Preventive Medicine 99 (2017) 264–268

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Analyzing the impact of public transit usage on obesity

Zhaowei She [a], Douglas M. King [b,*], Sheldon H. Jacobson [c]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States
[b] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States
[c] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, United States

ARTICLE INFO

Article history:
Received 15 August 2016
Received in revised form 27 February 2017
Accepted 12 March 2017
Available online 16 March 2017

Keywords:
Obesity
Transportation
Environment design

ABSTRACT

The objective of this paper is to estimate the impact of county-level public transit usage on obesity prevalence in the United States and assess the potential for public transit usage as an intervention for obesity. This study adopts an *instrumental regression* approach to implicitly control for potential *selection bias* due to possible differences in commuting preferences among obese and non-obese populations. United States health data from the 2009 Behavioral Risk Factor Surveillance System and transportation data from the 2009 National Household Travel Survey are aggregated and matched at the county level. County-level public transit accessibility and vehicle ownership rates are chosen as *instrumental variables* to implicitly control for unobservable commuting preferences. The results of this instrumental regression analysis suggest that a one percent increase in county population usage of public transit is associated with a 0.221 percent decrease in county population obesity prevalence at the $\alpha = 0.01$ statistical significance level, when commuting preferences, amount of non-travel physical activity, education level, health resource, and distribution of income are fixed. Hence, this study provides empirical support for the effectiveness of encouraging public transit usage as an intervention strategy for obesity.

1. Introduction

Recent studies show that people's commuting choices are associated with their obesity status; more driving is positively associated with obesity prevalence, while higher public transit usage is negatively associated with obesity prevalence (Edwards, 2008; Besser and Dannenberg, 2005; Behzad et al., 2013; Jacobson et al., 2011). As such, the Centers for Disease Control and Prevention (CDC) encourages public transit usage as a possible obesity intervention strategy (Centers for Disease Control and Prevention, 2009; Centers for Disease Control and Prevention, 2011). This strategy's effectiveness can be affected by confounding factors; if the obese population has significantly lower preference for public transportation, a potential increase in public transit usage may not translate into lower obesity prevalence, since this increase is less likely to come from the obese population. Therefore, to justify obesity interventions based on encouraging public transit usage, it is important to understand whether the negative association between public transit usage and obesity prevalence is independent of confounding factors.

Two commonly-discussed confounding factors in this association are selection bias and potential substitution effects between travel-related and non-travel physical activity. Selection bias refers to the possibility that unobservable differences in people's commuting preferences can affect the estimated association between public transit usage and obesity prevalence. For example, Eid et al. (2008) found that people who are obese tend to prefer living in more sprawling neighborhoods, while Plantinga and Bernell (2007) noted that public transportation is less viable in these neighborhoods. In this case, obesity could be a cause of lower public transit usage, rather than a result of lower public transit usage; a simple statistical model associating obesity and public transit usage would only estimate how less likely an obese individual commutes via public transit, instead of the impact of public transit usage on obesity. Another source of confounding is the possible substitution effect between travel-related and non-travel physical activity, such that increasing travel-related physical activity may reduce non-travel physical activity (Saunders et al., 2013). For example, when returning home from a bus ride, one may be either too tired or not have sufficient time for additional physical exercises. In this case, an overweight individual may prefer driving to taking public transit even if they wish to lose weight. As such, the impact of public transit usage on obesity is inconclusive if the negative association between public transit usage and obesity is due to confounding effects.

To address possible self-selection estimation bias, this study proposes an *instrumental regression*, or two-stage least squares (2SLS) regression approach to estimate the impact of public transit usage on obesity prevalence at the county population level. In the estimation,

* Corresponding author at: 117 Transportation Building, MC-238, 104 S. Mathews Ave., Urbana, IL 61801, United States.
E-mail address: dmking@illinois.edu (D.M. King).

http://dx.doi.org/10.1016/j.ypmed.2017.03.010
0091-7435/© 2017 Published by Elsevier Inc.

amount of non-travel physical activity, health resource, and distribution of income are explicitly controlled through data from multiple sources. Unobserved commuting preferences are implicitly controlled though two traffic-related instrumental variables: public transit accessibility and vehicle ownership rates. Hence, this approach focuses on people forced to use public transit due to traffic constraints. Therefore, variations in public transit usage due to commuting preferences have been statistically ruled out, and hence, should not bias the estimation. As such, this study addresses the limitations of earlier studies (Frank et al., 2007; Tiemann and Miller, 2013) and provides further evidence of the negative impact of public transit usage on obesity. By separating the impact of public transit usage on obesity from potential confounding effects, this study provides further support for the public health efforts to reduce obesity prevalence through encouraging public transit usage.

2. Methods

2.1. Data sets and data pre-processing

This study gathers and matches county-level aggregated health and transportation data from multiple sources. Health related variables are calculated from the 2009 Behavioral Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention, 2015). Surveys of BRFSS have been conducted annually since 1984 by the Centers for Disease Control and Prevention (CDC) and other federal agencies through a nationwide random sample (one per household) of adults (18 + years) in the United States. Health data capture obesity status, and its associated risk factors, with health variables defined as:

- *OBESE*: Percentage of county population with Body Mass Index (BMI) at least 30 kg/m^2; (Ogden et al., 2014)
- *LTPA*: Percentage of county population engaging in leisure time physical activity (e.g., running, calisthenics, golf, gardening, walking);
- *Employed%*: Binary variable, with 1 indicating data point collected from respondents who were employed/self-employed in 2009; 0 otherwise;
- *Education:* Percentage of county population with education above the high school level (at least one year of college education)
- *Healthcare*: Percentage of county population with health care coverage (e.g., health insurance, prepaid plans, or Medicare).

The 2009 National Household Travel Survey (NHTS) database provides variables related to transportation patterns (U.S. Department of Transportation, Federal Highway Administration, 2009). The National Household Travel Survey is conducted to examine travel behavior at the individual and household level in the United States, and is publicly accessible through a database published by Federal Highway Administration of the U.S. Department of Transportation. This study utilizes a special research version with more detailed geographic information; to remain consistent with the age limits of the BRFSS, all individuals with age below 18 years are excluded. Transportation data describe transportation patterns and transit mode choice, with transportation variables defined as:

- *Transit%*: Percentage of the county population using public transit at least eleven times per month (i.e., two or more days a week);
- *Transit_Important%*: Percentage of the county population ranking accessibility/availability of public transit as their most important transportation issue, compared to other issues like highway congestion, lack of walkways or sidewalks, price of travel, aggressive/distracted drivers and safety concerns;
- *AverageVehicle*: Average number of vehicles per household at county level;
- *Rail*: Binary variable, with 1 indicating data point collected from respondents residing in a metropolitan area with subway/rail; 0 otherwise;
- *Employed%*: Same as for Health data.

This study also includes data to control for social-economic factors and spatial correlations in the associations between obesity and public transit usage. To control for income, this study includes *Income* (county level median household income) and *Poverty* (percentage of county population that lives below the poverty threshold) (United States Census Bureau, 2015). The variable *Income* is obtained from the U.S. Census Bureau (United States Census Bureau, Small Area Estimates Branch, 2009), as median income statistics for each county cannot be computed from either BRFSS or NHTS, which only provides a range, instead of the exact number, of each interviewee's income level. The U.S. Census Bureau derived this *Income* estimate through combining the decennial census and the direct estimates from the American Community Survey. The variable *Poverty* is computed as the average of estimates from BRFSS and NHTS. The *Poverty* estimate is updated by the U.S. Census Bureau each year using the change in the average annual Consumer Price Index for All Urban Consumers. To control for possible spatial correlations between county observations, this study includes fixed effects for a vector $\overrightarrow{State}$, which describes the state in which each county is located; hence, possible confounding effects due to spatial closeness can be addressed.

These data sets are aggregated and matched based on two identifier types: At the county aggregate level (given the large sample size in each county), and according to their employment status (*Employed%*) (to control for the difference in occupational physical activities and leisure time physical activities). Each county-level statistic is a weighted average of at least 30 individual observations from the raw datasets, with 318 counties from 44 U.S. states represented. Table 1 summarizes the descriptive statics of relevant variables.

2.2. Statistical analysis

This study uses 2SLS regression to address the potential influence of self-selection bias. Self-selection bias cannot be controlled explicitly through an ordinary least squares model, because subjective motives (e.g., personal preferences) are often not evaluated in nationwide surveys. The advantage of 2SLS regression is its ability to control for potential confounding variables without direct estimations of these variables (Wooldridge, 2012). Conceptually, one can understand 2SLS regressions as “causal path analysis” (Angrist and Krueger, 2001). From Fig. 1, personal preference (PP) for a sedentary lifestyle can simultaneously influence transit mode choice (PT) and obesity (OB), and cannot be explicitly controlled with the available data. To address this confounding effect, a vector of instrumental variables (IV) would be needed, with variations in IV only associated with variations in OB through PT. For example, in a study on the causal effect of obesity on wages, Cawley (2004) uses maternal body weight as an instrumental variable to estimate the causal impact of females' body weight on their wages. Here Cawley assumes that maternal body weight can only associate with females' wages through body weight inheritance. By regressing mother's BMI on daughter's BMI, he obtained a predicted value of daughter's BMI in the first stage of the 2SLS model. In the second stage, he regresses wage outcomes against this predicted BMI and other control variables to obtain unbiased estimates of the impact of body weight on wage outcomes. In this case, he implicitly controlled for risk factors in obesity due to low wages, for example unhealthy food, because maternal body weight can only change females' inherited body weight and has no impact on other risk factors in obesity due to low wages. A similar 2SLS approach is adopted in this study.

In the first stage regression of our 2SLS model, *Transit_Important%* and *AverageVehicle* are the instrumental variables chosen to characterize a county's traffic constraints. Regardless of their commuting preferences, people living in a county with high *Transit_Important%* are more