# SOCI5012 ANALYSING DATA PROJECT: REPLICATION OF PIPPA NORRIS (1996)

Ben Bridge

06/01/2022

Word Count: 2,800

## INTRODUCTION

The paper chosen for this report is – Does Television Erode Social Capital? A reply to Putnam, by Pippa Norris (1996) – and it has been chosen as I have an interest in voter behaviour and its effects on the formation of society. The paper uses data from 'The American Citizen Participation Study 1990' this measures the level of political activity amongst people, drawing on a sample of a sub-sample of around 2,500 of an original 15,000. The paper looks to explore Putnam's assertion that higher media use has led to decreased 'social capital', which can be understood as a decline in contribution to collective active politics or general community activity, for example, local meetings or attending leisure centres. The methodology deploys inferential statistics, such as Correlations and Linear Regressions of Social Backgrounds and Media Use as independent variables, and Social Capital as the dependent. Overall the findings suggest that consideration of the types of channels people watch is an important intervention, as well as Demographics, in understanding Social Capital.

## APPROACH TO BE TAKEN

The Report performs multiple individual linear regressions using dependent variables representing 'Social Capital', including - Voting - our area of concern. The independent variables are constant for all the dependent variables, and are the following - Education, Gender, Employment Status, Race, Age, Income, Tv Public Affairs, Tv News, TV Hours, Paper and radio. A Linear Regresssion allows one to measure a Dependent variable against a number of selected independent variables, in order examine whether there is any explanatory power in the independent variables on the dependent.

## DESCRIPTIVE STATISTICS

The following section will identify the variables to be used in our later analysis, perform descriptive analysis on these variables, and outline what was done to clean these variables.

All of the following 'Media' variables are derived from Norris (1996) Appendix A:

**LIBRARY**

```
library(dplyr)

library(ggplot2)
```

```
library(ggthemes)

library(tidyverse)

library(coefplot)

library(olsrr)

library(lmtest)

library(car)

library(carData)
```

### WORKING DIRECTORY

```
setwd("~/1 University - Politics and IR/3rd Year/Analysing Data/REPORT")

d1 <- read.csv("tv_social_capital.csv")
```

### MEDIA

### 1. TV HOURS

TV Hours asks the question - Thinking back other the past 7 days, how many hours a day did you spend watching television? The question asks respondents to put down the amount of hours of TV they have watched in the past 7 days, rendering it a numerical and continuous variable.

```
d1$tvhrs<-as.numeric(d1$tvhrs)

# we change TV Hours to Numeric as it is wrongly attributed as a integer variable.
```

```
# table of frequencies

summary(d1$tvhrs)
```

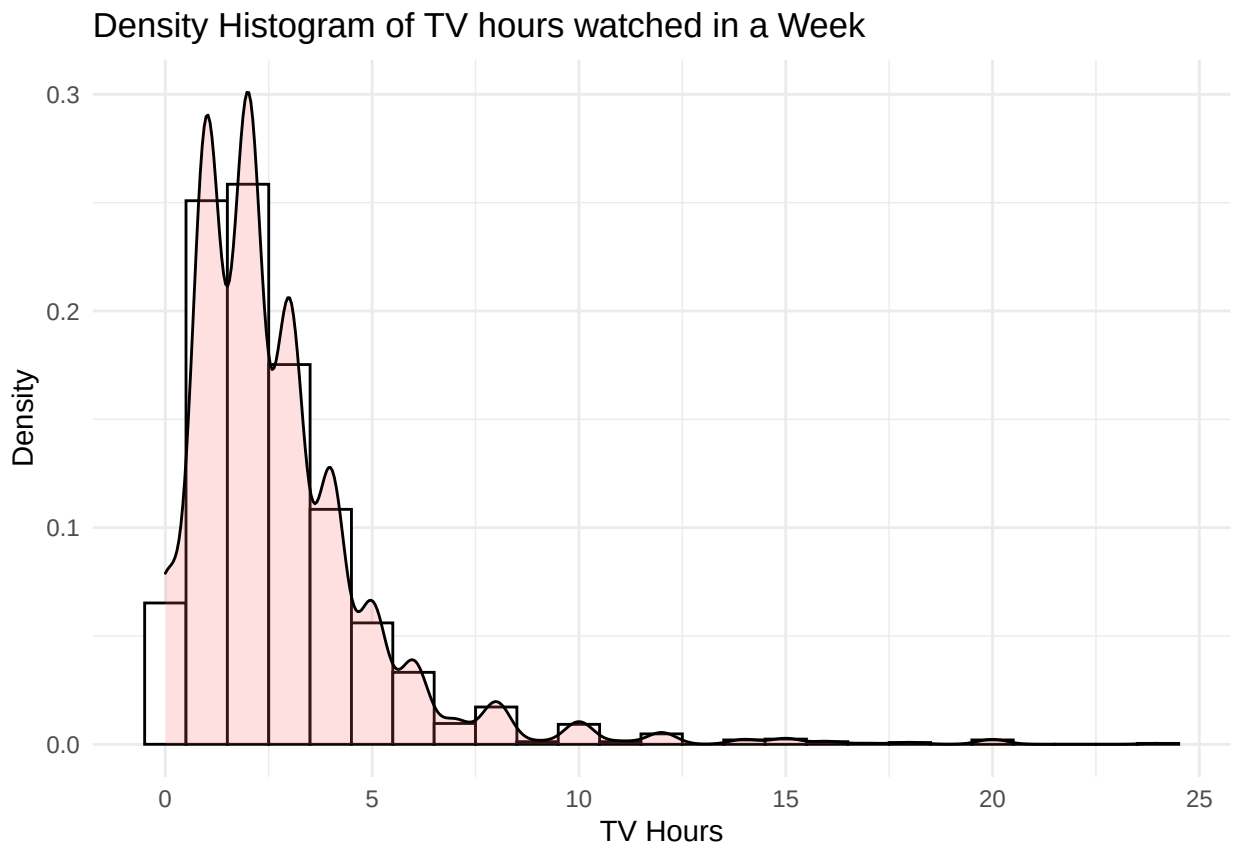
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   2.000   2.739   3.500  24.000      18
```

```
# standard deviation

sd(d1$tvhrs, na.rm=T)
```

```
## [1] 2.400578
```

The Median is 2, and Mean 2.8.

```
ggplot(d1, aes(x=tvhrs)) +
labs(x = 'TV Hours', y = 'Density', title = 'Density Histogram of TV hours watched in a Week') +
theme_minimal()+
geom_histogram(aes(y=..density..), binwidth = 1, colour="black", fill="white")+
geom_density(alpha=.20, fill="#FF6666")
```




A positive skew, and clustering around the mean of 2.7 shown in the low standard deviation (2.4), demonstrating a low variance of frequencies in the variable spread.

**2. TVNEWS**

The vairble - tvnews - measures how often one watches News channel TV in general. It is an high ordinal level variable, categorised in the following way: 1 = Never, 2 = less than once a month, 3 = once a month , 4 = several times a month , 5 = once a week , 6 = several times a week , 7 = every day.

```
# NAs have been coded as such previous to our cleaning in the dataset as shown below

table(d1$tvnews, useNA = "ifany")
```

```
##
##    1    2    3    4    5    6    7 <NA>
##   66   54   51   79  186  535 1377  169
```

```
# table as percentages

prop.table(table(d1$tvnews, useNA = "ifany"))
```

```
##
##          1          2          3          4          5          6          7
## 0.02622169 0.02145411 0.02026222 0.03138657 0.07389750 0.21255463 0.54707986
##       <NA>
## 0.06714342
```

In terms of the NAs, 6% is a small number and should not compromise the analysis.

```
summary(d1$tvnews)
```

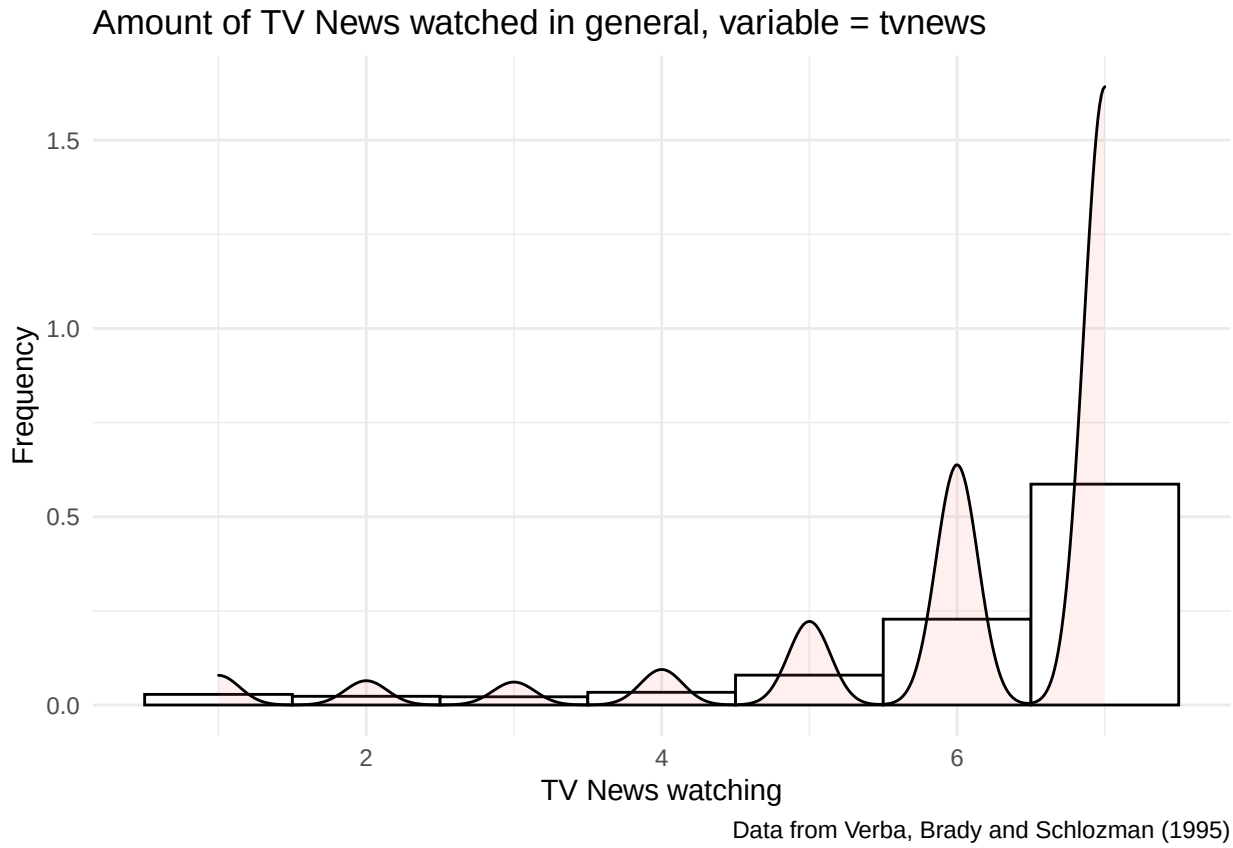
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   7.000   6.142   7.000   7.000     169
```

```
sd(d1$tvnews, na.rm=T)
```

```
## [1] 1.430761
```

TV News has a minimum value of 1 and maximum of 7, with a Mean of 6.1 and a Median of 7.

```
ggplot(d1,aes(x=tvnews)) +
labs(title = "Amount of TV News watched in general, variable = tvnews",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "TV News watching") +
theme_minimal()+
geom_histogram(aes(y=..density..), binwidth=1,colour="black", fill="white")+
geom_density(alpha=.10, fill="#FF6666")
```




The graph reflects the intense concentration of frequencies in the table at 7 - watches TV News every day - showing a clear negative skewness. A standard deviation of 1.4 suggests clustering around the mean of 6.1.

**3.NEWSPAPER**

The variables readnews - how often do you read newspapers, readnat - how much attention do you pay to national issues when reading, and readloc - attention paid to local issues when reading, are combined to create a 12 point scale, which starts with 1 equalling never reads Newspapers and ends with 12 equalling Reads News everyday and pays great attention to National, and Local issues when reading said paper.

```
# creates 12 point scale

d1$paper <- d1$readnews + d1$readnat + d1$readloc
d1$paper <- d1$paper-3
```

```
# table and satndard deviation

summary(d1$paper)
```

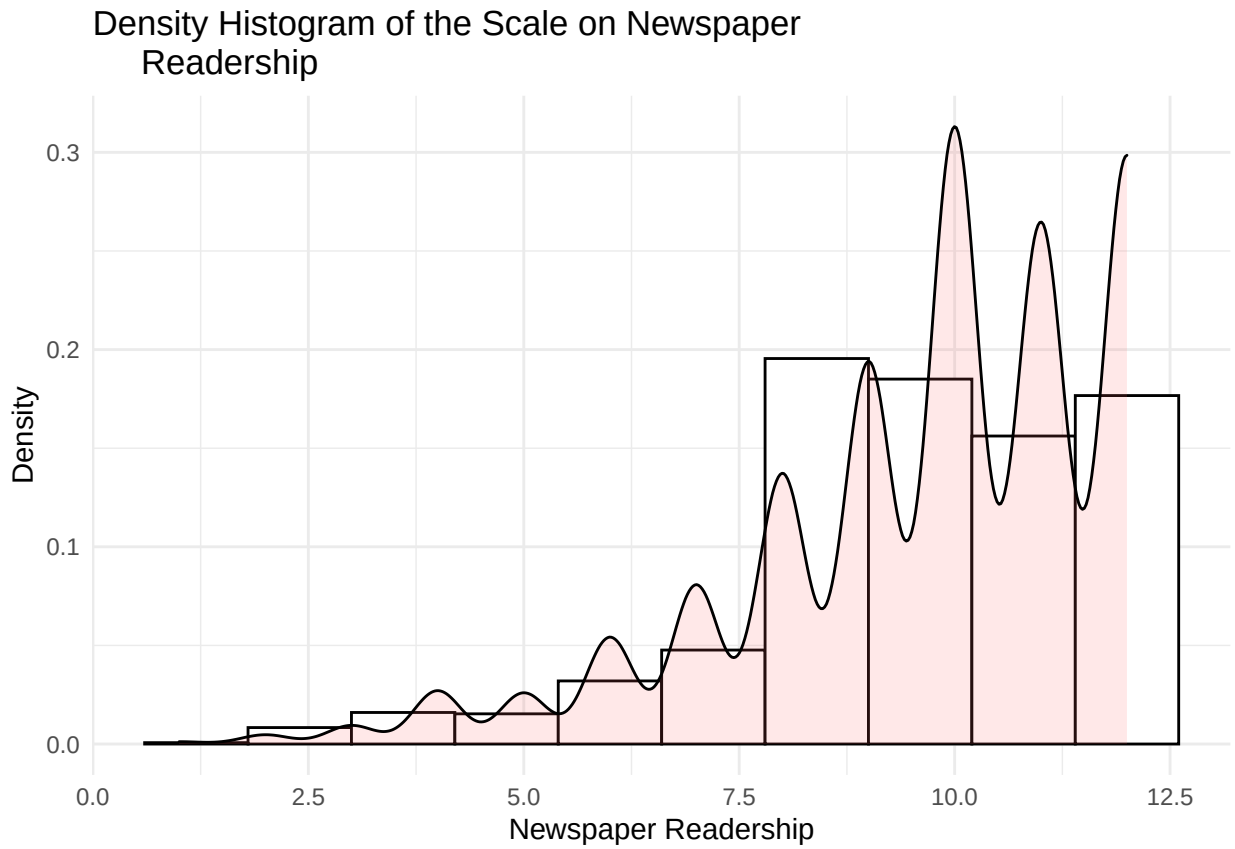
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   9.000  10.000   9.667  11.000  12.000     121
```

```
sd(d1$paper, na.rm=T)
```

```
## [1] 2.083292
```

The median is 10, Mean 9.6, and Standard Deviation is 2.1.

```
ggplot(d1, aes(x=paper)) +
labs(x = 'Newspaper Readership', y = 'Density', title = 'Density Histogram of the Scale on Newspaper
     Readership') +
theme_minimal()+
geom_histogram(aes(y=..density..), binwidth = 1.2,colour="black", fill="white")+
geom_density(alpha=.15, fill="#FF6666")
```




The Histogram shows a strong Negative Skew and Clustering around the Mean (9.6) with a Standard Deviation of 2.1, which suggests variance is limited, additionally a negative skewness toward the higher side of the scale is shown, Newspaper readership is high amongst respondents.

**4. RADIO**

The variable Radio is measured by combining radcall1 and radcall 2. Radcall1 measures how often people listen to call in political talk shows, and Radcall2, whether respondents actively call in to said political shows.

```
# recoding - creates a 0-6 scale

d1$radio <- d1$radcall1 + d1$radcall2-3
```

```
# raw table and percentages table of frequencies

table(d1$radio, useNA = "ifany")
```

```
##
##    0    1    2    3    4    5    6 <NA>
##  281  143  131  122  192  200   46 1402
```

```
prop.table(table(d1$radio, useNA="ifany"))
```

```
##
##           0           1           2           3           4           5           6
```

```
## 0.11164084 0.05681367 0.05204609 0.04847040 0.07628129 0.07945967 0.01827573
##       <NA>
## 0.55701232
```

The scale shows an alarming amount of missing values, accounting for 55% of respondents, which is a majority of respondents and is suspected to come from radcall2. Norris explicity states to use radcall 2 for the Radio variable so it is used despite the missing values.

```
summary(d1$radio)
```

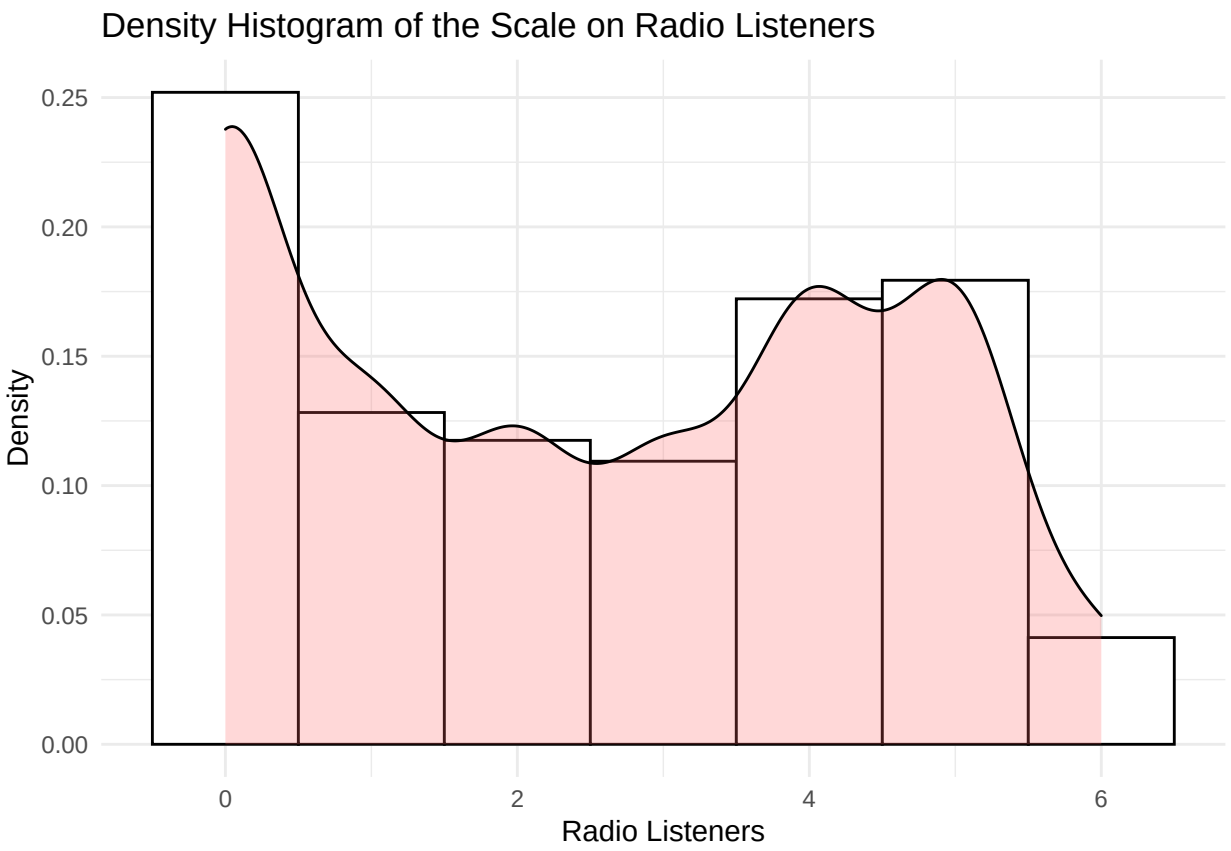
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   3.000   2.525   4.000   6.000    1402
```

```
sd(d1$radio, na.rm = T)
```

```
## [1] 1.984245
```

The Median is 3 and Mean is 2.525, with a Standard Deviation of 1.99.

```
ggplot(d1,aes(x=radio)) +
labs(x = 'Radio Listeners', y = 'Density', title = 'Density Histogram of the Scale on Radio Listeners')
theme_minimal()+
geom_histogram(aes(y=..density..), binwidth=1,colour="black", fill="white")+
geom_density(alpha=.25, fill="#FF6666")
```



The above Histogram shows there is no strong skewness to either side, with similar frequencies across all 6 values, with a peak at Never (1) listening to Radio or calling in.

### 5. PUBLIC AFFAIRS

The variable Public Affairs is derived from 'tvpubaff' and is a high ordinal variable made up of 7 categories, 1-7, in ascending order of frequency Public Affair TV channels watched. 1 represents Never, 2 Less than once a month, 3 Once a month, 4 several times a month, 5 once a week, 6 several times a week, 7 Every Day.

```
# median, mean, max and min

summary(d1$tvpubaff, useNA="ifany")
```

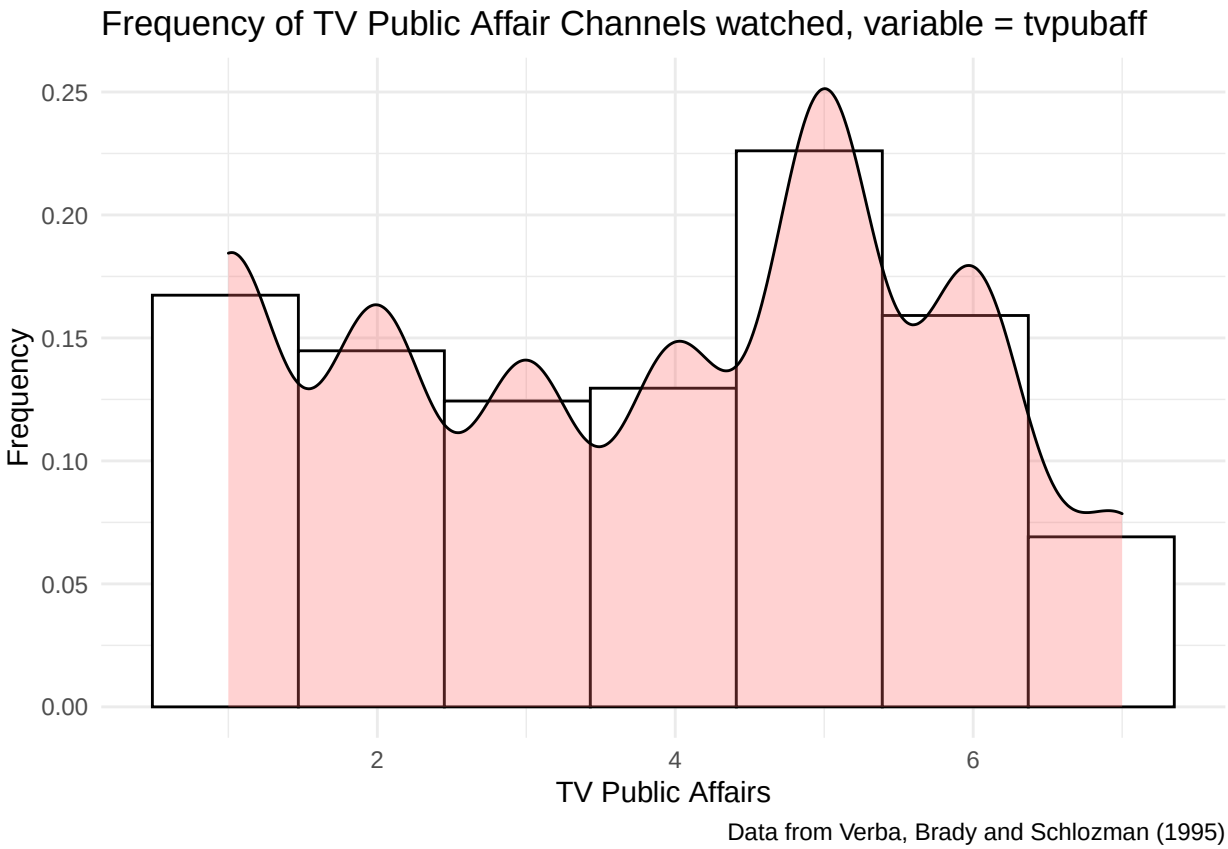
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   4.000   3.839   5.000   7.000     170
```

```
# standard deviation

sd(d1$tvpubaff, na.rm = T)
```

```
## [1] 1.896414
```

The median is 4, Mean is 3.9, and Standard Deviation is 1.9.

```
# histogram density

ggplot(d1,aes(x=tvpubaff)) +
labs(title = "Frequency of TV Public Affair Channels watched, variable = tvpubaff",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "TV Public Affairs") +
theme_minimal()+
geom_histogram(aes(y=..density..), binwidth=0.98,colour="black", fill="white")+
geom_density(alpha=.30, fill="#FF6666")
```




The graph shows a slight negaitve skewness, the mean is 3.8 with a standard deviation of 1.9 suggesting some clustering around the mean, which is reflected in the distrribution.

Moving onto the variables grouped under 'demographics', which are derived from Appendix B Sidney Verba et al (1995):

**DEMOGRAPHICS**

**1.EDUCATION**

The variable 'education', measures the highest schooling grades achieved by an individual and is an high ordinal variable, merging the variables 'edgrade' which asks for highest level obtained at lower years or high school and eddegree asking for highest level achieved at College/ University.

```
# The below merges the two variables edgrade and eddegree to create 'education' or Education variable.

d1$education[d1$edgrade %in% 0:8] <-1 #  representing "Grammar and less".
d1$education[d1$edgrade %in% 9:11] <- 2 #  representing "Some high school".
d1$education[d1$edgrade %in% 12] <- 3 # representing "GED" which is obtained in grade 12.
d1$education[d1$edgrade %in% 13] <- 4 # representing "Some college"
d1$education[d1$eddegree %in% 1] <- 5 # Eddegree 1, representing "Junior/Associate college degree", is
#coded into education.
d1$education[d1$eddegree %in% 2] <-6 # Representing "Bachelor's degree"
d1$education[d1$eddegree %in% 3] <- 7 # Representing "Masters degree"
d1$education[d1$eddegree %in% 4:7] <- 8 # representing several PHD/professional degrees
```

```
# median, mean
```

```
summary(d1$education)
```

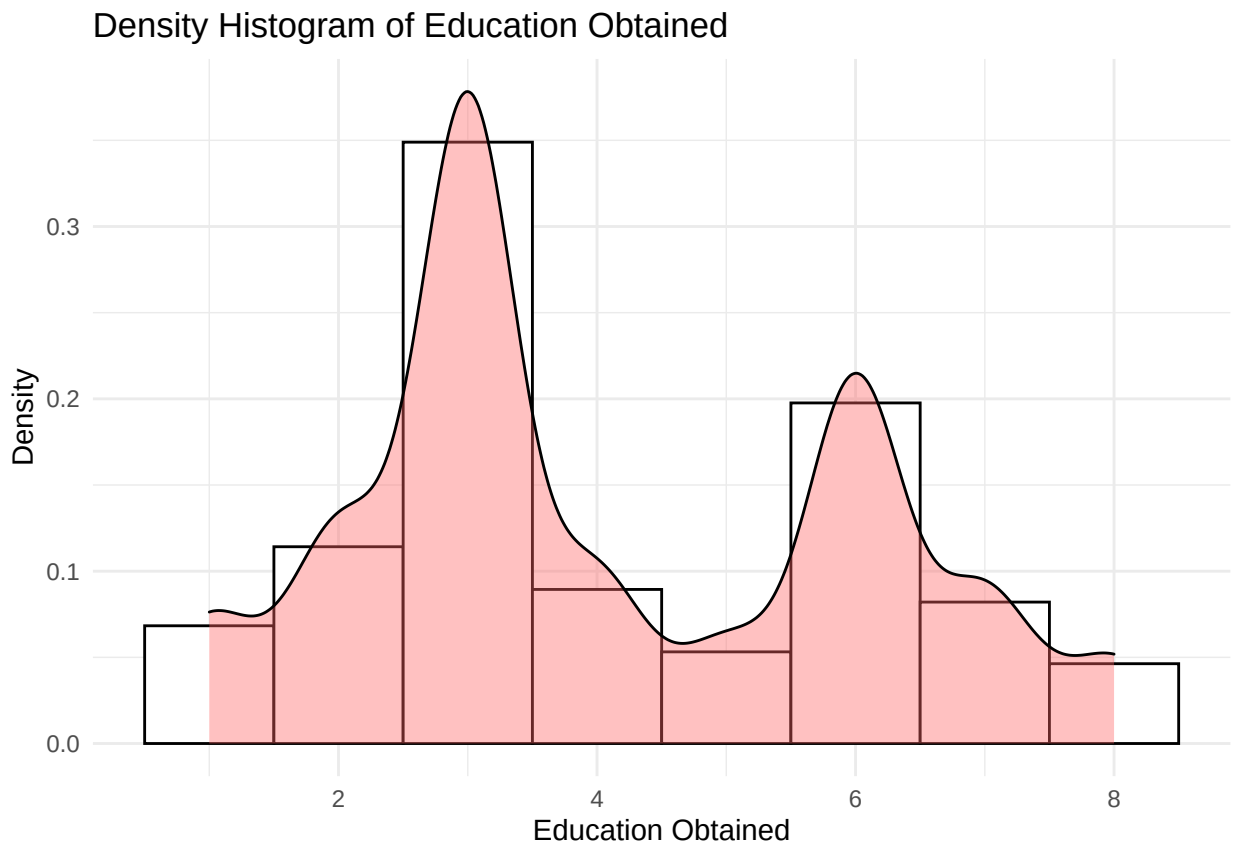
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   3.000   4.098   6.000   8.000     336
```

```
# standard deviation

sd(d1$education, na.rm = T)
```

```
## [1] 1.93284
```

The Median is 3, and Mean is 4.1. In addition, the Standard Deviation is 1.9, which suggests there is clustering around the Mean and Median, and subsequently low spread of the variables. As the graph below shows there is slight positive skewness to the graph, and with the highest frequency around the Mean of 4.1.

```
# density histogram

ggplot(d1,aes(x=education)) +
labs(x = 'Education Obtained', y = 'Density', title = 'Density Histogram of Education Obtained') +
theme_minimal()+
geom_histogram(aes(y=..density..), binwidth=1,colour="black", fill="white")+
geom_density(alpha=.40, fill="#FF6666")
```




## 2. INCOME

The following variable - faminc - measures a households combined Income, it is a high level ordinal variable with 16 ascending groups of income.

Below shows that Family Income has a Median of 7, Mean of 7.2, and Minimum of 1 and Maximum of 16.

```
summary(d1$faminc)
```

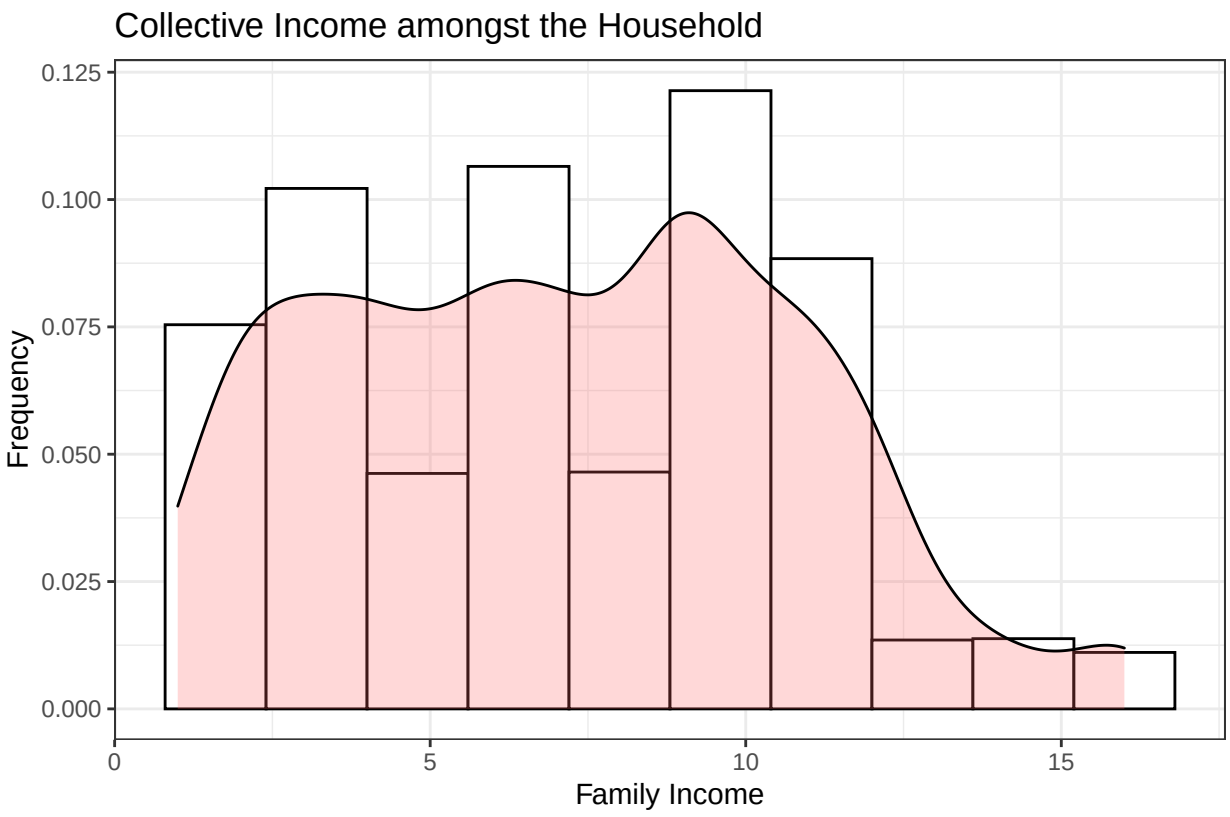
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   7.000   7.177  10.000  16.000     205
```

```
# standard deviation of Income

sd(d1$faminc, na.rm = T)
```

```
## [1] 3.653812
```

The standard deviation is 3.6.

```
# histogram

ggplot(d1, aes(x = faminc)) +
  theme_bw()+
  labs(title = "Collective Income amongst the Household",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "Family Income")+
  geom_histogram(aes(y=..density..), binwidth=1.6,colour="black", fill="white")+
geom_density(alpha=.25, fill="#FF6666")
```



The Standard Deviation is 3.65 indicating some spread of values and a broad frequency at different levels of incomes with values of 1 and 2 showing similar frequencies to 10 and 11, for example. This level of spread makes sense given the US' high amounts of inequality and high GDP.

### 3. EMPLOYMENT STATUS

Employment Status is measured by the variable - empstat, dervied from 'jblastwk' which asks about someones job status last week. The variable jblastwk is shortened to 3 categories from its original 10.

```
# recoding jblstwk to 'empstat' and 3 categories

d1$empstat[d1$jblastwk %in% 1] <-3 # full time
d1$empstat[d1$jblastwk %in% 2] <-2 # part time
d1$empstat[d1$jblastwk %in% 3:10] <-1 # trying to capture those not working for whatever reason
#- originally had seperate variables of retired, unemployed, disabled, full-time volunteer, in school,
```

Employment Status has the labels 1 - not working, 2 - part time, 3- full time.

```
table(d1$empstat, useNA="ifany")
```

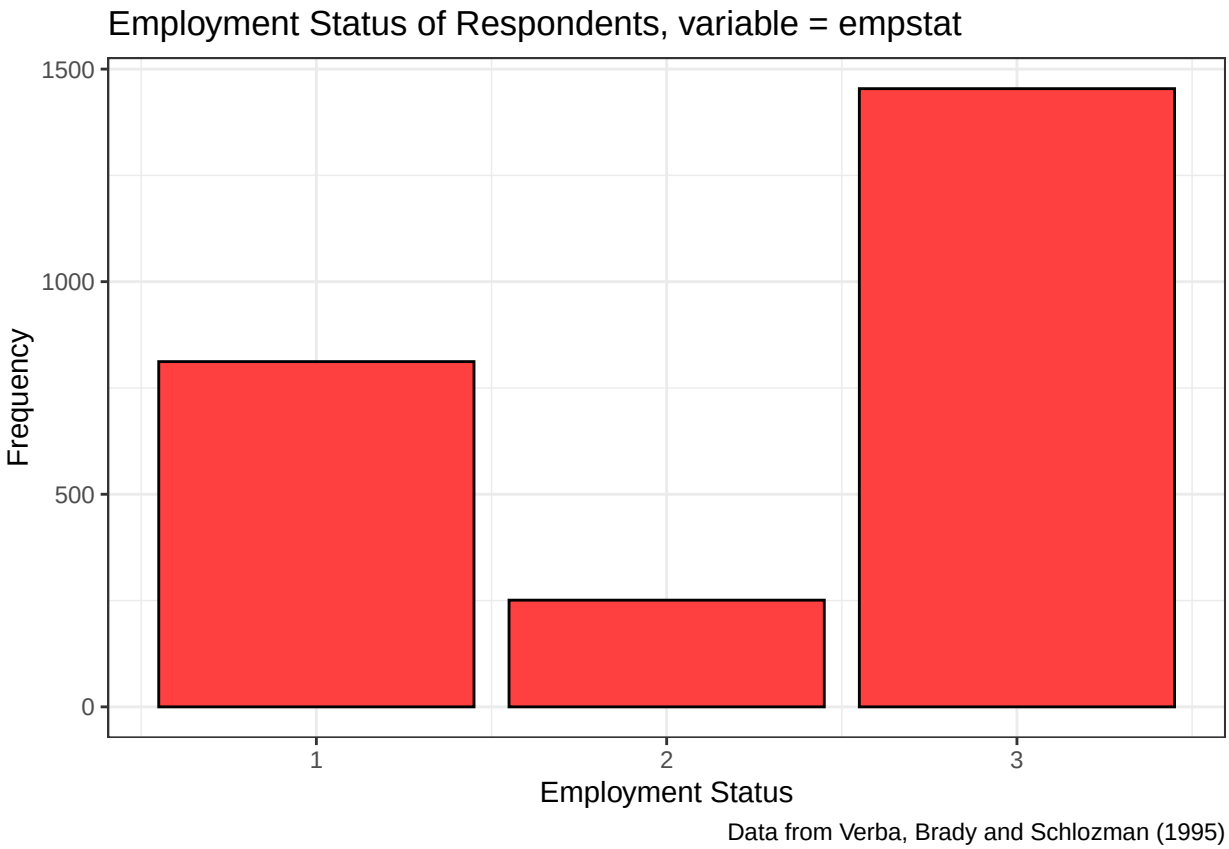
```
##
##    1    2    3
##  812  251 1454
```

```
prop.table(table(d1$empstat, useNA = "ifany"))
```

```
##
##          1          2          3
## 0.32260628 0.09972189 0.57767183
```

Those Not working make up 32%, part-time 9%, and full time 57%.

```
# barplot

ggplot(d1, aes (x=empstat))+
  geom_bar(colour="black", fill="brown1")+
  theme_bw()+
  labs(title = "Employment Status of Respondents, variable = empstat",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "Employment Status")
```




### 4. ETHNICITY

Ethnicity is derived from the variable 'flaprace' which asks for ones ethnic identification. It is assumed Asian and Native Americans are excluded due to their low totals.

```
# recoding ethnicity

d1$ethnicity <- d1$flaprace
d1$ethnicity[d1$ethnicity == 1] <- 1 #"Anglo-White"
d1$ethnicity[d1$ethnicity == 2] <- 2 #"African-American"
d1$ethnicity[d1$ethnicity == 3] <- NA # originally 'Asian'
d1$ethnicity[d1$ethnicity == 4] <- 3 #"Latino"
d1$ethnicity[d1$ethnicity == 5] <- NA #Originally Native American
d1$ethnicity[d1$ethnicity == 6] <- NA # Originally 'Other'
```

```
# frequency table

table(d1$ethnicity, useNA = "ifany")
```

```
##
##    1    2    3 <NA>
## 1614  478  356   69
```

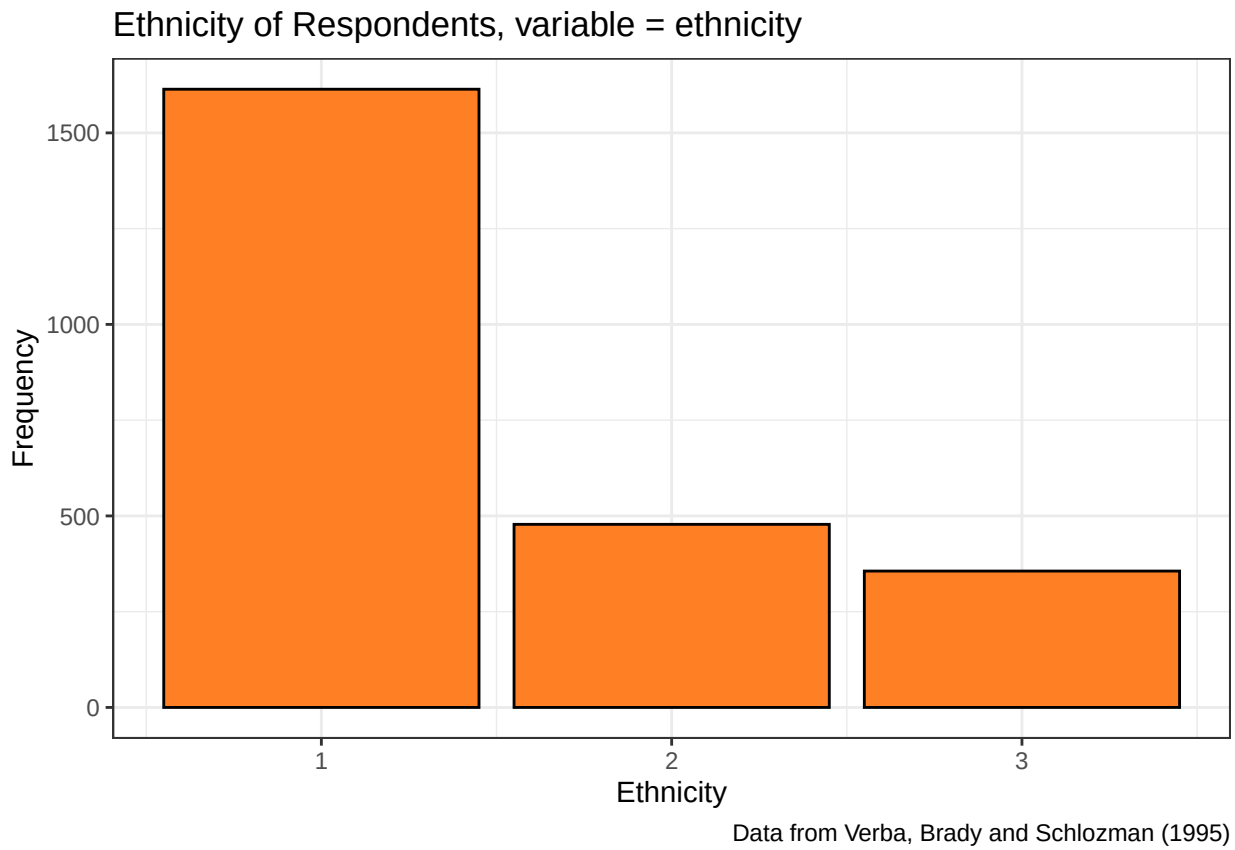
```
prop.table(table(d1$ethnicity, useNA = "ifany"))
```

```
##
```

```
##          1          2          3       <NA>
## 0.64123957 0.18990862 0.14143822 0.02741359
```

The above frequencies show Anglo-Whites (1) make up 64% of respondents, African Americans (2) 19%, Latinos (3) 14%, Missing values 2%. The missing values are quite low and not enough to deem the variable unusable.

```
# bar chart

ggplot(d1, aes(x=ethnicity))+
  geom_bar(colour="black", fill="chocolate1")+
  theme_bw()+
  labs(title = "Ethnicity of Respondents, variable = ethnicity",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "Ethnicity")
```




### 5. GENDER

Gender is measured by the respondents self-identification from respondents of whether they are male or female. We recoded Gender to be a Factor level variable as Gender is traditionally a Binary varible.

```
# switching male and female

d1$sex[d1$gender %in% 2] <-1 #becomes 'women'
d1$sex[d1$gender %in% 1] <-2 #becomes 'men'
```

```
# change sex to factor for purposes of descriptives and regression
d1$sex<-as.factor(d1$sex)

class(d1$sex)
```

```
## [1] "factor"
```

```
table(d1$sex)
```

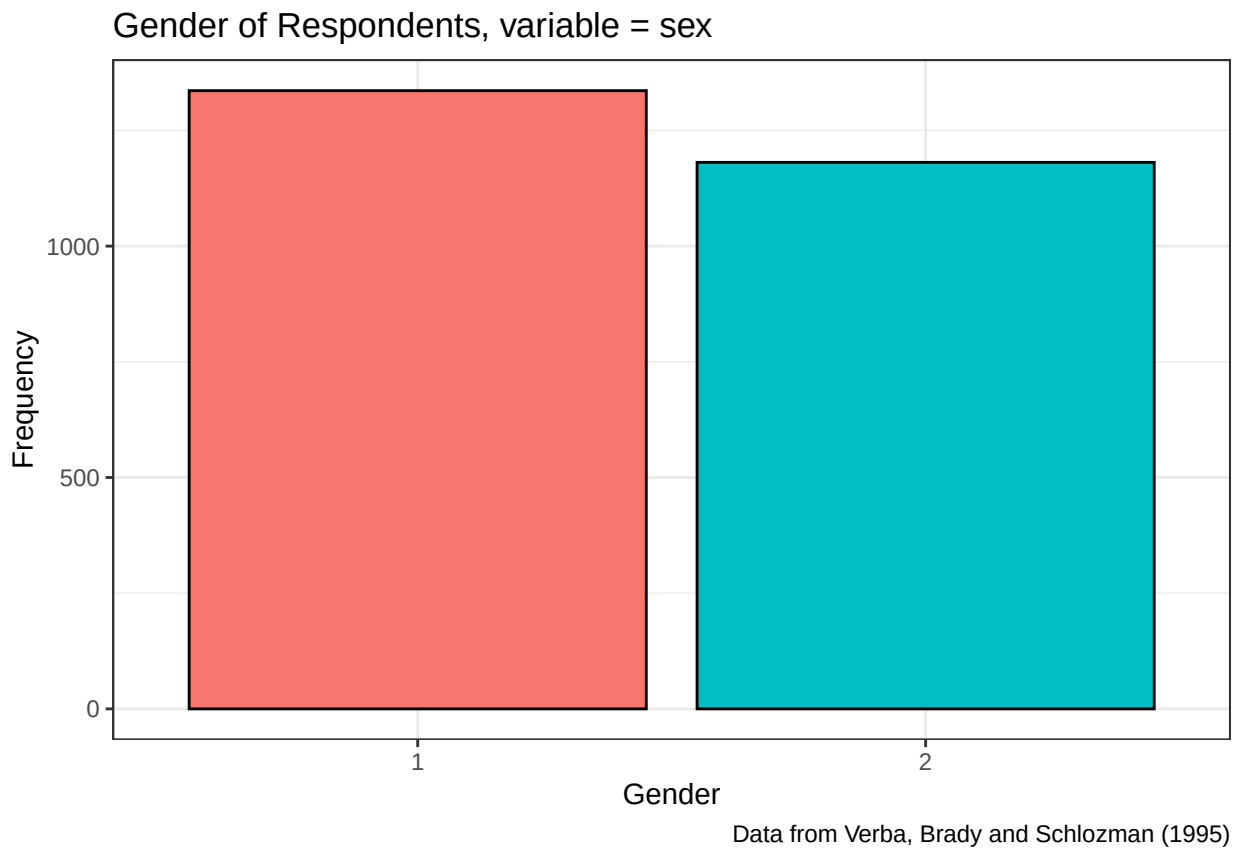
```
##
##    1    2
## 1336 1181
```

```
prop.table(table(d1$sex))
```

```
##
##         1         2
## 0.5307906 0.4692094
```

The frequencies above show that Women (1) make up 53% of respondents, and Men (2) 47%.

```
# bar plot

ggplot(d1, aes (x=sex, fill = sex))+
  geom_bar(colour="black")+
  theme_bw()+
  labs(title = "Gender of Respondents, variable = sex",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "Gender")+
  guides(fill=FALSE)
```




### 6. AGE

The variable Age is derived from the variable 'yearborn' which is a continuous variable, but for our analysis it will be recoded as a Factor variable with 6 ascending categories and 5 dummies with the category '45-54' as the reference category.

```
# recoding yearborn to Age as a Factor level variable

d1$age <- d1$yearborn - 1990
d1$age <- d1$age*-1
d1$age[is.na(d1$age)] <- -1
d1$age[d1$age < 25] <- "<25"
d1$age[d1$age >= 25 & d1$age <= 34] <- "25-34"
d1$age[d1$age >= 35 & d1$age <= 44] <- "35-44"
d1$age[d1$age >= 45 & d1$age <= 54] <- "45-54"
d1$age[d1$age >= 55 & d1$age <= 65] <- "55-65"
d1$age[d1$age > 65] <- ">65"
d1$age <- factor(d1$age)
d1$age <- relevel(d1$age, ref = 5)
```

```
# frequency table

table(d1$age, useNA = "ifany")
```

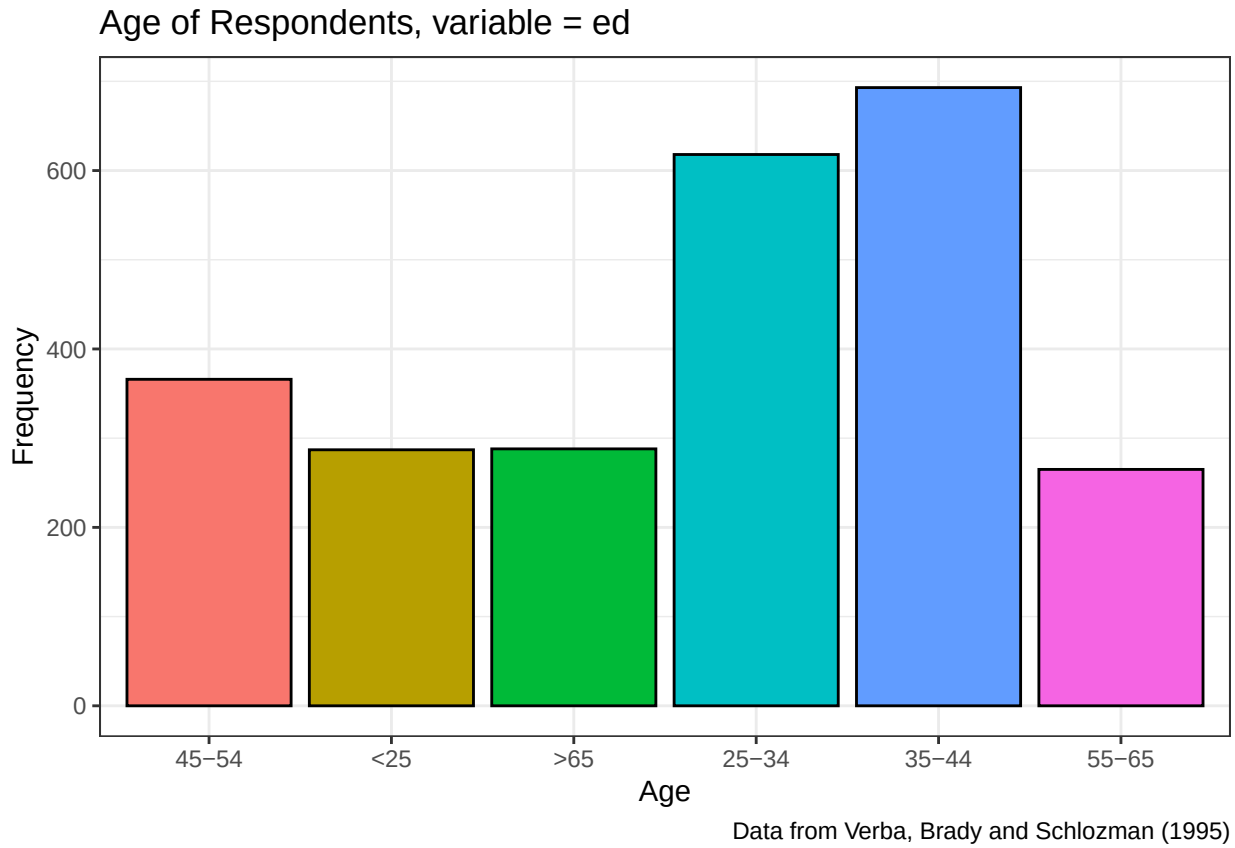
```
##
## 45-54   <25   >65 25-34 35-44 55-65
##   366   287   288   618   693   265
```

```
prop.table(table(d1$age, useNA = "ifany"))
```

```
##
##     45-54       <25       >65     25-34     35-44     55-65
## 0.1454112 0.1140246 0.1144219 0.2455304 0.2753278 0.1052841
```

The table above shows the <25 group makes up 11% of respondents, 25-34 - 24%, 35-44 - 27%, 45-54 - 14%, 55-65 - 10%, and >65 - 11%.

```
# bar plot

ggplot(d1, aes (x=age, fill=age))+
  geom_bar(colour="black")+
  theme_bw()+
  labs(title = "Age of Respondents, variable = ed",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "Age")+
  guides(fill=FALSE)
```




## DEPENDENT VARIABLE

## VOTING

"A vote scale is used as for the dependent variable in regression. It is constructed by adding the questions on national and local elections, with each question given a value of (0 ±never, 4 ± all). The scale runs from 0 to 8." (Verba et al: 1995:p.538). Using the variables vtpres - how many Presidential elections have you voted in the past 5 elections - and vtlocall - the same but local elections.

```
# creates the 8 point scale - 0-8

d1$vtpres[d1$vtpres > 5 ] <- NA
d1$vtlocall[d1$vtlocall > 5 ] <- NA
d1$vote <- d1$vtpres + d1$vtlocall
d1$vote <- d1$vote -2
```

```
# shows median, mean, max, min, NA's

summary(d1$vote)
```

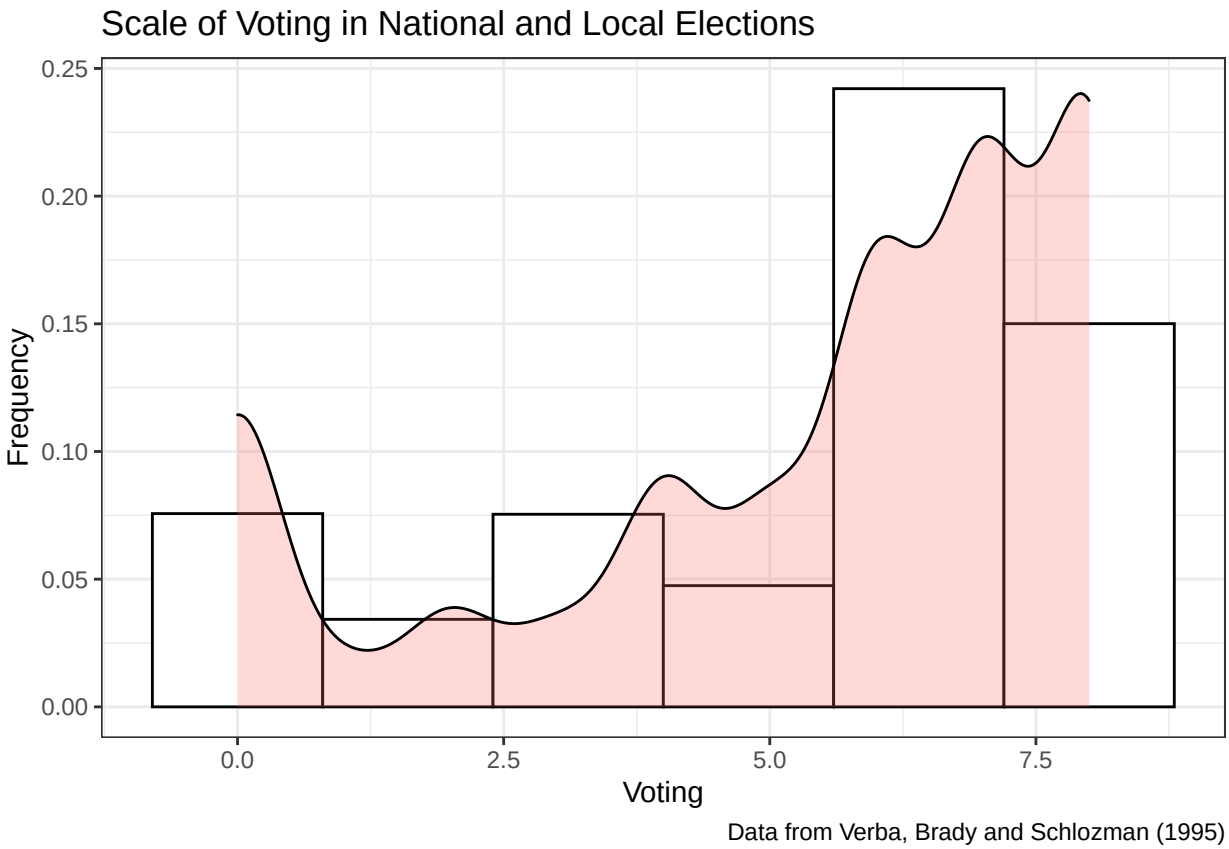
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   4.000   6.000   5.381   7.000   8.000     147
```

```
# standard deviation

sd(d1$vote, na.rm = T)
```

```
## [1] 2.625066
```

The Median of vote is 6, the mean 5.4 and the Standard Deviation 2.6.

```
# histogram

ggplot(d1, aes(x = vote)) +
  theme_bw()+
  labs(title = "Scale of Voting in National and Local Elections",
       caption = "Data from Verba, Brady and Schlozman (1995)",  y = "Frequency",
       x = "Voting")+
geom_histogram(aes(y=..density..), binwidth=1.6,colour="black", fill="white")+
geom_density(alpha=.25, fill="#FF6666")
```

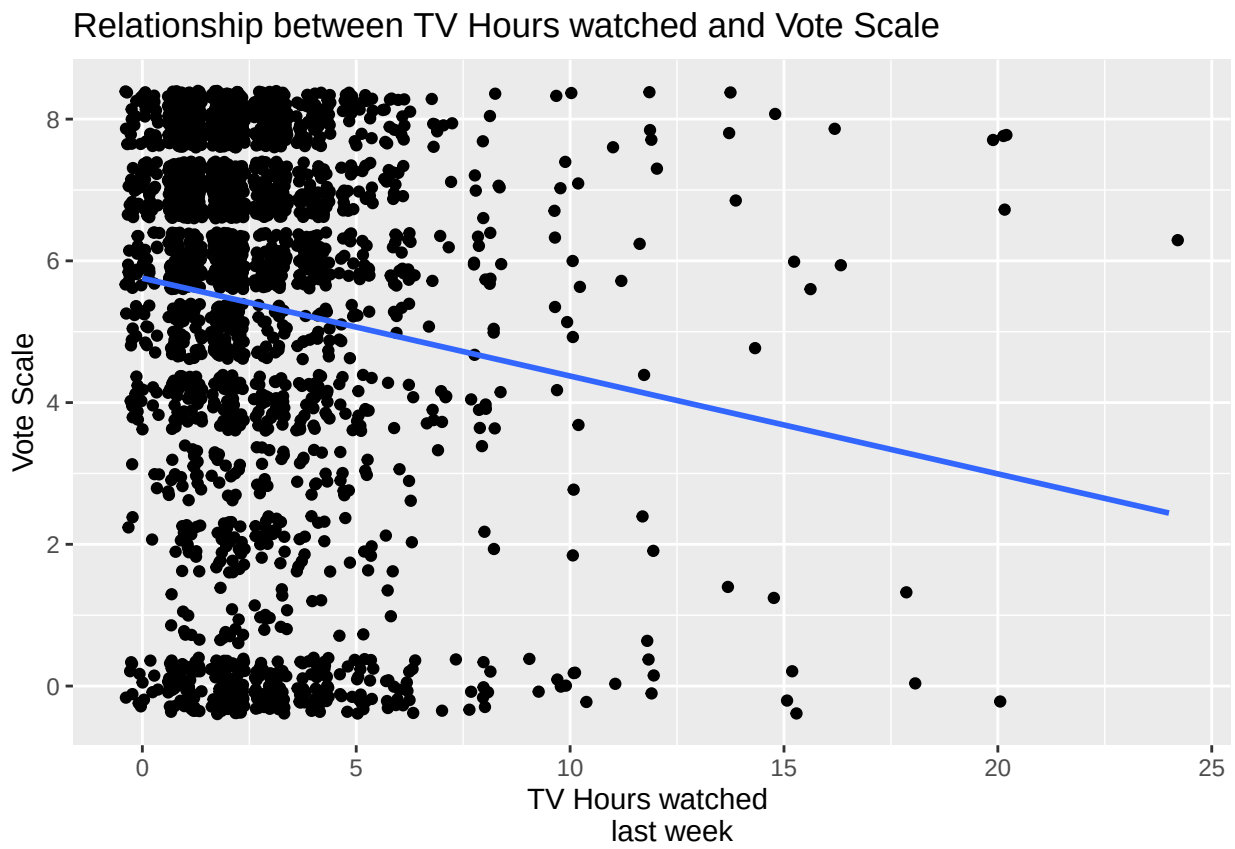



The histogram shows a skewness to the right hand side, and demonstrates low levels of variance in results when considering a Standard deviation of 2.6, compared to a mean of 5.4 , suggesting clustering around the higher end of the scale.

## REPLICATION

```
# removes missing values for Regression Analysis and Anovas.

d4 <- d1[which(complete.cases(d1[,c('tvhrs', 'tvnews', 'tvpubaff', 'paper', 'radio',
                                    'empstat', 'education', 'sex', 'ethnicity', 'age', 'faminc')])),]
```

```
# scatterplot of the relationship between TV Hours and Vote Scale

# As we are working with a lot of observations using "jitter" allowed more observations to be shown.

ggplot( data = d1,
aes(tvhrs, vote)) +
geom_point(position = "jitter") +
labs(title = "Relationship between TV Hours watched and Vote Scale", x = "TV Hours watched
     last week", y = "Vote Scale") +
geom_smooth(method=lm, se=FALSE)
```




The above graph is a visualization of the relationship between TV Hours and vote Scale, a central theme in Norris'piece (1996). The graph seems to reflect a weak negative linear correlation, those watching less TV seem to Vote more often.

Firstly, a linear regression of Tv Hours to Vote will be performed, then a Linear Regression with all Media variables, then a regression with all Media variables and Demographics to compare against Norris' model. This most aptly captures Norris' methodology within the limitations of the Project.

When 'in the population' is said it is meant as the US general population.

**MODEL 1 - TV HOURS**

```
# regression

lm1<-lm(vote~ tvhrs, data = d4)
summary(lm1)
```

```
##
## Call:
## lm(formula = vote ~ tvhrs, data = d4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0204 -1.0204  0.9796  1.9796  4.2653
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   6.1933     0.1456  42.548  < 2e-16 ***
## tvhrs        -0.1729     0.0449  -3.851 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.345 on 748 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.01944,    Adjusted R-squared:  0.01813
## F-statistic: 14.83 on 1 and 748 DF,  p-value: 0.0001275
```

```
# levels of confidence

confint(lm1)
```

```
##                 2.5 %      97.5 %
## (Intercept)  5.9075563  6.47907275
## tvhrs       -0.2610805 -0.08478506
```

Model 1 shows a statistically significant relationship at the 99.9% level. A one unit increase in TV Hours decreases Voting by -0.17. The null hypothesis can be rejected that the coefficient is 0 in the population, there is a statistically significant difference in the population. The R Squared is 1%, which suggests a very weak model and very small amounts of variance in Voting is explained through the differences in TV Hours.

**MEDIA MODEL - ALL MEDIA VARIABLES**

```
# regression

lm5<- lm(vote~ tvhrs+ tvnews + tvpubaff + paper + radio, data = d4)
summary(lm5)
```

```
##
## Call:
## lm(formula = vote ~ tvhrs + tvnews + tvpubaff + paper + radio,
##     data = d4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9765 -0.9966  0.4414  1.3625  4.9256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.35637    0.55896  -0.638 0.523952
## tvhrs       -0.16551    0.04062  -4.074 5.11e-05 ***
## tvnews       0.10425    0.06346   1.643 0.100851
## tvpubaff     0.17759    0.04630   3.836 0.000136 ***
## paper        0.51924    0.04534  11.452  < 2e-16 ***
## radio       -0.05220    0.03850  -1.356 0.175555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.088 on 744 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2221
## F-statistic: 43.77 on 5 and 744 DF,  p-value: < 2.2e-16
```

```r
# levels of confidence

confint(lm5)
```

```
##                   2.5 %      97.5 %
## (Intercept) -1.45369418  0.74094863
## tvhrs       -0.24525700 -0.08575746
## tvnews      -0.02033021  0.22882132
## tvpubaff     0.08670433  0.26848412
## paper        0.43022493  0.60825104
## radio       -0.12777301  0.02337921
```

Controlling for all other variables, TV Hours coefficient decreases from -0.17 in Model 1 to Model 5 of -0.16, a one unit increase in TV Hours relates to a -0.16 change in the amount someone Votes, statistically significant at the 99.9% level as in Model 1. A one unit change in TV News effects Voting by 0.10, controlled for all other variables, but is statistically insignificant in the population at the threshold 95% level. Controlled for all other variables, TV Public Affairs effects Voting by 0.18 and is statistically significant in the US population at the 99.9% level. Paper has a 0.52 effect on Voting when controlled for all other Media variables, statistically significant at the 99.9% level in the population. A one unit change in Radio , when controlled for all other variables, effects Voting by -0.05, but is statistically insignificant in the population, the confidence intervals further reflect its poor measurement of Voting as the intervals fall in between a negative and a positve, suggesting unreliability of the coefficient.

```r
anova(lm1, lm5)
```

```
## Analysis of Variance Table
##
## Model 1: vote ~ tvhrs
## Model 2: vote ~ tvhrs + tvnews + tvpubaff + paper + radio
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    748 4114.7
## 2    744 3242.5  4    872.17 50.03 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The R Squared is 22% in Model 5 compared to Model 1 at 1% which shows a good level of improvement in showing the variation in the dependent variable. The Anova tests suggest a 99.9% level p-value for Model 5, and the alternative hypothesis that there is a significant difference between the models is accepted. Overall, the inclusion of other Media variables to TV Hours seems to improve the predictive power seen in Model 1 with just TV Hours.

**NORRIS REGRESSION**

*Norris(1996: p.478)*

SOCIAL BACKGROUND

Education - .19**

Gender - -0.7*

Employment Status .04

Race - 0.4

Age - .30**

Income - .07*

MEDIA USE

TV Public Affairs - .08**

TV News - .04

TV Hours - -.08

Paper - .24**

Radio - .01

RSquared - .36 (36%)

**FINAL MODEL - DEMOGRAPHICS AND MEDIA VARIABLES**

```
# regression and confidence intervals

m12<- lm(vote~ education + sex + empstat + ethnicity + age + faminc + tvpubaff + tvnews
         + tvhrs + paper + radio, data = d4)
summary(m12)
```

```
##
## Call:
## lm(formula = vote ~ education + sex + empstat + ethnicity + age +
##     faminc + tvpubaff + tvnews + tvhrs + paper + radio, data = d4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.832 -1.039  0.183  1.202  5.494
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.91727    0.64058   1.432 0.152585
## education    0.23075    0.04420   5.221 2.32e-07 ***
## sex2        -0.49373    0.14889  -3.316 0.000958 ***
## empstat      0.09136    0.10043   0.910 0.363275
## ethnicity   -0.35214    0.12263  -2.871 0.004203 **
## age<25      -1.95818    0.34515  -5.673 2.02e-08 ***
## age>65       0.60108    0.29820   2.016 0.044194 *
## age25-34    -1.04874    0.24152  -4.342 1.61e-05 ***
## age35-44    -0.29103    0.22495  -1.294 0.196161
## age55-65     0.23933    0.27341   0.875 0.381662
## faminc       0.04065    0.02402   1.692 0.091036 .
## tvpubaff     0.11464    0.04337   2.643 0.008393 **
## tvnews       0.07204    0.05877   1.226 0.220725
## tvhrs       -0.08208    0.03936  -2.085 0.037412 *
## paper        0.36554    0.04417   8.275 6.05e-16 ***
## radio       -0.03469    0.03522  -0.985 0.324914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.9 on 734 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.3687, Adjusted R-squared:  0.3558
## F-statistic: 28.58 on 15 and 734 DF,  p-value: < 2.2e-16
```

```r
# levels of confidence
confint(m12)
```

```
##                     2.5 %      97.5 %
## (Intercept) -0.340311372  2.17485742
## education    0.143989646  0.31751958
## sex2        -0.786027932 -0.20143513
## empstat     -0.105798730  0.28851690
## ethnicity   -0.592893661 -0.11138551
## age<25      -2.635784034 -1.28057427
## age>65       0.015659314  1.18649502
## age25-34    -1.522900055 -0.57458698
## age35-44    -0.732643848  0.15059331
## age55-65    -0.297424229  0.77608975
## faminc      -0.006511021  0.08781703
## tvpubaff     0.029485860  0.19979217
## tvnews      -0.043348695  0.18741997
## tvhrs       -0.159355640 -0.00479536
## paper        0.278819116  0.45226676
## radio       -0.103823905  0.03444507
```

Adding the demographics to the Model impacts the TV Hours coefficient, controlled for all other variables, with an increase of 0.10 to -0.08, but the statistical significance in the population drops to 95% from Model 5 to the Final Model. Compared to Norris' coefficient, where TV Hours has a coefficient of -0.08, statistically significant at the 99% level. Controlled for Demographics and other Media variables, a one unit increase in TV Public Affairs effects Voting by 0.12, with statistical significance in the population at the 99% level. Controlled for Demographics and other Media variables, a one unit increase in TV News effects Voting by 0.12, and is statistically significant at the 99% level, and so the null hypothesis that there is no statistical difference in the population is rejected. Controlled for Demographics and other Media variables, Paper remains with a relatively strong coefficient of 0.36 with statistical significance in the population at the 99.9% level, a drop from Model 5's coefficient of 0.51. A one unit increase in Paper then effects Voting by 0.38 and the confidence interval suggests that this coefficient could be as low as 0.28, or high as 0.46, reflecting a robust coefficient. Similar to Model 5, Radio is still statistically insignificant in the population and the coefficient is tiny.

In terms of the overall Model in comparison to Norris' model it is somewhat reflective, especially amongst the Media variables and the Age variable. Despite treating Age as a Factor variable, the same logic Norris explains where older people seem to Vote more is reflected in our coefficients. <25 has a coefficient of -1.9 in reference to the 45-54 age group, which is a very big impact on Voting and is statistically significant in the population at the 99.9% level. The dummy variable >65 shows a 0.60 effect on Voting, statitsically significant at the 95% level, in reference to the 45-54 age group, suggesting that there is an important Age divide reflected which is also shown in Norris' Age coefficent of 0.30. This could point toward an incentive based interest in politics appealing to those who rely on the state for an income source and on state institutions as their health declines as Age increases.

In general, the demographics does not seem to have had much effects on the raw coefficients in 'Media', but has on their statistical significance in the population, either reducing their p-value level, or completely removing their statistical significance seen in Model 5. Demographics in our Final Model seems to have a much stronger effect on Voting than in Norris' Model, with much larger coefficients for Sex and Ethnicity. This is potentially due to not using weights where Latinos and African Americans are over sampled (Verba et al:1995). This could explain why ethnicity is so far off the Norris coefficient of 0.04 with no statistical significance, and our coefficient of -0.35 with a 99% level statistical significance.

```
anova(lm5, m12)
```

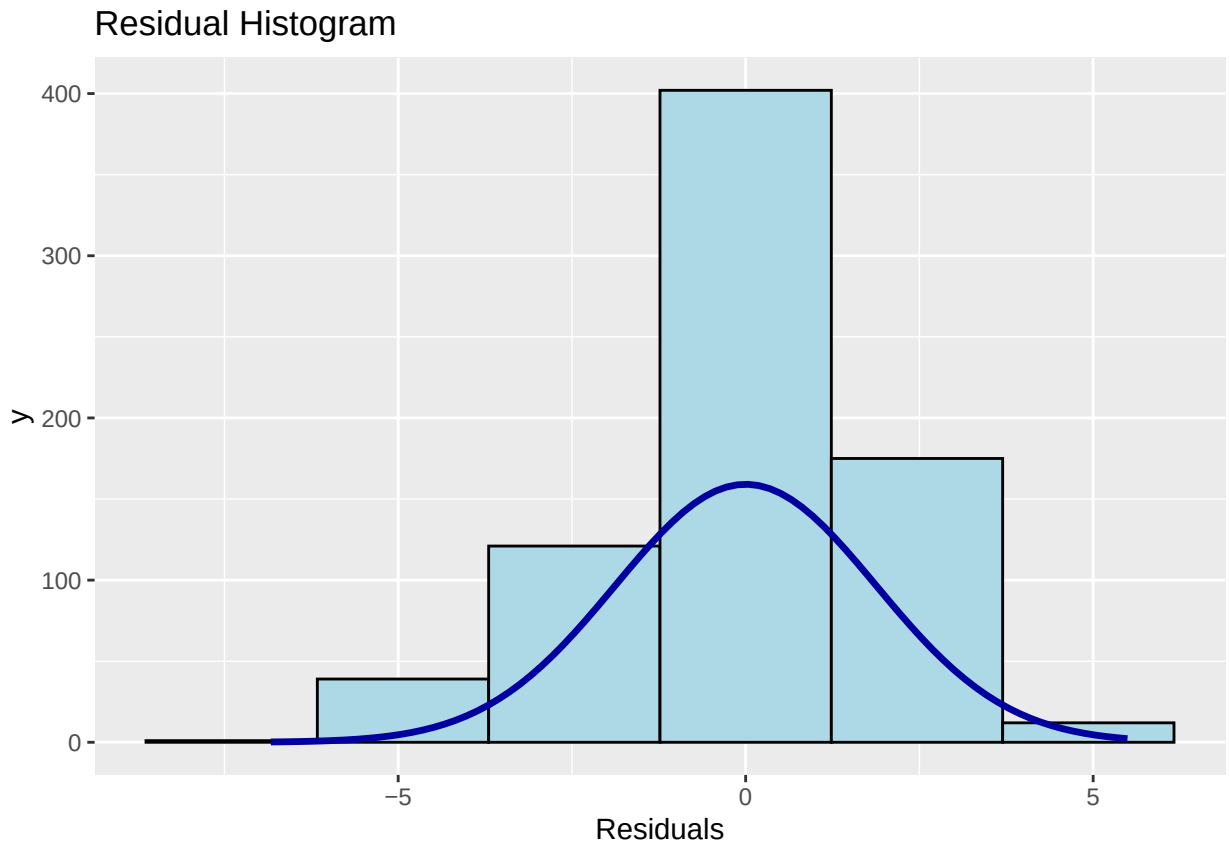
```
## Analysis of Variance Table
##
## Model 1: vote ~ tvhrs + tvnews + tvpubaff + paper + radio
## Model 2: vote ~ education + sex + empstat + ethnicity + age + faminc +
##     tvpubaff + tvnews + tvhrs + paper + radio
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    744 3242.5
## 2    734 2648.9 10    593.59 16.448 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The levels of confidence show that many coefficients have confidence intervals that sit between positive and negatives, for example, Employment Status, TV News, and Radio. However, there are some robust intervals, such as Education with a interval between 0.14 and 0.32, and Family Income at 0.002 and 0.08, and TV Hours between -0.15 and -0.4.

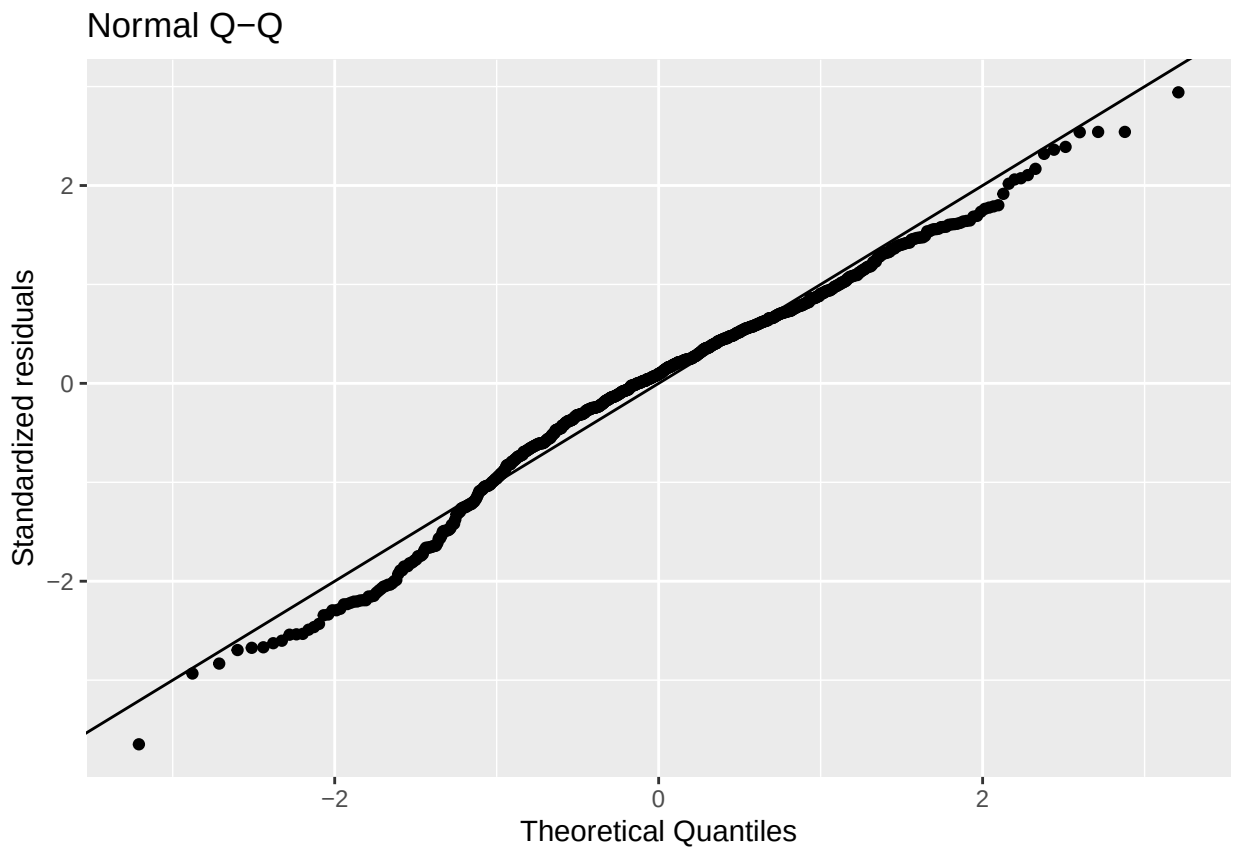
The Final Model has an R Squared of 36%, compared to 21% of Model 5, and 36% in Norris' Model. Thus, the Model explains 36% of the variance in Voting, level with Norris' and a good amount of extra variance in Voting is shown in this Model compared to Model 5. The Anova Test shows that there is a statistically significant difference between Model 5 and the Final Model in the population. The inclusion of demographics seems to have had a positive effect on the amount of variance the Model explains in Voting. The null hypothesis that there is no difference between the two models in the population is rejected.

## REGRESSION DIAGNOSTICS OF FINAL MODEL

```
# histogram of residuals of Final Model

ols_plot_resid_hist(m12)
```




```
# produces a Q-Q Plot to see whether our Residuals sit along a linear line

ggplot(m12) +
  stat_qq(aes(sample = .stdresid)) +
  geom_abline() +
  labs(title = "Normal Q-Q",
       y = "Standardized residuals",
       x = "Theoretical Quantiles")
```

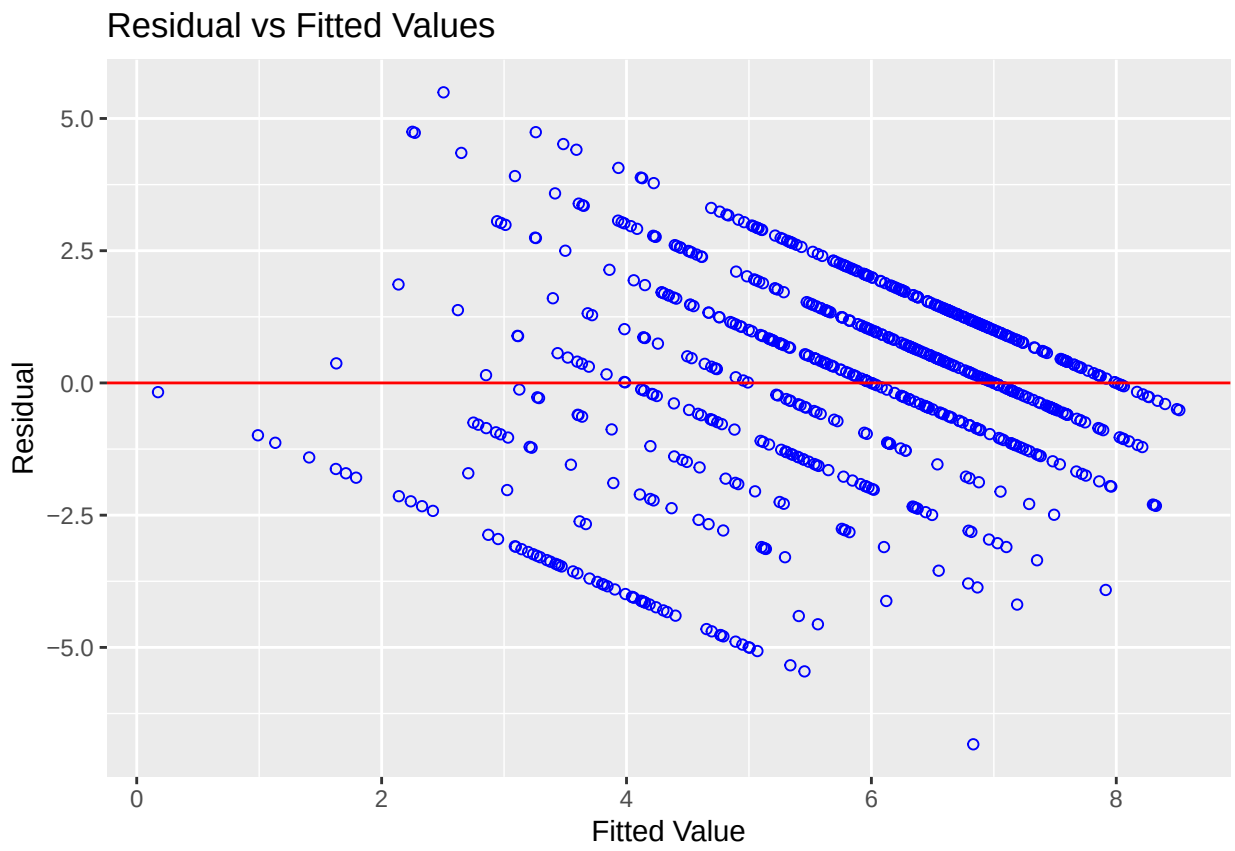



```
# group of normality tests - testing for whether there is normality in the residuals

ols_test_normality(m12)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9802         0.0000
## Kolmogorov-Smirnov        0.0675         0.0021
## Cramer-von Mises         40.2817         0.0000
## Anderson-Darling          5.1722         0.0000
## -----------------------------------------------
```

The residual histogram shows a normal distribution, and the Q-Q Plot seems to be close to the linear line showing a linear tendency but with wide tails. Reflecting on the Residuals in the Regression output the Min, Max, 1Q, 3Q are close together, and the Median at 0.24 is close to 0, together indicating a good model. In the Normality Tests all the p-values are under the 0.05 threshold meaning that the null hypothesis that the Models Residuals shows Normally Distributed (Homoskedasticity) residuals must be rejected, subsequently showing Heteroscedasticity (not normally distributed).

```
# testing for whether there is homoscedasticity - normal distribution - in the Models Residuals

ols_plot_resid_fit(m12)
```




```
bptest(m12)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m12
## BP = 128.27, df = 15, p-value < 2.2e-16
```

The Studentized Breusch-Pagan test shows a p-value below the 0.05 level and so the alternative hypothesis that there is Heteroscedasticity must be accepted, subsequently the heteroscedasticity assumption has not been fulfilled. In addition, the Fitted values vs Residuals plot shows a clear, non-random pattern of points on the plot suggesting the model can be improved.

```
# correlation test and inflation test to check for multi-collinearity

ols_coll_diag(m12)
```

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
##    Variables Tolerance      VIF
## 1  education 0.6814963 1.467359
## 2       sex2 0.8694869 1.150104
## 3    empstat 0.5869399 1.703752
## 4  ethnicity 0.8450625 1.183344
```

```
## 5     age<25 0.6876391 1.454251
## 6     age>65 0.4763884 2.099127
## 7   age25-34 0.4611237 2.168616
## 8   age35-44 0.4472403 2.235934
## 9   age55-65 0.6037881 1.656210
## 10    faminc 0.6529746 1.531453
## 11  tvpubaff 0.8249329 1.212220
## 12    tvnews 0.8854793 1.129332
## 13     tvhrs 0.8535960 1.171514
## 14     paper 0.7925848 1.261695
## 15     radio 0.9840553 1.016203
##
##
## Eigenvalue and Condition Index
## ------------------------------
##     Eigenvalue Condition Index    intercept    education         sex2
## 1  10.00482217        1.000000 1.147963e-04 8.850030e-04 2.437520e-03
## 2   1.08749320        3.033133 1.305832e-05 4.444561e-04 5.794289e-03
## 3   1.02037307        3.131304 6.535997e-07 2.124619e-04 2.340781e-04
## 4   1.00226630        3.159462 1.271479e-06 7.193802e-05 1.306698e-04
## 5   1.00077167        3.161820 9.253271e-07 3.538093e-06 9.701179e-05
## 6   0.47664163        4.581511 5.394156e-05 1.351190e-03 6.017822e-01
## 7   0.36865776        5.209463 4.037259e-05 2.206788e-04 1.299368e-03
## 8   0.32674447        5.533508 3.122813e-04 3.947972e-02 2.676096e-01
## 9   0.20935330        6.912972 3.301861e-04 4.107805e-02 2.656251e-02
## 10  0.14031688        8.444029 5.435068e-05 4.136967e-03 1.818402e-03
## 11  0.12019818        9.123380 1.369251e-04 3.445945e-04 3.744502e-03
## 12  0.08194731       11.049366 2.385273e-03 2.225596e-04 4.387740e-02
## 13  0.07706636       11.393897 6.530505e-04 8.705072e-01 2.591232e-02
## 14  0.04859010       14.349303 1.568671e-02 2.063278e-02 1.394194e-02
## 15  0.02547097       19.819011 1.323374e-02 1.888696e-02 1.489582e-03
## 16  0.00928665       32.822763 9.669825e-01 1.521891e-03 3.268618e-03
##         empstat    ethnicity       age<25       age>65     age25-34
## 1  6.762130e-04 1.177398e-03 0.0003795807 4.571861e-04 8.024415e-04
## 2  2.302215e-03 2.246089e-04 0.0396735537 2.071631e-01 5.460026e-03
## 3  2.506009e-04 1.302366e-03 0.3064591361 4.525791e-02 7.332372e-02
## 4  6.307295e-06 1.261640e-05 0.0320201986 6.165630e-02 2.226597e-02
## 5  4.707593e-06 1.170261e-06 0.1975707647 2.712369e-04 1.543925e-01
## 6  7.991073e-04 1.879229e-02 0.0083349753 1.053583e-02 3.226065e-03
## 7  1.255875e-05 4.564251e-03 0.0001396501 3.805651e-06 3.059485e-03
## 8  5.744303e-03 2.222033e-03 0.0231990719 2.119666e-02 3.984561e-04
## 9  4.260453e-03 1.284099e-01 0.0377592859 1.138508e-01 9.374229e-02
## 10 3.608915e-03 3.209767e-01 0.1518605432 1.888325e-01 2.691768e-01
## 11 4.942082e-02 1.803901e-01 0.0838008217 2.569154e-05 1.377151e-01
## 12 4.194797e-01 1.796364e-01 0.0339440005 1.739547e-02 1.169180e-01
## 13 6.100630e-02 3.911011e-02 0.0012031609 6.313885e-03 2.548778e-05
## 14 3.514339e-01 9.186439e-02 0.0131203330 2.155158e-01 8.890974e-03
## 15 1.533469e-02 1.253084e-04 0.0017534619 2.020905e-02 8.039263e-03
## 16 8.565923e-02 3.119032e-02 0.0687814619 9.131481e-02 1.025634e-01
##        age35-44     age55-65       faminc     tvpubaff       tvnews
## 1  0.0010223422 0.0006569451 9.855789e-04 1.063978e-03 3.246765e-04
## 2  0.0681508907 0.0078907305 1.082286e-03 2.642503e-04 1.200632e-04
## 3  0.0289240723 0.0369345552 1.269617e-04 5.715734e-04 8.367752e-06
## 4  0.0096880714 0.3726338086 1.212830e-04 4.407809e-06 1.744387e-08
```

```
## 5  0.0610188428 0.0304747423 9.510058e-07 5.758376e-07 1.038212e-06
## 6  0.0349191562 0.0005860875 3.235788e-03 5.861069e-06 2.038966e-04
## 7  0.0171723078 0.0021366253 7.629225e-04 5.773510e-04 3.298949e-04
## 8  0.0035912889 0.0005750232 3.765535e-02 7.720066e-04 7.498559e-04
## 9  0.1588507040 0.1493993278 1.016102e-01 9.801497e-03 4.302829e-04
## 10 0.3149026646 0.2125799070 8.953902e-03 1.730792e-01 2.377123e-03
## 11 0.1066889296 0.0270513032 6.221130e-02 5.832241e-01 4.963925e-05
## 12 0.1211781125 0.0501644192 4.161573e-01 5.053139e-04 1.411645e-02
## 13 0.0004385656 0.0076125213 3.068395e-01 3.434935e-02 3.745468e-03
## 14 0.0059557632 0.0570449010 2.528496e-02 1.918996e-01 2.757317e-01
## 15 0.0099566166 0.0135072748 3.383116e-02 1.392183e-03 5.168567e-01
## 16 0.0575416717 0.0307518280 1.140478e-03 2.488694e-03 1.849548e-01
##           tvhrs        paper        radio
## 1  1.986373e-03 2.263501e-04 2.477189e-03
## 2  1.148316e-02 1.620044e-05 7.364709e-04
## 3  7.131120e-04 2.619358e-05 4.568666e-05
## 4  5.050463e-05 2.110452e-07 4.987397e-07
## 5  2.163018e-06 1.378582e-06 2.193608e-04
## 6  5.864538e-02 1.514605e-06 3.942549e-02
## 7  1.701997e-01 1.477821e-05 7.573599e-01
## 8  2.985360e-01 1.166286e-03 1.498341e-01
## 9  2.734170e-01 2.596472e-06 6.662250e-03
## 10 7.186090e-02 6.218504e-05 1.943253e-02
## 11 2.134984e-02 4.897578e-08 1.197810e-04
## 12 1.377987e-02 3.358738e-03 5.504455e-04
## 13 4.697431e-03 1.287464e-03 3.710312e-03
## 14 5.180002e-02 6.928670e-02 3.192902e-03
## 15 4.730360e-05 5.167442e-01 7.216180e-03
## 16 2.143121e-02 4.078052e-01 9.016905e-03
```

Firstly, the Inflation Test does not show a value above or near 10 showing serious multicollinearity or a value over 4 which shows a area for concern, with the highest value belonging to Age 35-44 with the reference variable of 45-54 showing a value of 2.23. With every Variance inflation Factor (VIF) showing a value over 1 this suggests there is some level of correlation amongst the independent variables, but small amounts nevertheless.

## CONCLUSIONS

At times reproducing the Demographic variables where Norris said to refer to Verba, but did not always follow the advice of Verba, led to educated guesses from our group, which could have impacted the replication. The media variables were replicated well, with Norris' general point that Putnam's understanding of 'TV' being too simplistic, and that considering other Media outputs show higher levels of Voting. The Project has provided diagnostics and this goes beyond Norris in considering the robustness of the regression. However, the diagnostics would obviously be different when weighted, but we cannot be sure as they were not included in Norris' paper. We could not manage weights and standard errors, subsequently this seems to have heavily affected the replication of demographics especially where Ethnicity and Gender were particularly inaccurate, maybe also this could be attributed to their lower scales compared to the other variables in the model. An overall assessment suggests the paper could do with some work, the 'radcall2' inclusion seems to unnecessarily complicate the regression. Including an alternative model without radcall2 would be useful, a Project with a bigger scope could extend this Project's findings. In addition, the deductions in the paper may be a product of its time - media consumption has dramatically changed with the inclusion of social media - Voting tendencies could be eroded due to this?

## REFERENCES

Norris, P. (1996) *Does Television Erode Social Capital? A Reply to Putnam*, pp. 478-480. *Political Science and Politics, Vol.29, No.3.*

CODEBOOK - Sidney Verba, Kay Lehman Schlozman, Henry E. Brady, and Norman Nie, AMERICAN' CITIZEN PARTICIPATION STUDY, 1990 (ICPSR 6635)

Verba, S. (1995) *Voice and Equality.* Harvard University Press. Available at: https://www.perlego.com/book/2634073/voice-and-equality-pdf(Accessed: 1st January 2022)