The Influence of Individual Characterisitcs on Public Transportation Planning[*]

Iris Zhong

Abstract

xx

```
## Warning: package 'tidyverse' was built under R version 3.5.3

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'stringr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

## Warning: package 'lubridate' was built under R version 3.5.3

## Warning: package 'stargazer' was built under R version 3.5.2
```

[*] xx

1 Literature Review

Allen et al. (2016) study the reasoning of the failure of a referendum on a congestion charging scheme in Edinburgh. Instead of using direct voting data, they conduct a survey after the referendum, which allows them to ask more specific questions. Researchers can gain detailed data by surveying, because the unit of measurement is each individual; however, a possible disadvantage of surveying is that respondents who turn in the questionnaire tend to have stronger attitudes towards the proposal, generating sampling bias. They conclude that people who use cars as the primary transportation mean, demonstrate a misconception of the pricing plan, or question the effectiveness of the scheme at reducing congestion are more likely to oppose it. Their findings can give insights to the similar failure in the Gwinnett referendum. Voters against the proposal could be those who rarely use public transportation and those who are not convinced by the effectiveness of expanding public transit in alleviating the traffic.

Another crucial factor is the accessibility of the proposed transit system. Kinsey et al. (2010) examine the relationship between the distance to the scheduled railway station and voter turnout by studying the Seattle monorail referendum. They introduce the concept of diffused and concentrated benefit/cost. People who live far from the monorail enjoy the diffused benefit of less traffic congestion, and bear the diffused cost of increased tax. People living close to the rail experience the same diffused benefit and cost, but they also gain the concentrated benefit of easily accessing the public good. Finally, those who live very close to the railway have the same benefits and costs, but they also face the concentrated cost such as inconvenience during construction. Since "people are more strongly motivated to avoid losses than to approach gains," they expect a higher turnout rate in farther places with votes for "no," which is verified from their analyses. Besides distance, they also find out precincts with a higher percentage of people of lower socioeconomic status or young people have a lower turnout rate. Interestingly, there is a significant interaction between partisanship and distance, which would be also tested in my study. In essence, the effect of distance on turnout is weakened by partisanship, and vanishes beyond a threshold of distance. Even though my dependent variable is voters' responses rather than turnout, it can be inferred from Kinsey et al.'s findings that people farther away from the transit system would vote against the referendum more. However, the relationship might be non-linear and requires some form of transformation. Regarding the methods, they utilize the spatial lag model to correct for autocorrelation, which is proper to use in my project as well since both studies use precinct-level data.

Table 1: Variable definitions

Variable name	Description
GEOID	The geographic identifier of the census tract
medage	The median age of the population in the tract
medincome	The median income of the population in the tract
white_pct	The percentage of white population in the tract
public_pct	The percentage of people who go to work by public transportation (excluding taxi or cab)
time_pct	The percentage of people who travel more than an hour to work
trump_pct	The estimated percentage of votes for Donald Trump in that tract
voter_turnout	The estimated percentage of voters who voted in this referendum in the tract
yes_pct	The estimated percentage of voters who voted yes in this referendum in the tract
plan_yes	Whether the tract is covered by short-range plan. 1 stands for yes, 0 stands for no
current_yes	Whether the tract is covered by the existing public transportation. 1 stands for yes, 0 stands for no

2 Background

current transportation future plan referendum

3 Data & Methods

3.1 Conceptual model

3.2 Data

- **GEOID:** The geographic identifier of the census tract.

- **medage:** The median age of the population in the tract.
- **medincome:** The median income of the population in the tract.
- **white_pct:** The percentage of white population in the tract.
- **public_pct:** The percentage of people who go to work by public transportation (excluding taxi or cab).

- **time_pct:** The percentage of people who travel more than an hour to work.

- **trump_pct:** The estimated percentage of votes for Donald Trump in that tract.
- **voter_turnout:** The estimated percentage of voters who voted in this referendum in the tract.
- **yes_pct:** The estimated percentage of voters who voted yes in this referendum in the tract.

- **plan__yes:** Whether the tract is covered by the public transportation planned in the short-range (Y2020 – 2025), defined by whether any transportation is available within 500 meters. 1 stands for yes, 0 stands for no.

- **current__yes:** Whether the tract is covered by the existing public transportation, defined by whether any transportation is available within 500 meters. 1 stands for yes, 0 stands for no.

4 Results

model 1: all variables, no interaction, linear

```
load("processed_data/analysis_data.RData")
mod1 <- lm(data = final_data, yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_
summary(mod1)
```

Call: lm(formula = yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + plan_yes + current_yes, data = final_data)

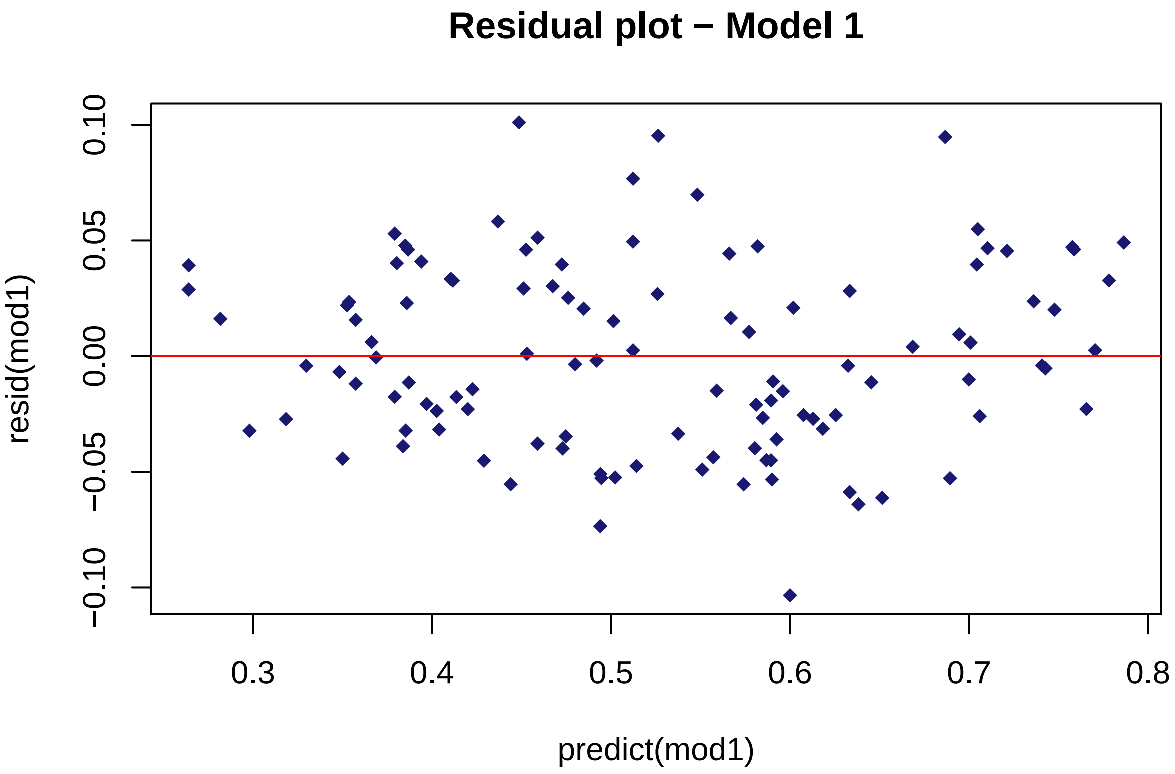
Residuals: Min 1Q Median 3Q Max -0.103384 -0.031723 -0.004037 0.030238 0.101018

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.016e-01 4.677e-02 17.138 < 2e-16 ***medage 2.733e-03 1.477e-03 1.851 0.067054 .***
medincome 1.458e-07 2.914e-07 0.500 0.617850
white__pct 1.177e-01 5.285e-02 2.226 0.028162
public__pct 8.393e-01 3.355e-01 2.501 0.013946 *
time__pct -3.029e-01 9.013e-02 -3.361 0.001090 trump_pct -8.884e-01 5.680e-02 -15.640 < 2e-16
voter__turnout -3.285e-01 1.257e-01 -2.613 0.010323
plan__yes1 -2.967e-02 1.153e-02 -2.573 0.011511 *
current__yes1 4.627e-02 1.209e-02 3.825 0.000224 * — Signif. codes: 0 ‘***’ 0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 0.04185 on 103 degrees of freedom Multiple R-squared: 0.917, Adjusted R-squared: 0.9098 F-statistic: 126.5 on 9 and 103 DF, p-value: < 2.2e-16

model 1 assumption checking

```r
plot(predict(mod1),resid(mod1),col="midnightblue",pch=18,main="Residual plot - Model 1")
abline(0,0,col="red")
```



collinearity:

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.5.3

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(mod1)
```

```
   medage     medincome     white_pct    public_pct      time_pct
 2.920936      3.221504      3.863617      1.151309      1.559837
trump_pct voter_turnout      plan_yes   current_yes
 4.512575      3.368751      1.741403      2.350032
```

all below 5: good, no collinearity problem

model 2: all variables, no interaction, logistic

```
mod2 <- glm(data = final_data, yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(mod2)
```

Call: glm(formula = yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + plan_yes + current_yes, family = "binomial", data = final_data)

Deviance Residuals: Min 1Q Median 3Q Max

-0.224114 -0.065598 -0.003135 0.067243 0.206940

Coefficients: Estimate Std. Error z value Pr(>|z|) (Intercept) 1.315e+00 2.311e+00 0.569 0.569 medage 1.068e-02 7.223e-02 0.148 0.882 medincome 6.880e-07 1.432e-05 0.048 0.962 white_pct 5.666e-01 2.627e+00 0.216 0.829 public_pct 3.642e+00 1.706e+01 0.213 0.831 time_pct -1.376e+00 4.462e+00 -0.308 0.758 trump_pct -3.790e+00 2.839e+00 -1.335 0.182 voter_turnout -1.453e+00 6.216e+00 -0.234 0.815 plan_yes1 -1.276e-01 5.671e-01 -0.225 0.822 current_yes1 1.891e-01 5.911e-01 0.320 0.749

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 9.05063  on 112  degrees of freedom
```

Residual deviance: 0.83282 on 103 degrees of freedom AIC: 133.92

Number of Fisher Scoring iterations: 4

model 3: all variables, no interaction, some transformation, linear

step 1: find the skewed variables

```
library(dlookr)
```

```
## Warning: package 'dlookr' was built under R version 3.5.3
```

```
## Loading required package: mice
```

```
## Warning: package 'mice' was built under R version 3.5.3
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
##
## Attaching package: 'dlookr'
```

```
## The following object is masked from 'package:base':
##
##     transform
```

```
find_skewness(final_data)
```

[1] 3 5 8

medincome, public_pct, voter_turnout

step 2: transform them

```r
data_tf <- final_data %>%
   mutate(medincome = log(medincome),
          public_pct = log(public_pct + 0.01),
          voter_turnout = (voter_turnout)^0.5)
find_skewness(data_tf)
```

integer(0)

step 3: model them

```r
mod3 <- lm(data = data_tf, yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct
summary(mod3)
```

Call: lm(formula = yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + plan_yes + current_yes, data = data_tf)

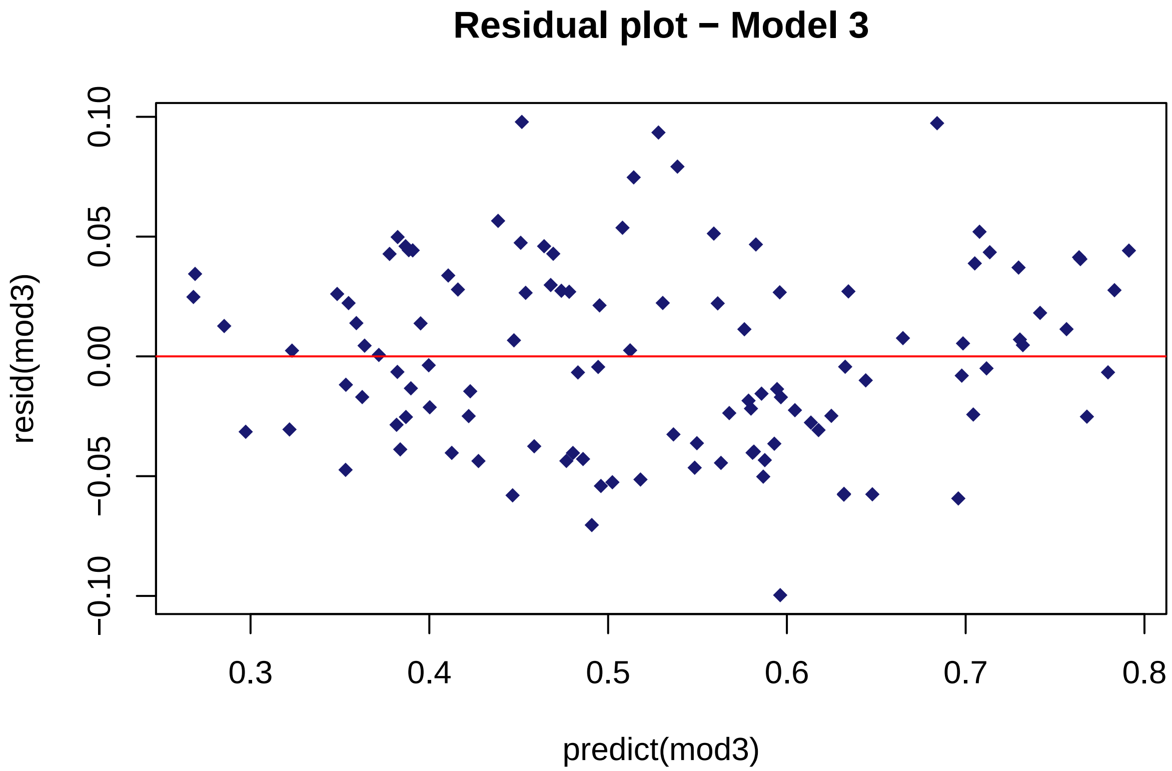
Residuals: Min 1Q Median 3Q Max -0.099662 -0.030760 -0.004331 0.027626 0.097867

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.022033 0.217430 4.701 8.07e-06 ***medage 0.003595 0.001506 2.386 0.01884***
medincome -0.007268 0.022156 -0.328 0.74356
white_pct 0.119648 0.050332 2.377 0.01929 *
public_pct 0.022875 0.008179 2.797 0.00616 time_pct -0.279306 0.089340 -3.126 0.00230 **
trump_pct -0.859270 0.058269 -14.747 < 2e-16 ***voter_turnout -0.321784 0.103026 -3.123 0.00232***
plan_yes1 -0.031542 0.011317 -2.787 0.00633 ***current_yes1 0.045541 0.011848 3.844 0.00021*** —
Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04112 on 103 degrees of freedom Multiple R-squared: 0.9199, Adjusted R-squared: 0.9129 F-statistic: 131.4 on 9 and 103 DF, p-value: < 2.2e-16

model 3 assumption checking

```r
plot(predict(mod3),resid(mod3),col="midnightblue",pch=18,main="Residual plot - Model 3")
abline(0,0,col="red")
```



collinearity:

```
vif(mod3)
```

```
   medage     medincome     white_pct    public_pct      time_pct
 3.149466      3.713809      3.629803      1.133155      1.587723
trump_pct voter_turnout      plan_yes   current_yes
 4.919508      3.621155      1.737295      2.336190
```

all below 5: no collinearity

model 4: all variables, interaction, no transformation, linear

```
mod4 <- lm(data = final_data, yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_
summary(mod4)
```

Call: lm(formula = yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + plan_yes * current_yes, data = final_data)

Residuals: Min 1Q Median 3Q Max -0.103714 -0.032077 -0.002399 0.030226 0.101740

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.041e-01 4.707e-02 17.085 < 2e-16 ***medage 2.738e-03 1.481e-03 1.849 0.06739 .***

medincome 1.326e-07 2.930e-07 0.453 0.65183

white_pct 1.143e-01 5.326e-02 2.147 0.03420

public_pct 8.173e-01 3.382e-01 2.417 0.01743 *

time_pct -3.057e-01 9.048e-02 -3.378 0.00103 trump_pct -8.859e-01 5.709e-02 -15.517 < 2e-16

voter_turnout -3.257e-01 1.262e-01 -2.582 0.01125

plan_yes1 -3.176e-02 1.201e-02 -2.645 0.00945 current_yes1 1.888e-02 4.399e-02 0.429 0.66866

plan_yes1:current_yes1 2.899e-02 4.476e-02 0.648 0.51864

— Signif. codes: 0 ‘***’ 0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 0.04197 on 102 degrees of freedom Multiple R-squared: 0.9174, Adjusted R-squared: 0.9093 F-statistic: 113.2 on 10 and 102 DF, p-value: < 2.2e-16

```
anova(mod1,mod4)
```

Analysis of Variance Table

Model 1: yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + plan_yes + current_yes Model 2: yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + plan_yes * current_yes Res.Df RSS Df Sum of Sq F Pr(>F) 1 103 0.18044

2 102 0.17970 1 0.00073903 0.4195 0.5186

Model w/ interaction doesn’t differ significantly from the one w/o interaction.