Impact of Bicycle Lanes on Bicycle Commuter Rates

Term Paper

submitted by

Shaurya Goyal (20HS20068)

For the subject

Econometrics Analysis 1 Lab

Department of Humanities and Social Sciences
Indian Institute of Technology Kharagpur
Spring Semester, 2021-22

Date of Submission: 06th April 2022

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Introduction and Literature Review:

The paper attempts to isolate the effects of increasing bike lanes and bike paths in U.S cities and establish a causal relationship with bike commuter rates. The application and methodology have been described in detail below. The data sources have also been mentioned in the appropriate sections below.

There are quite a few related works that study the use of bicycles and the facilities available. Jennifer Dill and Theresa Carr have built upon the work done by Nelson and Allen (1997), and in their paper 'Bicycle Commuting and Facilities in Major U.S. Cities: If You Build Them, Commuters Will Use Them'[1], they state that Research was conducted that affirms that cities with higher levels of bicycle infrastructure (lanes and paths) witnessed higher levels of bicycle commuting, by analyzing data from 43 large cities across the United States. Although their analysis has limitations, it does support the assertion that new bicycle lanes in large cities will be used by commuters.

Buehler and Pucher[2] have studied trends and policies with respect to walking and cycling in Western Europe and the United States. However, in studies conducted by both Dill and Carr, and Buehler and Pucher, the results do not imply causation. The direction of causation was not confirmed, but there was a positive correlation between bike facilities and bike commuters.

A study by Krenn, Oja, Titze[3] develops a bike-ability index for a mid-sized European city, to assess the bicycle-friendliness of urban environments. The study used Geographic Information Systems (GIS) data, and assessed the environmental characteristics of 278 bicycle trips in the city of Graz, Austria for the research. Another study by Aultman-Hall, Hall and Bates[4] used GIS to collect systematic information about the routes bicycle commuters choose to take. This implies that GIS data can be useful for further research and studies to augment existing research.

A study by Parkin, Wardman and Page[5] presents a model that relates the proportion of bicycle journeys to work for English and Welsh electoral wardsto certain socio-economic, transport and physical variables. This study uses UK 2001 census data, as opposed to USA-based data used in this study.

Model:

I hypothesise that the percentage of bicycle commuters in a city would depend on a varietyof intuitive factors like the facilities for cycling (length of bike lanes available in the city), perception of safety on roads (number of fatalities per 10000 bikers in a city), median income of the city, population without cars (percentage of population with no car), gas prices, and rain levels (average precipitation in the city).

However, there is a likelihood of endogeneity in this OLS model, which will distort the direction of causation: the miles of bike facilities in a city will also likely depend on the number of cyclists in that city and their collective demand for bike lanes. Since there is a potential for reverse causality, log (blane) might be correlated with the error term; and hence I use the two stage least squares method to first estimate this variable.

In the first stage, the area of the city was chosen as the instrumental variable. This is done by making a logical and intuitive assumption that the area of a city is uncorrelated with the percent of bicycle commuters in a city but it should have a direct impact on the miles of bike lanes in the city.

The first stage of the 2sls method is as follows:

$$log\ (blane) = \delta_1 log\ (area) + \delta_2 \text{safety}^2 + \delta_3 inc + \delta_4 inc^2 + \delta_5 nocar + \delta_6 \log(\text{rain}) + \delta_7 \text{gas}$$

Using this regression to predict values for log(blane) gives the variation in length of bike lanes exogenous to the level of bicycle commuting.

Now, in the second stage, I regress bicycle commuting such that the values used to represent the IV log(blane) are now the exogenous predicted values.

The second stage of the 2sls method is as follows:

$$log\ (com) = \pi_l log\ (\widehat{blane}) + \pi_2 safety^2 + \pi_3 inc + \pi_4 inc^2 + \pi_5\ nocar + \pi_6 log\ (rain) + \pi_7 gas + e$$

Application and Methodology:

In this paper, I have collected and analysed data from 69 cities in the United States of America. These 69 cities consist of the 50 biggest cities and 19 benchmark cities as defined by the League of American Bicyclists in the year 2016. The reason I use American data is multi-fold: first, data related to bicycle commuting and bicycle lanes is extensively collected and easily available for the USA as compared to India, second, I chose the US over other European nations where cycling is much more prevalent because the bicycle commuting trend is still developing in the United States, meaning bike lanes continue to be constructed every year and I are far from saturation, which is not the case for most major European cities, third, I chose the USA over other Asian nations such as India or Indonesia simply because cycling as a mode of commute has not yet emerged in these countries, there are barely any protected bike lanes except for a few in major cities, this would make the sample size far too small. I take city-level data instead of state-level data because bike commuters often cycle within a city and decisions related to the creation of bike lanes take place at the city level at municipalities.

The League of American Bicyclists collected data for multiple variables in this paper. The data for the percentage of bicycle commuters, bicycle fatalities, and the explanatory variable of interest, length of bike lanes per hundred thousand residents in miles were all directly collected from their 2016 report on the status of bicycles in the USA.

The data for the area of the city, median income, and percentage of residents without a car was collected from the American Community Survey 2016 (5-year estimates) and the US Bureau Data for 2015-2016. The data for gas prices were obtained from the United States Energy Information Administration. The data for the average precipitation for each state was obtained from the National Climatic Data Center.

I take a log transformation of certain variables for one or both reasons: firstly, intuitively I think the law of diminishing marginal returns holds for some of these variables. For instance, increasing the length of protected bike lanes may increase the number of bike ridersto an extent, but as I move closer to the saturation level, I shall get diminishing returns for each additional mile, which will yield skewed distributions, hence I take log transformations to normalise the values. Secondly, I have also taken log transform of certain variables in order to counter heteroskedasticity, unless previous theory tells otherwise.

I have also taken the square of certain variables for the following reasons: I take the square of safety because all previous research in this area indicates to the fact that safety is anextremely important part of people choosing to cycle (Morritz 1997; Handy, Xing, and Buehler 2010). Secondly, it also captures the fact that the relationship between fatalities and perception of safety is negatively sloped with decreasing slope, because as fatalities increase people’s fear for their safety will increase non-linearly.

I also include the income and squared median income variables despite the obvious issue of multicollinearity because of intuition and my understanding of previous literature. Firstly, it seems like at low levels of income the propensity to commute by bicycle should be high because cycles are cheaper than cars by a huge margin. However, I see empirically that the average cyclist in the Metropolitans of the United States often has income slightly above the median income (Morritz 1997). These people cycle due to environmental or health reasons, and it is seen that this privilege of caring about principles over comfort is for those that are slightly upper class. Hence, I expect that the number of bikers will increase with the increase in median income till a certain point. But at extremely high-income levels, I observe that most people can afford a luxury car and choose that comfort over other alternatives. What this means is that likely the percentage of commuters increases with income to a point and then falls down again, leading to a non-linear curve. To capture this relationship, I take both income and income squared.

Descriptive statistics for the variables are given in Table 1.

The largest limitation to the model is the small sample size, especially since the 2sls method of regression is generally meant for very large samples, as it tends to increase the variance of variables quite a bit.

The assumption of Larea as an IV also seems feasible since I see that correlation of Larea with Lcom is 0.1 and when I regress area and lcom I get a coefficient of 0.003, which indicates that lcom and area are not correlated.

Lastly, to prove that there exists endogeneity, I first obtain the residuals(zhat) after regressing lblane on the instrumental variable and the other controls. I then regress lcom onlblane and other controls plus the residuals, zhat. The purpose of this is to show that if the errors of bike lanes are correlated to the commuter rates, then there is reverse causality or endogeneity. I can see from Table 2 that the p-value of zhat is significant at 1% level meaning there is endogeneity, which is why I use the 2 stage least squares method.

It is important to note here that this is a potential problem in my analysis. The use of 2SLS is better suited for data with large samples, however since the data only has 69 observations this can lead to misleading results with high variance.

Statistical Tests:

1) Multicollinearity: Based on the model specification, it was clear that there was going to be multicollinearity between income and income squares, but I left that in as theory supported that idea and I needed a way to plot the non-linear relationship between income and biker commuter rates in a city.

   However, to statistically test this hypothesis, I run 2 tests for multicollinearity. Firstly, I just check the pairwise correlations between all explanatory variables and see that the only variables with high correlation coefficients are income and income squared. This can be seen in table 5.

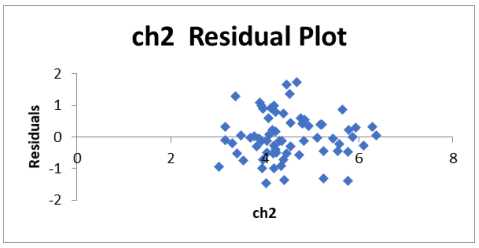
   Furthermore, I also compute the Variance Inflation Factors and further prove my assumption. I see that only inc and incsq have VIFs > 5, which is the usual threshold for multicollinearity.

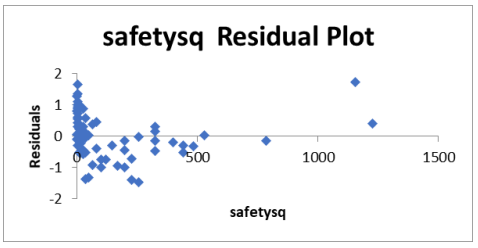
Variable	VIF	1/VIF
inc	242.55	0.004123
incsq	241.59	0.004139
ch2	4.83	0.207198
lrain	3.52	0.284205
gas	2.67	0.374784
nocar	2.32	0.431153
safetysq	1.52	0.657219
Mean VIF	71.29	

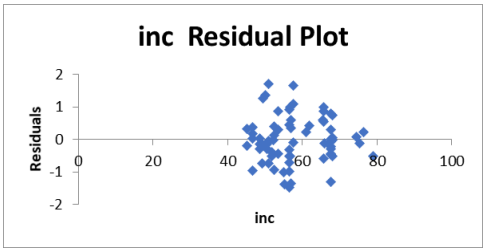
2) Test for heteroskedasticity:

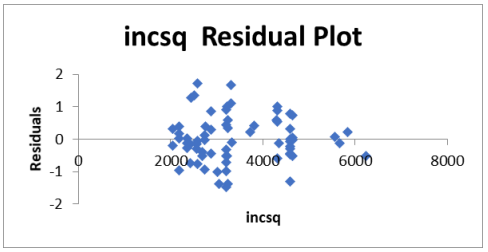
   a) Graphical method:

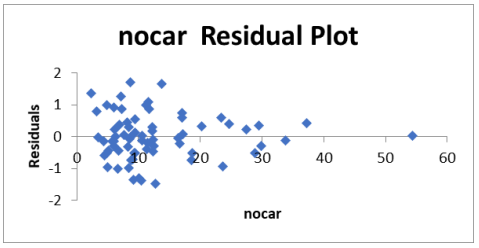
      I plotted the residuals of stage 2 against all the IVs to examine if heteroskedasticity was present.

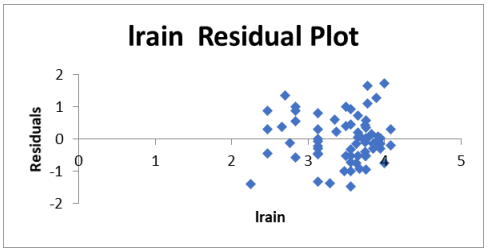


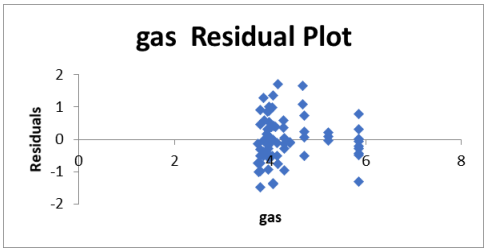












As it can be observed, there seems to be some presence of heteroskedasticity according to the plots, especially in the plots of variables like incsq (income squared) and nocar. To examine this further, I ran more formal tests like

b) Breusch-Pagan/Cook-Weisberg test

The results of the Breusch-Pagan/Cook-Weisberg testare as follows:

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lcom

         chi2(1)      =     0.02
         Prob > chi2  =   0.8984
```

c) White's test

The results of the White's test are as follows

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(34)      =     34.74
         Prob > chi2   =    0.4323
```

d) Cameron and Triverdi's decomposition of IM-test:

The results of Cameron and Triverdi's decomposition of IM-test are as follows:

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	34.74	34	0.4323
Skewness	11.16	7	0.1317
Kurtosis	0.11	1	0.7359
Total	46.02	42	0.3094

From all of these tests it can be concluded that there is presence of heteroskedasticity in the model. I tried reducing its effect by log transforming required variables as much as possible, but as can be seen fromthese tests, some presence of heteroskedasticity still remains. Some possiblemethods to resolve this can be to use a weighted regression method, or to perform the method of bootstrapping.

Analysis of Results:

The results of both Stage 1 and Stage 2 can be seen in table 4 at the end of the paper. I seethat the safety squared variable is not statistically significant (p=0.398), which goes againstthe hypotheses of previous research in this area and indicates that other variables have more explanatory power in deciding the bicycle commuter rates within a city.

I see that all other variables are statistically significant in my model, hence it becomesimportant to analyse the sign and magnitude of the coefficients of these variables.

I see that ch2 (which is the predicted value of log of blane) has a negative coefficient, which goes against my initial hypotheses of how an increase in bike lanes must encourage cyclists. There can be 2 reasons for this, it is possible that the other variables in the model have a much higher positive correlation with the commuter rates than the length of bike lanes,or it is possible that the 2 stage least squares model used increased or change the variance of the log of bike lane variable, hence changing the results in this manner. Before interpreting these results further, I investigate the statistical set-up and look at the possible impact of other variables.

All the other variables seem to have the expected sign of coefficients. I see that an increase in income does increase commuter rates at a coefficient of 0.3 and that it decreases at high levels because I see that income squared has a negative coefficient, which is exactly what my intuition told me.

I see that log rain has a highly negative coefficient (-2.27), indicating that with an increasein average precipitation there's a huge drop in bicycle commuter rates. I see that with an increase in gas prices, bicycle commuter rates increase, due to a positive regression coefficient (1.25), which is in tandem with my logic and intuition that as gas becomes more expensive people will switch to cheaper alternatives such as cycling.

Conclusion and Scope for Further Research:

The simplified/reduced form of the model that I set out to create was of the form:

$$log\ (com) = \pi_l log\ (\widehat{blane}) + \pi_2 safety^2 + \pi_3 inc + \pi_4 inc^2 + \pi_5\ nocar + \pi_6 log\ (rain) + \pi_7 gas + e$$

And the first stage of the 2SLS regression analysis for the total miles of bike lanes and paths(fac) was of the form:

$$log\ (blane) = \delta_1 log\ (area) + \delta_2 \text{safety}^2 + \delta_3 inc + \delta_4 inc^2 + \delta_5 nocar + \delta_6 \text{log (rain)} + \delta_7 \text{gas}$$

My initial hypothesis assumed that increasing the total bike lanes and paths in a city would increase the number of bike commuters. The log(blane) explanatory variable is highly significant for the number of bicycle commuters. The initial conclusion from the regression results of Table 4 is that the total bike lane length in a city would significantly impact the number of cyclists. Although the sign of the coefficient goes against my intuitive hypothesis,this could be impacted by the sample size and other regression factors that might impact the 2SLS method.

Although safety would appear to be a major issue for cyclists, the regression results indicate alow level of significance for this variable. The regression final results for the cyclist fatality rates from Table 4 clearly indicate that safety might not influence a cyclist's decision.
However, this variable is slightly more complex to draw conclusions from compared to theother regression variables used in the 2SLS model. For a moment, consider safety to be of paramount concern in an individual's decision to become a cyclist or not.

Safety is a factor that cannot be brought about into society by a single individual or establishment. It requires concentrated efforts of the government and policy-makers. Similarly, the likelihood of cyclist accidents occurring can be expected to sharply decline if the total miles of bike paths were to increase. Less contact with motorised traffic would drastically decrease the number of accidents that could potentially occur. Thus, safety is deeply connected to the other regression variables, and cleaning the data to incorporate theseparameters would improve the overall accuracy of the model.

Another point is that here safety is taken as a numerical value of accidents. A better and more accurate representation of safety when deciding the impact of bike lanes is perceived safety. What matters is how safe the (potential) commuters think the

bike lane will be. A survey where people are asked if they currently use a bike, their perceived safety on a scale of 1-5, their perceived safety on a scale of 1-5 if bike lanes are made, and would they use a bike if bike lanes were made.

Thus, the significance of factors such as public policy and the miles of bike paths in a city might ultimately explain some part of perceived cyclist safety. One prudent area of research would be to analyse how many accidents are caused by factors that can be entirely avoided byincreasing the distance of bike lanes and paths in the city. Correlating these two explanatory variables (blane and safety) would help improve the reliability of the model by negating the effects of factors that are completely outside the system.

On the other hand, the other explanatory variables appear to be very significant. For instance,owning a car, weather considerations for a city, the price of gasoline, and income are statistically significant at nearly 99% confidence. This can be explained perhaps from the changing economic setting that people live in, especially in the cities within which the surveyand sample data were taken. An improvement in weather-protective outerwear and gear would negate the effect of the number of rainy days in a year. Similarly, high income people may opt to cycle to be health-conscious while low-income individuals would be unable to buy a car.

For my level of exposure to regression models and economic factors that may determine an individual's likelihood to become a cyclist, expanding my sample size to include cities that have managed to implement public policies (in an effort to increase cyclists) would be helpful in drawing comparisons between successful and unsuccessful cities. While the initial findings from this 2SLS regression modelis that public policy is less relevant as compared to perceived commuter safety, a better understanding can be obtained by observing the economic aspects of this problem.

The single most important area for improvement and further research would be to expand the sample size of the study and aim to fully capture the effect of these explanatoryvariables in a 2SLS regression model. At present, it is likely that the data from a small U.S. city might not truly represent a state, let alone a nation. By increasing the sample data size and incorporating cities or nations that have successfully increased the percentage of cyclist commuters through economic and policy decisions, this model stands a chance at being moreaccurate. Increasing the sample size might also bring out new factors that could potentially affect the percentage of cyclist commuters. Increasing the sample size to include more cities and individuals from diverse demographics would also improve the overall model and resultsof the regression model.

Appendix:

Table 1.

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
lcom	69	.108	1.118	-1.623	3.05
com	69	2.136	3.126	.197	21.113
lblane	69	4.493	1.265	1.361	6.608
blane	69	163.4	161.965	3.9	740.8
larea	69	4.825	.953	2.292	6.616
lrain	69	3.469	.446	2.247	4.08
gas	69	4.362	.674	3.75	5.854
nocar	69	12.784	9.239	2.3	54.4
inc	69	58.198	8.537	45.146	78.945
incsq	69	3458.769	1028.953	2038.161	6232.313
safetysq	69	142.522	242.649	0	1225

Test for Endogeneity - Table 2.

```
. reg lcom lblane safetysq inc incsq nocar lrain gas zhat
```

Source	SS	df	MS		
Model	59.8302862	8	7.47878578	Number of obs	= 69
Residual	25.2053636	60	.420089394	F(8, 60)	= 17.80
				Prob > F	= 0.0000
				R-squared	= 0.7036
				Adj R-squared	= 0.6641
Total	85.0356498	68	1.25052426	Root MSE	= .64814

lcom	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lblane	-1.551343	.2161473	-7.18	0.000	-1.983702	-1.118984
safetysq	-.0003901	.0003996	-0.98	0.333	-.0011894	.0004091
inc	.2958312	.1440255	2.05	0.044	.0077374	.5839251
incsq	-.0025928	.0011924	-2.17	0.034	-.0049778	-.0002077
nocar	.0888466	.0129558	6.86	0.000	.0629312	.114762
lrain	-2.274961	.3309542	-6.87	0.000	-2.936968	-1.612954
gas	1.251407	.1905352	6.57	0.000	.8702799	1.632534
zhat	1.924857	.2314266	8.32	0.000	1.461935	2.387779
_cons	.1845352	4.461592	0.04	0.967	-8.739978	9.109048

The residuals (zhats) of the stage 1 regression are statistically significant at 1%levels, there is endogeneity between commuter rates and bike lanes.

Table 3: Stage 1 Results

```
. reg lblane larea safetysq inc incsq nocar lrain gas, noconstant
```

Source	SS	df	MS		
Model	1437.07406	7	205.296294	Number of obs	= 69
Residual	64.3973768	62	1.03866737	F(7, 62)	= 197.65
				Prob > F	= 0.0000
				R-squared	= 0.9571
				Adj R-squared	= 0.9523
Total	1501.47143	69	21.7604556	Root MSE	= 1.0192

lblane	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
larea	.4265973	.1446199	2.95	0.004	.1375063	.7156883
safetysq	.0002784	.0005899	0.47	0.639	-.0009008	.0014575
inc	.1119355	.0528112	2.12	0.038	.0063674	.2175036
incsq	-.00102	.0004851	-2.10	0.040	-.0019897	-.0000503
nocar	.0455702	.0142175	3.21	0.002	.0171499	.0739906
lrain	-1.165249	.2843364	-4.10	0.000	-1.73363	-.5968687
gas	.6577509	.2208616	2.98	0.004	.2162548	1.099247

Table 4: Stage 2 Results

```
. reg lcom ch2 safetysq inc incsq nocar lrain gas, noconstant
```

Source	SS	df	MS		
Model	51.6697517	7	7.38139311	Number of obs	= 69
Residual	34.1742301	62	.55119726	F(7, 62)	= 13.39
				Prob > F	= 0.0000
Total	85.8439819	69	1.24411568	R-squared	= 0.6019
				Adj R-squared	= 0.5570
				Root MSE	= .74243

lcom	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ch2	-1.550706	.2469591	-6.28	0.000	-2.04437	-1.057041
safetysq	-.0003895	.0004574	-0.85	0.398	-.0013038	.0005247
inc	.3014052	.0582063	5.18	0.000	.1850524	.4177579
incsq	-.0026382	.0005287	-4.99	0.000	-.0036952	-.0015813
nocar	.0888421	.0148399	5.99	0.000	.0591776	.1185066
lrain	-2.270575	.3591124	-6.32	0.000	-2.988431	-1.55272
gas	1.251236	.2182004	5.73	0.000	.8150595	1.687412

Table 5 - Multicollinearity (Matrix of correlations)

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) ch2	1.000						
(2) safetysq	0.044	1.000					
(3) inc	0.222	-0.384	1.000				
(4) incsq	0.204	-0.379	0.998	1.000			
(5) nocar	0.279	-0.093	0.077	0.090	1.000		
(6) lrain	-0.580	0.184	-0.258	-0.235	0.270	1.000	
(7) gas	0.507	-0.202	0.551	0.547	0.051	-0.175	1.000

Table 6: Explanation of variable

Variable name	Explanation
com	bicycle commuter rates in the city
Lcom	log of com
Blane	miles of bike lanes in city
Lblane	log of blane
Safetysq	square of number of fatalities per 10000 bikers in city
Inc	median income in thousands of dollars
Incsq	square of inc
Nocar	% of population with no car
Gas	avg regular gas prices
Lrain	log of average precipitation in a city in inches/year

References:

1. Dill, Jennifer, and Theresa Carr. "Bicycle Commuting and Facilities in Major U.S. Cities: If You Build Them, Commuters Will Use Them." Transportation Research Record 1828, no. 1 (January 2003): 116–23. https://doi.org/10.3141/1828-14.

2. Buehler, Ralph & Pucher, John. (2012). Walking and cycling in Western Europe and the United States: Trends, policies, and lessons. TR News. 280. 34-42. Link

3. Krenn, P. , Oja, P. and Titze, S. (2015) Development of a Bikeability Index to Assess the Bicycle-Friendliness of Urban Environments. Open Journal of Civil Engineering, 5, 451-459. doi: 10.4236/ojce.2015.54045.

4. Aultman-Hall, Lisa, Fred L. Hall, and Brian B. Baetz. "Analysis of Bicycle Commuter Routes Using Geographic Information Systems: Implications for Bicycle Planning." Transportation Research Record 1578, no. 1 (January 1997): 102–10. https://doi.org/10.3141/1578-13.

5. Parkin, J., Wardman, M. & Page, M. Estimation of the determinants of bicycle mode share for the journey to work using census data. Transportation 35, 93–109 (2008). https://doi.org/10.1007/s11116-007-9137-5