Abstract

Daily Travel time Budget (TTB) is an important element in Activity Based Travel Demand Modeling approach. Since Zahavi [1, 2] and his colleagues showed in the 80's that the average TTB is fixed across different time and geographic context, the debate over fixed TTB have been of many. TTB is claimed to be constant over most aggregate stage, but at disaggregate personal or household level, the TTB is variable, as mentioned in earlier researches. In modern Travel Demand Modeling, this variation can be utilized used to obtain travel pattern. Thus in this work, the variability of TTB for a person over a day has been explored and modeled.

Keywords: Travel Time Budget (TTB), Travel Demand Modeling, Activity Based Travel Demand Modeling.

Contents

Section 1. Introduction

Since Zahavi [1,2] tried to make the travel demand modeling process easier by finding that individual daily travel time is constant, the concept of TTB has been debated on. Thus [3] says, "travel time expenditures are not constant except, perhaps, at the most aggregate level". Since it is not constant at the person or household level, then it might be explained by the factors that influence it. [3] also mentions that with modern development of computing capacity small variations in person level can be harnessed and thus, instead of simplifying travel demand modeling, more rigorous modeling approach can be undertaken on the wake of Activity-Based Modeling approach. Thus modeling the daily travel time or individual trip time can enhance the understanding of the travel demand related behaviors. Thus this model may not be directly related to traditional four-step travel demand model, but can be used to explain the behavioral side of the travelers. In this project a suitable model is to be developed for the adult respondents of New England region. The project uses a linear regression approach to capture the data variability.

In [4], the change in individual travel time based on change in travel mode utility is discussed. But for modeling approach, [2] used a linear regression technique. Moreover, [5] discussed the various literatures on travel time modeling. In [6], set of log-linear equations were developed for modeling activity trip duration based on the order of the trip, j. The most relevant work that would match with this project is described by [7], where a OLS based Linear Regression Model is developed to model each activity type and frequency of that activity along the day. Given the relevant previous modeling framework, this project develops a similar MLR model to predict daily travel time for an individual person.

In this report,

- The second section provides brief description of the data

- Third section discusses about the methodology and the theoretical framework in brief
- Fourth section analyzes the data and shows the result of the model building process
- Fifth section provides the concluding remarks regarding the outcome of this project

Section 2. Description of Data

This research approach uses the data source of the National Household Travel Survey (NHTS). This dataset is the household travel survey information for the USA in 2009. From this dataset, the total daily travel time for each person is derived from the Day Trip file by aggregating the travel time for each person. This project considers the portion of the dataset that belongs to the adult population of the New England region, i.e. population with minimum age of 18years. It considers the data for all seven days of the week. The dataset for only those who traveled on the survey day are selected for building the model. There are nearly 4800 person cases in the dataset.

The variables are enlisted below. CNTTDTR or trip number is related to the duration of travel, more trips mean more time required for travel. Worker status and occupation are used individually and as an interaction variable to capture the notion that workers have variation in daily travel time given their fixed daily commute. Age of the travelers may affect their traveling attitude so this is captured in the dummy variable where middle aged people are taken as baseline. Since vehicle affects freedom of mobility, dummy variables are tested for possible effect on the model. Similarly, income can induce traveling behavior and thus included for modeling. If a

respondent is a driver, then her travel pattern may vary from those who do not drive. So this variable is included. Moreover, urban and rural household members may have different traveling pattern and different daily travel time.

As almost each of the variables is of categorical in nature, the detail explanation of these variables are repetitive and can be easily understandable. The CNTTDTR has a range of 1 to 25 (number of trips made on the day). Since transformations of variables are important, it is important to know that SMTRVLTIME has a range of 1 upto 841. For interaction variable DIST*WRKR, the lower limit is 0 (for non-workers) and for workers the lower limit is 0. This value indicates that for some workers the distance to work is below 1 mile, thus they report the value of DIST*WRKR as 0.

Table: Variables Considered

Variable	Description
SMTRVLTIME	Total time of travel for the respondent on the assigned travel day
CNTTDR	Count of trips for the respondent on the assigned travel day
WRKR	1 if respondent is a worker, 0 otherwise
DIST*WRKR	Distance to work if respondent is a worker
OLD	1 if respondent is above 60 years of age, 0 otherwise
OLD	1 if respondent is above 60 years of age, 0 otherwise
YOUNG	1 if respondent is 18 to 40 years of age, 0 otherwise
WRKTIMEFLEX	1 if respondent can fix or set his work time given respondent is a worker, 0 otherwise
URB	1 if household of the respondent is in urban area, 0 otherwise (As urban and rural area defined in NHTS 2009)
WEEKEND	1 if weekend (Saturday and Sunday), 0 if not
HIGHINC	1 if household income for the respondent is $70k or above, 0 otherwise
MEDINC	1 if household income for the respondent is $30 to below $70k 0 otherwise
DRVR	1 if respondent if a driver, 0 if not
WRKR_SALES	1 if occupation is sales and service, given respondent is worker, 0 otherwise

WRKR_ADMIN	1 if occupation is clerical and administrative, given respondent is worker, 0 otherwise
WRKR_MFGCONST	1 if occupation is manufacturing, construction, maintenance or farming, given respondent is worker 0 otherwise
WRKR_PROTECH	1 if occupation is professional, managerial and technical, given respondent is worker, 0 otherwise
ONEVEH	1 if household has only one vehicle, 0 otherwise
TWOVEH	1 if household has two vehicles, 0 otherwise
THREEVEH	1 if household has three or more vehicles, 0 otherwise

Section 3. Methodology

The whole data set is split into two, the larger with 65% data is used or building the model and the other 35% of the random data is used for making hold-out sample prediction. Multiple Linear Model approach is used in this investigation. The variable to predict is the daily travel time for each person. This model can be expressed as

$$\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where Y is the total daily travel time for the respondent. The covariates are the socio-economic and demographic characteristics of the respondent and the information related to the trips he makes.

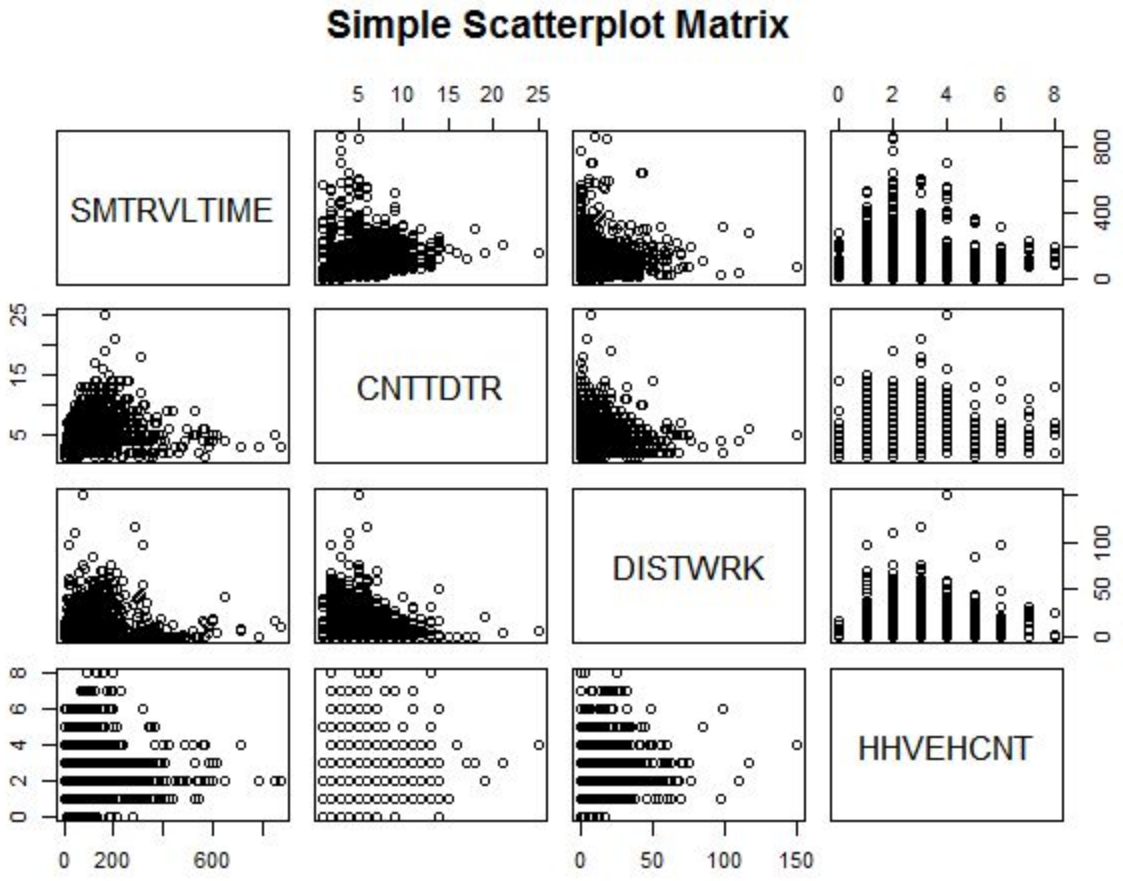
Here $\mathbf{Y}$ is a (nx1) column vector and $\mathbf{X}$ is a (nxp) vector and $\boldsymbol{\beta}$ is a (nx1) column vector which consists of the coefficients corresponding to the covariates in $\mathbf{X}$. Here $\boldsymbol{\varepsilon}$ is a (nx1) vector that is unknown and defines the distribution of Y_i, where i = 1,2,...n. In MLR framework, mean value of Y_i are predicted by $\mathbf{X}'\boldsymbol{\beta}$. The assumption of MLR is applicable here.

The data is first explored for possible covariates which has meaningful relationship with the response variable and then further explored for linear relationship. After the exploration of linearity and association then variables proper were selected using variable selection techniques, i.e. AIC, BIC, R^2 and Adjusted-R^2. For the large dataset, mallows Cp, PRESS and SSE were not used for simplicity and also due to large number of variables to be considered. After selection of variables then the MLR

model was built and analyzed. The diagnostic framework includes Externally Studentized Residual Plots, excluding High Leverage points. Thus influential outliers are found and discarded from the data set. The model is also checked for violation of MLR assumptions with Residual plots. Relevant Generalized Least Square method and Box-Cox Transformation were employed to check suitability of these methods. The model is also analyzed for its goodness-of-fit. Parameters and model were hypothesis-tested. Further deeper into the work, it also considers the Weighted Least Squares method. The variance is only transformed and the rest of the model is thus obtained through the variance stabilizing effect of WLS. The necessary application of the theory is explained at the corresponding subsection. For the hold-out sample prediction, the necessary equations are explained at the corresponding subsection.

Section 4. Data analysis

The variables have only a few continuous ones among them. So first the simple scatterplot matrix was constructed to find linear association between the variables.



The scatterplot matrix shows that CNTTDTR and DISTWRK have linear association with SMTRVLTIME while HHVEHCNT (vehicle count in the household) do not have linear association. So HHVEHCNT was converted into dummy variables as ONEVEH, TWOVEH and THREEVEH. Further. The Normal Q-Q plot shows that SMTRVLTIME is not linearly distributed, so the best transformation is obtained by Natural Log-Transformation which also has better Pearson's and Spearman coefficient. After this transformation, normality is followed in a better manner.

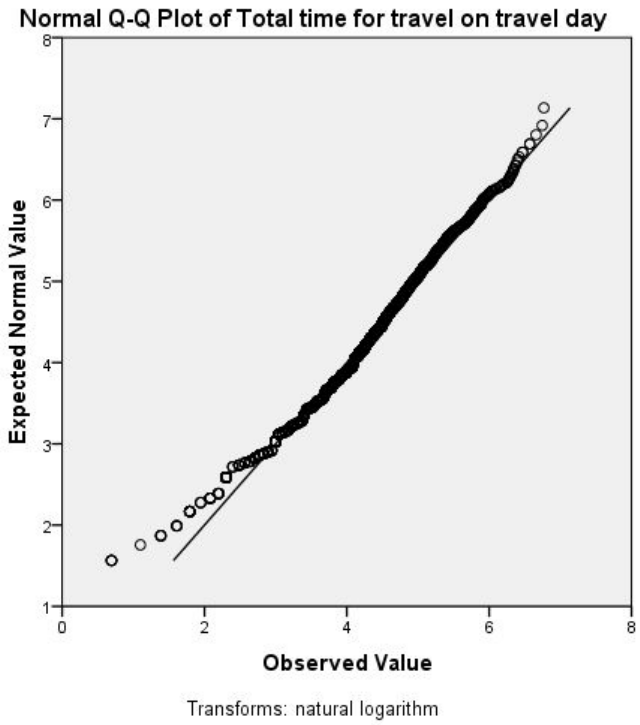


Table: Correlation Matrix

	Log(SMTRVLTIME)	DISTWRK	CNTTDTR	HHVEHCNT
Log(SMTRVLTIME)	1.000 .	.149** .000 .182** .000	.507** .000 .439** .000	.088** .000 .101** .000
DISTWRK		1.000 .	-.015 .288 -.032* .027	.232** .000 .123** .000
CNTTDTR			1.000 .	-.012 .421 -.026 .685
HHVEHCNT		*Pearson's Coeff* *Spearman's Coeff*		1.000 .

Variable Selection

Variables are selected based on their expected influence and their statistical significance. Trial-and-error process for selecting variables ensures important variables are not left out. So instead of automatic selection, variables are ordered based on their importance and then entered into the model. The outcomes are shown in the table below, the insignificant variables are marked with parentheses

Table: Preliminary Variable Selection

Explanatory Variables	AIC	BIC	R^2	Adj-R^2
CNTTDTR + DISTWRK	26274.04	26292.05	0.09667	0.09607
CNTTDR + DISTWRK + HIGHINC+MEDINC	26261.63	26291.66	0.1016	0.1004
CNTTDR + DISTWRK + HIGHINC+MEDINC + URB	26251.01	26287.04	0.1054	0.1039
CNTTDR + DISTWRK + HIGHINC + MEDINC + URB + Weekend	26241.87	26283.91	0.1087	0.1069
CNTTDR + DISTWRK + HIGHINC + MEDINC + URB + Weekend +MALE	26241.87	26283.91	0.1087	0.1069
CNTTDR + DISTWRK + HIGHINC + MEDINC +URB + Weekend + MALE + (WRKTIMEFLEX)	26241.87	26283.91	0.1087	0.1069
CNTTDR + DISTWRK + HIGHINC + MEDINC +URB + Weekend + MALE + (WRKTIMEFLEX) + (DRVR)	26241.87	26283.91	0.1087	0.1069
CNTTDTR + DISTWRK + HIGHINC +MEDINC + URB + Weekend + MALE + (WRKR)	26238.62	26292.68	0.1108	0.1085
CNTTDTR + DISTWRK + HIGHINC +MEDINC + URB + Weekend + MALE + (WRKR) + THREEVEH	26235.36	26295.42	0.1124	0.1097
NTTDR + DISTWRK + HIGHINC + MEDINC+ URB + Weekend+ (WRKR_SALES)+ (WRKR_ADMIN)+ (WRKR_MFGCONST)+ (WRKR_PROTECH)	26241.87	26283.91	0.1087	0.1069

Based on the variables obtained in the previous step, the MLR is developed. The initial was tested for constancy of error variance. The **Breusch-Pagan test** showed

that the variances are not constant as the p-value is nearly zero and thus null hypothesis of variance constancy is rejected.

Variance Stabilizing Transformation

a. Box-Cox Transformation

First Box-Cox transformation was used to obtain this model. The requirement that all responses are positive is satisfied as the minimum range for SMTRVLTIME is 1. The recommended λ value is positive fraction so the following equation was considered.

$$(\hat{SMTRVLTIME}^{Coefficient} - 1)/coefficient \sim CNTTDTR + URB + DISTWRK + MEDINC + MALE +$$

.... (Eq1)

Using Externally Studentized Residual plot, the figure on the next page shows that, the variances are not constant enough. With increasing predicted values, variances decrease. This systematic relationship is thus addressed by the Weighted Least Square model.

b. Weighted Least Square (WLS)

The next approach is to use Generalized Least Square method to stabilize the variances of the previous model. With trial and error, the natural log transformed response variable was found to make best display of linearity and homoscedasticity assumption. Also, in this model, the coefficients can be explained more easily. In travel demand behavioral practice, interpretation of the parameters is important to understand the mechanism of the model. In Box-Cox transformation the marginal effects are comparatively difficult to interpret. So log transformed response with simpler WLS model is decided on. In WLS, these variances are thus weighted based on their expected variance.

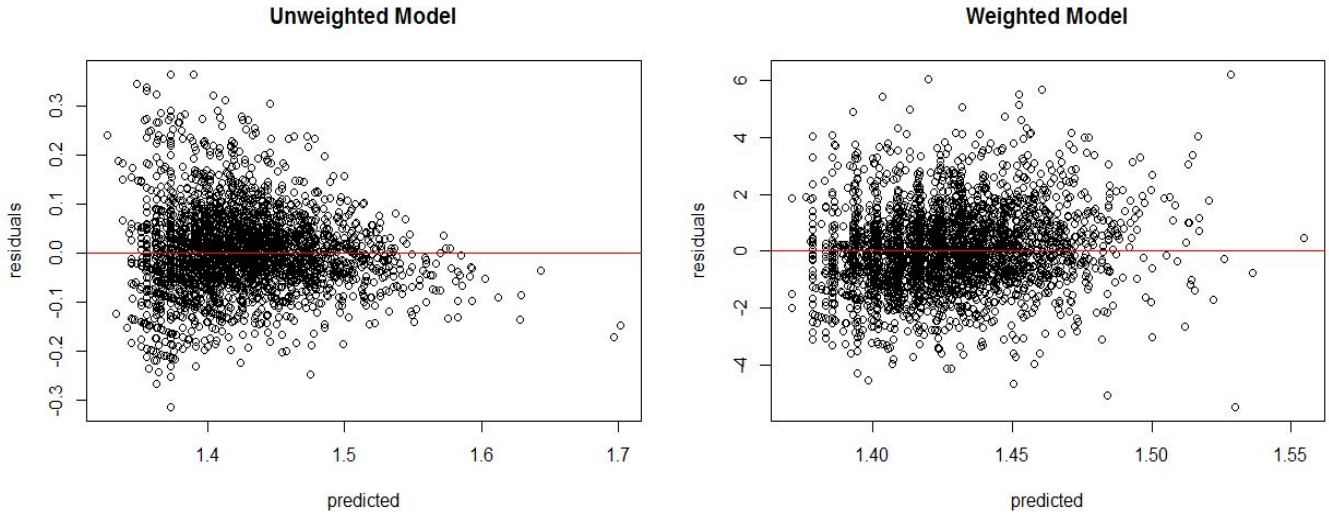
Here, the error variance is not $\sigma^2\mathbf{I}_n$, rather $\mathbf{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}^{-1}\sigma^2$

In the model, the error variance terms are assumed different from constant. The weight choice was based on the variances found by the estimation process.

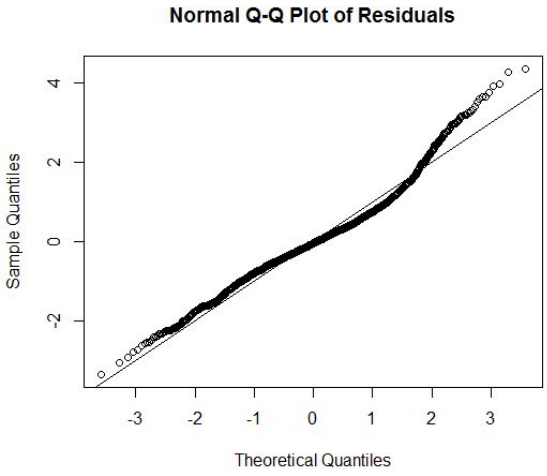
The figure below shows that the Box-Cox transformed models do not have homoscedastically distributed residuals. To tackle this problem, complex models

can be used, but WLS is simple yet capable enough to handle it. The WLS model is shown in later part of the project

Figure: (Left) Box-Cox Transformed Model (Eq1)
(Right) Weighted Least Square Model (Eq2)



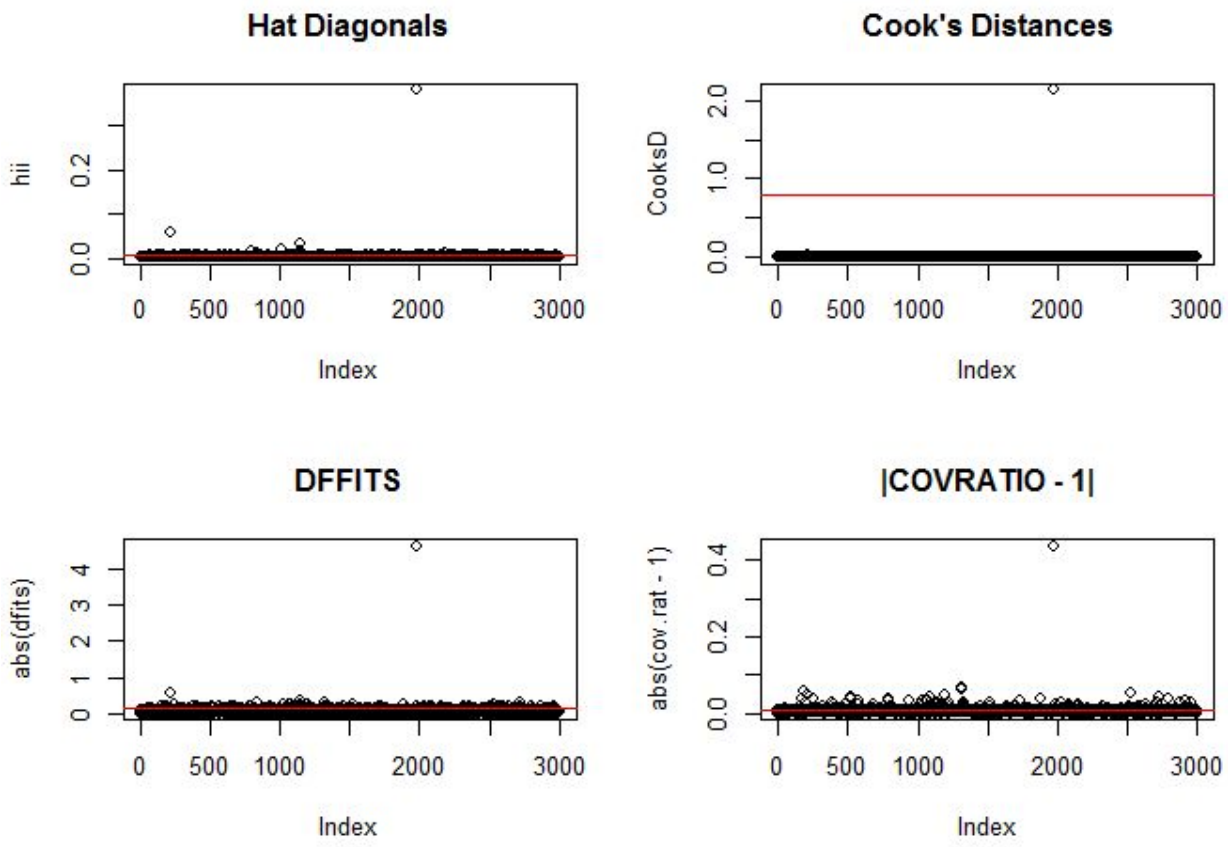
The WLS here assigns weight and the correction is displayed on the figure on the right above. The errors are now much more randomly distributed and homoscedastic. The Normal Q-Q plot for the model also shows that the variances for the WLS model are more aligned with normal distribution.



Detecting Outliers and Influential Cases

The Hat-diagonals or leverage values are plotted along with the COVRATIO, DFFITS and Cook's

Distance plots in the Figure below. Some points are distinctly outside the acceptable



limit marked by the red horizontal line. These plots are used to remove the outliers that are potential influencers of the model.

The final model

The WLS model that conforms to the MLR assumptions is presented below

$$Log(\hat{SMTRVLTIME}) \sim 3.479570 + \quad 0.151877\ CNTTDTR - 0.097403\ URB + 0.014379\ DISTWRK \quad - 0.084656\ MEDINC + \quad 0.0$$

Table: Model Coefficients and Error Values

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.479570	0.038307	90.834	< 2e-16 ***
CNTTDTR	0.151877	0.005357	28.351	< 2e-16 ***
URB	-0.097403	0.026695	-3.649	0.000268 ***

```
DISTWRK      0.014379   0.001205   11.934    < 2e-16 ***
MEDINC      -0.084656   0.030352   -2.789   0.005318 **
MALE         0.057195   0.026599    2.150   0.031617 *
WRKR        -0.092133   0.031349   -2.939   0.003318 **
THREEVEH     0.093728   0.028812    3.253   0.001154 **
---
Residual standard error: 0.7204 on 2981 degrees of freedom
Multiple R-squared:  0.2452
Adjusted R-squared:  0.2435
F-statistic: 138.4 on 7 and 2981 DF,  p-value: < 2.2e-16
AIC 26148.55, BIC 26196.57
```

Most often models are significant as variables are included. This model has a F-value of 138.4 which is highly statistically significant corresponding to its degrees of freedom, which is compared with F-distribution against the null hypothesis that all the model coeffcients are zero. P-value<0.05 indicates model is significant in 95% confidence level.

Moreover, the R^2 values in WLS models are not necessarily accurate. Thus these R^2 coefficients are taken with caution. So interpretation regarding its value is not done as R^2 value of 0.2452 in WLS may not actually indicate the degree of variability explained by this model. But the WLS model can be compared with the MLR models shown in Table in Variable Selection part based on its AIC and BIC value. Noticeably, the AIC and BIC values are much lower than all the other models in that table. So by stabilizing the variance, the model fits the data in a much better manner than the other MLR models and the model with Box-Cox Transformed SMTRVLTIME.

Coefficients interpretation

Since the response variable is natural log-transformed, so all the coefficients has to be interpreted as that one unit change of a variable will increase the SMTRVLTIME by

$$e^{b_j}, \qquad where,\ b_j = \ all\ coefficients\ in\ the\ model$$

So, for each unit change in any variable, daily travel time changes multiplicatively.

Table : Multiplicative Marginal Effects

Variable	Marginal Effects = Multiplicative increase in SMTRVLTIME

Constant	32.446
CNTTDTR	1.1640
URB	0.9071
DISTWRK	1.0145
MEDINC	0.9188
MALE	1.0589
WRKR	0.9120
THREEVEH	1.0983

For every another trip made on that day, daily travel multiplicatively increases by 1.1.6 times or by 16%, for urban areas, the daily travel time is less than in the rural areas by 9.29% (1-0.9071). For a worker, increasing distance to workplace by 1 mile, increases the daily travel time by 1.4%. Compared to lower income and higher income group, only medium income household members have statistically different travel-time characteristics. Compared to the other two groups they spend 8% less time for daily travel. The explanation can be that, the rich households have more dispensable money to spend for discretionary trips and the poor have higher travel time since they may have use public transit more which may take longer in general. The workers also spend 8.2% less time for daily travel as they may be busy in the job and has lesser time for other trips. Lastly, households with three or more vehicles have 9.8% higher daily travel time than the other household members who own none and up to 2 cars. Most likely more vehicles give them more freedom to travel. More importantly more vehicles may indicate more dispensable money which may be spent in some other activities.

On the other hand, occupations of the workers do not have statistically significance in any of the models. Weekdays and weekends also do not hold significant meaning. Household size also has no effect. More people in the household may not necessarily affect personal daily travel time.

ANOVA Table for Model Significance

From the ANOVA table, the type I variances are provided where the variances can be added up together for calculating the sequential SSR and Generalized Hypothesis testing.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
CNTTDTR	1	394.22	394.22	759.6813	< 2.2e-16	***
URB	1	12.98	12.98	25.0043	6.051e-07	***
DISTWRK	1	79.90	79.90	153.9675	< 2.2e-16	***
MEDINC	1	4.07	4.07	7.8490	0.005118	**
MALE	1	2.69	2.69	5.1863	0.022836	*
WRKR	1	3.31	3.31	6.3796	0.011595	*
THREEVEH	1	5.49	5.49	10.5829	0.001154	**
Residuals	2981	1546.94	0.52			

Since all these variables here are intuitively strong candidate for this model, so further testing for null hypothesis for determining whether any of the coefficients are statistically zero and has little extra sum of squares is not warranted. Each of the coefficients here are highly significant. So further testing is not required. The p-values from the table are all below 0.05.

Lack of Fit Test

Finally the model is tested for lack-of-fit on the basis of assumption that at least one observation is repeated in the dataset. For the large data set, this is often expected to be true. In this dataset, variables make repeated observations. The lack of fit portion is almost half of the Pure Error portion. So the ratio of MSLF and MSPE is very low. As a result F-value gains a high p-value in correspond F-distribution. So null hypothesis of perfect fitness of the model is rejected.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lack of fit	1486	735.77	0.50	0.9125	0.961285
Pure Error	1495	811.17	0.54		

So, as seen earlier, the R^2 and Adjusted- R^2 values are around 0.20, the lack of fit is thus present. Better model, transformation or more data is needed to gain a better fitness of the model.

Auto-correlated Variances

From WLS, it is seen that the variances are systematically controlled. So in WLS, variances can be auto-correlated. The **Durbin-Watson Test** assumes in null hypothesis that the auto-correlation do not exists. But in this model it is seen that the p-value is so low that null hypothesis is rejected, i.e. auto-correlation exists. But this is not a problem in WLS. Rather lack-of-fit is the problem here.

$$DW = 1.6738, \text{ p-value} < 2.2e\text{-}16$$

Hold-out Sample

Mean Absolute Deviation

The model is used to predict the daily travel time for the held-out 1803 cases. The MAD value is obtained by this equation -

$$MAD = \frac{\sum |actual - forecast|}{n}$$

From this model, the MAD = 46.2759 minutes. That is for the hold-out sample, the model on average makes error of 46.28 minutes. It indicates that this model is a poor fir for the data.

Mean Absolute Percentage Error (MAPE)

The coefficient can be easily understood from its name. The value obtained is 72.46%. So, the model on average makes 72.46% error. So the conclusion from the lack-of-fit test holds, that this model suffers from lack-of-fit problem.

Section 5. Concluding Remarks

In this project a model for predicting daily travel time is developed. This project tried to obtain the best possible Multiple Linear Regression model. But this model is not the best one rather it suffers from lack-of-fit. Balancing between, on one hand, model assumption, model-fitness, and on the other, actually available data, requires numerous iteration to obtain the desired powerful modeling solution. For the 35% hold-out sample randomly selected from total population this model shows very

high MAD and MAPE value. These values are manifestations of rejecting the null hypothesis of lack-of-fit test.

Yet, the model coefficients in the model are significant and make sense when interpreted. But the failure to obtain a significant fitness to the data indicates, there are room to make the model better by either/or

- **Selecting more cases for model building and testing**,
  in this dataset only New England region is considered. May be including more regions will provide a better picture
- **Trying more transformation for the variables,**
  In this project, a number of combinations of variables and transformation were tested. But to stick with Linear Modeling framework, necessary variable transformations would make the interpretation tougher. So this may be a possible solution, but it has its own drawback
- **Using some other modeling framework except linear regression**
  Future research can lead to selecting some Generalized Linear models that may better capture the variability present in the dataset, allowing the errors to take different shape other than Normal Distribution

From travel demand modeling point of view, no conclusion about Zahavi's claim can be made from this modeling approach. Though it is claimed that Travel Time Budget is constant over aggregate dataset, person and household level investigation of TTB or travel time models can enable the researcher to portray a better picture about the Activity-Based Travel Demand Modeling.

Appendix: *R*-codes

```
rm(list=ls())
ls()
a = read.csv("3000cases_NE_Adlt_AllDay_2.csv")
nrow(a)
attach(a)
data.frame(a$HHSIZE)
proj = data.frame(CNTTDTR
                        , DRVR
                        , HHSIZE
                        , HHVEHCNT
                        , ONEVEH
                        , TWOVEH
                        , THREEVEH
```

```
                        , DISTWRK
                        , WRKR
                        , WRKR_SALES
                        , WRKR_ADMIN
                        , WRKR_MFGCONST
                        , WRKR_PROTECH
                        , WRKTIMEFLEX
                        , HIGHINC
                        , MEDINC
                        , YOUNG
                        , OLD
                        , URB
                        , MALE
                        , Weekend, SMTRVLTIME)
proj = na.omit(proj)
nrow(proj)
head(proj)

pairs(~SMTRVLTIME+CNTTDTR+DISTWRK+ ONEVEH+TWOVEH+THREEVEH+URB+MEDINC+HIGHINC+
        YOUNG+OLD+URB+ MALE+ Weekend + WRKR_SALES
      + WRKR_ADMIN
      + WRKR_MFGCONST
      + WRKR_PROTECH
      + WRKTIMEFLEX,
      data=proj, main="Simple Scatterplot Matrix")
corrvariable= data.frame(log(SMTRVLTIME), CNTTDTR, DISTWRK, ONEVEH,TWOVEH,
                         THREEVEH,URB,MEDINC,HIGHINC, YOUNG
                         , OLD
                         , URB
                         , MALE
                         , Weekend)
k = cor(corrvariable, method="pearson")
as.matrix(k)
attach(proj)

null = lm(SMTRVLTIME ~ 1, data = proj)
full = lm(SMTRVLTIME ~ CNTTDTR
          + DRVR
          + HHSIZE
          + HHVEHCNT
          + ONEVEH
          + TWOVEH
          + THREEVEH
          + DISTWRK
          + WRKR
          + WRKR_SALES
          + WRKR_ADMIN
          + WRKR_MFGCONST
          + WRKR_PROTECH
          + WRKTIMEFLEX
          + HIGHINC
          + MEDINC
          + YOUNG
          + OLD
          + URB
          + MALE
          + Weekend)

library(leaps)
x  = nrow(proj)

step(null, scope = list(upper=full, lower=null),
     data=data, direction="both", nbest = 2)
step(null, scope = list(upper=full, lower=null),
     data=flu, direction="both", k = log(x))
```

```
testmod= lm((SMTRVLTIME) ~ (CNTTDTR) + DISTWRK + HIGHINC +MEDINC + URB + Weekend + MALE  +
(WRKR) + THREEVEH)
summary(testmod)
extractAIC(testmod)
extractAIC(testmod, k=log(2999))
library(lmtest)
bptest(testmod, data = proj)
abline(0,1)

qqnorm(resid(testmod))
plot(fitted(testmod), rstudent(testmod), main="Externally Studentized Residuals Plot Vs Fitted
Value",
            xlab= "Fitted Y", ylab="Ext Studentized Residuals" )
abline(0,0)
abline(2,0, col=4)
abline(-2,0, col=4)

#Leverage
p=length(coef(testmod))
n=nrow(proj)
hii = influence(testmod)$hat
par(mfrow=c(1,1))
plot(hii, main="Hat Diagonals")
abline(h=2*p/n, col="red")

#BC Transforation
library(car)
p1 <- powerTransform((SMTRVLTIME) ~  CNTTDTR + URB+ DISTWRK
                     +MEDINC + MALE  + WRKR
                     +THREEVEH, proj)
summary(p1)
p1$roundlam
m1 <- lm(bcPower(SMTRVLTIME, p1$roundlam) ~  CNTTDTR + URB+ DISTWRK
         +MEDINC + MALE  + WRKR
         +THREEVEH, data =proj)
summary(m1)
qqnorm(resid(m1), main="Normal Q-Q Plot of Residuals")
abline(0,1)
plot(fitted(m1), rstudent(m1),
     main="Residual Vs Fitted Values Plot",
     xlab="Fitted Values", ylab="Residuals")
abline(0,0, col=1)
abline(2,0, col=4)
abline(-2,0, col=4)
library(lmtest)
dwtest(m1,data=proj)
#INFLUNETIAL OUTLIERS
hii = influence(m1)$hat
p=length(coef(m1))
2*p/n
highleverage = hii[hii > 2*p/n]
highinfluentials

CooksD<-  cooks.distance(m1)
ri    <- rstandard(m1)
dfits  <- dffits(m1)
cov.rat<- covratio(m1)
p=length(coef(m1))
n=nrow(proj)

par(mfrow=c(1,1))
plot(hii, main="Hat Diagonals")
abline(h=2*p/n, col="red")

plot(CooksD, main="Cook's Distances")
abline(h=.8,col="red")
```

```
plot(abs(dfits), main="DFFITS")
abline(h=2*sqrt( p/n ), col="red" )
plot(abs(cov.rat-1), main="|COVRATIO - 1|")
abline(h=3*p/n, col="red")

highleverage = hii[hii > 2*p/n]
highleverage
highCooksD = CooksD[CooksD > 0.8]
highdffits = dffits[dffits > 2*sqrt( p/n )]
highcovratio = covratio[covratio > abs(cov.rat-1)]
sum.mat <- cbind( hii,CooksD, dfits, cov.rat)

proj = proj[-1145,]
nrow(proj)
pr= na.omit(pr)
sort(fitted(m1), decreasing=TRUE)
which.max( fitted(m1))
order(fitted(m1),decreasing=T)[1:10]
# unweighted regression
unwt.mod <- lm(log(SMTRVLTIME) ~ (CNTTDTR) + URB+ DISTWRK
                 +MEDINC + MALE  + WRKR
               +THREEVEH,data=proj)
res <- unwt.mod$residuals
yhat<- unwt.mod$fitted.values
# estimate sd
sd.mod <- lm(abs(res)~ (CNTTDTR) + URB+ DISTWRK
             +MEDINC + MALE  + WRKR
             +THREEVEH,data=proj)
sest <- sd.mod$fitted.values
wt <- 1/sest^2
# weighted regression
wt.mod <- lm(log(SMTRVLTIME) ~ (CNTTDTR) + URB+ DISTWRK
             +MEDINC + MALE  + WRKR
             +THREEVEH,data=proj)
resw <- wt.mod$residuals
ywhat <- wt.mod$fitted.values
par(mfrow=c(1,2))
plot(yhat, res, main="Unweighted Model", xlab="predicted", ylab="residuals")
abline(h=0, col="red")
plot(ywhat, sqrt(wt)*resw, main="Weighted Model", xlab="predicted", ylab="residuals")
abline(h=0, col="red")
qqnorm(resid(wt.mod), main="Normal Q-Q Plot of Residuals")
abline(0,1)
summary(wt.mod)
summary(m1)
sort(sqrt(wt)*resw,decreasing=F)[1:6]
sort(ywhat,decreasing=T)[1:10]
bptest(wt.mod, data=proj, studentize = FALSE)
library(alr3)
pureErrorAnova(wt.mod)
dwtest(wt.mod, data=proj)

#Hold-out
rm(list=ls())
ls()
hold = read.csv("1803cases_NE_Adlt_AllDay.csv")
attach(hold)
hold = na.omit(hold)
nrow(hold)
hold$predict = NULL
hold$predict = exp(3.479570 + 0.151877 *CNTTDTR - 0.097403* URB
                   + 0.014379 *DISTWRK  -0.084656 *MEDINC
                   +  0.057195  * MALE  -0.092133   *WRKR
                   +0.093728  * THREEVEH)
```

References

[1] Zahavi, Y., Talvitie, A., Regularities In Travel Time And Money Expenditures. Transportation Research Record 750, 1980, pp 13–19.

[2] Zahavi, Y., Ryan, J.M.,. Stability Of Travel Components Over Time. Transportation Research Record 750, 1980, pp19–26.

[3] Mokhtarian, P., Chen, C., TTB or not TTB, That Is The Question: A Review And Analysis Of The Empirical Literature On Travel Time (And Money) Budgets, Transportation Research Part A 38, 2004, 643–675

[4] Wee, B., Rietveld, P., Meurs, H, Is Average Daily Travel Time Expenditure Constant? In Search Of Explanations For An Increase In Average Travel Time, Journal of Transport Geography 14, 2006, pp 109–122

[5] Chen, C., Mokhtarian, P., A Review And Discussion Of The Literature On Travel Time And Money Expenditures, UCD-ITS-RR-99-25, November 1999

[6] Kitamura, R., Robinson, J., Golob, T. F., Bradley, M., Leonard, J., and van der Hoorn, T., A Comparative Analysis of Time Use Data in the Netherlands and California. Research Report Number UCD-ITS-RR-92-9, Institute of Transportation Studies, University of California, Davis, June, 1992

[7] Levinson, D. M., Space, Money, Life-stage, and the Allocation of Time. Transportation **26**, 1999 pp141-171.