2017 NHTS Data Challenge: Perceived Health Status and Travel Patterns

Sarah Grajdura, PhD Student, Department of Civil Engineering, UC Davis
Yunwan Zhang, MS Student, Department of Civil Engineering, UC Davis
Saurabh Maheshwari, PhD Student, Department of Civil Engineering, UC Davis
Zhenzi Liang, MS Student, Transportation, Technology, & Policy, UC Davis

Introduction

Transportation and physical health are undoubtedly linked through active transport, air quality, and traffic accidents (Raynault & Christopher, 2013; Mueller et al., 2015). Household location and its proximity to varying types and quality of infrastructure often affect mode choice, lifestyle, and in turn personal health (Sallis et al., 2006; Samimi et al., 2009). However, it is less clear how socio-demographics, travel patterns, and geographic location together affect individuals across the United States. Hence, we ask the question of what travel patterns and socio-demographic attributes impact personal health? To answer this question, we create an ordered logistic (logit) model in which the outcome of interest is self-reported health status, a new question that has been added to the NHTS. The covariates consist of a rich suite of socio-demographic, travel behavior, and geographic variables, many of which are new to the 2017 iteration of the NHTS.

Data

The data is from the 2017 National Household Travel Survey (NHTS), which is the source of the nations information about travel by US residents in all 50 States and the District of Columbia. The survey collects data of travel behavior from nationwide, which is valuable for transportation studies, such as traffic demand modeling, transportation planning and travel pattern analysis.

The original dataset was constructed by merging the 2017 persons and households files by the variable HOUSEID, resulting in 264,234 observations. We then limited the dataset to 85 socio-demographic and travel variables (tabulated in the Appendix) that we expected to impact health, along with HEALTH, our variable of interest. Next, we removed observations containing missing values in any of these variables from the dataset (i.e. those responses coded -9, -7, -77, -8, and -88), resulting in 199,489 remaining observations.

The 2017 NHTS User Guide indicates that variables with Appropriate Skip (coded as -1) are selected by the system when the question was appropriately skipped and no value exists. In order to avoid the removal of these observations from the dataset, we consulted the NHTS Main Study Retrieval Questionnaire to better understand which variables had been appropriately skipped and by whom. In many cases we recoded the Appropriate Skip values to 0; for example, WALK4EX was only asked if the person had taken a walk trip in the past 7 days, so if the person had not taken a walk trip in this time period, then we could assume that same person had not taken any walk trips specifically for exercise either.

After data cleaning, we conducted variable selection by using the following three methods:
1. Intuitive Literature-based selection: Independent variables are selected by importance in the literature of their relationships with health. For instance, count of walk trips will definitely influence the health condition. So, the variable NWALKTRP is chosen. In this way, 33 independent variables are selected.
2. Correlation analysis: Correlation analysis is implemented to find variables that are highly correlated. For two variables with high correlation, the one has lower correlation with HEALTH will be removed. The number of variables was reduced from 33 to 26 with the threshold set to 0.5.
3. Machine learning feature selection: To avoid omission of some important variables that are not obviously related to health, the machine learning model is applied to supplement the variable selection. According to the rank of the importance of the variables, 5 variables (*LPACT, MSASIZE, LK2SAVE, CDIVMSAR, and PLACE*) are added into the variable list. The principles of machine learning feature selection are introduced in the Methods section below.

Methods

In order to answer the research question, we create an ordered logit model with the dependent variable being health status. The ordered logistic model is an ordinal regression model for ordinal dependent variablesfirst considered by Peter McCullagh. It is suitable for our study because the dependent variable ($HEALTH$) is answered by a choice among poor, fair, good, very good and excellent, which can be thought of five ordered response categories. In order to ensure that the most important variables were selected for the ordered logit model, machine learning methods were next used on pre-selected variables.

Random Forest models are one of the famous machine learning models in terms of accuracy, robustness and ease of use. They also provide straightforward and efficient methodology for feature selection and most of the random forest libraries cater for this functionality. Random forest consists of average of many decision trees, each node of which splits the dataset into two parts with respect to a single feature, such that the similar response values end up in the same set. The feature whose split creates the most reduction in Gini impurity or Entropy is chosen at that node. Thus, more important features will be chosen for split before the less important once. To take into account the correlation between variables, another metrics called mean decrease accuracy in calculated. Higher the mean decrease accuracy, higher the feature importance. The feature importance of the top 30 variables can be seen in the Appendix.

Once variable selection was complete, we selected our final model:

$$H_{ik} = \beta_{0k} + \beta_{1k} X_i + \epsilon_{ik}$$

In the model above, H_{ik} is the outcome $HEALTH$, X_i is a vector of independent variables, and the subscript k refers to the health status groups, and i refers to an individual. The final ordered logistic model contains one dependent variable (HEALTH), and 31 independent variables (*R_AGE, CAR*, $BIKE$, $TAXI$, *TRAIN, WALK, URBRUR, R_SEX, R_HISP, CARSHARE, BIKESHARE, BUS, HBPPOPDN, PTUSED, NBIKETRP, NWALKTRP, EDUC, TIMETOWK, DELIVER, WALK4EX, YOUNGCHILD, R_AVGHHINC, HHVEHCNT, CNTTDHH, WRKCOUNT, PHYACT, LPACT, MSASIZE, WLK2SAVE, CDIVMSAR*, and $PLACE$). For the variable $CDIVMSAR$, we constructed 34 dummy variables representing each factor value. Dummies with less than 1,000 observations were dropped from the final model in order to avoid rank deficiency, which can originate from too little data.

Results

The results of the ordered logit model are shown in Table 2, and it indicates a good use of the aforementioned variables. Each coefficient of the above variables is statistically significant. Thus, its believed that these coefficients could be used to account for the prediction of health status across the population (199,489 observations).

Dependent Variable The dependent variable in the model is $HEALTH$ (self-reported health status), which is ordered from good to bad by 1 to 5. This order follows the code of health status in the codebook. Therefore, the higher the value of the dependent variable, the worse health status should be. In other words, a negative sign of coefficient indicates a probability of better health condition when the corresponding independent variable has an increase.

Mode-use Independent variables related to travel modes include $BIKE$, $TAXI$, BUS, etc. They represent the frequency of using certain travel mode by each respondent. These variables are categorical, and we treat them as ordered factors. Therefore, each mode variable in the model is transformed to 4 binary variables with one category as the base/reference. For example, the variable BUS (frequency of bus use for travel) has five categories, which in the model correspond to $BUS.L$, $BUS.Q$, $BUS.C$, BUS^4, and

reference level respectively. The coefficient tells us, for those whose $BUS.L$ is 1 (take bus a few times weekly) and all other variables hold constant, they are 0.225 times less likely to be in a higher category of health status, which also means those people tend to be healthier. Meanwhile, this probability decreases when bus use frequency decreases, from $BUS.L$ to BUS^4. This tendency holds in other transit use and human-powered modes but gets opposite sign for frequent car use. New modes such as carshare and bikeshare give little contribution to health status change, telling by their coefficients. This is interesting since one would think a mode like bikeshare which is good for health would ha ve a positive effect.

Socio-demographic Some of the variables that represent the socio-demographic status are numeric type, including R_AGE, $LPACT$ (times of light physical activities), etc. With one unit increase of $LPACT$, holding all other variables constant, a person is 0.049 times more likely to have a better health condition into the lower category. However, R_AGE, although its sign is positive which is in the direction of worse health condition, we believe it can be negatively related with the dependent variable in certain age ranges (e.g. age between 12-18). The model shows that females ($R\_SEX = 1$) are healthier than males, more children lead to higher possibility of good health, and wealthy of household ($AVGHHINC$) are correlated with better health status. Another interesting finding is the effect of racial variable. It is shown that Hispanic ($R\_HISP = 1$) tends to be healthier.

Geographic It is surprising to find the relationship between geographical difference and health status. For example, by looking at the set of dummy variables derived from $CDIVMSAR$, living in New England ($.data_1 3 = 1$) MSA or CMSA of 1 million or more population without heavy rail is 0.261 times more likely to bring a better health condition. This is plausible because transportation related activity can be varied across the land. Even urban and rural areas ($URBRUR$) have different direction of inclinations.

Miscellaneous People who dont think travel is a financial burden ($PLACE$) may have better health. More trips mean healthier. And frequent delivery service is in line with bad health status. Surprisingly, more vehicles in households ($HHVEHCNT$) could be a symbol for good health. This might be related to household income.

Conclusion

Our model shows that in general that the more frequent transportation activities, the better the health status. Travel, as an energy-exhausting activity, can bring people a better health condition. However, not all the modes are good for peoples health. Some socio-demographic data in the model indicates racial and gender-oriented differences often have impacts on health conditions, which might raise equity problems in transportation planning. The model tells us that across the nation, average health status is not constant in different areas.

*Future research points*Our study not only provides many meaningful findings but also raises some research questions for the future. Should we keep promoting bike share programs? As shown in the model, bikeshare use seems to not to improve peoples health status, which is quite opposite to popular belief.

Among all the respondents, health status varies by gender and race, and the model reflects this phenomenon statistically. To understand these differences, detailed surveys focus on specific group of people are helpful to study the transportation impact on peoples health condition. Besides, since our results show that health improvement is related to frequent transportation activities, its good to see whether equity problems in public health can be addressed by transportation.

References

Frank, L. D., Sallis, J. F., Conway, T. L., Chapman, J. E., Saelens, B. E., & Bachman, W. (2006). Many Pathways from Land Use to Health: Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality. *Journal of the American Planning Association, 72*(1), 75-87. doi:10.1080/01944360608976725

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

McCullagh, Peter (1980). "Regression Models for Ordinal Data". *Journal of the Royal Statistical Society. Series B (Methodological). 42* (2): 109142. JSTOR 2984952.

Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., Nazelle, A. D., Dons, E., Gerike, R., . . . Nieuwenhuijsen, M. (2015). Health impact assessment of active transportation: A systematic review. *Preventive Medicine, 76*, 103-114. doi:10.1016/j.ypmed.2015.04.010

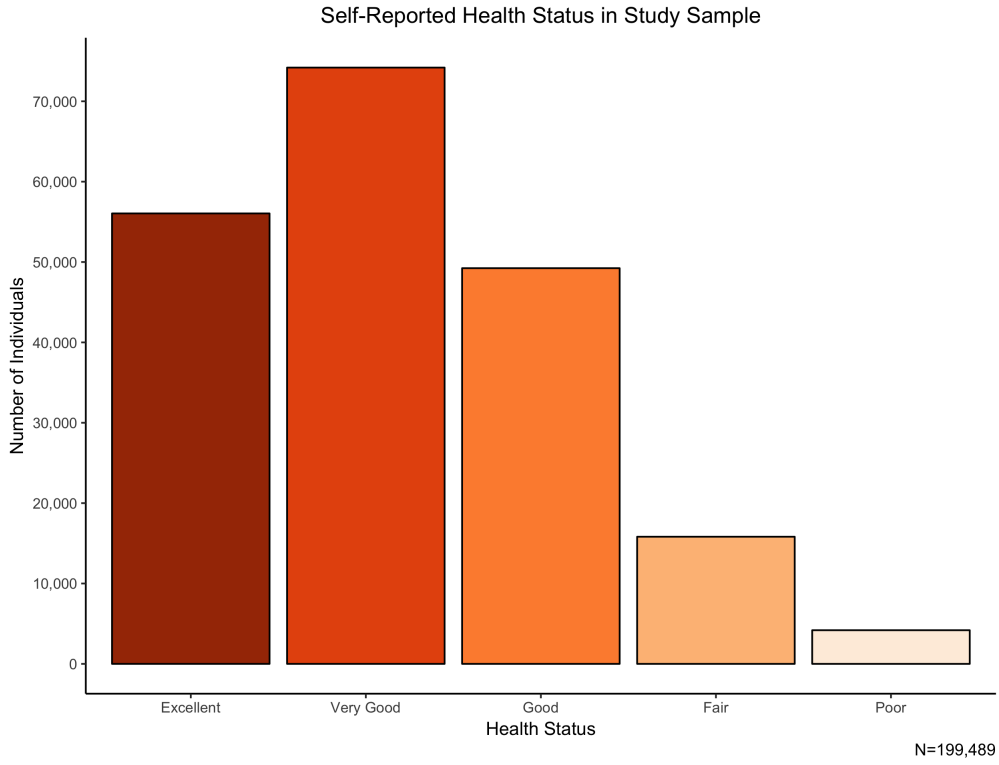
Raynault, E., & Christopher, E. (2013, May). How Does Transportation Affect Public Health? *Public Roads, 76*(6).

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

Samimi, A., Mohammadian, A., & Madanizadeh, S. (2009). Effects of transportation and built environment on general health and obesity. *Transportation Research Part D: Transport and Environment, 14*(1), 67-71. doi:10.1016/j.trd.2008.10.002

Table 1: **Variables Used in Analysis**

NHTS Variable	Definition
ALT_16	Alt. Mode of Transportation: Public Transport of Taxi
ALT_23	Alt. Mode of Transportation: Passenger to Friend/Family Member or Rental Car
ALT_45	Alt. Mode of Transportation: Bike or Walk
BIKE	Frequency of bicylce use for travel
BIKE_DFR	Reason for Not biking More: Infrastructure
BIKE_GKP	Reason for Not biking more: safety
BIKE4EX	Count of Bike Trips for Exercise
BIKESHARE	Count of Bike Share Program Usage
BORNINUS	Born in USA
BUS	Frequency of bus use for travel
CAR	Frequency of Personal Vehicle Use for Travel
CARSHARE	Count of Car Share Program usage
CDIVMSAR	Census division/MSA status/Presence of a subway system with pop. >1 million
CNTTDHH	Count of HH trips on travel day
CNTTDTR	Count of person trips on travel day
CONDNIGH	Medical condition- driving limited to daytime
CONDPUB	Medical condition- results in using bus/subway less frequently
CONDRIDE	Medical condition- results in asking others for rides
CONDRIVE	Medical condition- results in giving up driving
CONDSPEC	Medical condition- results in using special transportation services
CONDTAX	Medical condition- results in using a reduced fare taxi
CONDTRAV	Medical condition- results in reduced day-to-day travel
DELIVER	Count of Times Purchased Online for Delivery in Last 30 Days
DRVRCNT	Number of drivers in HH
EDUC	Education Attainment
GT1JBLWK	More than 1 job
HBHTNRNT	Percent renter occupied housing in census block group of HH
HBPPOPDN	Pop. density of census block of HH
HEALTH	Opinion of Health
HH_HISP	Hispanic status of HH
HH_RACE	Race of HH respondent
HHFAMINC	HH Family Income
HHSIZE	Count of HH members
HHSTATE	HH state
HHVEHCNT	Count of HH vehicles
HOMEOWN	Home Ownership
HTEEMPDN	Workers per mile in the census tract of HH
LIF_CYC	Life cycle classification for HH based on age and kids
LPACT	Count of times of light or moderate PA in past week
MCUSED	Count of moped/motorcycle trips
MEDCOND	Medical condition
MSACAT	Metro Statistical Area (MSA) for HH, including rail
MSASIZE	Population category of MSA
NBIKETRP	Count of bike trips
NUMADLT	Count of adult HH members over 18 years
NWALKTRP	Count of walk trips
OCCAT	Job category
PC	Frequency of accessing Internet with Laptop or Desktop
PHYACT	Level of Physical Activity
PLACE	Travel is a financial burden
PRICE	Price of gas affects travel
PRMACT	Primary Activity in Previous Week
PTRANS	Pub Trans to reduce Financial Burden
PTUSED	Count of public transit usage
R_AGE	Age of person
R_HISP	Hispanic or LatinX
R_RACE	Race
R_SEX	Gender
RESP_CNT	Count of responding persons in HH
SPHONE	Frequency of Smartphone to Access Internet
TAXI	Frequency of Taxi service or Rideshare for Travel
TIMETOWK	Trip Time to Work in Minutes
TRAIN	Frequency of Train Use for Travel
URBRUR	HH in Urban or Rural area
VPACT	Count of Times of Vigorous Physical Activity in Past Week
W_CANE	Medical device used: Cane
W_CHAIR	Medical device used: Wheelchair
W_CRUTCH	Medical device used: Crutches
W_DOG	Medical device used: Dog assistance
W_MTRCHR	Medical device used: Motorized wheelchair
W_NONE	Medical device used: None
W_SCOOTER	Medical device used: Motorized Scooter
W_WHCANE	Medical device used: White cane
W_WLKR	Medical device used: Walkwer
WALK	Frequency of walking for travel
WALK_DEF	Reason for Not Walking More: Infrastructure
WALK_GKQ	Reason for not walking more: Safety
WALK2SAVE	Walk to reduce financial burden of travel
WALK4EX	Count of walk trips for exercise
WEBUSE17	Frequency of internet use
WKFTPT	Full time or part time worker
WORKER	Worker status
WRK_HOME	Work from home
WRKCOUNT	Number of workers in HH
YOUNGCHILD	Count of persons age 0-4 in HH


Self-Reported Health Status in Study Sample
Number of Individuals
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Excellent
Very Good
Good
Fair
Poor
Health Status
N=199,489

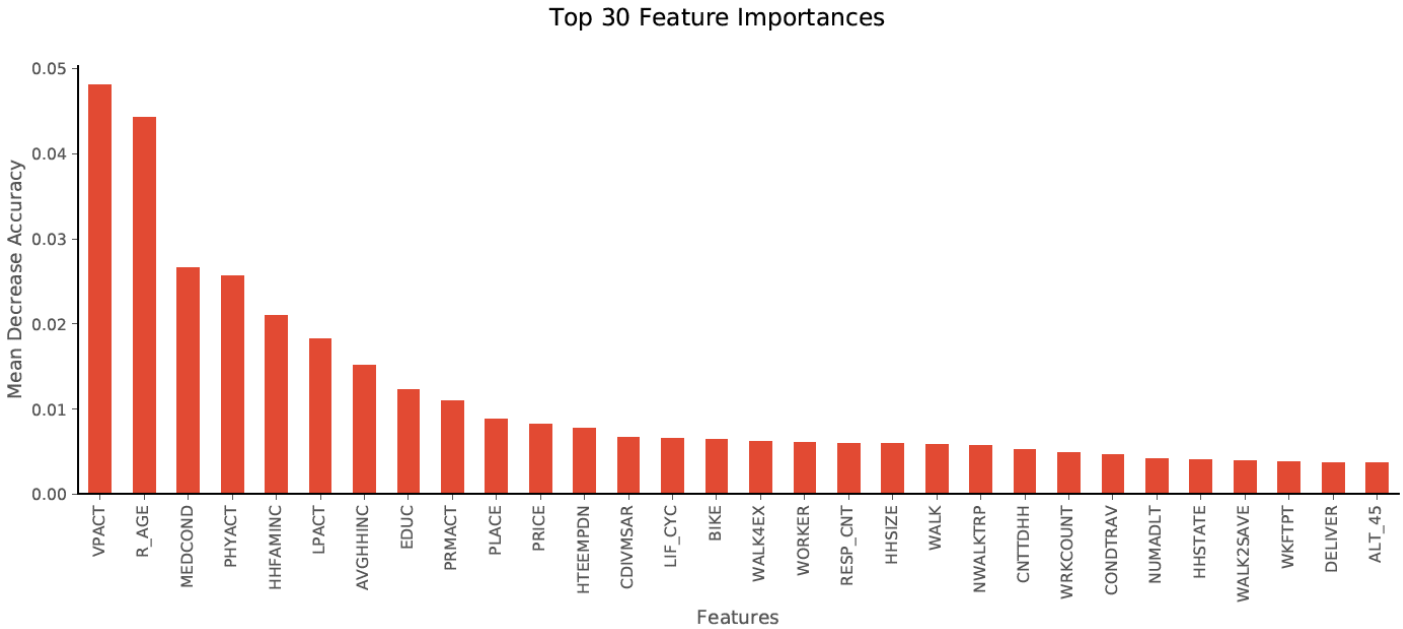

Top 30 Feature Importances
Mean Decrease Accuracy
0.05
0.04
0.03
0.02
0.01
0.00
VPACT
R_AGE
MEDCOND
PHYACT
HHFAMINC
LPACT
AVGHHINC
EDUC
PRMACT
PLACE
PRICE
HTEEMPDN
CDIVMSAR
LIF_CYC
BIKE
WALK4EX
WORKER
RESP_CNT
HHSIZE
WALK
NWALKTRP
CNTTDHH
WRKCOUNT
CONDTRAV
NUMADLT
HHSTATE
WALK2SAVE
WKFTPT
DELIVER
ALT_45
Features

Table 2: Ordered Logit Model Results: Predicting Health Status

	Dependent variable:
	HEALTH
BIKE.L	0.088^{***}
	(0.003)
BIKE.Q	0.095^{***}
	(0.005)
BIKE.C	-0.059^{***}
	(0.005)
BIKE^4	0.049^{***}
	(0.003)
TAXI.L	-0.062^{***}
	(0.002)
TAXI.Q	-0.010^{**}
	(0.004)
TAXI.C	0.058^{***}
	(0.005)
TAXI^4	-0.011^{***}
	(0.003)
TRAIN.L	0.152^{***}
	(0.002)
TRAIN.Q	0.119^{***}
	(0.005)
TRAIN.C	-0.048^{***}
	(0.006)
TRAIN^4	0.007^{**}
	(0.004)
BUS.L	-0.225^{***}
	(0.003)
BUS.Q	-0.121^{***}
	(0.005)
BUS.C	-0.045^{***}
	(0.006)
BUS^4	0.060^{***}
	(0.003)
WALK.L	0.126^{***}
	(0.005)
WALK.Q	-0.083^{***}
	(0.008)
WALK.C	0.006
	(0.009)
WALK^4	-0.005
	(0.010)
CAR.L	0.337^{***}
	(0.003)
CAR.Q	-0.061^{***}
	(0.005)
CAR.C	-0.025^{***}
	(0.006)
CAR^4	0.042^{***}
	(0.003)
EDUC.L	-0.215^{***}
	(0.006)
EDUC.Q	-0.378^{***}
	(0.008)
EDUC.C	0.260^{***}
	(0.009)
EDUC^4	-0.030^{***}
	(0.009)
PHYACT.L	-1.747^{***}
	(0.004)
PHYACT.Q	-0.151^{***}
	(0.009)
NWALKTRP	0.001^{*}
	(0.001)
R_AGE	0.031^{***}
	(0.0003)

DELIVER	0.003^{***}
	(0.001)
R_HISP	-0.025^{***}
	(0.005)
PTUSE	-0.003^{***}
	(0.001)
R_SEX	-0.179^{***}
	(0.008)
TIMETOWK	0.0004^{*}
	(0.0002)
NBIKETRP	-0.025^{***}
	(0.004)
CARSHARE	-0.011^{***}
	(0.0003)
BIKESHARE	0.002
	(0.005)
CNTTDHH	-0.012^{***}
	(0.001)
HBPPOPDN	0.00001^{***}
	(0.00000)
HHVEHCNT	-0.090^{***}
	(0.004)
URBRUR	0.003
	(0.005)
WRKCOUNT	-0.099^{***}
	(0.006)
YOUNGCHILD	-0.181^{***}
	(0.005)
CNTTDTR	-0.013^{***}
	(0.002)
PLACE.L	-0.475^{***}
	(0.003)
PLACE.Q	-0.009^{***}
	(0.003)
PLACE.C	-0.013
	(0.008)
PLACE^4	-0.019^{***}
	(0.007)
LPACT	-0.049^{***}
	(0.002)
WALK2SAVE.L	-0.049^{***}
	(0.004)
WALK2SAVE.Q	-0.093^{***}
	(0.006)
WALK2SAVE.C	0.025^{***}
	(0.006)
WALK2SAVE^4	-0.037^{***}
	(0.008)
AVGHHINC	-0.166^{***}
	(0.003)
MSASIZE.L	-0.124^{***}
	(0.008)
MSASIZE.Q	0.128^{***}
	(0.006)
MSASIZE.C	0.174^{***}
	(0.007)
MSASIZE^4	0.079^{***}
	(0.005)
MSASIZE^5	-0.035^{***}
	(0.008)
Observations	199,489
Note:	*p<0.1; **p<0.05; ***p<0.01