TBD*

TBD

Yingying Zhou, Yang Wu, Xinyi Xu, Hong Pan

16 March 2021

Abstract

With the growing ubiquity of deepfake technology, here comes one more channel for the spread of misinformation. We reproduce Soubhik Barari (2021) paper and identify several additional demographic traits contributing to the persuasion power of deepfake on political defamation through its material effect on feeling manipulation in a RCT field experiment on the 2020 Presidential election. We find little evidence that deepfake videos have a unique ability to shift viewers' perceptions on politicians. Apart from partisanship, education, and other socioeconomic factors identified by the original paper, we discover that income and internet usage would also significantly impact the candidate's affective appeal to the news feed viewer.

Contents

*Code and data are available at: https://github.com/yangg1224/Political_Deepfake_Videos.git.

1 Introduction

2 Data

2.1 EDA

2.1.1 treat distribution

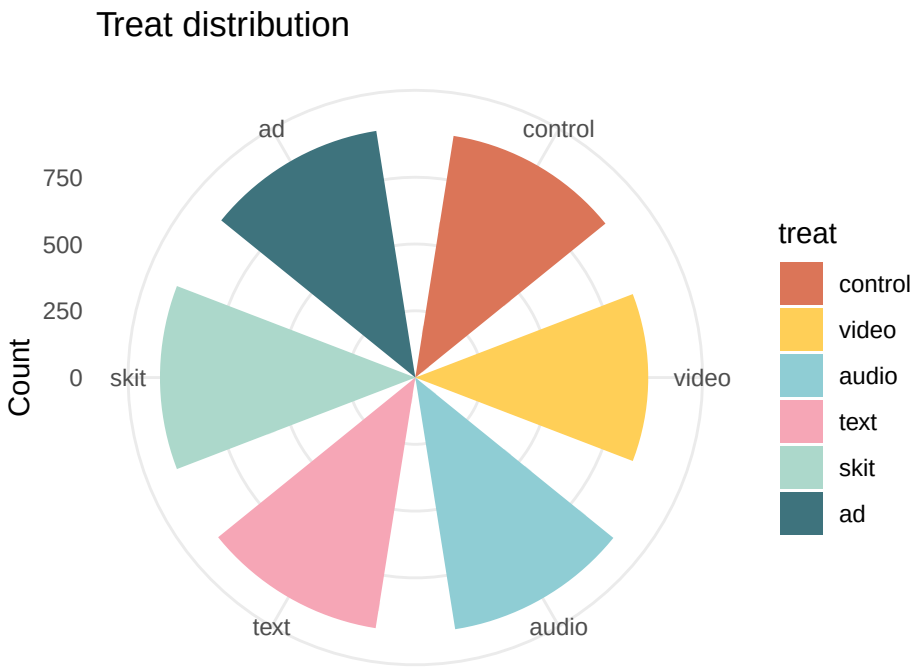


Figure 1: Employee numbers distribution

2.1.2 education level distribution by PID

2.1.3 sexism by education level

2.2 internet usage frequency by education level

2.2.1 post favor by treat

3 Model

$$\hat{Y} = \hat{\beta}_0 + \sum_{a=male}^{other} \hat{\beta}_a * Gender_a + \sum_{b=High\ school}^{Postgraduate} \hat{\beta}_b * educ_b + \sum_{c=HHI25K\ -HHI49K}^{HHI>150K} \hat{\beta}_c * HHI_c + \sum_{d=Ethnicity:white}^{Ethnicity:black} \hat{\beta}_d * Ethnicity_d$$

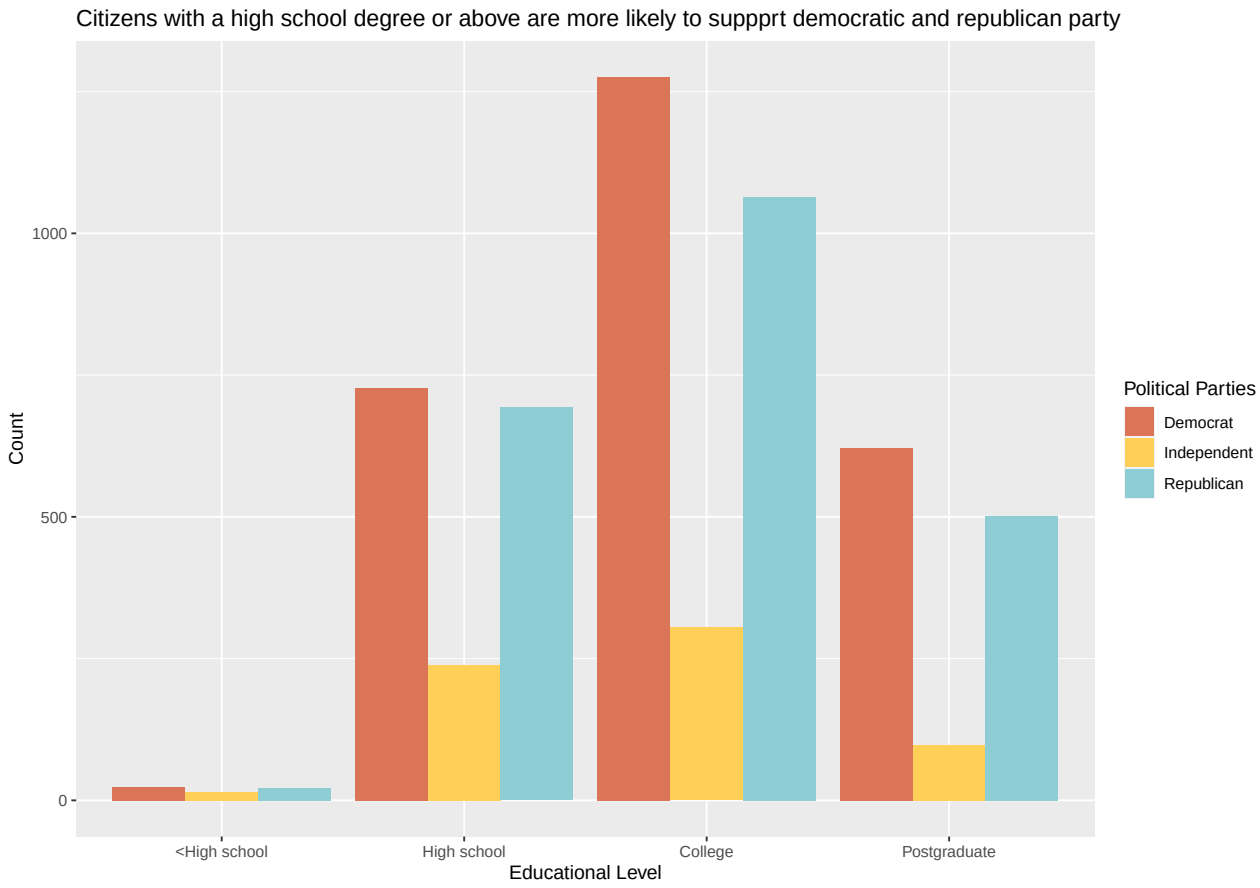


Figure 2: Educational level by PID

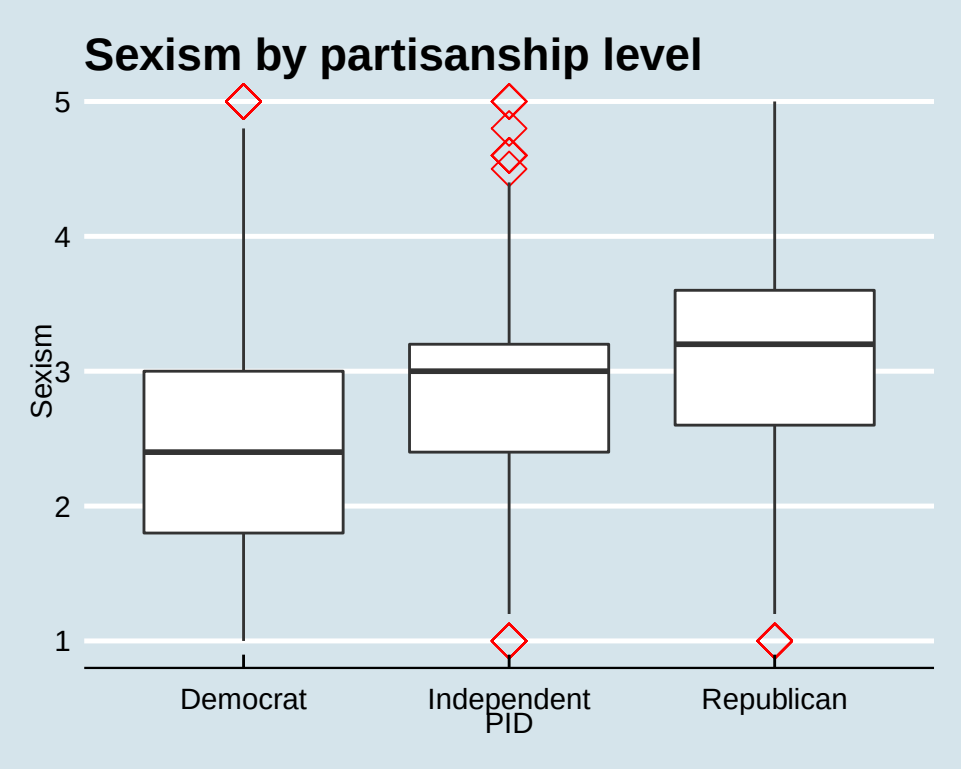


Figure 3: sexism by education level

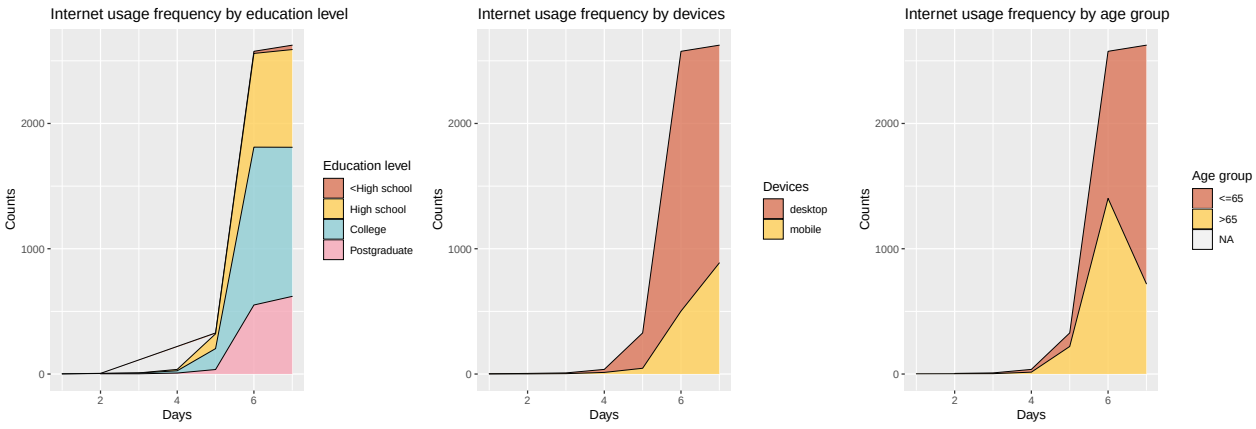


Figure 4: internet usages

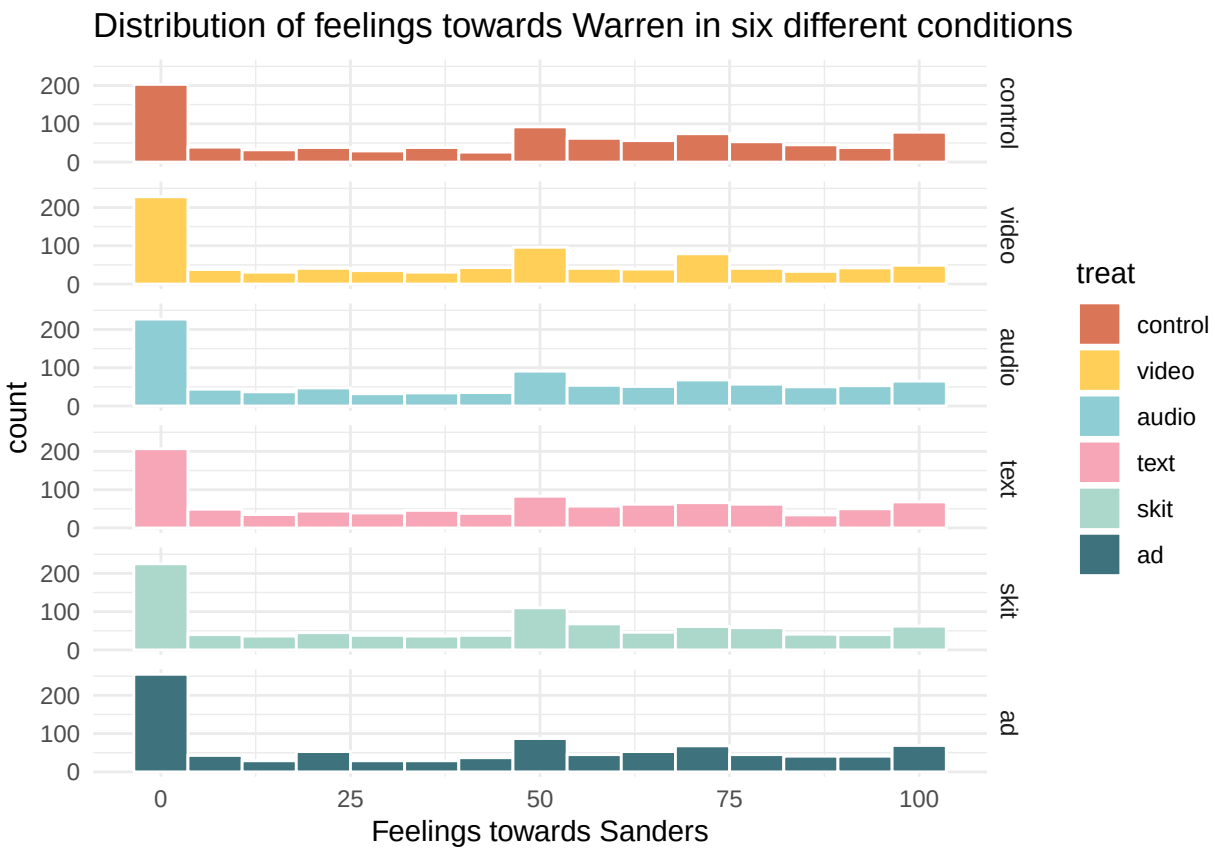


Figure 5: Distribution of feelings towards Sanders in six different situations

Table 1: Average deception level of each media format

treat	Average Deception Level
video	3.228438
audio	3.351178
text	3.305946
skit	2.569316
ad	2.989059

$$+\sum_{e=Region:Northeast}^{Region:West}\hat{\beta}_e*Region_e+\hat{\beta}_f*I_{WaveID}+\hat{\beta}_g*I_{Meta:OSmobile}+\hat{\beta}_h*I_{Age>65}+\sum_{i=PID:Independent}^{Republican}\hat{\beta}_i*PID_i$$

$$+\hat{\beta}_j*Ambivalent\ Sexism+\hat{\beta}_k*Polknow+\sum_{l=treat:video}^{treat:ad}\hat{\beta}_l*Treat_l+\sum_{m=script:bidenshit}^{script:lgbtq}\hat{\beta}_m*Script_m+\hat{\beta}_n*I_{exp\_1\_prompt:\ info}$$

$$+\hat{\beta}_o*post\_dig\_lit+\hat{\beta}_p*Internet\_usage+\hat{\beta}_q*CRT+\epsilon$$

4 Result

4.1 Two-Sample T-test results

4.1.1 Deception Level

In the descriptive analysis aspect, Table 1, shown below, illustrates the average deception level of each media format. The result shows that although deepfake videos have an average deception level of 3.23 out of 5, it is lower than the average level of audio(3.35) and text(3.30). Audio has the highest average deception level, and skit (2.57) has the lowest average deception level.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

In the statistical analysis aspect, unpaired two-sample t-tests were applied to test whether deepfake videos are statistically different from other media formats at the deception level. The results from Table 2 to Table 4 show that only the difference in the deception level between video and skit is significant ($p<0.01$). The p-values for comparing video and text, video and audio are larger than 0.05, which means there is not sufficient evidence to support that video is different from the audio or text. In other words, videos do not differ from audio or text significantly.

Table 2: T test: Deception level of video vs audio

AVG_deception_video	AVG_deception_audio	p.value	conf.low	conf.high	method	alternative
3.228438	3.348243	0.0538155	-0.2415774	0.0019682	Welch Two Sample t-test	two.sided

Table 3: T test: Deception level of video vs text

AVG_deception_video	AVG_deception_text	p.value	conf.low	conf.high	method	alternative
3.228438	3.304207	0.2244956	-0.1980699	0.0465321	Welch Two Sample t-test	two.sided

Table 4: T test: Deception level of video vs skit

AVG_deception_video	AVG_deception_skit	p.value	conf.low	conf.high	method	alternative
3.228438	2.574586	0	0.5024785	0.8052267	Welch Two Sample t-test	two.sided

4.1.2 Affect Level

The unpaired two-sample t-tests were utilized to investigate whether there is a different emotional impact on the target elite between deepfake videos and other conditions, including different deepfake formats and control groups that have no clip at all.

The result of comparing the deepfake video and the control group in Table 5 demonstrates that the video condition will cause a negative sentimental effect from respondents to Elizabeth Warren. The 95% confidence interval shows that the true difference in means is between -1.35 and -7.72. Given the p-value less than 0.05, the difference between the two groups is significant.

Table 5: T test: Affect level of video vs control

AVG_affect_video	AVG_affect_control	p.value	conf.low	conf.high	method	alternative
41.27797	45.81395	0.005278	-7.721219	-1.350748	Welch Two Sample t-test	two.sided

Similarly, the same analysis has been done for the rest unpaired two-sample t-tests. In our study, we used a 5% significance level. Table 6 shows the test result of how deepfake videos and texts impact audiences' feelings. The result shows that the difference in affect level between video (Mean = 41.28) and text (Mean=44.22) was not significant given the p-value is greater than 0.05.

4.1.3 T test2

Table 6: T test: Affect level of video vs text

AVG_affect_video	AVG_affect_text	p.value	conf.low	conf.high	method	alternative
41.27797	44.2234	0.0652461	-6.077025	0.1861569	Welch Two Sample t-test	two.sided

4.1.4 T test3

Table 7 shows the test result of how deepfake videos and audios impact audiences' feelings. The difference in affect level between video(Mean = 41.28) and audio (Mean=43.93) was not significant given the p-value is greater than 0.05.

Table 7: T test: Affect level of video vs audio

AVG_affect_video	AVG_affect_audio	p.value	conf.low	conf.high	method	alternative
41.27797	43.92593	0.0997404	-5.80127	0.5053586	Welch Two Sample t-test	two.sided

4.1.5 T test 4

Table 8 shows the test result of how deepfake videos and skit impact audiences feeling. The difference in affect level between video(Mean = 41.28) and skit (Mean=43) was not significant given the p-value is greater than 0.05.

Table 8: T test: Affect level of video vs skit

AVG_affect_video	AVG_affect_skit	p.value	conf.low	conf.high	method	alternative
41.27797	43	0.2772717	-4.829684	1.385625	Welch Two Sample t-test	two.sided

4.2 Feature selection

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following object is masked from 'package:here':
##
##     here

## The following object is masked from 'package:ggpubr':
##
##     mutate

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

4.2.1 Model Results

Multiple linear regression was applied to use the top 20 important explanatory variables generated from the random forest model to predict the affect level.

As shown in the summary table, only 9 variables are significant ($p<0.05$) to the affect level to Elizabeth Warren which are postgraduate education level is; income ranging in $100k to $150k; living in the Northeast of US; whether age older than 65; partisan; ambivalent sexism; and treat is advertisement.

In this model, the intercept represents the average affect level for the reference group which includes the following characteristics:

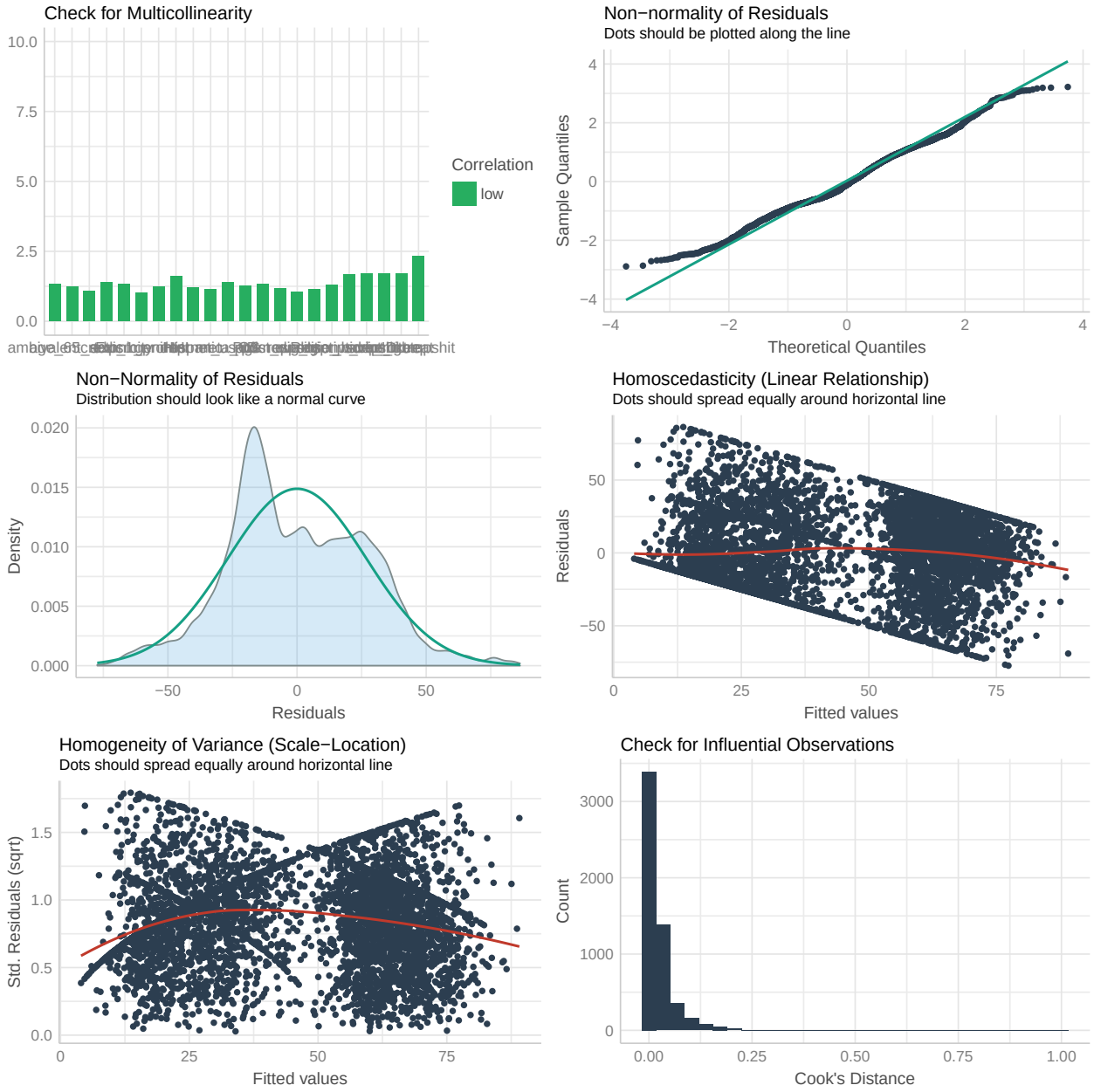
- Gender: Female
- Education Level: lower than high school degree
- Income: Less than $25K
- Ethnicity: Asian
- Region: Midwest
- Response_wave_ID: SV_OxlqWIOfO10wuYI
- Device: Desktop
- Age group: Less than 65 years old
- Partisan: Democrat
- Media condition: Control
- Prompt: Control

```
## Loading required namespace: qqplotr
```

```
## For confidence bands, please install ‘qqplotr‘.
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
## ‘geom_smooth()‘ using formula ’y ~ x’
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



The coefficient for Postgraduate is 9.32, suggesting that the average affect level from people whose education level is postgraduate is on average 9.32 units higher than people whose education level is less than high school level, holding other variables constant.

The coefficient for HHI$100k to $150k is 3.90, suggesting that the average affect level from people whose income are in the range of $100k to $150k is on average 3.90 units higher than people whose income are less than $25k, holding other variables constant.

The coefficient for HHI > $150K is 4, suggesting that the average affect level from people whose income are in the range of $100k to $150k is on average 4 units higher than people whose income are less than $25k, holding other variables constant.

The coefficient for RegionNortheast is 3.89. The average affect level from people who live in the Northeast in the US is 3.89 units higher than people who live in Midwest in the US, holding all other variables unchanged.

The coefficient for people older than 65 is -4.36. When other variables are unchanged, the average affect level from people who order than 65 is average 4.36 units lower than those less than 65.

The coefficient for the variable, PIDIndependent, is -26.69. Holding other variables constant, the Independent gives the feeling scores are on average 26.69 units lower than the scores given by the Democrats.

The coefficient for the variable, PIDRepublican, is -39.52. Holding other variables constant, the average feeling score from the people who are Republican is on average 39.52 units lower than the people who are Democrats.

The coefficient for Ambivalent Sexism is -4.06, suggesting that one unit increase in ambivalent sexism score is associated with 4.06 units decrease in affect level, holding other variables constant.

The coefficient for treataudio is -2.82, suggesting that the average affect level from people whose media condition is audio is on average 2.82 units lower than people whose media condition is control, holding other variables constant.

The coefficient for treatskit is -2.85, suggesting that the average affect level from people whose media condition is audio is on average 2.85 units lower than people whose media condition is control, holding other variables constant.

The coefficient for treated is -3.84, suggesting that the average affect level from people whose media condition is advertisement is on average 3.84 units lower than people whose media condition is control, holding other variables constant.

The coefficient for Internet Usage is 1.12, suggesting that one unit increase in internet usage is associated with 1.12 units increase in affect level, holding other variables constant.

4.2.2 Model Assessment Results

The summary table above shows that the R-square value for the multiple linear regression model is 0.39, which means about 39% of the variation in the dependent variable(affect level) can be explained by the multiple linear regression model.

In addition, in the Scale-Location plot (Figure 7 in Appendix), an approximately horizontal line is shown, which means the residuals are randomly distributed and have constant variance. The Normal QQ-plot (Figure 8 in Appendix) shows almost all the residuals match the diagonal line, meaning the residuals are normally distributed. The Residual versus Leverage plot (Figure 9 in Appendix) shows that there is no evidence of outliers, and none of the points come close to having both high residual and leverage.

Table 9: Regression results

	Model 1
(Intercept)	69.100***
	(6.702)
gender.L	-4.347
	(4.871)
gender.Q	-2.659
	(2.837)
educHigh school	-1.218
	(3.601)
educCollege	0.653
	(3.600)
educPostgraduate	9.126**
	(3.688)
HHI$100k-$150k	3.595**
	(1.760)
HHI>$150k	3.903**
	(1.936)
HHI$25k-$49k	-0.920
	(1.040)
HHI$50k-$74k	-0.256
	(1.170)
HHI$75k-$99k	0.104
	(1.149)
HHIN/A	-4.904*
	(2.916)
EthnicityBlack	3.136
	(2.456)
EthnicityOther	0.463
	(2.745)
EthnicityWhite	0.835
	(1.919)
HispanicNot Hispanic	3.483*
	(1.814)
RegionNortheast	3.902***
	(1.105)
RegionSouth	0.460
	(0.988)
RegionWest	0.967
	(1.168)
response_wave_IDSV_eyxdeXOuISXzakt	-1.304
	(0.944)
qualitypassed all quality screens	-4.604***
	(1.172)
meta_OSmobile	-1.115
	(0.987)
age_65>65	-4.246***
	(0.851)
PIDIndependent	-26.707***
	(1.221)
PIDRepublican	-39.633***
	(0.857)
ambivalent_sexism	-4.100***
	(0.476)
polknow	1.050
	(1.724)
treatvideo	-2.574*
	(1.549)
treataudio	-2.805*
	(1.527)
treattext	-1.655
	(1.553)
treatskit	-2.802*
	(1.525)
treatad	-3.722***
	(1.272)
script_bidenshit	-1.526
	(1.405)
script_trumpshit	-2.016
	(1.393)
script_cherokee	-1.099
	(1.399)
script_lgbtq	0.774
	(1.400)
script_loans	NA
	()
exp_1_promptinfo	0.716
	(0.730)
post_dig_lit	-3.640
	(2.772)
internet_usage	1.123*
	(0.581)
crt	-1.161
	(1.638)
Num.Obs.	5468
R2	0.388
R2 Adj.	0.384
AIC	51567.1
BIC	51838.0
Log.Lik.	-25742.561
F	88.289

* p < 0.1, ** p < 0.05, *** p < 0.01

5 Results

6 Discussion

6.1 First discussion point

6.2 Second discussion point

6.3 Third discussion point

6.4 Weaknesses and next steps

A Appendix

A.1 missing value

```
## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep

## Warning in plot.aggr(res, ...): not enough vertical space to display frequencies
## (too many combinations)
```

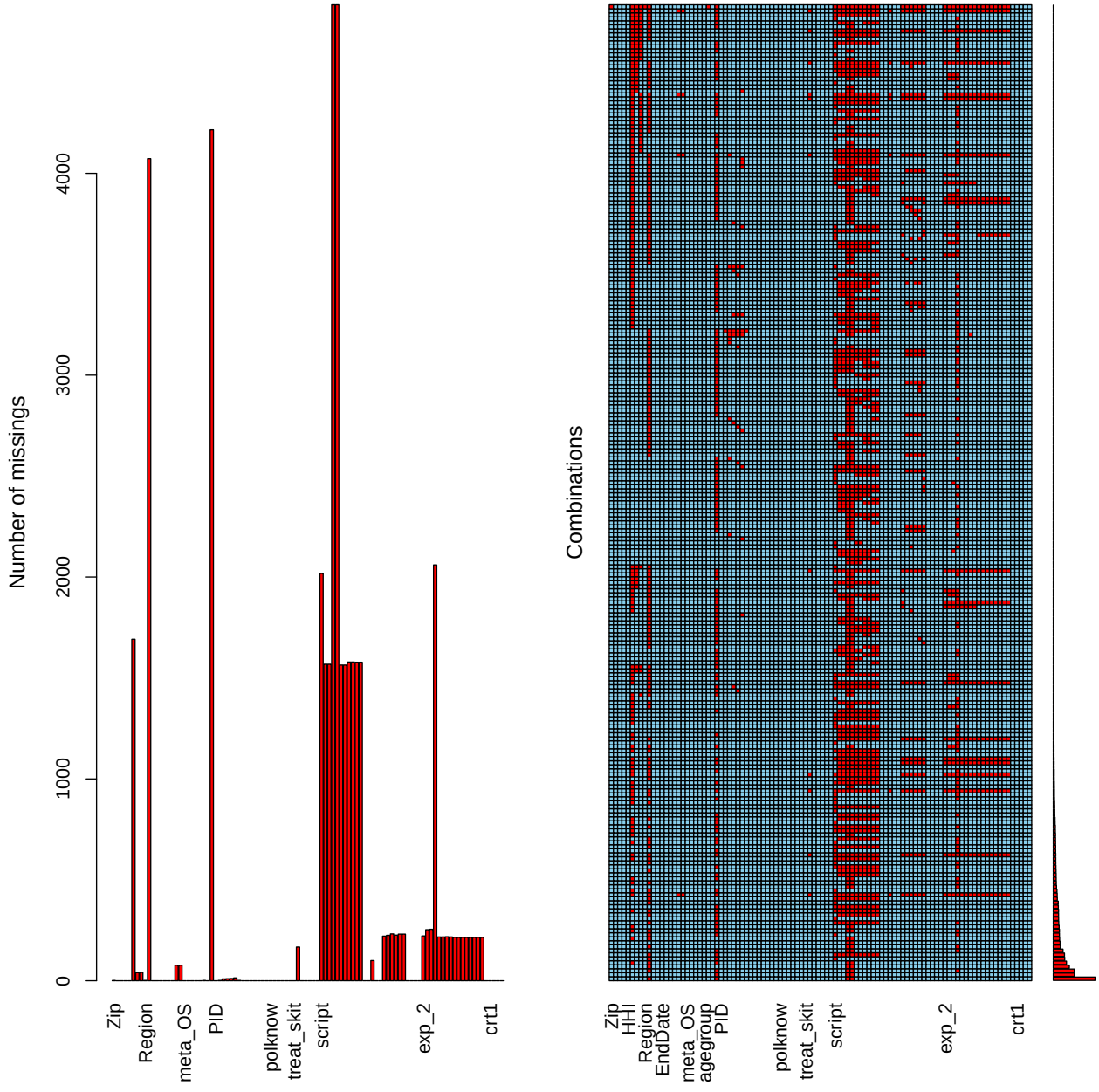


Figure 6: Missing value Visualization

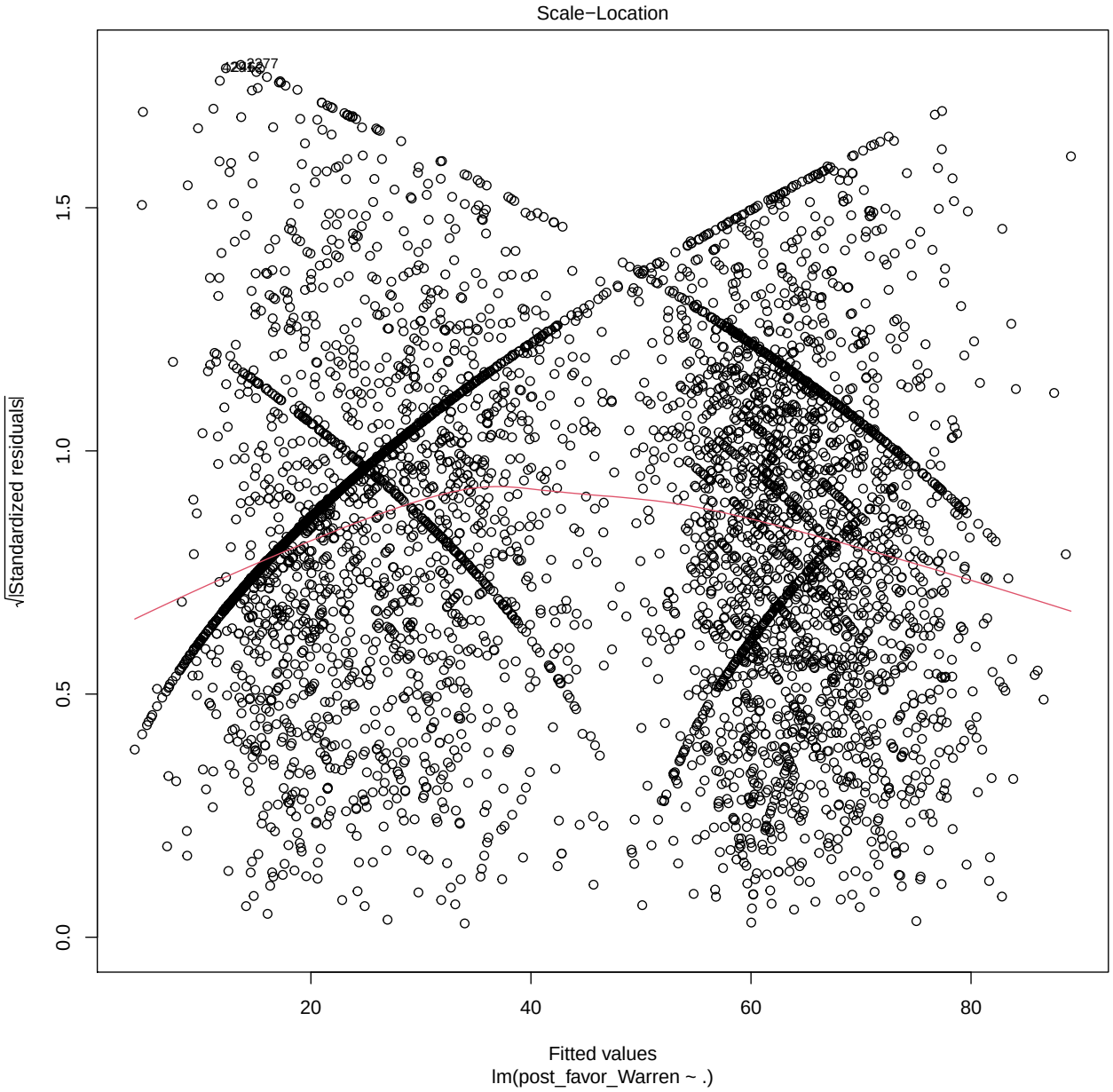


Figure 7: Scale-Location plot

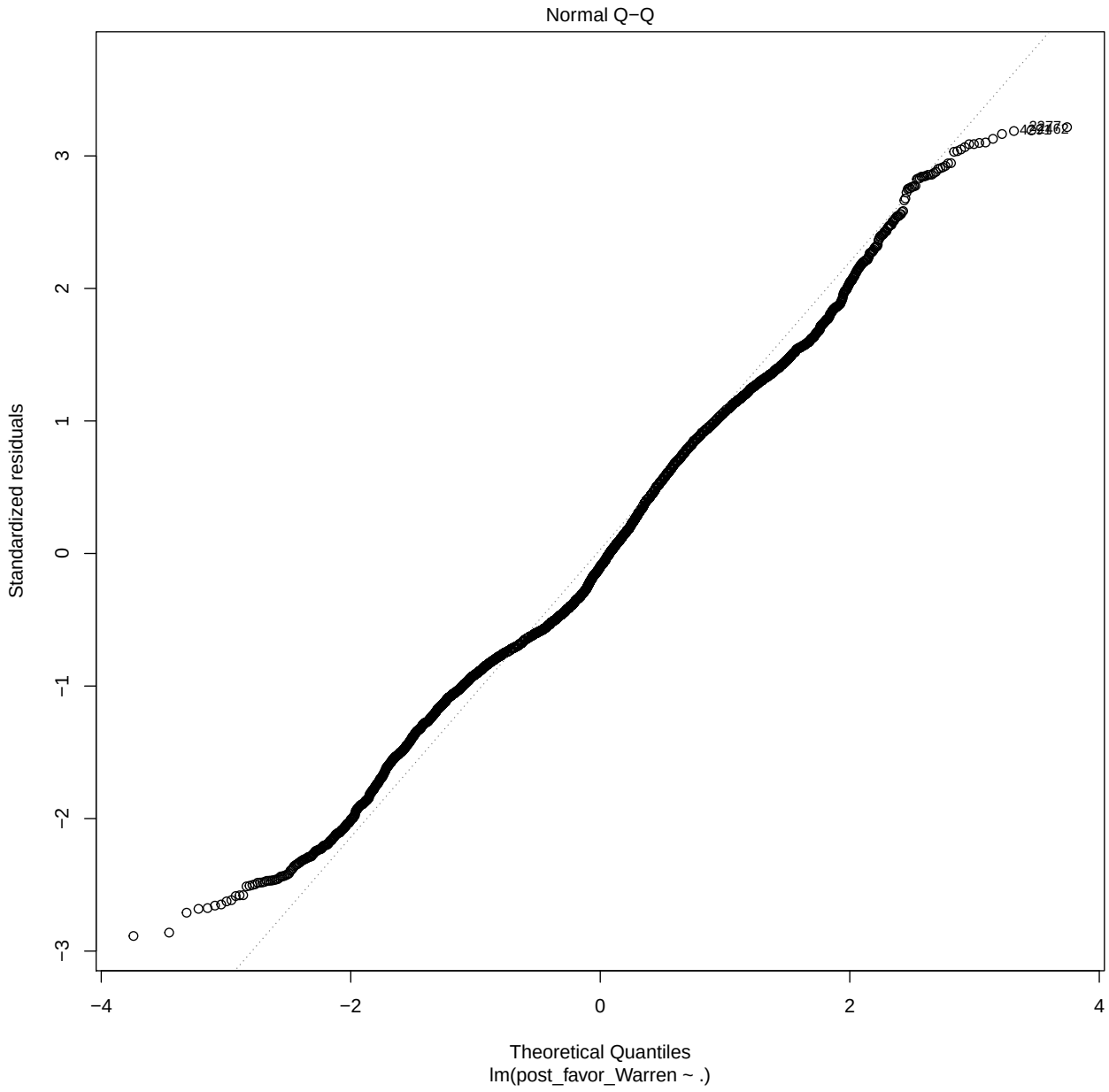


Figure 8: Normal QQ-plot

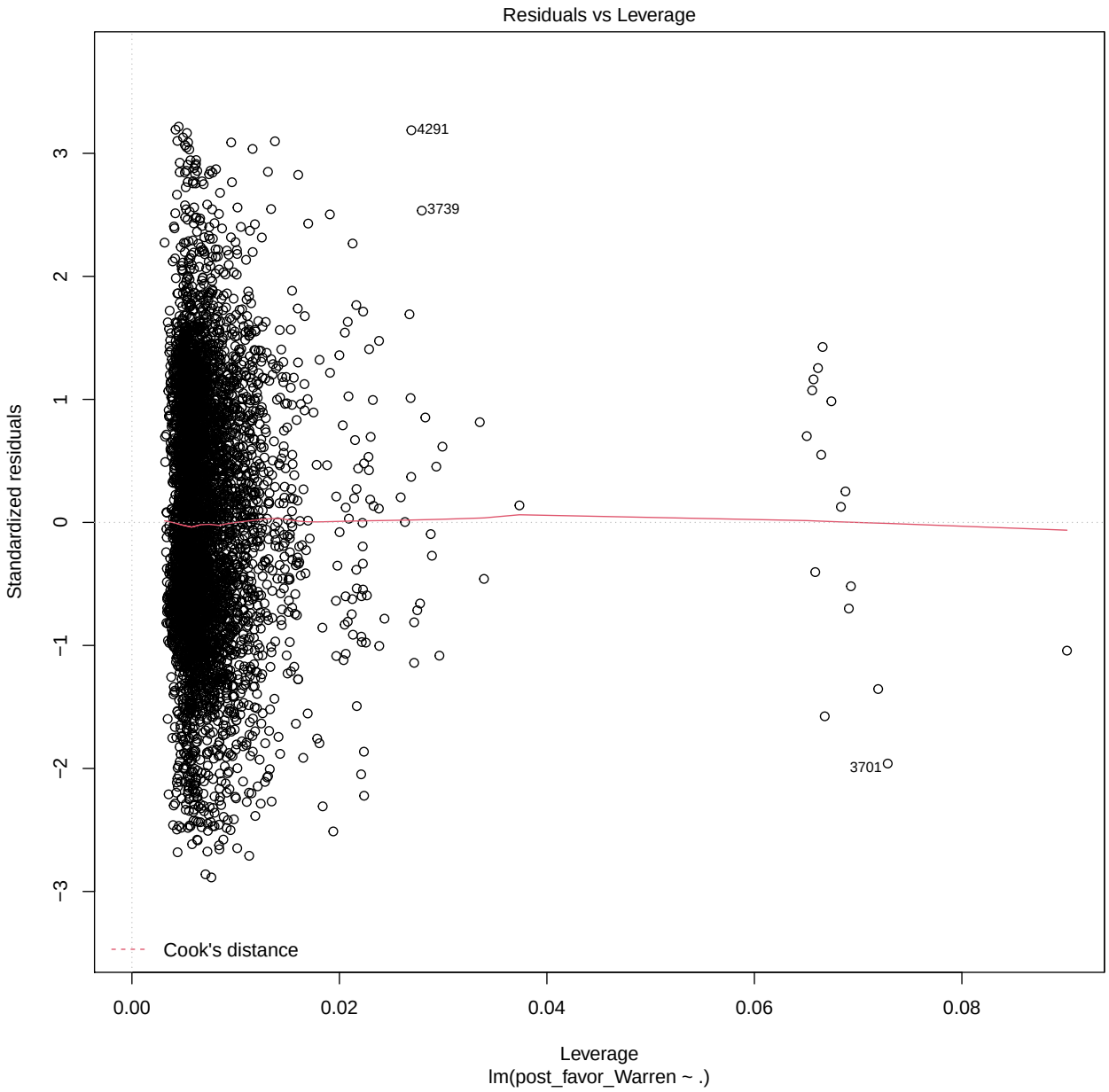


Figure 9: Residual versus Leverage plot

B References