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

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

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

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

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

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.

2 Background

current transportation future plan referendum

3 Data & Methods

3.1 Conceptual model

According to previous research, sociodemographic elements can influence people's voting decisions in the referendum. For example, the effect of income is mixed: on the one hand, people with higher income will pay a smaller portion of their earnings for the implementation of the plan; on the other hand, they will pay a larger amount of tax. Bollino (2008) finds a positive correlation between income and people's willingness to pay for renewable resources. Burkhardt and Chan (2017) separate the influence of income from tax, and discover their opposite effects on voting. Therefore, it is worth considering the relationship between income and percentage of supporters in this referendum. Voters' partisanship attachment is found to be a significant factor as well in Burkhardt and Chan's (2017) paper. Areas with higher proportions of Republicans are less supportive of fiscally costly propositions. In my project, it can be hypothesized that tracts that have a higher proportion of Trump supporters tend to have a lower percentage of agreement to the proposal.

In addition, some factors related to transportation can intuitively shape people's attitudes towards public transit. For example, the areas in which people do not use public transit at all might have a higher percentage of refusal of the proposal. People who have to travel a long time to work are more likely to support the extension plan if it helps save time.

Finally, people favor the proposition if it benefits them. Specifically, tracts that are not covered by public transport at present but will be covered in the expansion plan are predicted to support the proposal more.

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 the public transportation now and in the short-range (Y2020 – 2025), defined by whether any transportation is available within 500 meters. 1 stands for the tract doesn't have 2 stands for the tract has transit now but not in the short-range plan. 3 stands for the tract that doesn't 4 stands for the tract that has public transit both now and in the future.

3.2 Data

- **current__plan:** Whether the tract is covered by the public transportation now and in the short-range (Y2020 – 2025), defined by whether any transportation is available within 500 meters. 1 stands for the tract doesn't have transit both now and in the short-range plan. 2 stands for the tract has transit now but not in the short-range plan. 3 stands for the tract that doesn't have transit now and will have in the future. 4 stands for the tract that has public transit both now and in the future.

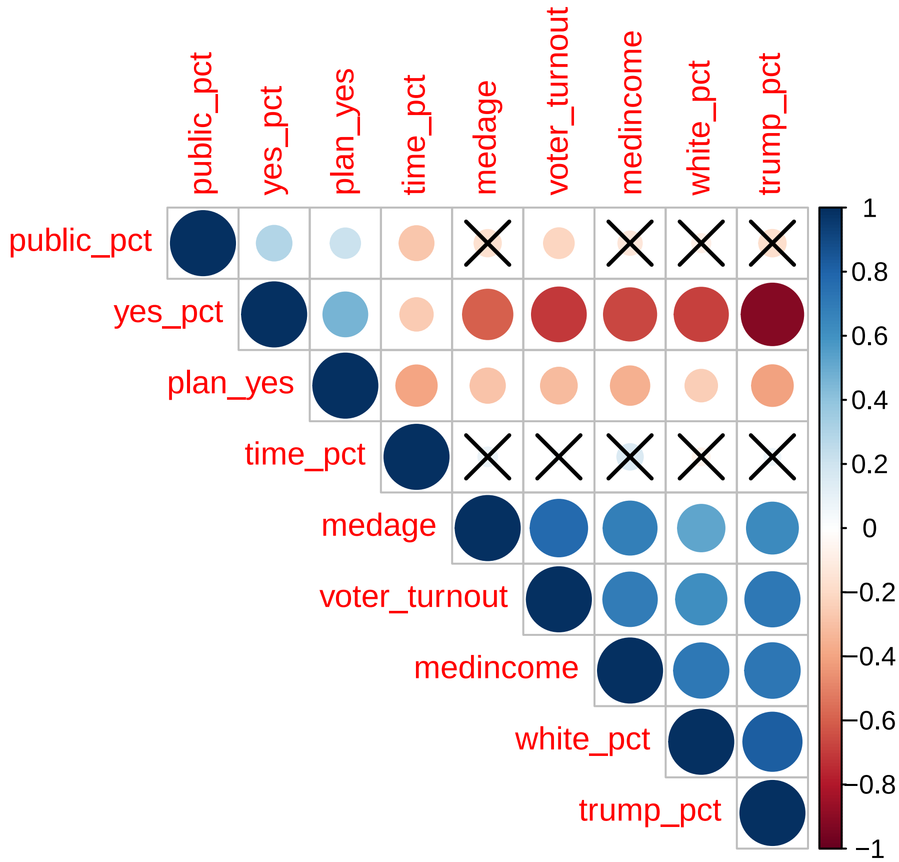
Table 2: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
medage	113	35.56	4.58	26	32.8	38.8	52
medincome	113	69,439.24	24,358.44	33,020	51,429	82,845	156,136
white__pct	113	0.48	0.15	0.17	0.38	0.61	0.89
public__pct	113	0.01	0.01	0	0.002	0.02	0
time__pct	113	0.16	0.05	0.04	0.12	0.20	0.31
trump__pct	113	0.40	0.15	0.11	0.27	0.52	0.69
voter__turnout	113	0.16	0.06	0.05	0.13	0.18	0.37
yes__pct	113	0.53	0.14	0.27	0.42	0.61	0.84

```r
data_numeric <- final_data %>%
  mutate(plan_yes = as.numeric(current_plan)) %>%
  select(-c(current_plan,GEOID))


data_cor = cor(data_numeric)
```

```r
data_cor_1 <- rcorr(as.matrix(data_numeric))
M <- data_cor_1$r
p_mat <- data_cor_1$P
corrplot(M, type = "upper", order = "hclust",
         p.mat = p_mat, sig.level = 0.05)
```



3.3 Model specification

Model 1: $yes\_pct = \beta_0 + \beta_1 * medage + \beta_2 * medincome + \beta_3 * white\_pct + \beta_4 * public\_pct + \beta_5 * time\_pct + \beta_6 * trump\_pct + \beta_7 * voter\_turnout + \beta_8 * current\_plan + \epsilon$

4 Results

model 1: no interaction, linear

```r
mod1 <- lm(data = final_data, yes_pct ~ medage + medincome + white_pct
          + public_pct + time_pct + trump_pct + voter_turnout +
             current_plan)
summary(mod1)
```

Call: lm(formula = yes_pct ~ medage + medincome + white_pct + public_pct + time_pct + trump_pct + voter_turnout + current_plan, 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

current_plan2 1.888e-02 4.399e-02 0.429 0.66866

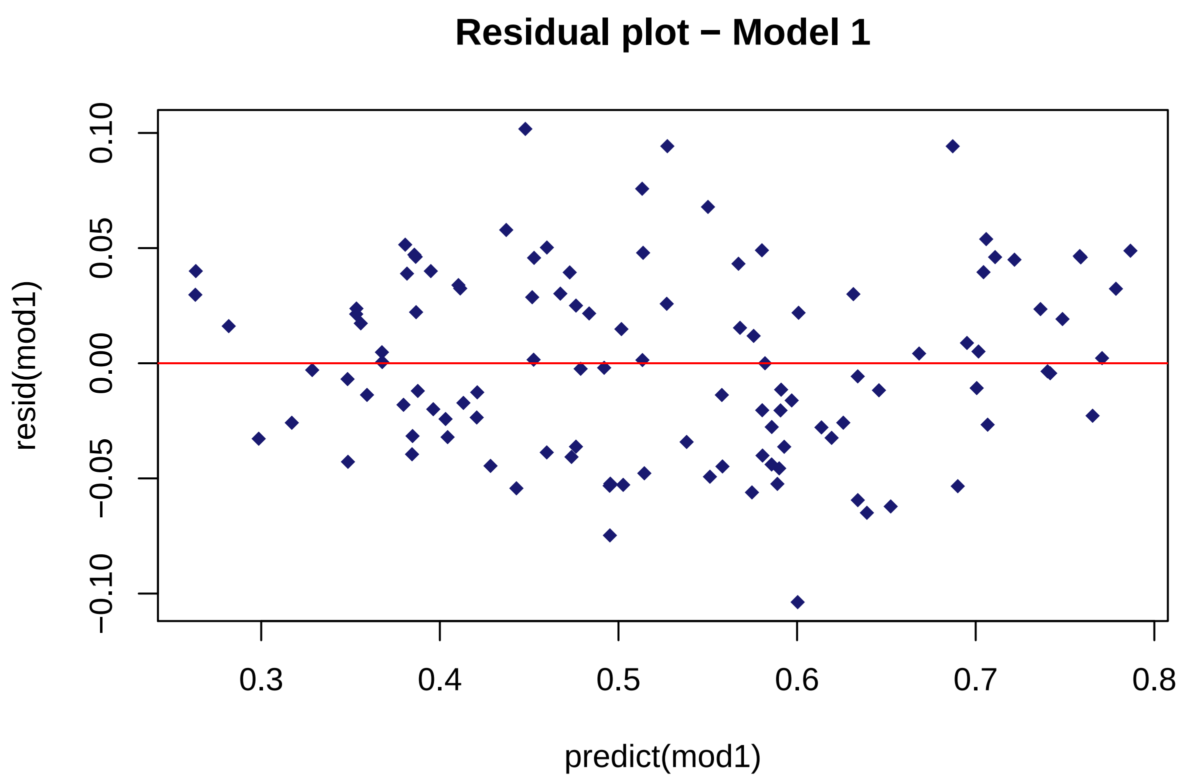
current_plan3 -3.176e-02 1.201e-02 -2.645 0.00945 current_plan4 1.612e-02 1.155e-02 1.395 0.16606

— 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

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)
```

```
              GVIF Df GVIF^(1/(2*Df))
```

medage 2.921005 1 1.709095 medincome 3.237296 1 1.799249 white_pct 3.900578 1 1.974988 public_pct 1.163004 1 1.078427 time_pct 1.563279 1 1.250312 trump_pct 4.533128 1 2.129114 voter_turnout 3.372750 1 1.836505 current_plan 1.853504 3 1.108321

all below 5: good, no collinearity problem

model 2: 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 + current_plan, family = "binomial", data = final_data)

Deviance Residuals: Min 1Q Median 3Q Max

-0.224855 -0.066351 -0.001643 0.067049 0.208799

Coefficients: Estimate Std. Error z value Pr(>|z|) (Intercept) 1.328e+00 2.320e+00 0.572 0.567 medage 1.070e-02 7.223e-02 0.148 0.882 medincome 6.272e-07 1.435e-05 0.044 0.965 white_pct 5.506e-01 2.639e+00 0.209 0.835 public_pct 3.528e+00 1.715e+01 0.206 0.837 time_pct -1.390e+00 4.468e+00 -0.311 0.756 trump_pct -3.779e+00 2.844e+00 -1.329 0.184 voter_turnout -1.441e+00 6.219e+00 -0.232 0.817 current_plan2 5.722e-02 2.128e+00 0.027 0.979 current_plan3 -1.378e-01 5.889e-01 -0.234 0.815 current_plan4 5.908e-02 5.632e-01 0.105 0.916

(Dispersion parameter for binomial family taken to be 1)

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

Residual deviance: 0.82869 on 102 degrees of freedom AIC: 135.96

Number of Fisher Scoring iterations: 4

model 1 & 2 table

model 3: no interaction, some transformations, linear

step 1: find the skewed variables

```r
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:Hmisc':
##
##     describe
```

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

Table 3: Initial regression results

	Dependent variable:	
	yes_pct	
	OLS	*logistic*
	(1)	(2)
medage	0.003*	0.011
	(0.001)	(0.072)
medincome	0.00000	0.00000
	(0.00000)	(0.00001)
white_pct	0.114**	0.551
	(0.053)	(2.639)
public_pct	0.817**	3.528
	(0.338)	(17.153)
time_pct	−0.306***	−1.390
	(0.090)	(4.468)
trump_pct	−0.886***	−3.779
	(0.057)	(2.844)
voter_turnout	−0.326**	−1.441
	(0.126)	(6.219)
current_plan2	0.019	0.057
	(0.044)	(2.128)
current_plan3	−0.032***	−0.138
	(0.012)	(0.589)
current_plan4	0.016	0.059
	(0.012)	(0.563)
Constant	0.804***	1.328
	(0.047)	(2.320)
Observations	113	113
R^2	0.917	
Adjusted R^2	0.909	
Log Likelihood		−56.978
Akaike Inf. Crit.		135.955
Residual Std. Error	0.042 (df = 102)	
F Statistic	113.224*** (df = 10; 102)	
Note:	*p<0.1; **p<0.05; ***p<0.01	
	Initial linear and logistic regression results	

```r
find_skewness(final_data)
```

[1] 3 5 8

medincome, public_pct, voter_turnout

step 2: transform them

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

integer(0)

step 3: model them

```r
mod3 <- lm(data = data_tf, yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + t
summary(mod3)
```

Call: lm(formula = yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + trump_pct + sqrt_voter_turnout + current_plan, data = data_tf)

Residuals: Min 1Q Median 3Q Max -0.099920 -0.031591 -0.003479 0.027983 0.098568

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

(Intercept) 1.031840 0.218778 4.716 7.64e-06 ***medage 0.003603 0.001511 2.384 0.01899***

log_medincome -0.008331 0.022302 -0.374 0.70952

white_pct 0.116703 0.050746 2.300 0.02350 *

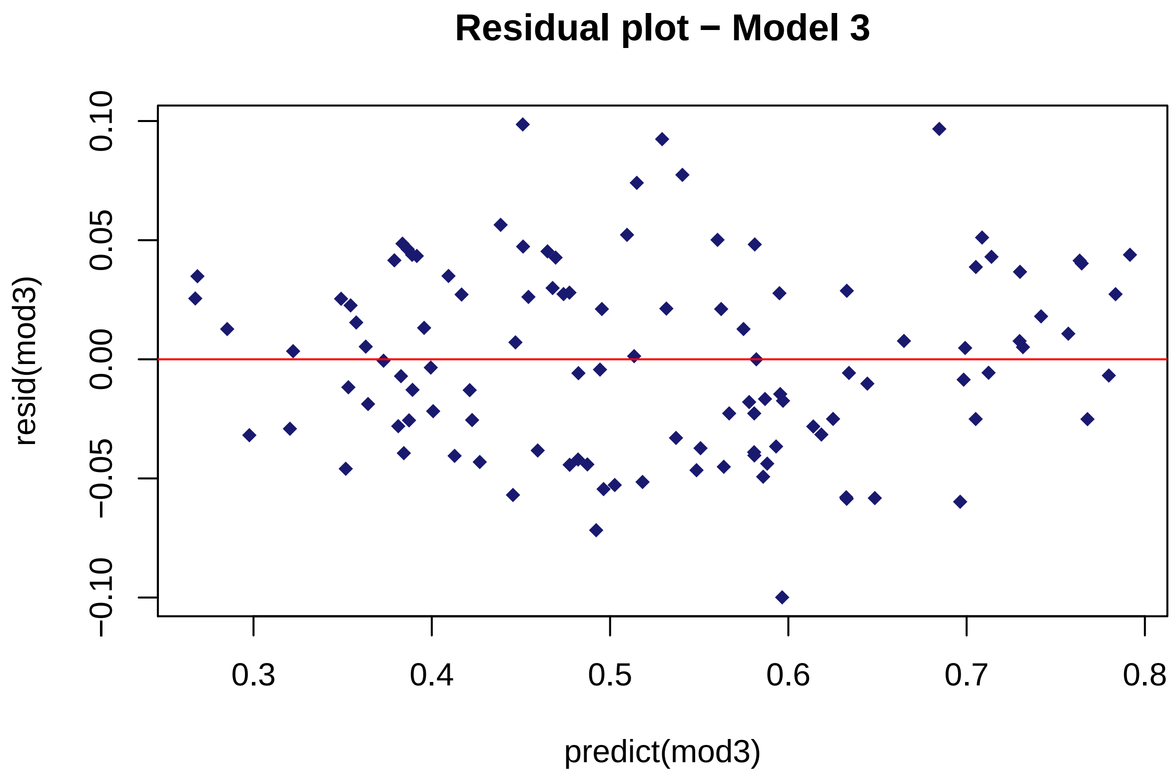
log_public_pct 0.022304 0.008263 2.699 0.00814 time_pct -0.281374 0.089698 -3.137 0.00223 ** trump_pct -0.856963 0.058590 -14.626 < 2e-16 * **sqrt_voter_turnout -0.319107 0.103460 -3.084 0.00263** current_plan2 0.021184 0.043430 0.488 0.62677

current_plan3 -0.033367 0.011777 -2.833 0.00555 ** current_plan4 0.013583 0.011459 1.185 0.23864

— Signif. codes: 0 '*' ***0.001*** ''* *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04125 on 102 degrees of freedom Multiple R-squared: 0.9202, Adjusted R-squared: 0.9123 F-statistic: 117.6 on 10 and 102 DF, p-value: < 2.2e-16 model 3 assumption checking

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



collinearity:

```
vif(mod3)
```

```
GVIF Df GVIF^(1/(2*Df))
```

medage 3.149713 1 1.774743 log_medincome 3.738783 1 1.933593 white_pct 3.666112 1 1.914709 log_public_pct 1.149252 1 1.072032 time_pct 1.590210 1 1.261035 trump_pct 4.942057 1 2.223074 sqrt_voter_turnout 3.628296 1 1.904809 current_plan 1.883541 3 1.111295

all below 5: no collinearity

model 4: transformation, logistic

```r
mod4 <- glm(data = data_tf, yes_pct ~ medage + log_medincome + white_pct
          + log_public_pct + time_pct + trump_pct + sqrt_voter_turnout +
            current_plan, family = "binomial")
```

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

```r
summary(mod4)
```

Call: glm(formula = yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + trump_pct + sqrt_voter_turnout + current_plan, family = "binomial", data = data_tf)

Deviance Residuals: Min 1Q Median 3Q Max

-0.21662 -0.06950 0.00000 0.06082 0.20908

Coefficients: Estimate Std. Error z value Pr(>|z|) (Intercept) 2.30500 10.89739 0.212 0.832 medage 0.01489 0.07514 0.198 0.843 log_medincome -0.03485 1.11172 -0.031 0.975 white_pct 0.56326 2.55987 0.220 0.826 log_public_pct 0.09511 0.41908 0.227 0.820 time_pct -1.28743 4.50172 -0.286 0.775 trump_pct -3.64687 2.95956 -1.232 0.218 sqrt_voter_turnout -1.45781 5.21572 -0.280 0.780 current_plan2 0.06583 2.13728 0.031 0.975 current_plan3 -0.14496 0.58781 -0.247 0.805 current_plan4 0.04787 0.56805 0.084 0.933

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 9.0506  on 112  degrees of freedom
```

Residual deviance: 0.7965 on 102 degrees of freedom AIC: 136.15

Number of Fisher Scoring iterations: 4

model 3 & 4 table

Table 4: Data-transformed regression results

	Dependent variable:	
	yes_pct	
	OLS	*logistic*
	(1)	(2)
medage	0.004**	0.015
	(0.002)	(0.075)
log_medincome	−0.008	−0.035
	(0.022)	(1.112)
white_pct	0.117**	0.563
	(0.051)	(2.560)
log_public_pct	0.022***	0.095
	(0.008)	(0.419)
time_pct	−0.281***	−1.287
	(0.090)	(4.502)
trump_pct	−0.857***	−3.647
	(0.059)	(2.960)
sqrt_voter_turnout	−0.319***	−1.458
	(0.103)	(5.216)
current_plan2	0.021	0.066
	(0.043)	(2.137)
current_plan3	−0.033***	−0.145
	(0.012)	(0.588)
current_plan4	0.014	0.048
	(0.011)	(0.568)
Constant	1.032***	2.305
	(0.219)	(10.897)
Observations	113	113
R^2	0.920	
Adjusted R^2	0.912	
Log Likelihood		−57.076
Akaike Inf. Crit.		136.153
Residual Std. Error	0.041 (df = 102)	
F Statistic	117.561*** (df = 10; 102)	
Note:	*p<0.1; **p<0.05; ***p<0.01	
	Data-transformed linear and logistic regression results	