내생성을 고려한 지하철 이용자수 추정에 관한 연구

Study on subway ridership estimation with the consideration of endogeneity

이정정*, 홍정열*, 박동주*

Tingting Li*, Jungyeol Hong*†, Dongjoo Park*

Abstract The objective of this study is to develop a consistent model for predicting subway ridership. Previous studies have focused on the relationship between subway ridership and factors such as socio-economy and land-use. However, there was a lack of research on the endogenous impacts of alternative transit demand and exogenous trade area. Therefore, this study incorporated the effects of bus ridership and retail properties within 300m radius from 248 subway stations in Seoul in the modeling process. In a methodological aspect, we applied two-stage least squares (2SLS) approach to capturing the endogeneity issue. As a result, bus ridership and attributes for trade area were found to be significant, and the methodological approach for the model development was plausible to estimate subway ridership.

***Keywords* :** Subway ridership, Trade area, Endogenity, Two-stage least squares

초 록 본 연구는 지하철 이용자 수요예측을 위하여 내·외생적 영향요인들을 분석하고, 각 요인들이 일반선형회귀모형 상에서 내재하고 있는 편의를 통제하여 보다 설명력이 높은 예측 방법론을 제시하는데 주요 목적이 있다. 지하철 이용수요와 관련하여 기존 연구들은 주로 사회경제 및 토지이용 등의 요인들과의 상관관계 위주의 연구를 수행해 왔으나, 버스 등의 타 대중교통 수요의 내생성 및 지하철역 주변상업용도의 외생적 영향에 대한 연구는 고려하지 않았다. 따라서 본 연구는 이러한 문제들을 개선하기 위하여 2단계 최소자승법 (2SLS)을 적용하였으며, 연구결과를 통하여 버스 이용자수와 상업용도 요인들의 내 · 외생성은 2SLS 추정에 의해 지하철 이용수요 예측을 보다 고도화 할 수 있는 것으로 나타났다.

주요어 : 지하철 이용자수, 상업시설, 2단계 최소자승법, 내생성

1. Introduction

Traffic jams and low efficiency of land use are the most serious problems in society. As we all know, the subway has lots of advantages, such as reducing traffic congestion, reducing expenditure, and eliminate the problem of parking difficulty, also reduce the parking land use form. The objective of this study is to develop a consistent model for predicting subway ridership. Previous studies had focused on the relationship between the subway ridership and factors of socio-economy and land-use. However, no study has considered endogenous problems in a modeling process. Therefore, this study incorporated the effects of a bus ridership as an endogenous variable applied from 248 subway stations in Seoul in the to discover the relationship applied two-stage least squares (2SLS) approach. Besides, we collected data of retail properties within 300m radius facility for developing the model of subway ridership predictions.

* 서울시립대학교 도시과학대학 교통공학과

2. Literature Review

Previous studies about a topic of transit ridership impact factors can be divided into two categories. One is travelers' attitudes; the other is a factor of land use, behavior characteristics. The analysis for the first category is based on a collection of data from the survey and interviews of transit operators to identify factors that consider affecting ridership. However, the information is highly subjective and depends on the respondents' views and assumptions about the internal and external factors. Therefore, the data is biased with limited or incorrect information. Secondly, most previous studies were used the multiple regression models for methodology and selected the main factors of socio-economic factors, land use, proportion density, and so on. In term of methodological weakness for the previous studies, generalizability is limited due to small sample sizes, multicollinearity and endogeneity problems among independent variables are ignored. (Taylor and Camille 2003)

3. Methodology

The model structure of this study is

$$y = Y\beta_1 + X_1\beta_2 + u = X\beta + u \qquad (1)$$

$$Y = X_1\Pi_1 + X_2\Pi_2 + v = Z\Pi + V$$

where y is an N×1 vector of the left-hand-side variable; N is the sample size; Y is an N×p matrix of p endogenous regressors; $X_1$is an $N \times k_1$ matrix of k1 included exogenous regressors; X2 is an $N \times k_2$ matrix of $k_2$excluded exogenous variables, X = [Y X_1], Z = [X_1 X_2]; u is an N×1 vector of errors; V is an N×p matrix of errors; β=[β_1 β_2] is a k = $(p + k_1) \times 1$ vector of parameters; and Π is a $(k_1 + k_2) \times p$ vector of parameters. If a constant term is included in the model, then one column of X1 contains all ones. Define the k-class estimator of β as $b = \{X'(I - kM_z X)X\}^{-1}X'(I - kM_z)y$

where $M_Z = I - Z(Z'Z)^{-1}Z'$. The 2SLS estimator results from setting k= 1. The LIML estimator results from selecting k to be the minimum eigenvalue of $(Y'M_Z Y)^{-\frac{1}{2}}Y'M_{X_1}Y(Y'M_Z Y)^{-1/2}$, where $M_{X_1} = I - X_1(X'_1X_1)^{-1}X'_1$.

4. Data Analysis

The spatial scope of this study is Seoul in South Korea, which has 25 districts, and each district has high population density. All 248 subway stations in Seoul were included in the analysis, representing different types of social forms. The temporal scope is from January 2015 to September 2016.

We use socio-economic, land use, transit facility, geometry and safety factors within 300m radius from 248 subway stations as the independent variables for predictive models, and environment factors, the number of bus stations and bus lines as instrument variables of bus ridership which is endogenous variable with subway ridership. The variables used in the models were obtained by the Statistics Department of Seoul. In particular, types of land use in this paper were categorized based on the proportion of ten retail facilities around each subway station throughout Seoul city are shown relatively special form compared with existing papers. Moreover, subway and bus ridership data, whether to transfer, the number of bus stations and bus lines in Seoul were obtained from the smart card and geographic information system (GIS) graphics and tables provided by the transportation bureaus of local governments.

Mass transit ridership was affected by a variety of internal and external factors. Subway ridership was the dependent variable and factors can directly or indirectly explain transit ridership. Table 1 describes dependent, explanatory and instrument variables used in the models.

Table 1 Descriptive statistics

Category	Variables
Demography	the number of families, employment density, number of parking, area of parking (m^2) vehicle registration, school-age population, vehicle speed(km/h)
Environment	sulfur dioxide gas, carbon monoxide, Nitrogen dioxide, fine dust, fine particulate matter(PM2.5), ozone
Geometry	road length(m), road area(m^2), sidewalk length(m), sidewalk area(m^2)
Safety	number of traffic accidents, traffic Safety Index
Transit facility	whether to transfer, number of bus stops, number of bus line or route
Retail facility	the rate of living, medical, food, education, cultural, finance, welfare, shopping, public, child-care within 300 m radius around subway station
Subway ridership	7~11 subway get on ridership 11~17 subway get on ridership 17~21 subway get on ridership
Bus ridership	7~11 bus get on ridership 11~17 bus get on ridership 17~21 bus get on ridership

5. Results

5.1 Endogeneity test

The study emphasizes the endogeneity issue in the model, and the variable of bus ridership was treated as endogeneity variable for the subway ridership prediction models. It is essential to confirm whether the selected endogeneity variable is significant. Therefore, we conducted a test of endogeneity, namely Durbin (1954) and Wu-Hausman(Wu 1974; Hausman 1978) test, with the 2SLS estimation. The null hypothesis of the Durbin and Wu-Hausman tests is that the variable is exogenous. In this study, the test statistics are highly significant, therefore we rejected the null hypothesis of exogeneity and it showed that we should treat the bus ridership as the endogenous variable.

5.2 Instrument variables test

To ensure that the endogenous variable interacted with strong instrument variables, we conducted Stock and Yogo’s (2005) tests. The null hypothesis of each of Stock and Yogo’s tests is that the set of instruments is weak. If we are willing to tolerate only a relative bias of 5%, then we can reject the null hypothesis and concluded that the instruments are not weak. Since the test statistic of 429.026 far exceeds the critical value of 19.86 in our model, we concluded that our instruments are strong.

5.3 Model results

The result of the bus ridership variable was statistically significant and showed consistently positively signs in all models. It means the bus ridership and subway ridership had a complementary relationship instead of a competitive relationship. The variable of the number of families is statistically significant in the morning, afternoon and evening, but not in the evening. Showing a positive sign indicates that ridership is increasing when the number of families is more in the morning and afternoon. A higher rate in medical and food facility resulted in greater transit ridership. A higher rate in culture, public and child-care facility resulted in lower transit ridership. For the rate in living, Education, shop, finance and welfare facility showed different influences in different time periods.

Table 2 Predictive models of boarding ridership using 2SLS estimation

Category	Dependent var. Subway ridership	morning		afternoon		evening	
		Coef	Robust Std. Err	Coef	Robust Std. Err	Coef	Robust Std. Err
Log(bus ridership)		0.1399***	0.0132 10.63	0.1868***	0.0122 15.27	0.2395***	0.0127 18.84
Socio -economic	Log(number of families)	1.2715***	0.0581 21.88	0.0758	0.0498 1.52	-0.3781***	0.0561 -6.75
	Log(number of workers)	0.4058***	0.0590 6.87	0.3196***	0.0526 6.07	0.5296***	0.0593 8.94
	Log(parking spots/veh)	-0.4065***	0.0825 -4.93	NA	NA	NA	NA
	Log(education /family)	-0.1291**	0.0579 -2.23	-0.3757***	0.0565 -6.65	-0.7697***	0.0636 -12.10
Geometry	Log(road areas)	-0.7820***	0.0766 -10.20	0.3351***	0.0733 4.57	1.0107***	0.0822 12.30
Safety	Log(number of accidents)	-0.9428***	0.0705 -13.37	-0.5408***	0.0685 -7.89	-0.5510***	0.0767 -7.18
	Safety index	0.0103***	0.0032 3.23	0.0105***	0.0030 3.46	0.0161***	0.0034 4.71
Transit facility	transfer	0.5744***	0.0188 30.57	0.5313***	0.0186 28.63	0.5102***	0.0210 24.26
Retail facility	Rate of living facility	NA	NA	-0.01***	0.00 -3.45	0.01***	0.00 3.20
	Rate of medical facility	0.0487***	0.0036 13.71	0.0525***	0.0038 13.65	0.0352***	0.0043 8.17
	Rate of food facility	0.0716***	0.0066 10.93	0.1005***	0.0056 17.88	0.1071***	0.0064 16.78
	Rate of education facility	NA	NA	-0.01**	0.00 -1.99	-0.02***	0.01 -3.46
	Rate of cultural facility	-0.0294***	0.0044 -6.71	NA	NA	NA	NA
	Rate of financial facility	-0.0020	0.0013 -1.51	0.0106***	0.0013 8.19	0.0197***	0.0015 13.46
	Rate of welfare facility	0.0069*	0.0038 1.82	-0.0333***	0.0040 -8.39	-0.0319***	0.0045 -7.16
	Rate of shop facility	0.0046***	0.0014 3.26	0.0101***	0.0013 7.71	0.0071***	0.0015 4.80
	Rate of public facility	-0.0090***	0.0024 -3.83	-0.0064***	0.0023 -2.77	-0.0055**	0.0026 -2.10
	Rate of day care facility	NA	NA	-0.04***	0.00 -8.82	-0.06***	0.00 -12.45
constant		8.4346***	0.9621 8.77	2.9679***	0.9349 3.17	-5.4043***	1.0443 -5.18
R-squares		0.5301		0.6161		0.6616	

1) ***,**and* indicates the coefficient is statistically significant at 1%, 5%, 10% of confidence level, respectively.

2) Instruments: bus stop, line, sulfurousacidgasppm, carbonmonoxideppm, nitrogenmonoxide, finedust, ozoneppm

Table 3 shows the results of the accuracy test for three models. The MAE(Mean Absolute Error), MSE(Mean Square Error), RMSE(Root Mean Square Error), MAPE(Mean Absolute Percentage Error) were estimated for each period model, and it was found that the model for the evening had higher errors than other two models for morning and afternoon time periods.

Table 3 Model accuracy tests

No	Model	MAE	MSE	RMSE	MAPE
1	2SLS model for morning	0.4132	0.2919	0.5402	0.0360
2	2SLS model for afternoon	0.5267	0.4490	0.6701	0.0465
3	2SLS model for evening	1.8469	4.3332	2.0816	0.1986

6. Conclusion

This study aims to develop a consistent model for predicting subway ridership incorporating the endogenous variable of bus ridership. It was found that there was the endogenous relationship between the variables of bus ridership and subway ridership, and the instrument variables such as bus lines, bus stops and environmental five factors were statistically significant and strong. Model accuracy tests are high, and the coefficients of independent variables have a consistent sign in three time periods. Five retail facilities variables (living, education, shop, finance, and welfare) have different signs in different time periods, and it was shown that the subway ridership during the morning, afternoon, and evening is likely to change sensitively with the types of retail facility.

Reference

[1] B.D. Taylor and N.Y. Camille (2003) The Factors Influencing Transit Ridership: A Review and Analysis of the Ridership Literature, *University of California Transportation Center*, pp.4 -18.

[2] Y.C. Chiou, R. Jou, and C. Yang (2015) Factors affecting public transportation usage rate: Geographically weighted regression, *Journal of Transportation Research Part A*, pp. 5-17.

[3] H.M. Ashley (2014) Correlation Between Land Use and Metro Rail Ridership in Los Angeles, *Master Thesis*, Columbia University.

[4] Natalie L. Stiffler (2011) The Effect of Transit-Oriented Development on Vehicle Miles Traveled: A Comparison of a TOD versus a non-TOD Neighborhood in Carlsbad, CA, Master Thesis, California Polytechnic State University

[5] B.D. Taylor, D. Miller, H. Iseki, C. Fink (2003) Analyzing the Determinants of Transit Ridership Using a Two-Stage Least Squares Regression on a National Sample of Urbanized Areas, *University of California Transportation Center*, pp. 8-27.

[6] J.A. Hausman (1978) Specification tests in econometrics. *Econometri*ca 46: pp.1251-1271.

[7] J. H. Stock, J. H. Wright, and M. Yogo (2002) A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20, pp. 518–529.

[8] D.M. Wu (1974) Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica* 42, pp. 529–546.