Transportation Research Part A 119 (2019) 162–169

Contents lists available at ScienceDirect

Transportation Research Part A

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

Is promoting public transit an effective intervention for obesity? A longitudinal study of the relation between public transit usage and obesity

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

[a] H. Milton Stewart School of Industrial & 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

ABSTRACT

There is increasing evidence on the association between public transit usage and obesity. To further understand the causal impact of changes in county public transit usage on county obesity rates, this paper presents a longitudinal study on this topic. Annual health data from the Behavioral Risk Factor Surveillance System (BRFSS) and transportation data from the National Household Travel Survey (NHTS) were aggregated and matched at the county level, to create a panel data set with 227 counties from 45 states across two time periods, 2001 and 2009. Annual public transit funding, obtained from the National Transit Database (NTD), is chosen as an *instrumental variable* to simulate changes in public transit usage caused by exogenous changes in public policies. Possible confounding variables such as amount of leisure time physical activity, health care coverage and distribution of income are explicitly controlled. All time-invariant county level heterogeneities are implicitly controlled using first difference estimators. This study shows that promoting public transit in a county can effectively decrease the county obesity rate. Specifically, a one percentage point increase of frequent public transit riders in a county population is estimated to decrease the county population obesity rate by 0.473% points. This result supports findings in previous research that the extra amount of physical activity involved in public transit usage can have a statistically significant impact on obesity. In addition, this study also provides empirical evidence for the effectiveness of encouraging public transit usage as a public health intervention for obesity.

1. Introduction

Since World War II, the United States has witnessed a spiraling growth of obesity rates and automobile travel. In the past five decades, the national obesity rate in the United States has increased more than 20%, reaching 35.1% among the population with age over 20 years in 2012 (Cutler et al., 2003; Ogden et al., 2014). In about the same period, annual vehicle miles traveled (VMT) of all vehicle types in the United States experienced a steady increase, both in gross amount and per capita (Puentes and Tomer, 2008). Jacobson et al. (2011) and Behzad et al. (2013) document a high correlation between obesity rates and VMT per licensed driver from 1985 to 2007 at the national level, with R^2 above 90%. Public transit usage, in contrast, is shown to be negatively correlated with obesity rates (Besser and Dannenberg, 2005; Edwards, 2008; Flint et al., 2014; Frank et al., 2007; Tiemann and Miller, 2013; She et al., 2017). These associations naturally lead to the following question: will obesity rates decrease if more people choose to commute with public transit instead of their own vehicles?

Current literature provides no straightforward answer. It is possible that the association between transit mode choice and obesity is a spurious correlation in which other variables simultaneously influence both transportation patterns and obesity rates, so that change of transit mode choice cannot cause changes in obesity rates. For example, there is evidence that individuals who are obese

* 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).

https://doi.org/10.1016/j.tra.2018.10.027
Received 30 June 2017; Received in revised form 20 August 2018; Accepted 22 October 2018
Available online 21 November 2018
0965-8564/

tend to prefer driving over other transit modes (Plantinga and Bernell, 2007; Eid et al., 2008). Therefore, a preference for sedentary lifestyle may simultaneously cause high VMT per licensed driver and high obesity rates. However, it is difficult to measure people's subjective preference in lifestyle using nationwide data. Consequently, simple models such as ordinary least squares cannot answer this question. Hence, despite the various associations documented, few studies consider the causal effect of public transit usage on obesity directly.

This study investigates causality between public transit usage and obesity through changes in a longitudinal setting. The hypothesis that higher public transit usage causes lower obesity rates is tested with longitudinal data from 2001 and 2009. A first difference model is used to implicitly control for all time invariant omitted variables (e.g., weather patterns). This type of model is ideal for this investigation, since geographic specific transit and lifestyle choice are most likely to be time invariant at the county population level. For example, Tucker and Gilliland (2007) review evidence that weather patterns can impact physical activity in some regions. Hence, the first difference model is capable of addressing the omitted variable problems in previous cross sectional studies, such as Flint et al. (2014) and She et al. (2017).

Moreover, this study adopts the latent class instrumental variables framework to design a quasi experiment, which provides causal evidence for the effectiveness of public health interventions for obesity based on changes in transit mode choice. Specifically, there is an ongoing debate about the impacts of public transit service availability on physical activity (Chang et al., 2017; Cao et al., 2010). The challenge is that studies of this type are best conducted in a longitudinal setting, because cross sectional study often suffers from the self selection bias (Cao et al., 2010). However, the previous longitudinal studies are mostly at the local level. For example, Chang et al. (2017) conducted a local longitudinal study, and showed that the implementation of bus rapid transit (BRT) in Mexico City significantly increased the local residents' time spent in physical activity, especially walking. To generalize the existing local longitudinal studies to a national scope, this paper creates a panel data set with 227 counties from 45 states, and partitions them into four groups, based on the presence of frequent transit riders in the county in year 2001 and 2009. This partition seeks to explain the between group difference of obesity rates in counties where the intervention (public transit riding) was utilized in only one year of 2001 or 2009 (but not both). Furthermore, to exploit exogenous variations of public transit usage within these groups, this study uses the change in public transit funding in each county between 2001 and 2009 as an *instrumental variable* to simulate policy induced transit behavior changes. In summary, to estimate the effectiveness of frequent transit riding as a public health intervention for obesity, this model uses both the between group difference (existence of frequent transit riders in a county) and within group difference (changes of frequent transit riders as a percentage in overall county population) in public transit usage to explain the difference of obesity rates in these counties between 2001 and 2009. Therefore, this study can help policy makers to assess the potential public health impacts of providing frequent transit service in a county.

This paper is organized as follows. Section 2 introduces the data sources used in the first difference model, and explains the construction and meanings of the dependent and independent variables. Section 3 illustrates the regression model of this study, emphasizing the structure of the first difference approach and the selection of instrumental variables. Section 4 presents the estimation results, while Section 5 and Section 6 discuss the implications of these results and conclude this study.

2. Data sets and preprocessing

Following from previous research studying the relation between public transit usage and obesity rates with U.S. data, a national level analysis of this relation can only be performed at an aggregate level (Tiemann and Miller, 2013; She et al., 2017). Specifically, since health and transportation data need to be drawn from different sources and may not possess a default common identifier, they need to be matched at a geographically aggregated level for further analysis. This study uses three data sources as in She et al. (2017), as well as public transit data from the 2016 National Transit Database Annual Data Products (NTD). Data measuring obesity rates and common risk factors for obesity were compiled from the 2001 and 2009 annual surveys of the Behavioral Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention, 2016). Notably, since BRFSS only surveys the adult population, this study only analyzes observations from individuals with age at least 18 years. Transportation data were collected from the 2001 and 2009 National Household Travel Survey (NHTS) (United States Department of Transportation, Federal Highway Administration, 2001; United States Department of Transportation, Federal Highway Administration, 2009). Data from the 2000 and 2010 national census provide the necessary demographic and geographic data for this analysis (United States Census Bureau, 2015b; United States Census Bureau, 2015a; United States Census Bureau, 2010; United States Census Bureau, 2002). The county level public transit funding data in 2001 and 2009 are calculated from the 2016 National Transit Database Annual Data Products (NTD) (Federal Transit Administration, 2017). These data are matched using the same matching strategy proposed in She et al. (2017). Observations from these three sources are aggregated and matched at the county level. In the aggregated dataset, 227 counties from 45 states in the United States are included, with each county having at least 30 individual observations from both BRFSS and NHTS to balance the representativeness of the sample space and efficiency of estimates obtained, as explained in She et al. (2017).

To obtain a panel structure, each observation is indexed over time. As a result, each statistic in this analysis has a time dimension $t \in T$ and a spatial dimension $i \in I$. The time index set T has 2 elements, year 2001 and year 2009. The spatial index set I has 227 elements, representing the counties which have records in both year 2001 and year 2009 in the data set. The final panel dataset used in this study contains the following statistics:

- ΔX_i^{HC}, defined as $X_{i2009}^{HC} - X_{i2001}^{HC}$, is the difference in percentage of health care coverage in that county between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here X_{it}^{HC} measures the percentage of county i's population with some health care coverage, for example health insurance, prepaid plans or government plans such as Medicare, during year $t \in T = \{2001, 2009\}$.

- ΔX_i^{MHI}, defined as $X_{i2009}^{MHI} - X_{i2001}^{MHI}$, is the difference in median household annual income in that county between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here X_{it}^{MHI} measures the median household annual income in county i during year $t \in T = \{2001, 2009\}$.
- ΔX_i^{LTPA}, defined as $X_{i2009}^{LTPA} - X_{i2001}^{LTPA}$, is the difference in percentage of population engaging in some kind of leisure time physical activity in that county between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here X_{it}^{LTPA} measures the percentage of county i's population that engages in some kind of leisure time physical activity on at least a monthly basis during year $t \in T = \{2001, 2009\}$.
- ΔY_i^{Obese}, defined as $Y_{i2009}^{Obese} - Y_{i2001}^{Obese}$, is the difference in county population obesity rates between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here Y_{it}^{Obese} measures the percentage of county i's population with Body Mass Index (BMI) at least 30 kg/m^2, where BMI is the individual's weight in kilograms divided by the square of their height in meters during year $t \in T = \{2001, 2009\}$.
- ΔX_i^{Pov}, defined as $X_{i2009}^{Pov} - X_{i2001}^{Pov}$, is the difference in percentage of population that live below the poverty threshold in that county between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here X_{it}^{Pov} measures the percentage of county i's population that live below the poverty threshold during year $t \in T = \{2001, 2009\}$. The poverty threshold is based on United States Census Bureau (2015b).
- ΔX_i^{Pub}, defined as $X_{i2009}^{Pub} - X_{i2001}^{Pub}$, is the difference in percentage of the population that use public transit at least 11 times a month or at least two days a week in that county between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here X_{it}^{Pub} measures the percentage of county i's population that use public transit at least 11 times a month or at least two days a week during year $t \in T = \{2001, 2009\}$.
- $\widehat{\Delta X_i^{Pub}}$ is a fitted value of ΔX_i^{Pub}, where $i \in I = \{1, ..., 227\}$. Its derivation is explained in Section 3.
- I_{it}^{Pub} is a categorical variable, that takes the value 1 if there is a positive percentage of county i's population that use public transit at least 11 times a month or at least two days a week during year $t \in T = \{2001, 2009\}$, and takes the value 0 otherwise, where $i \in I = \{1, ..., 227\}$. In other words, I_{it}^{Pub} measures the existence of frequent transit riders in county i during year t.
- I_i^{Fund} is a categorical variable, that takes the value 1 if the base year (2001) public transit funds are nonzero, and takes the value 0 otherwise
- ΔX_i^{Fund}, defined as $\frac{X_{i2009}^{Fund} - X_{i2001}^{Fund}}{X_{i2001}^{Fund}}$, is the proportional difference of public transit fund available to county transit agencies between year 2009 and year 2001, where $i \in I = \{1, ..., 227\}$. Here X_{it}^{Fund} measures the amount of public transit fund available to county transit agencies during year $t \in T = \{2001, 2009\}$. For those counties with zero X_{it}^{Fund} in both periods, ΔX_i^{Fund} is set to 0. If a county have nonzero public transit funding in year 2009 but zero public transit funding in the base year 2001, the ΔX_i^{Fund} of this county is set to 1.

In the rest of the paper, the dependent variables of interest are Y_{it}^{Obese} and ΔY_{it}^{Obese}, while other variables are considered as independent variables. Specifically, county population obesity rates are measured by Y_{it}^{Obese}, county population public transit usage is measured by X_{it}^{Pub}, daily energy expense in non travel related physical activity is approximated by X_{it}^{LTPA}, local accessibilities of health resources are reflected in X_{it}^{HC} and income distributions are controlled by X_{it}^{MHI} and X_{it}^{Pov}. When expressed as temporal differences between 2001 and 2009 (i.e. ΔY_{it}^{Obese}, ΔX_{it}^{Pub}, ΔX_{it}^{LTPA}, ΔX_{it}^{HC}, ΔX_{it}^{MHI} and ΔX_{it}^{Pov}), these variables are used to estimate the longitudinal model. Details of this *first difference* approach are discussed in Section 3. Instrumental variable I_{it}^{Pub} indicates whether frequent transit riders exists (= 1) in county i of year t. Instrumental variable I_i^{Fund} denotes whether public transit funds are nonzero (= 1) in county i of year 2001. Instrumental variable ΔX_i^{Fund} measures the proportional increase in public transit fund in year 2009, compared with the amount of public transit fund in year 2001. The purpose and role of the instrumental variables are discussed in Section 3.

3. Methodology

This section derives an estimator of the causal impact of public transit usage on obesity. The estimator is derived in a longitudinal framework, and hence, is statistically independent of all time invariant omitted variables by construction. The derivation of this estimator is based on the following key assumption: the linear association between the dependent variable of interest Y_{it}^{Obese} and independent variables X_{it}^{Pub}, X_{it}^{LTPA}, X_{it}^{HC}, X_{it}^{MHI} and X_{it}^{Pov} does not change over time. In other words, this derivation assumes that, for a time invariant county fixed effect, α_i,

$$Y_{i2001}^{Obese} = \beta_{2001} + \beta_{Pub} X_{i2001}^{Pub} + \beta_{LTPA} X_{i2001}^{LTPA} + \beta_{HC} X_{i2001}^{HC} + \beta_{MHI} X_{i2001}^{MHI} + \beta_{Pov} X_{i2001}^{Pov} + \alpha_i + \epsilon_{i2001}, \quad (1)$$

and

$$Y_{i2009}^{Obese} = \beta_{2009} + \beta_{Pub} X_{i2009}^{Pub} + \beta_{LTPA} X_{i2009}^{LTPA} + \beta_{HC} X_{i2009}^{HC} + \beta_{MHI} X_{i2009}^{MHI} + \beta_{Pov} X_{i2009}^{Pov} + \alpha_i + \epsilon_{i2009}, \quad (2)$$

share the same set of parameters $\beta^* = \{\beta_{Pub}, \beta_{LTPA}, \beta_{HC}, \beta_{MHI}, \beta_{Pov}\}$, where $i \in I = \{1, ..., 227\}$. Under this assumption, a first difference estimator will be derived to estimate this common set of parameters β^*. Specifically, in this step, one would estimate the linear association presented in (1) and (2) with the temporal difference variables (i.e. ΔY_{it}^{Obese}, ΔX_{it}^{Pub}, ΔX_{it}^{LTPA}, ΔX_{it}^{HC}, ΔX_{it}^{MHI} and ΔX_{it}^{Pov}) instead of their cross section counterparts (i.e. Y_{it}^{Obese}, X_{it}^{Pub}, X_{it}^{LTPA}, X_{it}^{HC}, X_{it}^{MHI} and X_{it}^{Pov}). Notably, this estimator implicitly controls for all time invariant omitted variables because the county fixed effect, α_i, is time invariant and therefore is eliminated in the temporal differences. Moreover, the latent class instrumental variable method is employed to design a quasi experiment and evaluate

the effectiveness of encouraging public transit usage as a public health intervention for obesity. Specifically, the 227 counties are partitioned into four latent classes based on the existence of frequent transit riders in year 2001 and 2009. Here the existence of frequent transit riders can be interpreted as a treatment assigned to obesity prevalence: counties where frequent transit riders existed are analogous to the patients who received the treatment (treatment group), while counties where frequent transit riders did not exist are analogous to the patients who did not received the treatment (control group). By this interpretation, these four classes can be labeled following the standard terminology in biostatistics (see Baker et al. (2016)):

1. *Always-Receiver*: Counties where frequent transit riders exist in both year 2001 and year 2009 ($I_{i2001}^{Pub} = 1$ and $I_{i2009}^{Pub} = 1$). The dataset contains 168 county observations in this group, in which the District of Columbia had the largest increase in the percentage of frequent transit riders.
2. *Consistent-Receiver*: Counties where frequent transit riders exist in year 2009 but not in year 2001 ($I_{i2001}^{Pub} = 0$ and $I_{i2009}^{Pub} = 1$). The dataset contains 33 county observations in this group, in which Kanawha County in West Virginia had the largest increase in the percentage of frequent transit riders.
3. *Inconsistent-Receiver*: Counties where frequent transit riders exist in year 2001 but not in year 2009 ($I_{i2001}^{Pub} = 1$ and $I_{i2009}^{Pub} = 0$). The dataset contains 18 county observations in this group, in which East Baton Rouge Parish in Louisiana had the largest decrease in the percentage of frequent transit riders.
4. *Never-Receiver*: Counties where frequent transit riders exist in neither year 2001 nor year 2009 ($I_{i2001}^{Pub} = 0$ and $I_{i2009}^{Pub} = 0$). The dataset contains 8 county observations in this group. These counties are mostly in low population density area such as Ada County in Idaho and Yellowstone County in Montana.

Moreover, to explore the variations of ΔX_{it}^{Pub} within each of these four classes, this study uses the proportional change in public transit funds in each county between year 2001 and 2009, ΔX_i^{Fund}, as another instrumental variable to simulate the policy induced changes in public transit usage in between these periods.

To ensure that the simulated changes in public transit usage are exogenous, the instrumental variables I_{i2001}^{Pub}, I_{i2009}^{Pub}, I_i^{Fund}, ΔX_i^{Fund} and $I_i^{Fund} \times \Delta X_i^{Fund}$ should only associate with obesity rates ΔY_i^{Obese} though public transit usage ΔX_i^{Pub}. Specifically, here two assumptions are made:

1. *Relevance Assumption*: Changes the existence of frequent transit riders and amount of public transit funding (i.e., the instrumental variables) should be associated with changes in public transit usage (i.e., the dependent variable in the first stage regression).
2. *Exogenous Assumption*: Emergence/disappearance of frequent transit riders and public transit funding decisions (i.e., the instrumental variables) should not be directly associated with obesity concerns (i.e., the dependent variable in the second stage regression).

The relevance assumption is empirically supported by the weak instrument test performed in Section 4. While the exogenous assumption cannot be tested directly, Section 4 provides partial empirical support to this assumption through the Wu-Hausman test and the Sargan test. Moreover, the specific instrumental variables chosen in this study play a key role in limiting the potential sources of endogeneity. While other trends may have influenced county-level obesity rates between 2001 and 2009, the exogeneous assumption is violated only if these trends lack independence with the instrumental variables. Based on the instrumental variables chosen in this study, the exogenous assumption therefore requires that not all frequent transit riders in a county choose public transit as their main transit mode due to these other trends (i.e., these other trends are independent of I_{i2001}^{Pub} and I_{i2009}^{Pub}), and that transit funding decisions are not made based on these other trends (i.e., these other trends are independent of I_i^{Fund}, ΔX_i^{Fund} and $I_i^{Fund} \times \Delta X_i^{Fund}$).

Beyond the empirical assessments of the exogenous assumption to be presented in Section 4, the following points provide rationale for the two key components of the exogenous assumption.

- The emergence/disappearance of frequent transit riders, represented by I_{i2001}^{Pub} and I_{i2009}^{Pub}, should be exogenous to obesity concerns in population. This is logical, since individuals may choose to ride public transit for many reasons. Hence, from a county population perspective, it is unlikely that *all* frequent transit riders in a county choose public transit as their main transit modes out of obesity concerns.
- *Some* frequent transit riders in a county, on the other hand, were likely to choose public transit as their main transit modes due to health considerations. However, it is not necessary to assume that all within-group variations in public transit usage are exogenous. Instead, this study only requires public transit funding decisions, represented by I_i^{Fund}, ΔX_i^{Fund}, and $I_i^{Fund} \times \Delta X_i^{Fund}$, to be exogenous to obesity concerns among county population.

In summary, this study uses policy induced exogenous changes in public transit funds to simulate exogenous changes in public transit usage within the four groups: Always-Receiver, Consistent-Receiver, Inconsistent-Receiver and Never-Receiver, and only assumes that changes in public transit usage across these groups were exogenous. Therefore, the estimated effects of public transit usage on obesity can be interpreted as the treatment effect of a public health intervention in a quasi experiment setting.

The main model of interest in this study is a two stage least squares model. The first stage regression uses instrumental variables I_{it}^{Pub}, I_i^{Fund} and ΔX_i^{Fund} to simulate variations in ΔX_{it}^{Pub},

$$\Delta X_i^{Pub} = \beta_0' + \beta_1' I_{i2001}^{Pub} + \beta_2' I_{i2009}^{Pub} + \beta_3' I_i^{Fund} + \beta_4' \Delta X_i^{Fund} + \beta_5' I_i^{Fund} \times \Delta X_i^{Fund} + \beta_6' \Delta X_i^{LTPA} + \beta_7' \Delta X_i^{HC} + \beta_8' \Delta X_i^{MHI} + \beta_9' \Delta X_i^{Pov} + \epsilon_i', \tag{3}$$

where $i \in I = \{1, ..., 227\}$. The simulated temporal differences in public transit usage, $\widehat{\Delta X_i^{Pub}}$, are then used in the second stage to estimate the parameter of interest, β_{Pub}, as in

$$\Delta Y_i^{Obese} = \Delta\beta + \beta_{Pub} \widehat{\Delta X_i^{Pub}} + \beta_{LTPA} \Delta X_i^{LTPA} + \beta_{HC} \Delta X_i^{HC} + \beta_{MHI} \Delta X_i^{MHI} + \beta_{Pov} \Delta X_i^{Pov} + \epsilon_i, \tag{4}$$

where $i \in I = \{1, ..., 227\}$. In this two stage least squares model (3) and (4), the dependent variable ΔY_i^{Obese} measures the difference in obesity prevalence in county i between year 2009 and year 2001. Independent variables, ΔX_i^{Pub}, ΔX_i^{LTPA}, ΔX_i^{HC}, ΔX_i^{MHI} and ΔX_i^{Pov}, measure the temporal differences of X_{it}^{Pub}, X_{it}^{LTPA}, X_{it}^{HC}, X_{it}^{MHI} and X_{it}^{Pov}, as introduced in Section 2. Notably, these difference variables have the same set of coefficients as their counterparts in (1) and (2). The intercept term, $\Delta\beta$, is the difference between β_{2009} and β_{2001}, namely $\Delta\beta = \beta_{2009} - \beta_{2001}$. It measures the average difference in county-level obesity prevalence in the United States when there is no change in X_{it}^{Pub}, X_{it}^{LTPA}, X_{it}^{HC}, X_{it}^{MHI} and X_{it}^{Pov}. Other parameters have the same interpretations as in (1) and (2):

- β_{Pub} measures the percentage point change in county population obesity rates caused by a one percentage point increase of frequent public transit riders in the county population. The objective of this study is to provide further evidence for the potential causal effect of public transit usage on obesity.
- β_{LTPA} measures the percentage point change in county population obesity rates associated with a one percentage point increase of individuals who engage in some kind of leisure time physical activity at least on a monthly basis in the county population.
- β_{HC} measures the percentage point change in county population obesity rates associated with a one percentage point increase of health care coverage in the county population.
- β_{MHI} measures the percentage point change in county population obesity rates associated with a one dollar increase in county median annual household income.
- β_{Pov} measures the percentage point change in county population obesity rates associated with a one percentage point increase in the poverty rate in the county population.

The first difference estimator derived from the model in (3) and (4) can better address possible omitted variable bias in the ordinary least squares model (1) and (2). To obtain an unbiased estimate of β_{Pub} from (1) and (2), one needs to ensure that omitted variables measured by α_i are independent of other regressors X_{it}, where $X_{it} \in \{X_{it}^{Pub}, X_{it}^{LTPA}, X_{it}^{HC}, X_{it}^{MHI}, X_{it}^{Pov}\}$. However, this is a very strong assumption. For example, transit mode preference can be a potential omitted variable which is simultaneously associated with obesity Y_{it}^{Obese} and public transit usage X_{it}^{Pub}. Therefore, a first difference model (3) and (4) is necessary for this study.

4. Results

Table 1 presents the estimation results of the parameters in model (3) and (4). The overall model is statistically significant at the $\alpha = 0.01$ level, with a p-value of 3.12×10^{-7}. [1] There are three main findings in these results.

- *The estimation result of β_{Pub} confirms the causal impact of public transit usage on obesity rates, and suggests that a one percentage point increase of frequent public transit riders in a county population can decrease the county population obesity rate by* 0.473% *points.* This result is significant at the $\alpha = 0.05$ level. This estimation result is also consistent with other studies with cross sectional data (e.g. (Flint et al., 2014; Tiemann and Miller, 2013; She et al., 2017)).
- *Encouraging public transit usage and leisure time physical activity can both effectively reduce obesity rates. Taking public transit more frequently may have a larger impact on obesity rate, though the difference is not statistically significant.* The joint hypothesis test comparing the impacts of these two factors (H_0: $\beta_{Pub} = \beta_{LTPA}$, H_1: $\beta_{Pub} \neq \beta_{LTPA}$) cannot reject the null hypothesis at the $\alpha = 0.1$ level (p-value = 0.15). Similar results are also found in Flint et al. (2014) with data from the United Kingdom. Therefore, when aiming to reduce obesity rates, policy makers will face less individual heterogeneity if they choose to reduce obesity rates through encouraging more population level leisure time physical activity, while the impact on obesity rates may vary if they choose to do so through encouraging more population level public transit usage.
- *The estimated value of $\Delta\beta$ suggests that obesity remains a public health concern in the United States, as the average obesity rates in the United States increased by* 8.50% *from 2001 to 2009 when there was no change in public transit usage (X_{it}^{Pub}), levels of leisure time physical activity (X_{it}^{LTPA}), health care coverage (X_{it}^{HC}), annual median household income (X_{it}^{MHI}) and poverty rates (X_{it}^{Pov}).* The result is statistically significant at $\alpha = 0.01$ level and its 95% confidence interval contains the 7.386% increase reported in a similar study by Dwyer-Lindgren et al. (2013), where a first difference estimation is conducted on obesity rate in the United States in the 2001 to 2009 period with BRFSS data with a similar set of covariates. The difference between this study and Dwyer-Lindgren et al. (2013) is that public transit usage is controlled in this study but not in Dwyer-Lindgren et al. (2013); hence, the obesity trend in the United States in the 2001–2009 period found in this study is consistent with other studies analyzing the same dataset.

From Section 3, the validity of instrumental variable design in (3) and (4) depends on two assumptions, namely the *Relevance*

[1] The corresponding Wald statistic is 38.41, which follows a Chi-Squared distribution with 5 degrees of freedom.

Table 1
The estimate values represent the percentage point change in county population obesity rates independently associated with a one unit increase in each factor of model (3) and (4). The unit in every factor in this table is percentage points, except *Income*, whose unit is dollars.

Factor	Parameter	Estimate	95% Confidence Interval	*p* value
(Intercept)	$\Delta\beta$	8.00	(6.17, 9.83)	1.99×10^{-15}
Public	β_{Pub}	−0.473	(−0.906, −0.0399)	0.0334
LTPA	β_{LTPA}	−0.314	(−0.449, −0.178)	9.89×10^{-6}
Healthcare	β_{HC}	−0.106	(−0.258, 0.0470)	0.176
Income	β_{MHI}	-2.75×10^{-4}	$(-4.44 \times 10^{-4}, -1.07 \times 10^{-4})$	0.00158
Poverty	β_{Pov}	0.0781	(−0.00254, 0.159)	0.0590

Assumption and the *Exogenous Assumption*. Based on the data analyzed in this study, these assumptions can be at least partially empirically justified as follows.

- Relevance Assumption: The *F*-test of the first stage regression (H_0: $\beta_1' = \beta_2' = \beta_3' = \beta_4' = \beta_5' = 0$ in (3), H_1: At least one parameter is non-zero) yields a *p*-value of 5.08×10^{-5}, suggesting that the coefficients of instrumental variables are jointly significant in the first stage at the $\alpha = 0.05$ significance level. The *F*-statistic of the instrumental variables in (3) is 5.7675, which implies that the maximum bias in instrumental variable estimators is less than 20%. In other words, if one is willing to accept the maximum bias in instrumental variable estimators to be less than 20%, instead of the 10% rule of thumb suggested by Staiger and Stock (1997), the relevance assumption of our instrumental variable choices holds (Yamano, 2010). To further rule out the weak instrument problem, this study examines the significance of each parameter individually (i.e., H_0: $\beta_i' = 0$, H_1: $\beta_i' \neq 0$ for each $i = 1, 2, 3, 4, 5$), yielding *p*-values of 0.000185, 0.000819, 0.00871, 0.0165 and 0.0145, respectively. Taken together, these tests suggest that the instrumental variables are both jointly and individually significant at the $\alpha = 0.05$ significance level. These results confirm that the instrumental variables chosen are indeed correlated with the dependent variable ΔX_i^{Pub} in the first stage regression (3), and therefore satisfy the relevance assumption.
- Exogenous Assumption: To investigate whether the selected instrumental variables could resolve a potential endogeneity problem, the Sargan test is used to assess whether there exists a set of parameter values in the second stage regression (4) that makes the regressors in the first stage regression (3) independent of the resulting residual errors in the second stage regression (i.e., H_0: Such parameter values exist and H_1: Such parameter values do not exist). This test yields a *p*-value of 0.123, suggesting that the instrumental variables chosen in this study have the potential to address an endogeneity problem. Moreover, if it is assumed that the instrumental variable approach used in this study avoids endogeneity issues, the Wu-Hausman test can assess whether an ordinary least squares approach to the regression model presented in (4) can do the same (i.e., H_0: $\vec{\beta}_{2SLS} = \vec{\beta}_{OLS}$ and H_1: $\vec{\beta}_{2SLS} \neq \vec{\beta}_{OLS}$, where $\vec{\beta}_{2SLS}$ and $\vec{\beta}_{OLS}$ are vectors containing the 2SLS estimators and OLS estimators of the regression parameters in (4), respectively). This test yields a *p*-value 0.0425, suggesting that the ordinary least squares would not be able to avoid an endogeneity problem. Hence, when taken together, the results of these tests suggest (at the $\alpha = 0.05$ significance level) that there is potential for the two-stage least squares approach proposed in this study to avoid endogeneity problems, while an ordinary least squares approach would not be able to do so.

Hence, all of these tests support the instrumental variable choice in this study.

5. Discussion

Though there is abundant evidence of an association between public transit usage and obesity, relatively little is known about the causal relation between public transit usage and obesity. Previous studies on this topic are based on cross sectional models with limited controls for possible confounding effects (Flint et al., 2014; Frank et al., 2007; Tiemann and Miller, 2013; She et al., 2017). In particular, no data directly document individuals' preferences for specific transit modes and whether shifting from people's preferred transit modes to public transportation can lead to other outcomes which increase obesity rates, such as increased food consumption or decreased non travel related physical activity (Saunders et al., 2013; Plantinga and Bernell, 2007; Eid et al., 2008). Consequently, hypotheses of this kind cannot be directly tested with a cross sectional model when there is not enough data to explicitly control for the aforementioned possible confounding effects. Therefore a longitudinal study, which can implicitly control for all time invariant omitted variables, is needed to understand the causal effect of increased public transit usage on obesity.

This study provides evidence for a negative causal relation from public transit usage to obesity with longitudinal data in a quasi experiment framework. Since all time invariant confounding variables are differenced out in the estimation process, individual preferences in transit mode, as a time invariant factor in the aggregate level, cease to be confounders in the model. Therefore omitted variable bias in cross sectional models is better addressed through the longitudinal design of this study. Moreover, the latent class instrumental variables design provide evidence for a causal effect of public transit usage on obesity in a quasi experiment framework, and confirms the effectiveness of encouraging public transit usage as a public health intervention for obesity. In fact, most of the counties in the Consistent-Receiver group had made considerable investment in public transit infrastructure during the 2001–2009 period. For example, Kanawha County in West Virginia started a weekday bus route between Huntington and Charleston under a

three-year federal grant in 2009, which provided more than 13,000 rides in 2009 (The Herald-Dispatch, 2015). As such, the NHTS dataset used in this study shows that the county had a 23.92% emergence of frequent transit riders during the 2001–2009 period. Overall, estimation results presented in Section 4 provide strong evidence for the hypothesis that an increase in county population public transit usage will cause a decrease in county population obesity rates.

Lastly, it should be noted that the treatment effect of public transit usage on obesity is estimated based on simulated variations in public transit usage. As discussed in Section 3, this study only uses the fitted value of public transit usage, simulated by the emergence/disappearance of frequent transit riders and changes in public transit funds, to derive the estimate of the treatment effect. Nevertheless, though the estimated impact of public transit usage on obesity is larger in magnitude compared to that documented in She et al. (2017), their −0.221% point estimate falls in the 95% confidence interval for β_{Pub} reported in Table 1. Therefore, the estimated impact of public transit usage on obesity does not contradict the estimate reported by She et al. (2017), who conducted a cross sectional study using the same data sources.

6. Conclusion

This paper uses aggregated county level panel data to identify causal relations between public transit usage and obesity. Specifically, all time invariant omitted variables which can potentially influence this relation are implicitly controlled and differenced out in the estimation process. This study provides longitudinal evidence for the causal impact of county population public transit usage on county population obesity rates. The estimated impact is consistent with those from previous studies. Therefore, this study suggests that encouraging public transit usage is indeed an effective public health intervention for obesity.

As an observational study, this research still leaves some questions open regarding the causality relation between public transit usage and obesity. For example, this study does not rule out the possibility of time variant confounding effects in this relation. In fact, causal relation is best understood through randomized controlled trial (RCT). Ideally, the increase in public transit usage should be randomly assigned to different subpopulations, to ensure that the treatment group and control group are from the same population and only differ in their amounts of public transit usage. Though rare, RCT data indeed exist. For example, to limit the number of vehicles on the road, the Traffic Management Bureau of Beijing, China, requires a lottery for each citizen who wishes to purchase a new private vehicle in Beijing after January 1, 2012 (Traffic Management Bureau of Beijing, 2011). This is essentially a random assignment of transportation methods among populations who are otherwise the same population. The obesity rate difference between the group who win the lottery and the group who does not win is thus a result of randomized control trial. As these RCT data become available, a more direct estimate of the causality relation between public transit usage and obesity can be obtained.

Acknowledgments

The computational work was conducted with support from the Simulation and Optimization Laboratory and the Bed Time Research Institute at the University of Illinois.

References

Baker, S.G., Kramer, B.S., Lindeman, K.S., 2016. Latent class instrumental variables: a clinical and biostatistical perspective. Stat. Med. 35 (1), :147–160.
Behzad, B., King, D.M., Jacobson, S.H., 2013. Quantifying the association between obesity, automobile travel, and caloric intake. Prev. Med. 56 (2), :103–106.
Besser, L.M., Dannenberg, A.L., 2005. Walking to public transit: steps to help meet physical activity recommendations. Am. J. Prev. Med. 29 (4), :273–280.
Cao, X.J., Xu, Z., Fan, Y., 2010. Exploring the connections among residential location, self-selection, and driving: propensity score matching with multiple treatments. Transport. Res. A: Pol. Pract. 44 (10), :797–805.
Centers for Disease Control and Prevention, 2015. Behavioral Risk Factor Surveillance System. Available at http://www.cdc.gov/brfss/annual_data/annual_data.htm (Accessed October 24, 2016).
Chang, A., Miranda-Moreno, L., Cao, J., Welle, B., 2017. The effect of BRT implementation and streetscape redesign on physical activity: A case study of Mexico city. Transport. Res. A: Pol. Pract. 100, 337–347.
Cutler, D., Glaeser, E., Shapiro, J., 2003. Why have Americans become more obese? J. Econ. Perspect. 17 (3), :93–118.
Dwyer-Lindgren, L., Freedman, G., Engel, R.E., Fleming, T.D., Lim, S.S., Murray, C.J., Mokdad, A.H., 2013. Prevalence of physical activity and obesity in US counties, 2001–2011: a road map for action. Popul. Health Metrics 11 (7).
Edwards, R.D., 2008. Public transit, obesity, and medical costs: assessing the magnitudes. Prev. Med. 46 (1), :14–21.
Eid, J., Overman, H.G., Puga, D., Turner, M.A., 2008. Fat city: questioning the relationship between urban sprawl and obesity. J. Urban Econ. 63 (2), :385–404.
Federal Transit Administration, 2017. 2016 NTD Total Funding Time Series. Available at https://www.transit.dot.gov/ntd/data-product/ts11-total-funding-time-series-2 (Accessed November 27, 2017).
Flint, E., Cummins, S., Sacker, A., 2014. Associations between active commuting, body fat, and body mass index: population based, cross sectional study in the United Kingdom. Brit. Med. J. 349 g4887.
Frank, L.D., Saelens, B.E., Powell, K.E., Chapman, J.E., 2007. Stepping towards causation: do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity? Soc. Sci. Med. 65 (9), :1898–1914.
Jacobson, S.H., King, D.M., Yuan, R., 2011. A note on the relationship between obesity and driving. Transp. Pol. 18 (5), :772–776.
Ogden, C.L., Carroll, M.D., Kit, B.K., Flegal, K.M., 2014. Prevalence of childhood and adult obesity in the United States, 2011–2012. J. Am. Med. Assoc. 311 (8), :806–814.
Plantinga, A.J., Bernell, S., 2007. The association between urban sprawl and obesity: is it a two-way street? J. Reg. Sci. 47 (5), :857–879.
Puentes, R., Tomer, A., 2008. The road ··· less traveled: an analysis of vehicle miles traveled trends in the U.S. Technical report, Brookings Institution, Washington, DC.
Saunders, L.E., Green, J.M., Petticrew, M.P., Steinbach, R., Roberts, H., 2013. What are the health benefits of active travel? A systematic review of trials and cohort studies. PLoS ONE 8 (8), e69912.
She, Z., King, D.M., Jacobson, S.H., 2017. Analyzing the impact of public transit usage on obesity. Prev. Med. 99, 264–268.
Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), :557–586.
The Herald-Dispatch, 2015. Funding to end for Charleston-Huntington commuter bus. Available at http://www.herald-dispatch.com/news/funding-to-end-for-charleston-huntington-commuter-bus/article_cce488b2-1072-5517-8b89-4e967a92d7db.html (Accessed May 22, 2017).

Tiemann, T.K., Miller, P., 2013. Reducing sprawl, riding the bus and losing weight in America. Spaces Flows: Int. J. Urban Extra Urban Stud. 3 (3), :103–113.
Traffic Management Bureau of Beijing, 2011. Beijing small passenger car passenger quota management. Available at http://www.bjhjyd.gov.cn/person/ (Accessed October 24, 2016).
Tucker, P., Gilliland, J., 2007. The effect of season and weather on physical activity: a systematic review. Public Health 121 (12), :909–922.
United States Census Bureau, 2015a. Patterns of Metropolitan and Micropolitan Population Change: 2000 to 2010. Available at http://www.census.gov/population/metro/data/c2010sr-01patterns.html (Accessed October 24, 2016).
United States Census Bureau, 2015b. Poverty Thresholds by Size of Family and Number of Related Children Under 18 Years. Available at https://www.census.gov/hhes/www/poverty/data/threshld/ (Accessed October 24, 2016).
United States Census Bureau, Small Area Estimates Branch, 2002. 2001 Poverty and Median Income Estimates – Counties. Available at https://www.census.gov/did/www/saipe/data/statecounty/data/2001.html (Accessed October 24, 2016).
United States Census Bureau, Small Area Estimates Branch, 2010. 2009 Poverty and Median Income Estimates – Counties. Available at https://www.census.gov/did/www/saipe/data/statecounty/data/2009.html (Accessed October 24, 2016).
United States Department of Transportation, Federal Highway Administration, 2001. 2001 National Household Travel Survey. Available at http://nhts.ornl.gov (Accessed October 24, 2016).
United States Department of Transportation, Federal Highway Administration, 2009. 2009 National Household Travel Survey. Available at http://nhts.ornl.gov (Accessed October 24, 2016).
Yamano, T., 2010. Lecture notes on advanced econometrics. Available at http://www3.grips.ac.jp/~yamanota/Lecture%20Note%208%20to%2010%202SLS%20&%20others.pdf (Accessed July 27, 2018).