# TBD*

TBD

Yingying Zhou, Yang Wu, Xinyi Xu, Hong Pan

**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

Political misinformation stems the spread of proper knowledge and threatens the electorate's ability to evaluate politicians' public credibility (Jerit and Zhao 2020). Since the advent of popular deep learning technology that enables the fabrication of deepfake videos, concerns have been aroused about their defamation effect on the target politicians.

Soubhik Barari (Barari, Lucas, and Munger 2021) examines the news feed intervention through two Randomized Controlled Trial field experiments on the 2020 U.S. Democratic presidential election to investigate the persuasion power of deepfake. In the first experiment, Soubhik Barari (Barari, Lucas, and Munger 2021)exposes participants to one of five different forms of false media treatment including deepfaked videos and compares viewer's deceptive level and affective response. In the second experiment, Soubhik Barari (Barari, Lucas, and Munger 2021) measures the viewer's ability to discriminate between a feed of 8 real and fake videos. Soubhik Barari (Barari, Lucas, and Munger 2021) finds that the marginal deceptive and affective effects of a deepfaked political scandal are insignificant relative to other false media. In addition, subgroups aged above 65, with partisanship, holding ambivalent sexist views, being highly educated, constrained by cognitive resources, or with little political knowledge are more susceptible to deepfake misinformation.

Using replication dataset and code provided by Soubhik Barari (Barari, Lucas, and Munger 2021), we re-implement the first experiment to further study the persuasion power of deepfake through the channel of its affective manipulation on viewer's political perception. We review the media heterogeneity via t-tests between alternative false media with the deepfake video. We analyse the heterogeneity in affective response by viewer traits with a multiple linear regression model. Feature selection is performed via the Random Forest classifier to select the top 20 relevant variables into our model input.

Our findings about differential effects by media are consistent with Soubhik Barari (Barari, Lucas, and Munger 2021) that the degree of detriment from deepfake video is no different from false audio or text in terms of deception and affective impression. For the linear regression statistical inference, in addition to the socioeconomic features identified by the original paper (age, party identification, sexism, cognitive ability, and education), we find that demographic traits of household income, internet usage, and region play an important part in influencing viewer's affective response to the involved politician as well. Ethical concerns about deepfake experiments are addressed by an extensive post-experiment debrief to educate participants about deepfake media. During modelling design, location information is proxied by region and zip code was excluded from the model to avoid privacy leakage. Internal validity is ensured by stratified sampling and randomized intervention in the experiment as well as having controlled confounders and sufficient sample

size for modelling. The external validity of the study is enhanced by the use of customized deepfake contents and running two waves of experiment in different time periods. Due to the limitation in the topic itself, U.S. presidential election, the research result cannot be generalized to outside of the U.S.

Deepfake technology has the potency to sway viewer's belief on their political understanding and in turn exert real impact in altering viewer's sentiment towards the targeted politician. Consequently, deepfake and other false media can dispute politician's credibility and even manipulate election results. Given that heterogeneity in affective response is observed for subgroups with different levels of internet usage and cognitive reflection, which are intervenable characteristics in information processing for viewers, we can reduce deepfake's deceptive potential by intervening via information provision and promoting public digital literacy.

The remainder of the paper is constructed as follows. Section 2 describes the dataset, experiment design, and exploratory data analysis on feature visualization. Section 3 outlines the reproduced experiment models, which is designed to discover relationships between features and the target variable. Section 4 summarizes the model results according to evaluation criteria. Finally, Section 5 discusses our research findings and provides directions for future research.

We replicate this paper using R statistical programming language (R Core Team 2020). In particular, we use packages `tidyverse` (Wickham et al. 2019), `here` (Muller 2020), `ggpubr` (Kassambara 2020) to manipulate data and packages, `kableExtra` (Zhu 2020) to generate tables, and `ggplot2` (Wickham 2016), `ggthemes` (Arnold 2021) to adjust diagrams themes.

### 1.1 Intervention

## 2 Data

### 2.1 Dataset features

The paper intends to replicate the experimental report by Soubhik Barari(Barari, Lucas, and Munger 2021), Christopher Lucas, and Kevin Munger. We are going to analyze the dataset collected by their experiment. The survey dataset records 5,750 observations with 100 variables. In this case, we focus on the top 20 most important features related to the favorability towards the Democratic politician Elizabeth Warren. It covers viewer traits on demographic, socioeconomic status, and political related positions that are immutable or intervenable characteristics to deepfakes' misinformation. The authors hypothesized several risky subgroups that may be differentially susceptible to deepfakes. That included the following categories of media consumers:

- Age Group
- Directional motivated reasoning
- Evaluation driven by negative stereotypes
- Constraints on cognitive resources or knowledge
- Political knowledge
- Digital literacy

In particular, they divide respondents into two `age groups`: below 65 and above 65. `Ambivalent Sexism` evaluates negative stereotypes towards women on a scale of 1 to 5. `PID` shows the party identification of Republican, Independent, or Democrat. Polknow is scaled to be within 0 to 1, reflecting respondents' political knowledge that might influence public attitudes and opinions. `Treat` records one of the six treatment conditions in the experiment: deepfake video, false media in audio, text, skit, campaign attack ad, or control. `Exp_1_prompt` collects information about the intervention: received information and received no information on deepfake prior to the newsfeed experiment. `Internet_usage` aims to collect information about the frequency of using the internet per week. `Meta_OS` indicates the platforms of internet access,

namely either mobile devices or desktop. The dataset also includes additional demographic information such as gender, income and education. `Educ` identifies subjects into four categories by educational level including <High school, High school, College, and Postgraduates. Similarly, `HHI` divides participants into six groups on the basis of their income including <$25k, $25k-$49k, $50k-$74k, $75k-$99k, $100k-$150k, and >$150k. In addition, they recorded the `post_favor_Warren`, ranging from 0 to 100 as the effective response to Warren after subjects receive the news feed treatment.

## 2.2 Population, frame, and sample

The sampling population included all citizens in the US. A total of 17,501 national representatives were recruited in the survey experiment on the Lucid survey research platform. To qualify for the experiment, all observations should complete quality checks which consist of randomly dispersed attention checks and technology checks. 629 respondents failed the front-end pre-treatment attention check; in order words, the gender or age that they entered did not match up with the demographic characteristics. Finally, only 5,750 of 17,501 subjects passed a series of quality checks and completed the survey experiment.

## 2.3 Experiment in the paper

They first created a collection of realistic deepfake with industry partners. At the early stage of the experiment, some of the participants received a brief informational message reminding them of the existence of deepfake. During the experiment, all participants obtained a news feed in a natural environment setting where the experience of scrolling on the Facebook news feed was replicated. They watched or listened to posts about Elizabeth Warren, one of the 2020 democratic primary candidates. With the fixed order and the content of these media, the news feed scenario design primarily minimized the influence of alternative characteristics other than the intervention on the final results. Participants accessed two conditions before and three conditions after watching a video. During the experiment, participants were randomly exposed to one of the five different fictitious scandals, namely that respondents were able to watch or listen to five possible defamation strategies: incivility toward an in-party member, incivility toward an out-party member, a past controversy, a novel controversy, or political insincerity. They chose Elizabeth Warren to be the target for the political scandal of the 2020 democratic election because she is a salient politician who is more valid than those politicians with low profiles. Also, she is not slated for re-election until 2024 which minimized the influence of the 2020 election. Finally, as a female candidate, she is more likely to be involved in non-political deepfakes.

## 2.4 What is good and bad about the data

The experiment is very well-designed in experiment treatment settings. To receive a rich answer to the experiment, they conducted the experiment in a realistic setting. Deepfake was combined with authentic campaign news to replicate a natural news-browsing experience. They provided a series of realistic deepfake videos with a sufficient sample size. In terms of the external validity, the deepfakes used in the experiment are high-quality face-swap videos of a single elite depicted in a number of slightly different scandals that they have produced to be maximally deceptive. Since the experiment involved five media conditions, it can also be used for analyzing the interaction effect across a number of comparable media. In terms of how to gauge the persuasion power of deepfake, they isolated the measurement of deception and affect from a single clip. Furthermore, they compared the attitudinal effects of a single deepfake to its related textual, audio, and un-deepfake video counterparts in order to investigate the relative influence deepfake has on political misinformation compared to other traditional media. One of the major drawbacks is that the original survey dataset consists of too many missing values. According to Appendix A, more than 40 variables involve missing data. To be more precise, there are 15 attributes containing more than 1,000 missing values such as script, believed_funny, believed_offensive, and PID_presurvey, four of which have over 4,000 incomplete data such as comments and PID_learner. In addition, although the experiment was anonymous, the geographical information such as zip and regions of the participants was still recorded in the dataset.

## 2.5 Methodology

They used the Lucid survey research platform which is one of the world's biggest survey panels. It is a platform aiming to provide data-driven knowledge in the workplaces. It delivers powerful answers, inspired by the sentiments of real people (Lucid 2021).

As for the treatment conditions, the experiment randomly paired informational messages about deepfake to one of six conditions. The six conditions included video, audio, text, skit, campaign attack ad, and control which is no clip. In particular, the audio condition consists of the audio recording of the actor making a scandalous announcement. The video condition exploits a deepfake created by the face-swap algorithm in the skit condition. The text condition only keeps the title and subtitle describing the event captured on video. A skit displays an audio recording accompanying a video of the actor impersonates a campaign event in a realistic setting. Campaign attack advertisement subjects are exposed to a real negative campaign ad title, which is an actual campaign stimulus used in the primary election to activate negative emotions towards Warren. The control means no clips presented.

## 2.6 Sampling and experiment approach

According to the article, the authors used a post-stratification sampling method to correct demographic skews in the sample. Post-stratification refers to the weights being adjusted so that the weighted totals equal to the known population totals. They applied post-stratification weights estimated from the US Census in aspects of education, age, household income, gender, race, and hispanic, and consequently compared the demographic traits of their sample with the demographic trait in the most recent Current Population Survey (CPS). A weighted regression was implemented to guard against measurement error from possible demographic skews (Matias 2018). To adjust for remaining discrepancies, they generate post-stratification weights via raking to match the CPS marginal population totals.

## 2.7 Intervention

In this experiment, the intervention is receiving a concise informational message about the existence of deepfakes prior to the news feed treatment. The authors would like to analyze if receiving an information prompt about deepfakes before the news feed will affect the reliability and trustworthiness of all fake news clippings Through this experiment, participants had the opportunity to detect fake media, enclosing a debrief which demonstrates how the deepfake process works. We believe that this experiment will help to educate the public about the growing ubiquity of deepfakes and the potential threat they might pose to information processing for the general public. By understanding how deepfakes function from their low-cost interventions, citizens can effectively prevent the harm caused by the misinformation.

## 2.8 EDA

### 2.8.1 Treat distribution

According to the experiment, there are six treatment conditions including deepfake video, audio, text, skit, campaign attack ad, and control. From Figure 1 , we can see that the six types of treatments are equally sampled. In other words, in the process of exposure, all media have the same chance of being viewed.

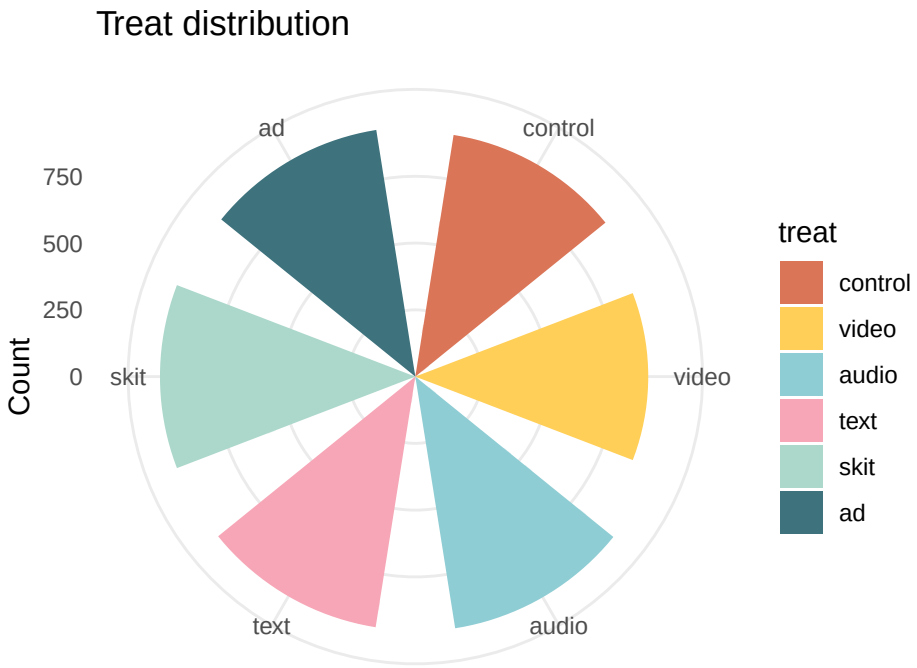



Figure 1: Employee numbers distribution

### 2.8.2 Education level distribution by PID

The grouped bar chart (Figure 2) below shows the distribution of political parties for four educational levels. We notice that most participants identified themselves as Democrats, followed by the Republicans. Therefore, there are grounds to consider Democrat and Republican as the two major political parties in American society. Among the citizens with college degrees, more than 1,250 citizens identify themselves as Democrats and over 1,000 citizens have a disposition to Republican.

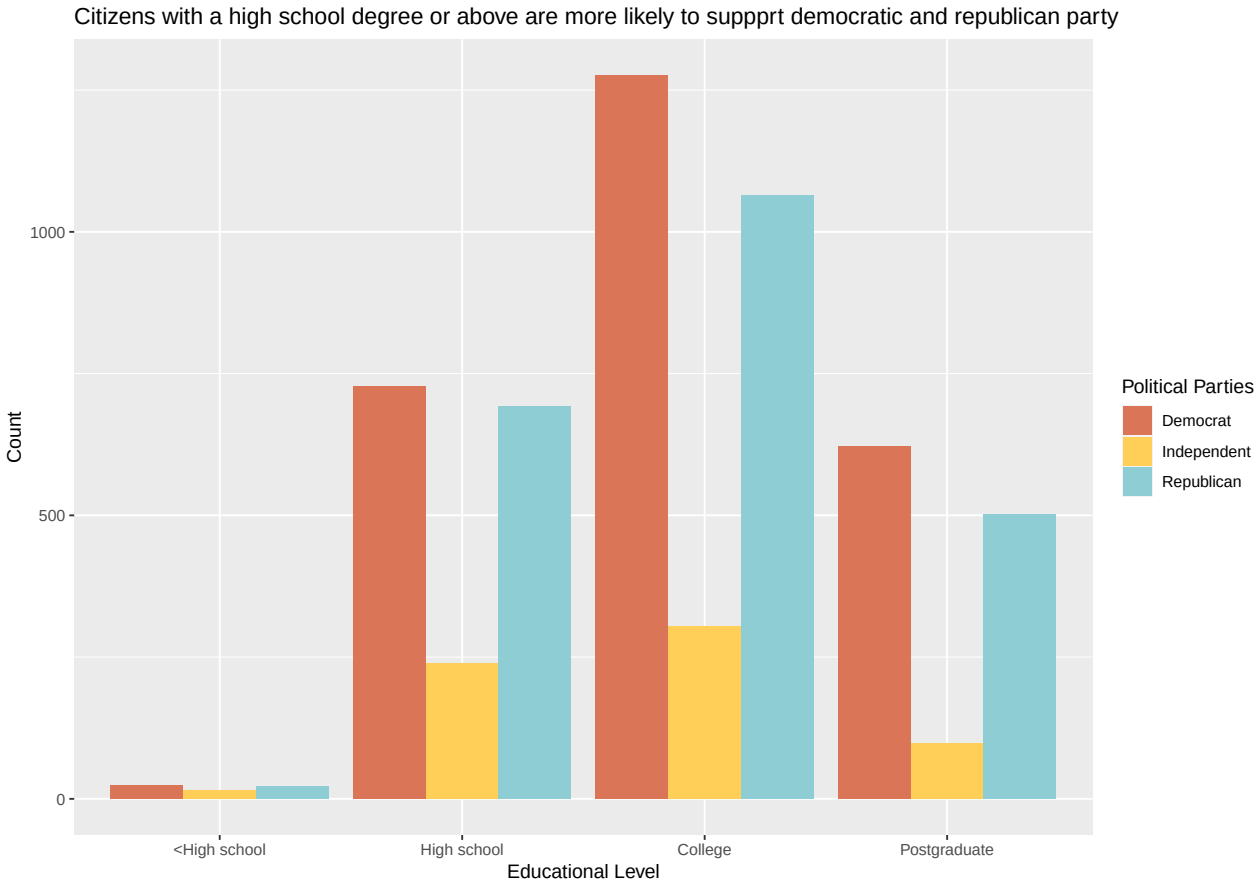



Figure 2: Educational level by PID

### 2.8.3 Sexism by education level

Figure 3 uses the box plot method to show sexism by partisanship. As we can see from the chart, voters in the Republican party have the most number of sexists, which also explains one potential reason why their support rate is relatively lower for Warren. Next comes the Independent party. People's sexism level is approximately level 3. At last, people in the Democratic party mostly hold the fairest attitude about sexism.

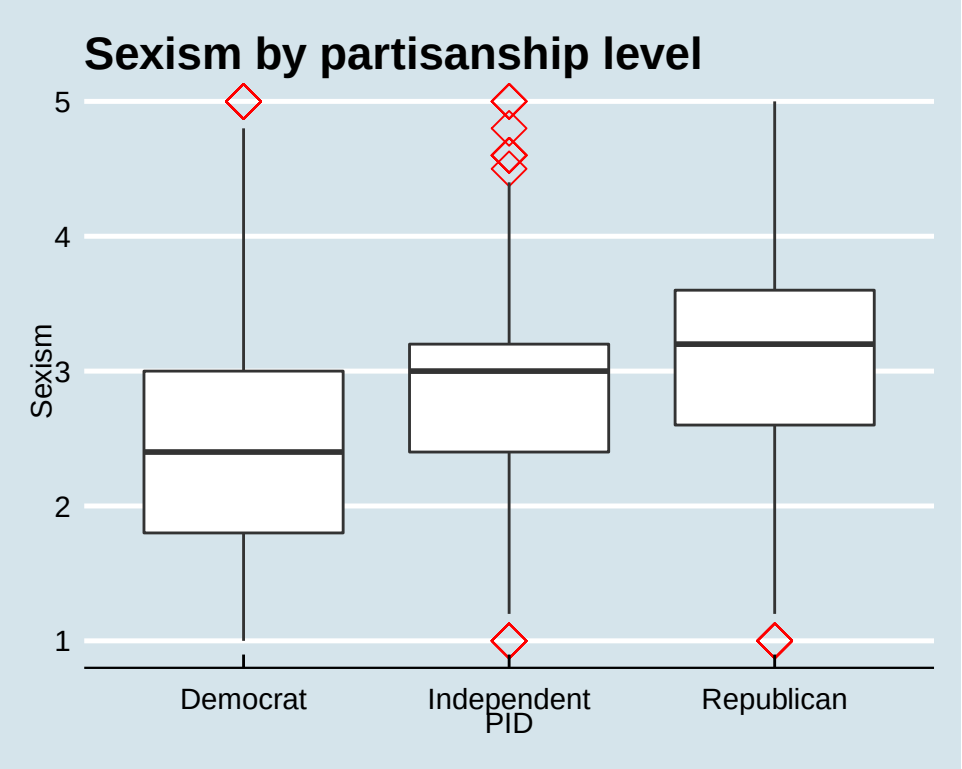



Figure 3: sexism by education level

### 2.8.4 Internet usage frequency by education level

Figure 4 demonstrates internet usage frequency by educational level, devices, and age groups respectively. Participants with college education level are the most frequent internet users (more than 5 days per week), followed by people with high school education and people with postgraduate education. In addition, the dataset also shows desktop devices are more popular than mobile devices in terms of how participants' access the internet. People younger than 65 will use the internet more often than those older than 65.

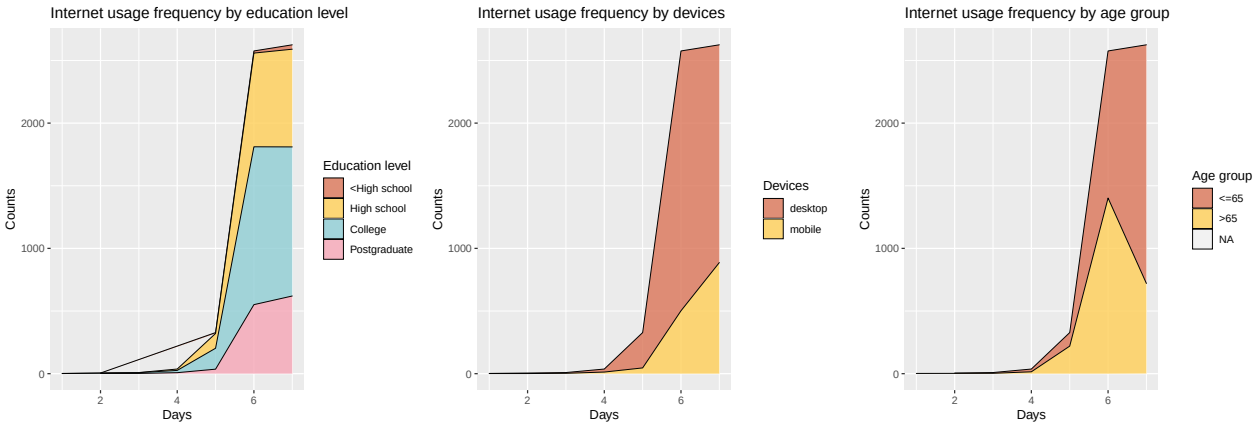



Figure 4: internet usages

#### 2.8.5 Post favor after treatments

Figure 5 illustrates the distribution of feeling towards Elizabeth Warren after receiving different news feed treatments. We realize that there is no significant difference in the distribution of affective response. The x-axis stands for the affective response towards Elizabeth Warren, ranging from 0 to 100. No matter what treatment subjects were exposed, a certain number of participants rated their feeling towards Warren as 0, along with a number of people who voted 50. It is worth mentioning that deepfake videos do increase the negative sentiment towards Elizabeth Warren compared to the control group.

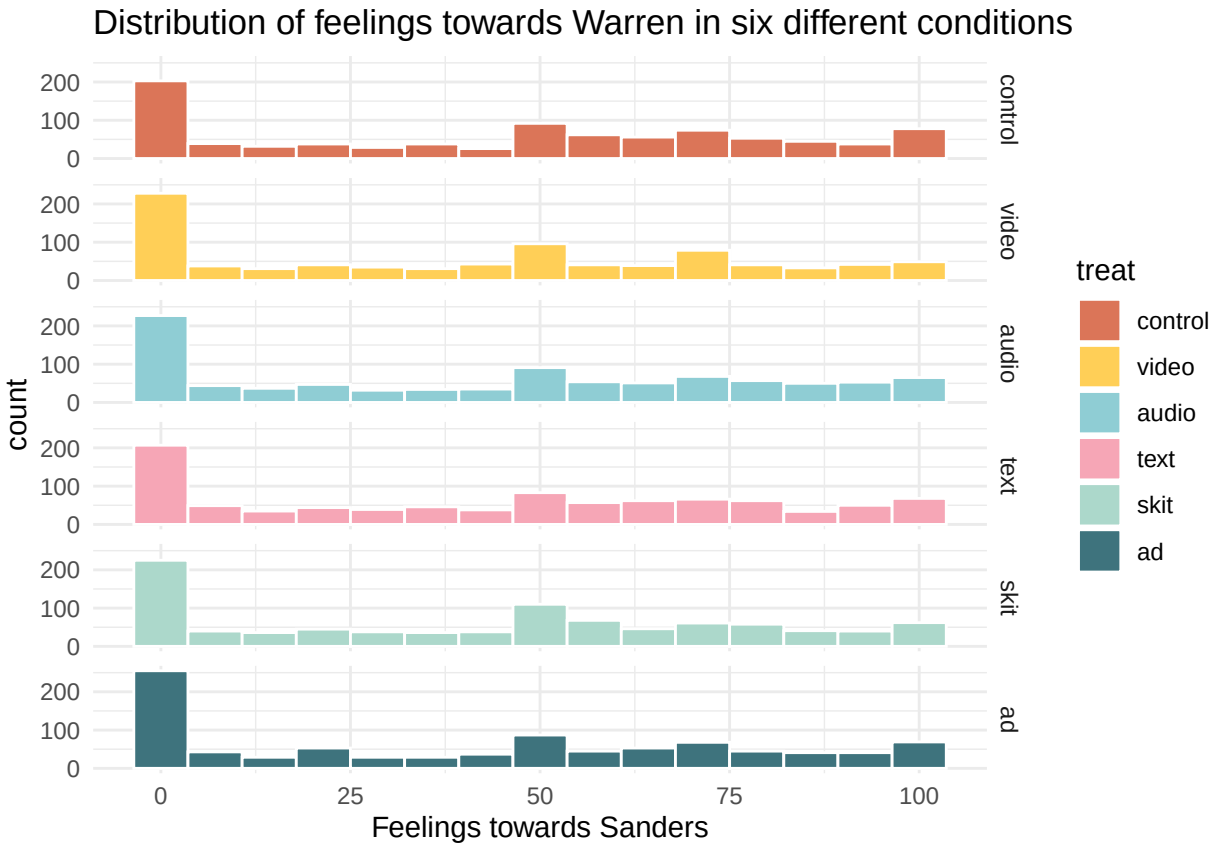



Figure 5: Distribution of feelings towards Sanders in six different situations

# 3 Model

## 3.1 Feature Selection

Before building our model, feature selection is conducted on the selected dataset variables. In general, feature selection has the following advantages:

- Enhance the understanding between features and feature importances
- Reduce features overfitting and features dimensionality to improve the model accuracy

Feature importance plays a vital role in predictive modelling projects. It can provide insights on data, models, and how to reduce dimensionality and select features, thereby improving efficiency and effectiveness of the model prediction.

Random Forest is the chosen method for our feature selection model because it can handle multi-dimensional data and output feature importance according to maximized information gain in node splitting. The random forest algorithm forms a series of classification methods that rely on a combination of several decision trees. According to the guideline from Liaw, the random forest algorithm is optimized via three parameters in our project: we set ntree=500 (the number of trees), mtry=all the variables in FS dataset divided by 3 (the number of predictors randomly tested at each node) and node size=5 (the minimal size of the terminal node).

The slip plot [crossref] shows the feature selection results. As we can see, `PID` partisanship is the most vital variable to affect the sentiment score towards Warren. As we know, partisanship is one of the most important factors affecting voters' voting. In us, there is a large group of fanatic political supporters. They don't care about who is the candidate, or what those candidates' ideas are. They just vote for the right party. Warren is a representative of the Democratic Party, so voters who are optimistic about the Democratic Party are willing to give higher sentiment scores. On the other hand, people who are in the Republican party and the Independent party would not support Warren.

Next important variable comes to `sexism`. Lovenduski mentioned that "Women remain significantly underrepresented in political life." A big reason behind this is that people still associate toughness and leadership qualities with men. "Female marginalisation is hardwired into the traditional institutions within which politics takes place." Of course, things have two sides. American feminists also let Warren have a large number of loyal female supporters.

We are interested to find the variable "HHI" (household income) plays an important role in Warren's sentiment score. Before we go to see the regression model result, we are guessing people with higher annual income care more about the candidate's attitudes towards economic development. Because they might have their investment in different industries which will be affected. In addition, income is usually correlated with education level. People with higher income tend to have more general knowledge and thus are less likely to be fooled by deepfake.

The last two top five variables which have high feature importance are `age over 65` and `political knowledge`. The year 65 is like a threshold. While younger people would have more open ideas, older people usually have more conservative political ideas and slow to follow technology trends. Therefore, they might be more susceptible to false information fabricated by new technology such as deepfake. Older people generally stand in different positions and hold different expectations, therefore their sentiment score might vary a lot.

Finally, we decided to drop variables with lower than feature importance value of 0.5 to reduce dimensions for the regression model. So, `quality`, `script_loans`, and `hispanic` would be dropped and the remaining 20 features will go into the regression model.

## 3.2 Multiple Linear Regression

We are using RStudio to run the multiple linear regression model. The following reasons explain why we decide to choose multiple linear regression:

- It account for all of the potentially important factors in one model
- It leads to a more precise understanding of the relationship between dependent variables and independent variables
- It is able to identify outliers or anomalies in the dataset.

One-hot encoding needs to be implemented to transform categorical textual variables to dummies since multiple linear regression can only take in numeric variables. All these steps are implemented by the [lm] package.
The results of the regression model will be detailed in the next section. We are focusing on three model parameters for variable interpretation and performance evaluation.

R squared :

R-squared is a measure of the goodness of fit of the model. A larger R-squared indicates a closer fit of the model to the data. It is used as an optimality criterion in parameter selection and model selection.

P-value:

P-value is used to describe the occurrence possibility of the extreme outcome when the null hypothesis is true. If the p-value is small, it means that the probability of occurrence of the null hypothesis to be true is very small. And if it does occur, we have a reason to reject the null hypothesis. In short, the smaller the p-value, the more significant the result. Usually the threshold for significant p value is set to 0.05.

Regression coefficient:

The sign of the regression coefficient describes whether there is a positive or negative correlation between each feature variable and the dependent variable (Warren affective score). A positive coefficient means that as the value of the independent variable increases, the average value of the affective score also tends to increase. A negative coefficient indicates that as the independent variable increases, the affective score tends to decrease.

$$\hat{Y} = \hat{\beta}_0 + \sum_{a=male}^{other} \hat{\beta}_1 \cdot Gender_a + \sum_{b=High\ school}^{Postgraduate} \hat{\beta}_2 \cdot educ_b + \sum_{c=25K\ -49K}^{>150K} \hat{\beta}_3 \cdot HHI_c + \sum_{d=white}^{black} \hat{\beta}_4 \cdot Ethnicity_d$$

$$+ \sum_{e=Northeast}^{West} \hat{\beta}_5 \cdot Region_e + \hat{\beta}_6 \cdot I_{WaveID} + \hat{\beta}_7 \cdot I_{Meta:OSmobile} + \hat{\beta}_8 \cdot I_{Age>65} + \sum_{i=Independent}^{Republican} \hat{\beta}_9 \cdot PID_i$$

$$+\hat{\beta}_{10} \cdot Ambivalent\ Sexism + \hat{\beta}_{11} \cdot Polknow + \sum_{l=video}^{ad} \hat{\beta}_{12} \cdot Treat_l + \sum_{m=bidenshit}^{lgbtq} \hat{\beta}_{13} \cdot Script_m + \hat{\beta}_{14} \cdot I_{exp\_1\_prompt:\ info}$$

$$+\hat{\beta}_{15} \cdot post\_dig\_lit + \hat{\beta}_{16} \cdot Internet\_usage + \hat{\beta}_{17} \cdot CRT + \epsilon$$

## 3.3 Two-Sample T-test results

### 3.3.1 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.

Table 1: Average deception level of each media format

| Average Deception Level |
|---|
| 3.135187 |

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

| estimate1 | estimate2 | 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

| estimate1 | estimate2 | 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

| estimate1 | estimate2 | p.value | conf.low | conf.high | method | alternative |
|---|---|---|---|---|---|---|
| 3.228438 | 2.574586 | 0 | 0.5024785 | 0.8052267 | Welch Two Sample t-test | two.sided |

#### 3.3.2 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

| estimate1 | estimate2 | 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.

#### 3.3.3 t-test 2

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

| estimate1 | estimate2 | 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 |

#### 3.3.4 t-test 3

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

| estimate1 | estimate2 | 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 |

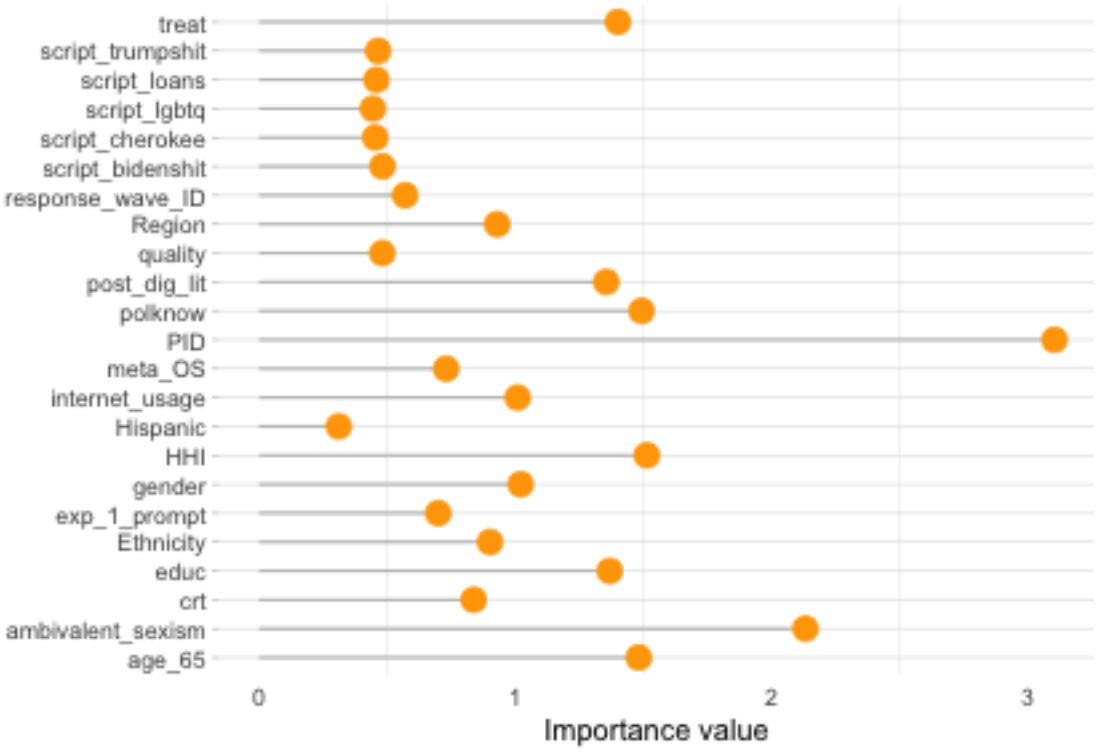
#### 3.3.5 t-test 4

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

| estimate1 | estimate2 | p.value | conf.low | conf.high | method | alternative |
|---|---|---|---|---|---|---|
| 41.27797 | 43 | 0.2772717 | -4.829684 | 1.385625 | Welch Two Sample t-test | two.sided |

### 3.4 Feature selection




#### 3.4.1 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:

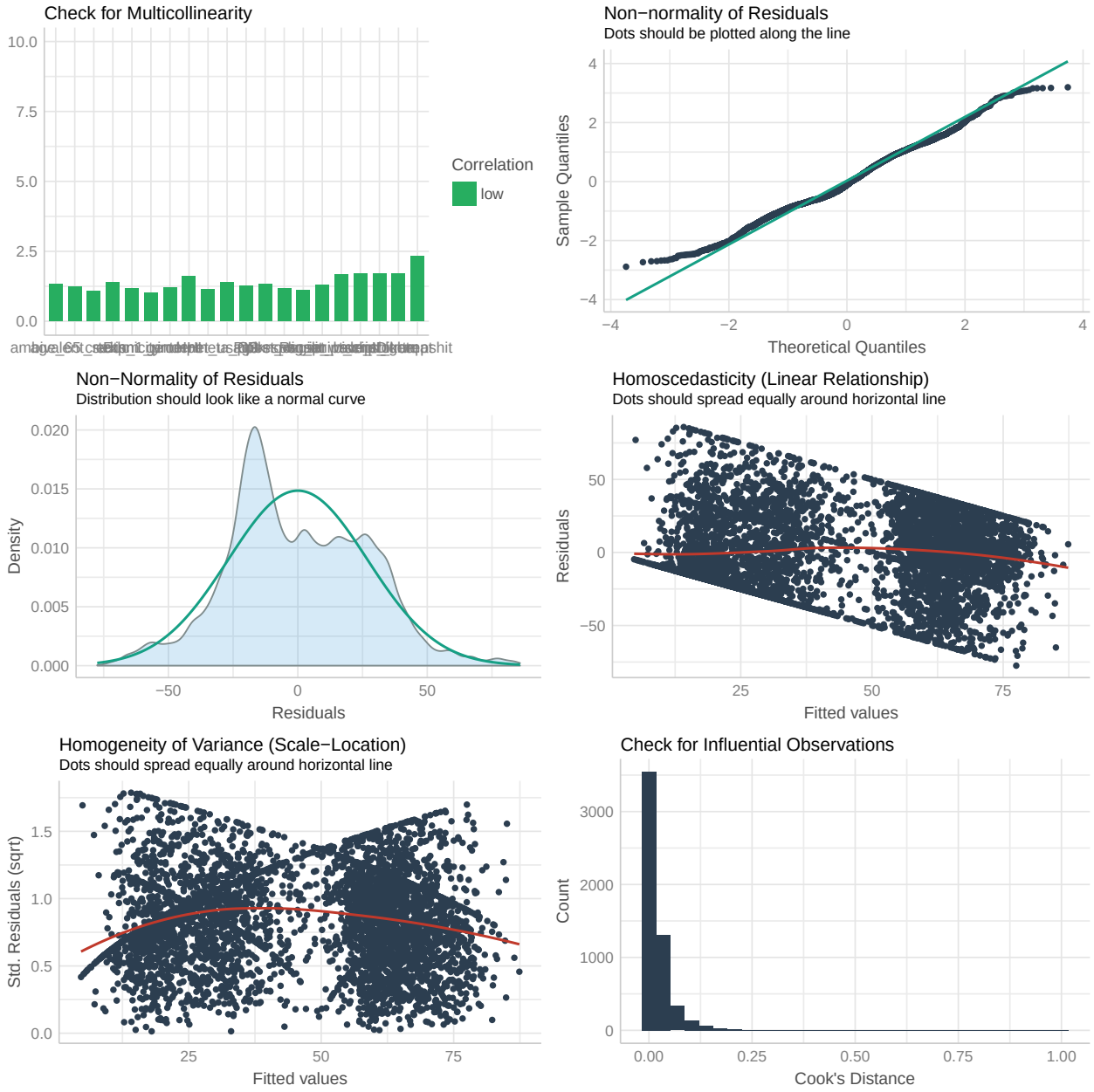
- Gender: Female
- Education Level: lower than high school degree
- Income: Less than $25K
- Ethnicity: Asian
- Region: Midwest
- Response_wave_ID: SV_OxlqWIOfO10wuYI
- Device: Desktop
- Age group: Less than 65 years old
- Partisan: Democrat
- Media condition: Control
- Prompt: Control

```
## 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.

### 3.4.2 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) | 69.872*** |
| | (6.448) |
| gender.L | -1.560 |
| | (4.824) |
| gender.Q | -0.969 |
| | (2.806) |
| educHigh school | -1.207 |
| | (3.606) |
| educCollege | 0.709 |
| | (3.605) |
| educPostgraduate | 9.317** |
| | (3.693) |
| HHI$100k-$150k | 3.904** |
| | (1.757) |
| HHI>$150k | 4.007** |
| | (1.939) |
| HHI$25k-$49k | -0.919 |
| | (1.041) |
| HHI$50k-$74k | -0.208 |
| | (1.172) |
| HHI$75k-$99k | 0.057 |
| | (1.151) |
| HHIN/A | -4.947* |
| | (2.920) |
| EthnicityBlack | 3.113 |
| | (2.459) |
| EthnicityOther | -1.064 |
| | (2.643) |
| EthnicityWhite | 0.563 |
| | (1.921) |
| RegionNortheast | 3.891*** |
| | (1.106) |
| RegionSouth | 0.453 |
| | (0.988) |
| RegionWest | 0.923 |
| | (1.164) |
| response_wave_IDSV_eyxdeXOuISXzakt | -1.306 |
| | (0.945) |
| meta_OSmobile | -1.269 |
| | (0.987) |
| age_65>65 | -4.357*** |
| | (0.847) |
| PIDIndependent | -26.689*** |
| | (1.222) |
| PIDRepublican | -39.521*** |
| | (0.856) |
| ambivalent_sexism | -4.064*** |
| | (0.477) |
| polknow | 1.333 |
| | (1.724) |
| treatvideo | -2.531 |
| | (1.551) |
| treataudio | -2.817* |
| | (1.530) |
| treattext | -1.725 |
| | (1.555) |
| treatskit | -2.850* |
| | (1.527) |
| treatad | -3.838*** |
| | (1.274) |
| script_bidenshit | -1.680 |
| | (1.406) |
| script_trumpshit | -2.132 |
| | (1.394) |
| script_cherokee | -1.145 |
| | (1.401) |
| script_lgbtq | 0.647 |
| | (1.401) |
| exp_1_promptinfo | 0.763 |
| | (0.731) |
| post_dig_lit | -3.950 |
| | (2.775) |
| internet_usage | 1.120* |
| | (0.582) |
| crt | -0.893 |
| | (1.639) |
| Num.Obs. | 5468 |
| R2 | 0.386 |
| R2 Adj. | 0.382 |
| AIC | 51582.2 |
| BIC | 51839.8 |
| Log.Lik. | -25752.077 |
| F | 92.263 |

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

# 4 Discussion

Deepfake technology can be used against candidates to synthesize false videos, which would potentially disrupt the elections. The purpose of the research is to inform the public of the existence of deepfake technology and consequently to improve the resilience of democratic politics to this AI threat.

## 4.1 Bias and ethical concerns

Some modelling bias and ethical concerns are addressed:

### 4.1.1 Privacy problem on zip code

To protect the privacy of respondents in the experiment, detailed address information such as zip code was excluded from the linear regression model. Instead, location was proxied by region. Thereby, it would be impossible to reverse engineer the identity of the respondent by any ill-intended third parties. Furthermore, postal code is a proxy for socioeconomic status, which correlates with HHI (household income), causing multicollinearity and eventually biasing the coefficient (Tu 2005).

### 4.1.2 Choice of candidate - election result disruption & pre-existing gender prejudice

The one-month experiment was run prior to the 2020 U.S. presidential election, to minimize the risk of influencing the election outcome and in the meantime to have discernible effect for the experiment, a high-profile candidate who was not finalisted to be the democratic nominee should be selected. In addition, we would like to test for the impact of pre-existing sexism against women among subjects on the effect of the deepfake. Elizabeth Warren came naturally to be our choice of the candidate who meets all the conditions above.

### 4.1.3 Active debrief - post-experiment potential change in participants' political behavior

At the completion of the survey, an extensive debrief was conducted on all subjects which instructed respondents to affirm that the treatment media about Elizabeth Warren was false. The active debrief was done to educate participants on deepfake media. And thus it would be unlikely for the treated group to change their political behavior later on as a post-exposure consequence.

## 4.2 Model results

### 4.2.1 Heterogeneity in different socioeconomic groups

As for the regression model result on affective response towards Elizabeth Warren, heterogeneous effects in respondents from different socioeconomic groups were observed. All else held constant, on average people with a postgraduate degree significantly favor Warren by 9.3 points more than people with lower than high school education. High-incomers with annual income above $100k favor Warren by 4 points more than people who earn less than $25,000 a year. In contrast, people who were unwilling to disclose their salaries disliked the candidate by 4.94 points compared to the low-incomers. Elderly people aged over 65 deduct feeling marks by 4.35 points towards Warren after being exposed to the deepfake media about Warren.

#### 4.2.2 Impact from partisan motivated reasoning and sexism discrimination

Partisanship accounts for a large part of the sentiment score towards Warren. For people supporting rival parties to the Democratic as the Republican and Independent, they substantially dislike Warren by about 40 and 27 points on average respectively compared with people identifying with the Democratic. Additionally, the map would also justify the positive coefficient for the Northeast region, where Warren is from; she is home based in Massachusetts (Merrilees, Kaji, and Szabo 2020). For people holding ambivalent sexism views against women, a higher level in their hostile attitude would deduct 4 more marks from their feeling scales for the female candidate Warren on average.

#### 4.2.3 Deepfake and other false media, not that different in fooling the public

Although there is significant influence from the deepfake video on altering people's sentiment towards the candidate, the two-sample t-test results between deepfake and other false media assure us that the its degree of detriment is no different from false audio or text in terms of deception and affective impression.

#### 4.2.4 Political knowledge and internet usage

It is worth our attention that compared to political knowledge, internet usage has a more significant effect on respondents' affective impression. Perhaps that the more frequent interest users gain a more holistic understanding of political positions and current trends in the world thus they are more immune to the infamatory deepfake technologies.

### 4.3 Real world implication

Based on the observed effects in affective response from different subject subgroups in the deepfake experiment, we conclude that deepfake can have real impact on the candidate through altering people's perceptions and sentiment on the politicians. Once the participants believe the deepfake video content to be true, it is likely that they would negatively impact their impression on the politician. Elderly people, people holding hostile gender prejudice and those identifying with rival political parties are susceptible to the deepfake information; the impact can be observed through deducted affective score on the candidate. However, for people with a higher education degree, high-incomers, and people from the candidate's home state, they tend to favor the politician more than other groups. Interestingly, political knowledge has less significant influence on participants' affective response to the candidate than internet usage.

Nevertheless, the effect of deepfake videos on the election is no different than other false media such as audio and text in terms of affective manipulation done to tarnish the politician's public image.

With the growing ubiquity of deepfake technology, here comes one more channel for the spread of misinformation. Given the deeply-rooted views in partisan motivated reasoning and gender discrimination, along with uneven distributed soicioeconomic resources for people to access proper knowledge and digital technology, the barriers to a democratic society largely come from people's cognitive characteristics. Therefore, informing the public about the deepfake and digital technology is of top priority to curb the spread of misinformation in terms of deepfake technology.

### 4.4 Potential caveats

Although Warren has long been delisted as a Democratic presidential candidate, she is still a salient politician. During the experiment time window from Sep. 29th to Oct. 29th 2020, Warren delivered a speech in Wisconsin on Oct. 16th to denounce Donald Trump (Glauber 2020). Her public appearance might cause confusion to the experiment participants who happen to know about this news and potentially disturb the deepfake experiment result if the deepfake script content was somewhat related to the actual news.

## 4.5 Internal validity & external validity of model

### 4.5.1 Internal validity

The internal validity of the model can be justified by the procedures of sampling, randomization in experiment intervention, controlling for confounders and having sufficient sample size in the regression model, so that alternative explanations but the treatment can be eliminated from the cause of effect on the outcome.

**Sampling**: Subjects were stratified sampled at a nationally representative weight in this field experiment, and sample's demographic traits were cross-compared with the demographic traits in the most recent Current Population Survey (CPS) to ensure the randomization in sampling which simulates real world demographic distributions.

**Randomization in intervention**: The intervention of whether receiving information about deepfakes before watching the newsfeed treatment was randomized to avoid this knowledge biasing their perceptions and behaviors and thus the experiment outcome.

**Controlling for confounders**: Potential confounders were controlled for in the regression model to avoid the omitted variable bias. For example, education, age, and household income (HHI) are assumed to correlate with deepfake deception and affective appeal through their correlation with digital literacy, internet usage, and political knowledge. Race, gender, and ethnicity correlate with partisanship. These variables were included into the model as predictors for measured affective responses.

**Sufficient sample size**: There were 5,750 subjects participated in the experiment and after cleaning of missing values, 4,131 rows of observations data were fed to the multiple linear regression model. The sample size was sufficient to bring about statistical power in conducting inference on variable significance.

### 4.5.2 External validity

External validity of conclusion on deepfakes was enhanced by the use of 5 customized deepfake contents instead of recycling one well-recognized deepfake example in the treatment experiment. The sampling population is from the U.S. national level, so that the study conclusion and subject trait summary are applicable to the domestic level. The study consists of two waves between Sep. 29th and Oct. 29th. The repetition of experiment conducted at different time periods is to smooth out some possible time effects in the result. However, the model result cannot be generalized to outside the U.S. due to the limitation in the topic itself - U.S. presidential election.

## 4.6 Weaknesses and opportunities for future work

### 4.6.1 Weakness

**Discussion on race and ethnicity**: Despite the extensive investigation into the factors of perception bias in partisanship and gender prejudice on the impact of respondent's sentiment score for the candidate under the deepfake scenario, the paper skips the discussion on race and ethnicity. Race and ethnicity are confounders that correlate with people's party identification and affect people's perception of the candidate as well (McDaniel and Ellison 2008).

**Democratic party only**: The paper chose the candidate to be a politician in the Democratic party, and focuses the topic on the Democratic election. It is possible that the experiment outcome was only valid for the analysis on Democratic party, especially supporter traits on socioeconomic and demographic characteristics.

**Within the U.S. only**: The nature of the U.S. domestic election topic limits the conclusion only applicable to regions within the U.S. on the study of political misinformation impact by deepfake media.

### 4.6.2 Opportunities

The study investigates the impact of deepfake media on political misinformation through two field experiments conducted within a month period through the topic of 2020 Democratic presidential election. The external validity of the study can be further improved in order for the result to be generalizable to other political parties, time periods and situations. **Time window**: for the experiment time window, it can be prolonged or randomized in a series of experiment repetitions as long as they are conducted after the time point when the in-party presidential nominee has been finalized.

**Candidate from other parties**: The same experiment could be repeated using candidates from other parties (Republican, or Independent) to test the robustness and generalizability of the model result.

**Bayesian modelling**: The persuasion effect of deepfake on politics misinformation is measured through two channels, deception level and affective score. When interpreting the regression result on respondent's affective response to the candidate in the deepfake experiment, the assumption is made that the viewer forms their belief on the media content before they change their mind on the involved candidate. We assume that the direction of change in affective score in the experiment is caused by whether the viewer believed the newsfeed content. A negative coefficient indicates that they were fooled by the content and thus further dislike Warren based on their belief in the false content. Bayesian inference modelling ("Bayes Regression: How Is It Done in Comparison to Standard Regression?" 2016) can be used to incorporate the conditional probability of whether the viewer believed the media content to be true into the analysis on affective response.

## 4.7 Differences and difficulties

### 4.7.1 Differences

**Feature selection**: Before building the multiple linear regression model, feature engineering was used to visualize feature importance and select the top 20 most important features from a total of 100 variables with the Random Forest classifier (Menze 2009). The tree-based algorithm splits nodes based on how well they improve the purity of the node through maximizing information gain. The most informative features are at the start of the trees, and less relevant features are places at lower nodes. This process naturally ranks the importance of features, therefore it can be used for feature selection.

**Additional features**: Apart from the treatment variable and regressors describing demographic and socioeconomic traits (age, education, gender, party identification, political knowledge, metrics on internet usage, .etc) in the original model, the random forest algorithm identified household income (`HHI`), region, script type, and ethnicity as relevant features. Income confounds with deepfake deception and affective appeal through its correlation with digital literacy and internet usage. Region and ethnicity are correlated with partisanship. The five types of script contents were included to control for any differential effect of persuasion ability by deepfake.

**A different target variable, affective appeal**: The original model emphasized more on the study of deceptive level of deepfake technology. The persuasion power of deepfake can be measured through the channel of deception into believing and negative affect towards the politician. Furthermore, it is the consequence of the change in people's affective response that influences the election outcome. Therefore, this paper decided to study the deepfake persuasive ability through the second channel, its impact on a candidate's affective appeal.

Based on the regression model result of adjusted R-squared, the model fit outperformed the regression on deceptive levels from the original paper (38% vs. 8%). Our model can explain more of the variations in a candidate's affective appeal to a great extent.

### 4.7.2 Difficulties

**Missing values**: Every variable in the survey data was cleaned for missing or extreme values. Some columns have a large proportion of missing values and thus were removed. As for columns with a few missing records,

these rows were omitted.

**Outliers and model fit**: When inspecting the residual vs. leverage plot obtained from the regression result, some extreme cases and influential points exist in the data, which might skew the overall linear model fit. To improve the model fit, a non-linear regression model such as random forest regression or Xgboost can be considered.

# A Appendix

## A.1 Appendix A

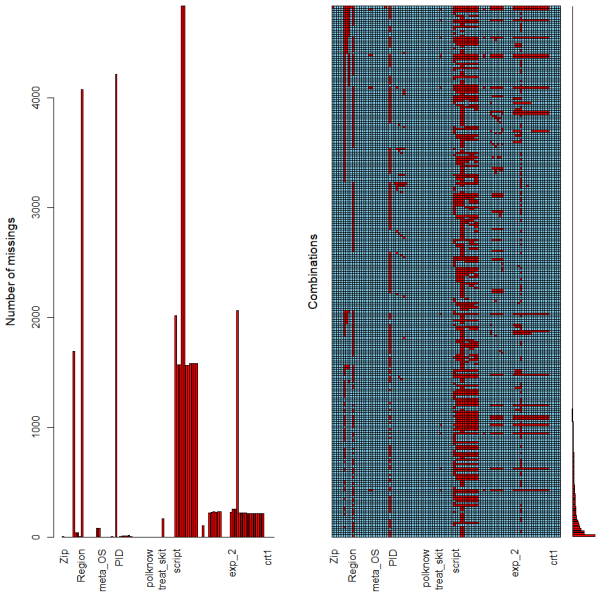



Figure 6: Missing value Visualization

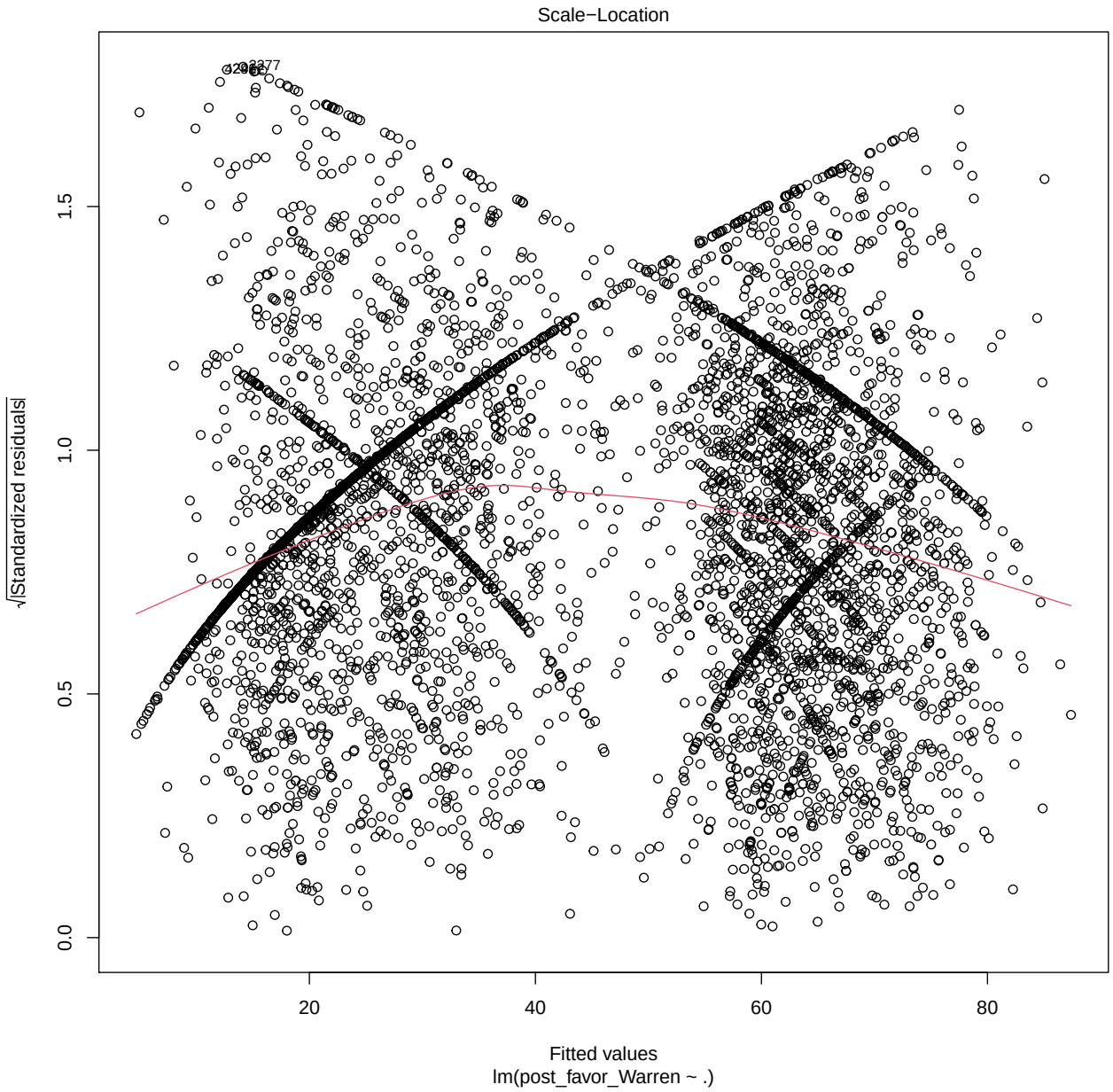



Figure 7: Scale-Location plot

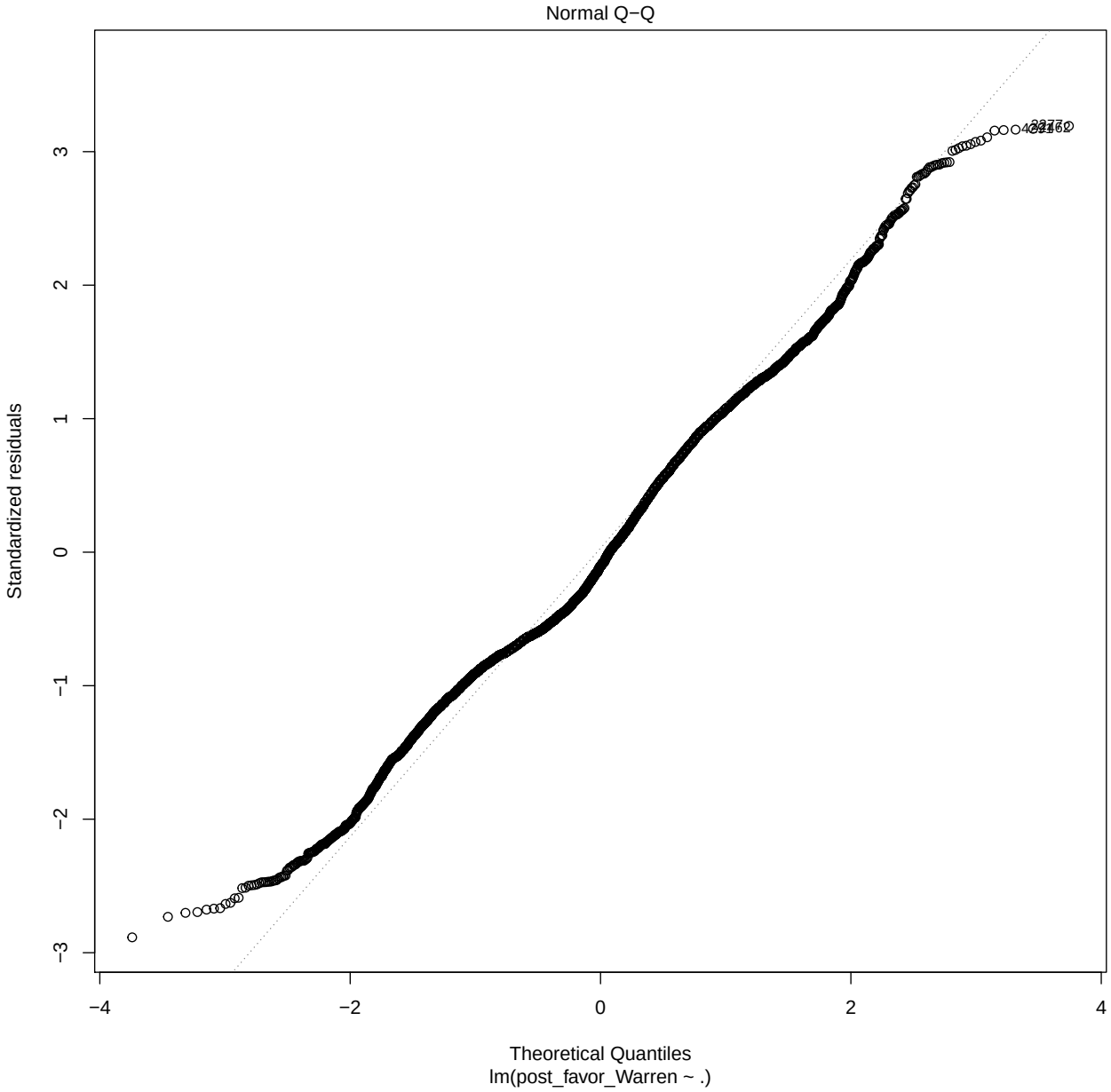



Figure 8: Normal QQ-plot

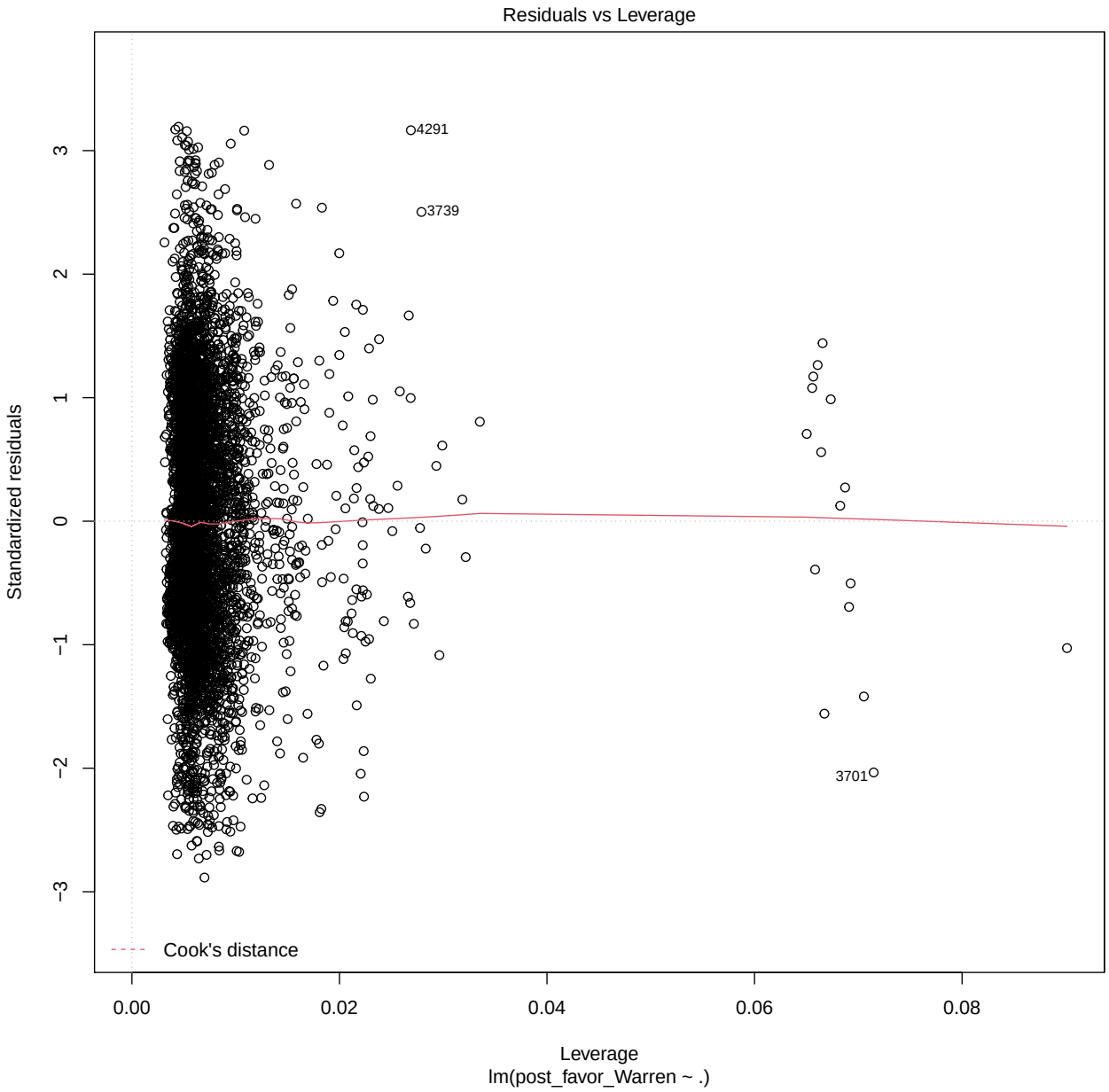



Figure 9: Residual versus Leverage plot

# References


Arnold, Jeffrey B. 2021. *Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'.* https://CRAN.R-project.org/package=ggthemes.

Barari, Soubhik, Christopher Lucas, and Kevin Munger. 2021. "Political Deepfake Videos Misinform the Public, but No More Than Other Fake Media." OSF Preprints. https://doi.org/10.31219/osf.io/cdfh3.

"Bayes Regression: How Is It Done in Comparison to Standard Regression?" 2016. StackExchange. https://stats.stackexchange.com/questions/252577/bayes-regression-how-is-it-done-in-comparison-to-standard-regression.

Glauber, Bill. 2020. "Elizabeth Warren Charges into Wisconsin, Saying It's Time to Hold Trump Accountable." *Milwaukee Journal Sentinel*. jsonline. https://abcnews.go.com/Politics/elizabeth-warren-senator-massachusetts/story?id=60522652.

Jerit, Jennifer, and Yangzi Zhao. 2020. "Political Misinformation." *Annual Reviews*. https://www.annualreviews.org/doi/10.1146/annurev-polisci-050718-032814.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots.* https://CRAN.R-project.org/package=ggpubr.

Lucid. 2021. "Research Technology for Sampling and Media Measurement 2021." https://luc.id/.

Matias, J. Nathan. 2018. "The Obligation to Experiment." *Medium*. MIT MEDIA LAB. https://medium.com/mit-media-lab/the-obligation-to-experiment-83092256c3e9.

McDaniel, Eric L., and Christopher G. Ellison. 2008. "God's Party? Race, Religion, and Partisanship over Time." *Political Research Quarterly* 61 (2): 180–91. https://doi.org/10.1177/1065912908314197.

Menze, Kelm, B. H. 2009. "A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data." *BMC Bioinformatics* 10 (213). https://doi.org/10.1186/1471-2105-10-213.

Merrilees, Mina Kaji, and Christine Szabo. 2020. "Elizabeth Warren: Everything You Need to Know About the 2020 Presidential Candidate." abc News. https://abcnews.go.com/Politics/elizabeth-warren-senator-massachusetts/story?id=60522652.

Muller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tu, Kellett, YK. 2005. "Problems of Correlations Between Explanatory Variables in Multiple Regression Analyses in the Dental Literature." *British Dental Journal*, no. 199: 457–61. https://doi.org/10.1038/sj.bdj.4812743.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.