people will use them less and use public transport more.
- *Median Income*: The theory is that public transport is an inferior good, such that people consume less as their incomes increase. The coefficient should be negative.
- *Firms per capita*: Firms per capita is the number of all registered firms within the official city area divided by the population. The theory is that if a city has a higher density of firms within its city limits, more people will be drawn into the city for work and leisure. Therefore, it is expected that if the firms per capita increases, so will the ridership level.
- *Educational attainment*: Educational attainment is the percentage of residents holding a bachelor's degree or higher. The theory here is that people with higher educational attainment have greater choice, and those with choice will opt out of using public transportation. In this case the coefficient would be negative.
- *Rail service*: This will be used through a dummy variable with "1" indicating the availability of rail service in the city. The coefficient of this is expected to be positive.

If data were absent then the entirety of the corresponding observation was removed. This resulted in the deletion of 31 rows, leaving a total of 105.

First Results

Summary statistics and the results of the first regression are reported below:

Variable	Mean	Variation	Variance	Std. Dev.	Upper	Lower
Ridership Rate	25.3441	90673.0079	871.8558	29.52721	199.83	2.74
Distance per Trip	4.65919	205.206403	1.973138	1.40468	12.60	1.84
Metropolitan Area	4088.316	2853755053	27439952	5238.31580	39370.3	395.8
Metro. Density	365.6871	13235200.1	127261.5	356.73735	2361.37	33.38
Median Income	35888.49	5252235907	50502268	7106.49480	$66,384	$23,483
Commute Time	22.43973	1938.40236	18.63848	4.31723	43.9	16.6
Poverty Rate	0.178567	0.29388052	0.002826	0.05316	30.6%	7.3%
Bachelor's Rate	0.253778	0.68076697	0.006546	0.08091	48.2%	8.6%
Firms per Person	0.037509	0.04933903	0.000474	0.02178	0.10749	0.00784
Gas Price Index	1.001302	0.59431081	0.005715	0.07559	1.239561587	0.893528
Service Availability	0.000501	1.1647E-05	1.12E-07	0.00033	0.00177	0.00009

Regression Statistics	
R Square	0.746697157
Adjusted R Square	0.719750046
Standard Error	15.63130067
Observations	105

	F	*Significance F*
Regression	27.70972954	7.478E-24

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-82.9183793	29.0851337	-2.8508853	0.005358961
Metro Pop Density	0.021750321	0.005743529	3.78692607	0.000268707
Metro Area	0.001076186	0.000351436	3.06225187	0.002864926
Distance Per Trip	-3.4030236	1.217265504	-2.7956297	0.006280417
Transit Cars/Person	46496.39671	6195.786941	7.50451834	3.44881E-11
Gas Index	***27.6754902***	***25.78276326***	***1.0734106***	**0.28583482**
Mean Work Commute Time	1.233180873	0.676692416	1.82236544	0.071579069
Poverty Rate	***57.4856548***	***66.07690017***	***0.8699811***	**0.386526906**
Median Income	***0.00031717***	***0.00061846***	***0.5128448***	**0.609263007**
Firms/Person	***-14.8368233***	***94.45073946***	***-0.1570853***	**0.875514533**
Bachelor's Degree Rate	47.29981838	26.79666438	1.76513829	0.080788171

This regression as a whole is extremely significant under the F-test and most of the coefficients are significantly different from zero under the t-tests. The adjusted R-Square for the model is reasonably high at almost 0.72.

The Park test (Gujarati, 2006; 402), however, indicated that the regression residuals are heteroscedastic.[2] Heteroscedasticity biases the standard errors and makes the t-tests, F-test, and R-squared unreliable.

[2] The natural logs of the squared residuals from the first regression are regressed against the predicted values of ridership. A significant slope coefficient (in this case, a p-value of less than 0.01) indicates heteroscedasticity.

To address the heteroscedasticity, the variables were converted to natural logarithms. The Park test on the residuals of the log-linear form of the first regression could not reject the null hypothesis of no heteroscedasticity (p-value of .993). Consequently, the backwards selection was performed on the log-linear equation producing the following result:

Regression Statistics	
R Square	0.830323594
Adjusted R Square	0.821754079
Standard Error	0.384014779
Observations	105

	F	*Significance F*
Regression	96.89271236	1.5215E-36

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6.434251744	1.024552438	6.280061	9.06E-09
LnDensity	0.214369411	0.07144169	3.000621	0.003409
LnArea	0.340488498	0.065235453	5.219378	9.91E-07
LnDistance	-0.650181955	0.154879017	-4.198	5.89E-05
LnAvailability	1.093673729	0.07402253	14.77488	8.78E-27
LnCommute	0.630514974	0.326274575	1.932467	0.05616

Somewhat surprisingly, the adjusted R-Square in this log-linear regression is higher at 0.82 than the 0.72 of the original linear regression.[3] The model obtained from this regression analysis is:

Ln(Ridership) = 6.434 + 0.214*Ln(Density) + 0.340*Ln(Area) - 0.650*Ln(Distance)

+ 1.094*Ln(Availability) + 0.631*Ln(Commute)

Further Analysis

A referee suggested that the price of public transportation had been left out and should be included. Data on fare revenue per unlinked trip were available for 102 of the cities. Including the natural logarithm of this price variable (LnP) among the explanatory variables produced the following result:

Regression Statistics	
R Square	0.865596696
Adj. R Square	0.849169625
Standard Error	0.358347148
Observations	102

[3] An F-test comparing the log-linear version of the first regression to that reported here resulted in an F statistic of 0.727, which is not significant at the 0.10 level.

	F	*Significance F*
Regression	52.69330857	2.18186E-34

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-3.26071421	4.808313864	-0.67814088	0.499422224
LnP	0.139746246	0.082260825	1.698818916	0.092807818
LnAvail	0.964385568	0.07966288	12.10583357	1.37723E-20
LnDist	-0.60584290	0.149233404	-4.05970036	0.000104612
LnDensty	**-0.07895075**	**0.066377366**	**-1.18942281**	**0.237401828**
LnInc	1.213959462	0.520481941	2.33237576	0.02191162
LnTtime	**-0.10834237**	**0.412516569**	**-0.26263763**	**0.793430151**
LnPov	0.896332121	0.268601936	3.337027776	0.001232307
LnBach	**0.019760397**	**0.155307994**	**0.127233614**	**0.899039345**
LnFperC	0.708435222	0.142184294	4.982513897	3.01763E-06
LnGasP	1.179287777	0.592875657	1.98909799	0.049727071
LnFirms	-0.63472259	0.117240165	-5.41386641	5.07501E-07

Although this increased the Adjusted R Square, the price variable is positive and significant. This is a classic indicator of an identification problem: the price variable has the wrong sign – it should be negative in a demand equation – suggesting that the regression fails to identify the demand equation for public transportation.

The remedy is to perform a two-stage least squares (2SLS) regression. To do so requires an instrument for the first stage price equation that does not appear in the second stage ridership equation. Since LnDensity is not significant above, it is deleted from the ridership equation and used as an instrument for the price equation. This makes sense as an instrument, because costs of providing trips should rise as density falls – vehicles must travel farther to pick up the same number of riders.

The 2SLS procedure performed in SAS produced an Adjusted R Square of 0.7904 and the coefficient estimates shown below:

Parameter	Estimate	Std Err	t Value	Pr > |t|
Int	-7.48035	5.8315	-1.28	0.2028
LnP	**-0.3605**	**0.4563**	**-0.79**	**0.4315**
LnAvail	1.00824	0.0989	10.20	<.0001
LnDist	-0.53142	0.2017	-2.63	0.0099
LnInc	1.479599	0.5278	2.80	0.0062
LnPov	0.805217	0.2820	2.86	0.0053
LnFperC	0.621974	0.1987	3.13	0.0023
LnFirms	-0.67034	0.1198	-5.59	<.0001
LnGasP	2.179802	1.0459	2.08	0.0399

The 2SLS results yield a correct negative sign for LnP, but it is not significant. Of the variables that were initially significant in our first analysis without the price variable, only LnAvailability and LnDistance remain, but display the expected signs. Interestingly, the income variables (LnInc, LnPov) are now significant with positive signs, suggesting that public transportation is a normal good, but that this effect may be more pronounced at low income levels. The log of the gas price index and the log of the number of firms per capita are also positive as expected.

Conclusion

The first results seemed fairly successful. The insignificant variables were poverty rate, bachelor's degree rate, median income, gas price index, rail availability, and firms per capita, while the independent variables that turned out to be significant were metropolitan density, metropolitan area, average distance traveled, average commute time, and service availability. The significant variables had the correct signs.

The addition of price as an explanatory variable, however, changed many of these results. Initial results here indicated an identification problem and a two-stage least squares procedure was performed in response. The 2SLS results confirmed ridership's positive relationship to availability and its negative relationship to distance traveled, but the other relations in the first results were overturned.

The income variables became positive and significant, as one would expect in a demand equation. Gasoline prices and firms per capita also became positive and significant, as was expected. The price variable's coefficient was negative, as one would expect in a demand equation, but it was not significant. This seeming unimportance of price may be due to the already heavy subsidization of fares with tax revenues in most cities.

Overall, if one is interested in encouraging ridership of public transportation, the results suggest that availability trumps price as a policy variable. In the log-linear form, a 1% change in the independent variable causes a 1% change times the coefficient in the dependent variable. The coefficient on the log of availability is approximately 1 in all of these results: an increase, say, of 10% in availability would be expected to increase ridership by 10%.

The research possibilities on this topic certainly have not yet been exhausted. Tourism data could be included in the regression analysis to reflect those who visit a city and have little choice but to use public transportation while on vacation or business. The number of automobiles per capita and transportation expenditures also may correlate with the public transportation ridership level.

References

American Public Transportation Association. 2009. "2008 Annual Report." Accessed 20 Mar. 2009 <http://www.apta.com/about/annual_report_08.cfm>.

American Public Transportation Association. 2008. *2008 Public Transportation Fact Book*. June.

Glaeser, Edward L., Matthew E. Kahn, and Jordan Rappaport. 2008. "Why Do the Poor Live in Cities? The Role of Public Transportation." *Journal of Urban Economics*, 63, 1-24.

Glaeser, Edward L. and Jesse M. Shapiro. 2003. "Urban Growth in the 1990s: Is City Living Back?" *Journal of Regional Science,* 43(1), 139-165.

Gujarati, Damodar N. 2006. *Essentials of Econometrics*. 3rd ed., Boston: McGraw-Hill/Irwin.

Murray, Alan T., Rex Davis, Robert J. Stimson, and Luis Ferreira. 1998. "Public Transportation Access." *Transportation Research,* 3(5), 319-328.

U.S. Census Bureau. 2006. Statistics about Metropolitan Areas. Accessed 15 Mar. 2009 <http://www.census.gov/epcd/www/metros.htm>.

U.S. Department of Transportation, Bureau of Transportation Statistics. 2002. National Household Travel Survey. Washington, DC.