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 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

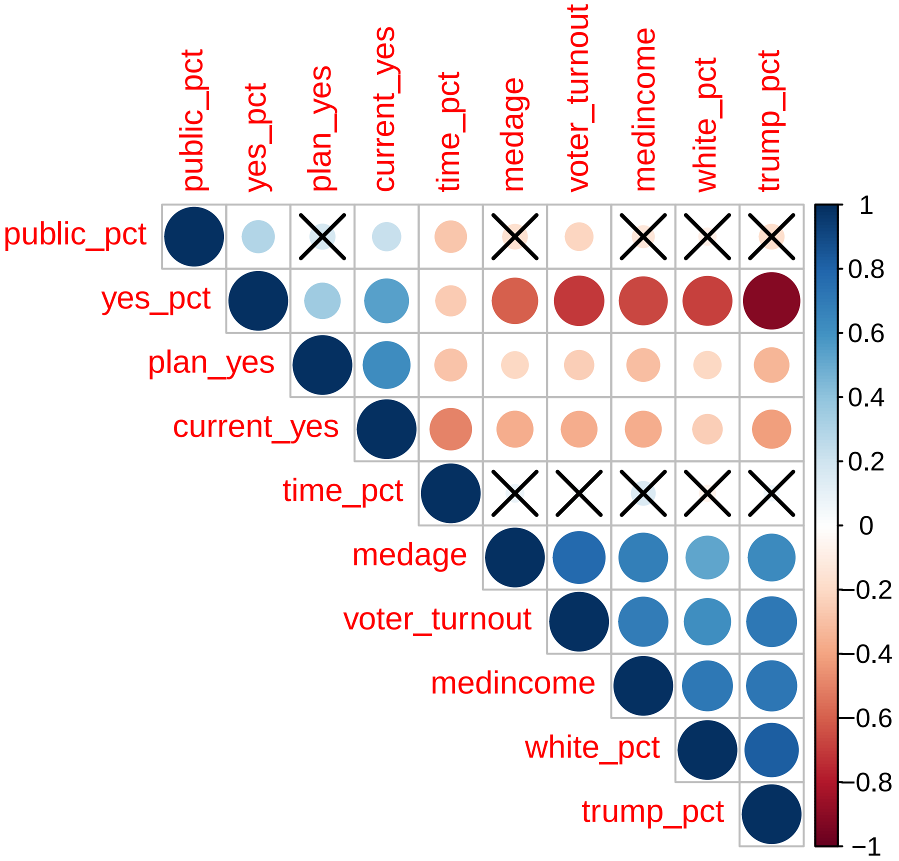
3.2 Data

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(plan_yes),
          current_yes = as.numeric(current_yes)) %>%
   select(-GEOID)
data_cor = cor(data_numeric)

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: $yes\_pct = \beta_0 + \beta_1 * medage + \beta_2 * medincome + \beta_3 * white\_pct + \beta_4 * public\_pct + \beta_5 * time\_pct + \beta_6 * trump\_pct + \beta_7 * voter\_turnout + \beta_8 * plan\_yes + \beta_9 * current\_yes + \epsilon$

4 Results

model 1: no interaction, linear

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

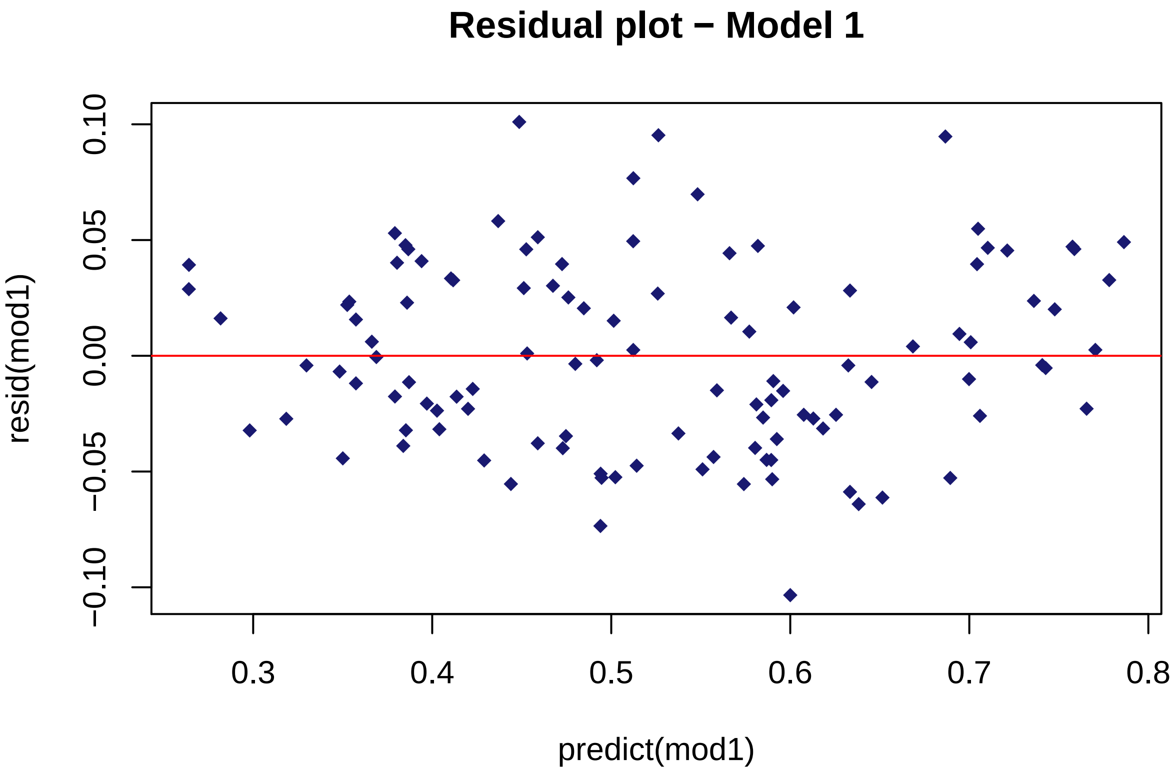
plan_yes1 -2.967e-02 1.153e-02 -2.573 0.011511 *

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:

```
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: 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 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	0.118**	0.567
	(0.053)	(2.627)
public_pct	0.839**	3.642
	(0.336)	(17.064)
time_pct	−0.303***	−1.376
	(0.090)	(4.462)
trump_pct	−0.888***	−3.790
	(0.057)	(2.839)
voter_turnout	−0.328**	−1.453
	(0.126)	(6.216)
plan_yes1	−0.030**	−0.128
	(0.012)	(0.567)
current_yes1	0.046***	0.189
	(0.012)	(0.591)
Constant	0.802***	1.315
	(0.047)	(2.311)
Observations	113	113
R^2	0.917	
Adjusted R^2	0.910	
Log Likelihood		−56.958
Akaike Inf. Crit.		133.915
Residual Std. Error	0.042 (df = 103)	
F Statistic	126.471*** (df = 9; 103)	
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 + ti
summary(mod3)
```

Call: lm(formula = yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + trump_pct + sqrt_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***

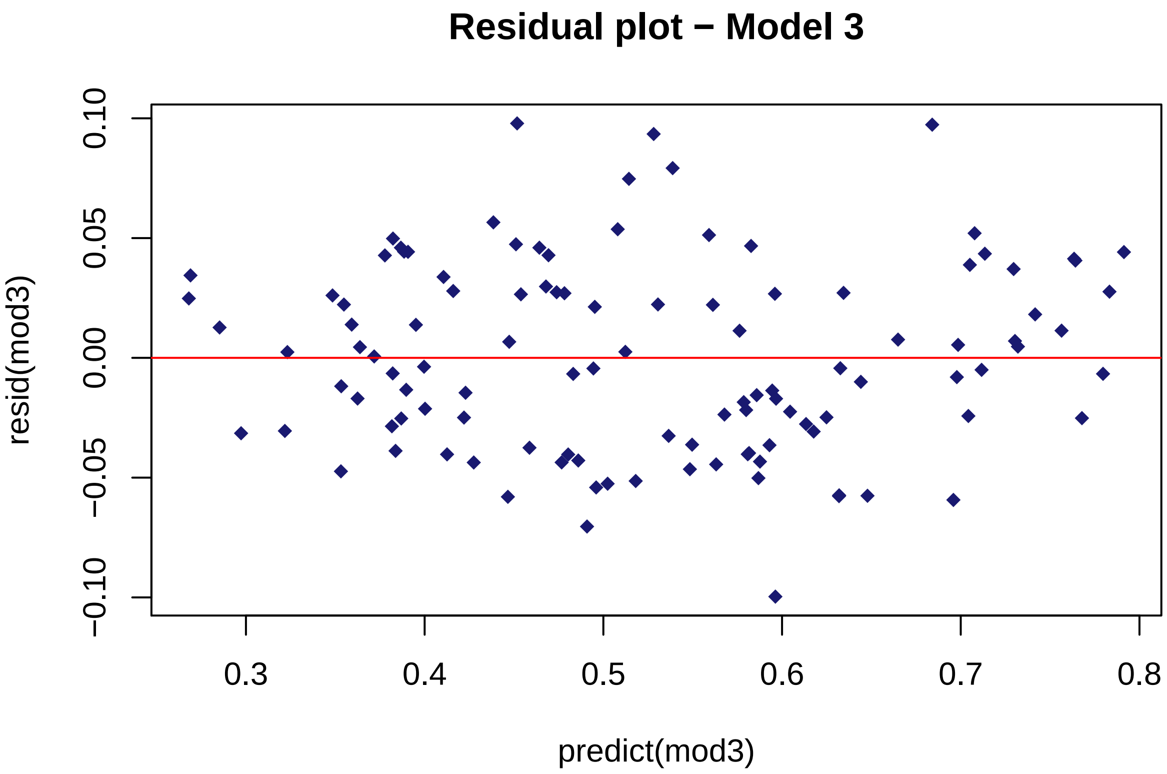
log_medincome -0.007268 0.022156 -0.328 0.74356

white_pct 0.119648 0.050332 2.377 0.01929 *

log_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 ***sqrt_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:

```r
vif(mod3)
```

```
         medage       log_medincome           white_pct       log_public_pct
       3.149466            3.713809            3.629803             1.133155
       time_pct           trump_pct sqrt_voter_turnout             plan_yes
       1.587723            4.919508            3.621155             1.737295
    current_yes
       2.336190
```

all below 5: no collinearity

model 4: transformation, logistic

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

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

```
summary(mod4)
```

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

Deviance Residuals: Min 1Q Median 3Q Max

-0.215991 -0.067982 -0.000552 0.062025 0.210621

Coefficients: Estimate Std. Error z value Pr(>|z|) (Intercept) 2.25723 10.86650 0.208 0.835 medage 0.01485 0.07514 0.198 0.843 log_medincome -0.02970 1.10820 -0.027 0.979 white_pct 0.57743 2.54810 0.227 0.821 log_public_pct 0.09802 0.41609 0.236 0.814 time_pct -1.27653 4.49740 -0.284 0.777 trump_pct -3.65734 2.95408 -1.238 0.216 sqrt_voter_turnout -1.46993 5.21101 -0.282 0.778 plan_yes1 -0.13583 0.56655 -0.240 0.811 current_yes1 0.18584 0.58959 0.315 0.753

(Dispersion parameter for binomial family taken to be 1)

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

Residual deviance: 0.79989 on 103 degrees of freedom AIC: 134.11

Number of Fisher Scoring iterations: 4

model 3 & 4 table

model 5: interaction, transformation, linear

```
mod5 <- lm(data = data_tf, yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + t
summary(mod5)
```

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

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

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	-0.007	-0.030
	(0.022)	(1.108)
white_pct	0.120^{**}	0.577
	(0.050)	(2.548)
log_public_pct	0.023^{***}	0.098
	(0.008)	(0.416)
time_pct	-0.279^{***}	-1.277
	(0.089)	(4.497)
trump_pct	-0.859^{***}	-3.657
	(0.058)	(2.954)
sqrt_voter_turnout	-0.322^{***}	-1.470
	(0.103)	(5.211)
plan_yes1	-0.032^{***}	-0.136
	(0.011)	(0.567)
current_yes1	0.046^{***}	0.186
	(0.012)	(0.590)
Constant	1.022^{***}	2.257
	(0.217)	(10.867)
Observations	113	113
R^2	0.920	
Adjusted R^2	0.913	
Log Likelihood		-57.057
Akaike Inf. Crit.		134.115
Residual Std. Error	0.041 (df = 103)	
F Statistic	131.428^{***} (df = 9; 103)	
Note:	*p<0.1; **p<0.05; ***p<0.01 Data-transformed linear and logistic regression results	

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** plan_yes1 -0.033367 0.011777 -2.833 0.00555 ** current_yes1 0.021184 0.043430 0.488 0.62677 plan_yes1:current_yes1 0.025766 0.044187 0.583 0.56111

— 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

```
anova(mod3,mod5)
```

Analysis of Variance Table

Model 1: yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + trump_pct + sqrt_voter_turnout + plan_yes + current_yes Model 2: yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + trump_pct + sqrt_voter_turnout + plan_yes * current_yes Res.Df RSS Df Sum of Sq F Pr(>F) 1 103 0.17417

2 102 0.17360 1 0.00057868 0.34 0.5611

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

mod 6: only interaction, transformation, linear

```
mod6 <- lm(data = data_tf, yes_pct ~ medage + log_medincome + white_pct + log_public_pct + time_pct + t
summary(mod6)
```

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

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

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.045422 0.216525 4.828 4.85e-06 ***medage 0.003603 0.001511 2.384 0.018988***

log_medincome -0.008331 0.022302 -0.374 0.709522

Table 5: Linear regression with interaction results

	Dependent variable:
	yes_pct
medage	0.004^{**}
	(0.002)
log_medincome	-0.008
	(0.022)
white_pct	0.117^{**}
	(0.051)
log_public_pct	0.022^{***}
	(0.008)
time_pct	-0.281^{***}
	(0.090)
trump_pct	-0.857^{***}
	(0.059)
sqrt_voter_turnout	-0.319^{***}
	(0.103)
plan_yes1	-0.033^{***}
	(0.012)
current_yes1	0.021
	(0.043)
plan_yes1:current_yes1	0.026
	(0.044)
Constant	1.032^{***}
	(0.219)
Observations	113
R^2	0.920
Adjusted R^2	0.912
Residual Std. Error	0.041 (df = 102)
F Statistic	117.561^{***} (df = 10; 102)
Note:	*p<0.1; **p<0.05; ***p<0.01 Linear regression with interaction result

white_pct 0.116703 0.050746 2.300 0.023501 *

log_public_pct 0.022304 0.008263 2.699 0.008137 time_pct -0.281374 0.089698 -3.137 0.002232 ** trump_pct -0.856963 0.058590 -14.626 < 2e-16 ***sqrt_voter_turnout -0.319107 0.103460 -3.084 0.002625*** *plan_yes0:current_yes0 -0.013583 0.011459 -1.185 0.238638 plan_yes1:current_yes0 -0.046950 0.012129 -3.871 0.000192* ** plan_yes0:current_yes1 0.007601 0.042599 0.178 0.858746

plan_yes1:current_yes1 NA NA NA NA

— 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

Table 6: Linear regression with interaction-only results

	Dependent variable:
	yes_pct
medage	0.004^{**}
	(0.002)
log_medincome	-0.008
	(0.022)
white_pct	0.117^{**}
	(0.051)
log_public_pct	0.022^{***}
	(0.008)
time_pct	-0.281^{***}
	(0.090)
trump_pct	-0.857^{***}
	(0.059)
sqrt_voter_turnout	-0.319^{***}
	(0.103)
plan_yes0:current_yes0	-0.014
	(0.011)
plan_yes1:current_yes0	-0.047^{***}
	(0.012)
plan_yes0:current_yes1	0.008
	(0.043)
plan_yes1:current_yes1	
Constant	1.045^{***}
	(0.217)
Observations	113
R^2	0.920
Adjusted R^2	0.912
Residual Std. Error	0.041 (df = 102)
F Statistic	117.561^{***} (df = 10; 102)
Note:	*p<0.1; **p<0.05; ***p<0.01 Linear regression with interaction-only result