

# W271 Section 3 Lab 1

*Daghan Altas, Zhaoning Yu, Hoang Phan*

*9/23/2017*

```
knitr::opts_chunk$set(cache=TRUE)
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

## 1) Problem statement

In this lab, we are going to investigate the relationship between voters' age, party affiliation, gender, race and preference for Bernie Sanders over Hillary Clinton.

Our objective is to derive an accurate model to help target marketing efforts on the likely buyers for Bernie Sanders T-shirts.

## Dataset

The dataset comes from the 2016 American National Election Survey.

```
library(dplyr)
library(ggplot2)
library(Hmisc)
library(GGally)
library(data.table)
library(stargazer)

if (dir.exists("/Users/daghanaltas/Hacking/Berkeley/W271/Labs/w271_lab1/")) {
  setwd("/Users/daghan/Hacking/Berkeley/W271/Labs/w271_lab1/")
} else if (dir.exists("/Users/daghan/Hacking/Berkeley/W271/Labs/w271_lab1/")) {
  setwd("/Users/daghan/Hacking/Berkeley/W271/Labs/w271_lab1/")
} else {
  setwd("~/Desktop/w271/Lab1")
}

df <- read.csv("./public_opinion.csv")
dt <- data.table(df)
head(dt)
```

```
##      sanders_preference party race_white gender birthyr
## 1:                    1      1          1      1    1960
## 2:                    0      2          1      2    1957
## 3:                    1      3          1      1    1963
## 4:                    1      1          1      1    1980
## 5:                    1      2          1      1    1974
## 6:                    1      2          1      1    1958
```

```
describe(dt)
```

```

## dt
##
## 5 Variables      1200 Observations
## -----
## sanders_preference
##      n missing distinct      Info      Sum      Mean      Gmd
##    1191      9        2    0.733      686    0.576    0.4889
##
## -----
## party
##      n missing distinct      Info      Mean      Gmd
##    1200      0        3    0.875      1.851    0.8309
##
## Value      1      2      3
## Frequency  459  461  280
## Proportion 0.382 0.384 0.233
## -----
## race_white
##      n missing distinct      Info      Sum      Mean      Gmd
##    1200      0        2    0.592      875    0.7292    0.3953
##
## -----
## gender
##      n missing distinct      Info      Mean      Gmd
##    1200      0        2    0.748      1.525    0.4992
##
## Value      1      2
## Frequency  570  630
## Proportion 0.475 0.525
## -----
## birthyr
##      n missing distinct      Info      Mean      Gmd      .05      .10
##    1200      0        73      1      1968      19.53      1940      1946
##      .25      .50      .75      .90      .95
##    1955      1968      1982      1991      1994
##
## lowest : 1921 1924 1925 1926 1927, highest: 1993 1994 1995 1996 1997
## -----

```

## Description of the data

The dataset contains 1200 samples with 5 variables:

- **sanders\_preference**: A categorical variable with 2 levels, denoting whether the voter prefers Bernie Sanders (=1) or Hillary Clinton (=0).
- **party**: A categorical variable with 3 levels, denoting whether the voter prefers is affiliated with the Democratic Party (=1), Independent (=2), or Republican Party (=3).
- **race\_white**: A categorical variable with 2 levels, denoting whether the voter is White (=1), or not (=0).
- **gender**: A categorical variable with 2 levels, denoting whether the voter is male (=1), or female (=2).
- **birthyr**: A numerical variable, denoting the birthyear of the voter.

## Observations:

- There are 9 missing values (NAs) for the **sanders\_\_preference** variable.
- There is no direct **age** variable. We'll derive it from the **birthyr** variable.
- None of the other variables have missing values.

## Clean-up

- We are going to create a new variable, **age** based on the **birthyr** variable through the following formula:

$$age = 2015 - birthyr$$

- Why 2015?  
We will use 2015, as the data was collected in 2016, since we are interested in the age of the voters *when* the data was collected. (All people born in 2015 will be counted as age=1, but this is only true if the survey was taken on 12/31/2016; suppose the survey was taken on 01/01/2016, then none of the people born in 2015 actually reached age 1, they are all actually age 0 when being surveyed, in this case “age = 2015 - birthyr” is the correct formula. We don't know when the survey was taken, but by changing the calculation to “age = 2015 - birthyr”, the minimum age is 18 now). If our assumption is wrong, there may be a significant omission in the data (the 18 year olds).
- We'll convert all categorical variables to R factor variables.

```
dt$sanders_preference <- as.factor(dt$sanders_preference)
dt$party <- factor(x = dt$party, levels = c("1", "2", "3"), labels = c("D",
  "I", "R"))
dt$race_white <- as.factor(dt$race_white)
dt$gender <- factor(x = dt$gender, levels = c("1", "2"), labels = c("M",
  "F"))
dt$age <- 2015 - dt$birthyr
```

- We are going to remove the 9 observations without the **sanders\_\_preference** value. A possible way to impute these **NA** values could be to use a logistic regression but for this work, we'll investigate if we can simply remove these observations (9 out of 1200 observations is less than 1% of the data). A table of the NA observations are detailed below:

```
dt_missing = dt[is.na(sanders_preference)]
dt_notmissing = dt[!is.na(sanders_preference)]
table = data.table(Data = c("NA's", "Remaining Data"), PartyDem = c(round(nrow(dt_missing[party ==
  "D"])/nrow(dt_missing), 2), round(nrow(dt_notmissing[party ==
  "D"])/nrow(dt_notmissing), 2)), PartyInd = c(round(nrow(dt_missing[party ==
  "I"])/nrow(dt_missing), 2), round(nrow(dt_notmissing[party ==
  "I"])/nrow(dt_notmissing), 2)), PartyRep = c(round(nrow(dt_missing[party ==
  "R"])/nrow(dt_missing), 2), round(nrow(dt_notmissing[party ==
  "R"])/nrow(dt_notmissing), 2)), Race_white = c(round(nrow(dt_missing[race_white ==
  1])/nrow(dt_missing), 2), round(nrow(dt_notmissing[race_white ==
  1])/nrow(dt_notmissing), 2)), GenderFemale = c(round(nrow(dt_missing[gender ==
  "F"])/nrow(dt_missing), 2), round(nrow(dt_notmissing[gender ==
  "F"])/nrow(dt_notmissing), 2)), MeanAge = c(round(dt_missing[,
  mean(age)], 2), round(dt_notmissing[, mean(age)], 2)))
table
```

##	Data	PartyDem	PartyInd	PartyRep	Race_white	GenderFemale
## 1:	NA's	0.44	0.33	0.22	0.67	0.89
## 2:	Remaining Data	0.38	0.38	0.23	0.73	0.52
##	MeanAge					

```
## 1: 49.89
## 2: 47.04
```

```
dt <- dt[!is.na(dt$sanders_preference), ]
```

We observe that 8 out of 9 NAs are female and this could be problematic. However, i) given that gender is not a meaningful explanatory variable for preference\_bernies (see relevant section below in univariate analysis) and ii) the relatively low number of data points with missing values (9 out of 1200), and iii) that all other variables deviate very little between the NAs subset and the rest of the data, we believe we can justify our decision to remove NA subset from the original dataset.

## 2) Exploratory Data Analysis

### Univariate analysis

#### Sanders vs. Hillary Preference

```
# xtabs( ~ sanders_preference, data=dt)
c.table <- array(data = c(sum(dt$sanders_preference == 1), sum(dt$sanders_preference ==
1)/length(dt$sanders_preference), sum(dt$sanders_preference ==
0), sum(dt$sanders_preference == 0)/length(dt$sanders_preference),
sum((dt$sanders_preference == 0) | (dt$sanders_preference ==
1)), sum((dt$sanders_preference == 0) | (dt$sanders_preference ==
1))/length(dt$sanders_preference)), dim = c(2, 3), dimnames = list(Count = c("Voter Count",
"pi.hat"), Preference = c("Bernie", "Hillary", "Total")))

round(c.table, 2)

##              Preference
## Count      Bernie Hillary Total
## Voter Count 686.00  505.00 1191
## pi.hat      0.58    0.42    1
```

In this survey, 58% of voters prefer Bernie over Hillary. This is a larger than expected Bernie preference (since Hillary Clinton won the Democratic nomination, one would expect to see a higher ratio for Clinton than Bernie). One possible explanation is that Bernie is more popular among the Independents and Republicans. So we are going to assume that this sample does not exhibit a selection bias (sample set is random).

#### Party affiliations

```
c.table <- array(data = c(sum(dt$party == "D"), sum(dt$party ==
"D)/length(dt$party), sum(dt$party == "I"), sum(dt$party ==
"I)/length(dt$party), sum(dt$party == "R"), sum(dt$party ==
"R)/length(dt$party), sum((dt$party == "D") | (dt$party ==
"I") | (dt$party == "R")), sum((dt$party == "D") | (dt$party ==
"I") | (dt$party == "R))/length(dt$party)), dim = c(2, 4),
dimnames = list(Count = c("Voter Count", "pi.hat"), Party_affiliation = c("Democrat",
"Independent", "Republican", "Total")))

round(c.table, 2)

##              Party_affiliation
```

```
## Count      Democrat Independent Republican Total
## Voter Count 455.00      458.00      278.00 1191
## pi.hat      0.38        0.38        0.23   1
```

We observe that 38% of the voters in the dataset are affiliated with the Democratic Party, whereas only 23% are affiliated with the Republican Party. The 0.6 Democratic to Republican ratio is noteworthy. The data appears to be skewed towards Democratic voters (perhaps a specific region of the country). Our model may not be applicable to the entire country. A further analysis of how the dataset was sampled from the entire population would be very useful.

## Race

```
c.table <- array(data = c(sum(dt$race_white == 1), sum(dt$race_white ==
1)/length(dt$race_white), sum(dt$race_white == 0), sum(dt$race_white ==
0)/length(dt$race_white), sum((dt$race_white == 1) | (dt$race_white ==
0)), sum((dt$race_white == 1) | (dt$race_white == 0))/length(dt$race_white)),
dim = c(2, 3), dimnames = list(Count = c("Voter Count", "pi.hat"),
Race = c("White", "Non White", "Total")))

round(c.table, 2)
```

```
##          Race
## Count      White Non White Total
## Voter Count 869.00    322.00 1191
## pi.hat      0.73      0.27   1
```

The 73% white / non-white ratio is inline with the overall US population (according to the 2016 US Census, whites made up 72.4% of the population). The dataset does not appear to have a selection bias with respect to the voter race.

## Gender

```
c.table <- array(data = c(sum(dt$gender == "M"), sum(dt$gender ==
"M)/length(dt$gender), sum(dt$gender == "F"), sum(dt$gender ==
"F)/length(dt$gender), sum((dt$gender == "M") | (dt$gender ==
"F")), sum((dt$gender == "M") | (dt$gender == "F))/length(dt$gender)),
dim = c(2, 3), dimnames = list(Count = c("Voter Count", "pi.hat"),
Gender = c("Male", "Female", "Total")))

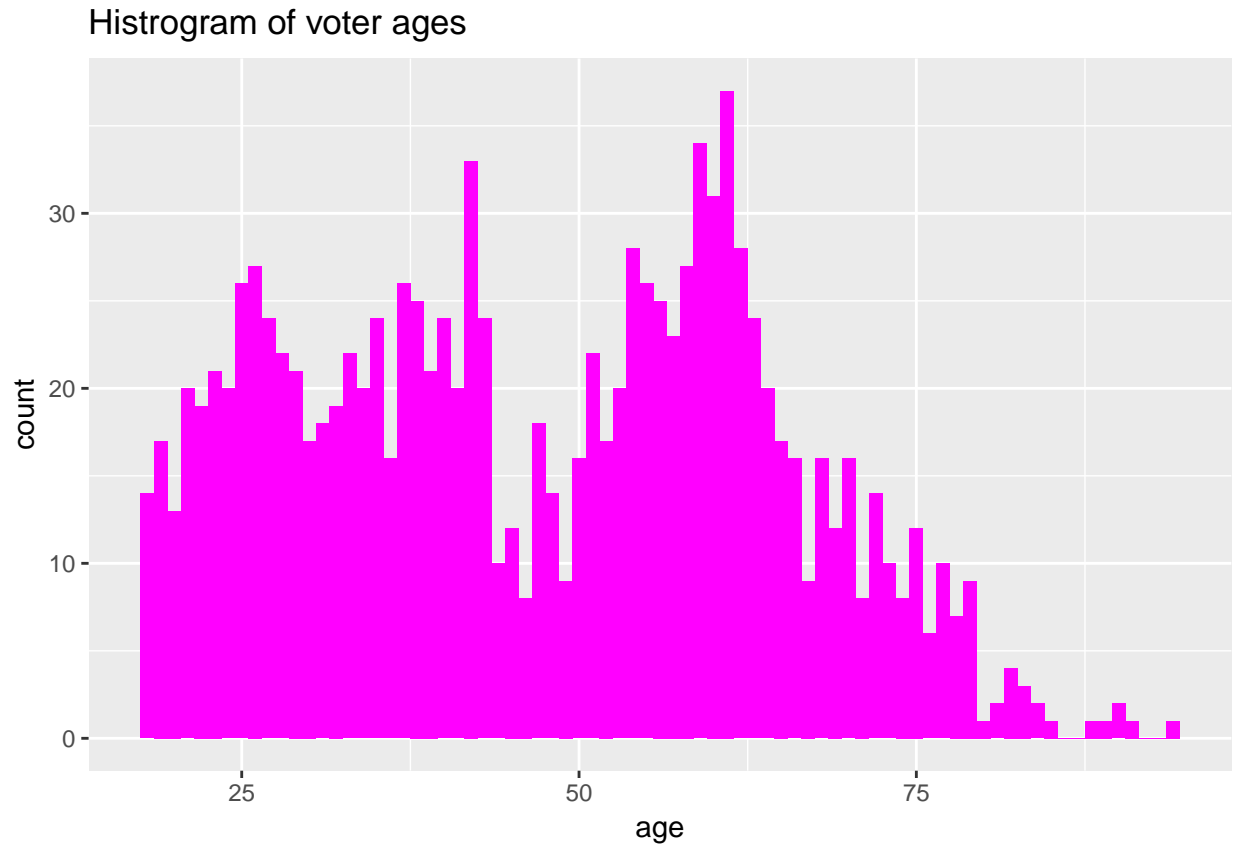
round(c.table, 2)
```

```
##          Gender
## Count      Male Female Total
## Voter Count 569.00 622.00 1191
## pi.hat      0.48  0.52   1
```

The female / male ratio is 1.10 in the dataset and the population ratio (according to Wikipedia) is 1.05. So the sample data doesn't appear to have a strong skew.

## Age

```
ggplot(dt, aes(age)) + geom_histogram(binwidth = 1, fill = "magenta") +
  ggtitle("Histogram of voter ages")
```



```
c.table <- array(data = c(range(dt$age)), dim = c(1, 2), dimnames = list(c("Voter Age"),
  c("Youngest", "Oldest")))
round(c.table, 2)
```

```
##           Youngest Oldest
## Voter Age       18     94
```

We observe that age data for the voters in the survey is within the norms (i.e, there is no one below 18) and as the age variable moves beyond 70, there is a rapid decline. However the data appears to be tri-modal. There is not an obvious explanation for that. This may point to a selection bias (i.e, the sample isn't trully random). We'll note the shortcoming in our final analysis as a caution.

## Multivariate analysis

At this section, we are going to focus on the relationship between the outcome (**sanders\_preference**) and possible explanatory variables including:

- preference vs. party
- preference vs. race
- preference vs. gender

## preference vs. party

```
c.tabs <- xtabs(~sanders_preference + party, data = dt)
c.tabs <- rbind(c.tabs, colSums(c.tabs))
c.tabs <- cbind(c.tabs, rowSums(c.tabs))

c.table <- array(data = as.array(c.tabs), dim = c(3, 4), dimnames = list(Preference = c("Hillary",
"Bernie", "Total"), Party = c("Democratic", "Independent",
"Republican", "Total")))
round(c.table, 2)
```

```
##           Party
## Preference Democratic Independent Republican Total
##   Hillary          249          156          100   505
##   Bernie           206           302          178   686
##   Total            455           458          278  1191
```

```
round(c.table/dim(dt)[1], 2)
```

```
##           Party
## Preference Democratic Independent Republican Total
##   Hillary          0.21          0.13          0.08  0.42
##   Bernie           0.17          0.25          0.15  0.58
##   Total            0.38          0.38          0.23  1.00
```

```
chisq.test(c.tabs)
```

```
##
## Pearson's Chi-squared test
##
## data:  c.tabs
## X-squared = 46.046, df = 6, p-value = 2.898e-08
```

We have further proof that Bernie is popular with the wrong group (i.e, Independents and Republicans), which is a point we touched on the univariate analysis section for the preference variable. While he enjoys roughly 2x the popularity of Hillary Clinton among the Independents and Republicans, he is less popular among the Democrats. Our intuition to include party affiliation as an explanatory variable is backed by the Chi-Square independence test. With the p-value close to 0, we can accept the alternative hypothesis that party and preference\_bernies are dependent and therefore **party variable should be added to the model.**

## preference vs. race\_white

```
c.tabs <- xtabs(~sanders_preference + race_white, data = dt)
c.tabs <- rbind(c.tabs, colSums(c.tabs))
c.tabs <- cbind(c.tabs, rowSums(c.tabs))

c.table <- array(data = as.array(c.tabs), dim = c(3, 3), dimnames = list(Preference = c("Hillary",
"Bernie", "Total"), Race = c("Non whites", "Whites", "Total")))
round(c.table, 2)
```

```
##           Race
## Preference Non whites Whites Total
##   Hillary          190          315   505
```

```
##      Bernie      132    554    686
##      Total      322    869   1191
```

```
round(c.table/dim(dt)[1], 2)
```

```
##              Race
## Preference Non whites Whites Total
##      Hillary      0.16    0.26  0.42
##      Bernie       0.11    0.47  0.58
##      Total        0.27    0.73  1.00
```

```
chisq.test(c.tabs)
```

```
##
##  Pearson's Chi-squared test
##
## data:  c.tabs
## X-squared = 49.823, df = 4, p-value = 3.932e-10
```

Bernie gets a 40% preference from non\_white voters but gets 65% from White voters. Also the Chi-square independent test statistic has a very low p-value, so there is strong empirical evidence to accept the alternative hypothesis that race and bernie\_preference are not independent. **We will add race as a dependent variable to our model.**

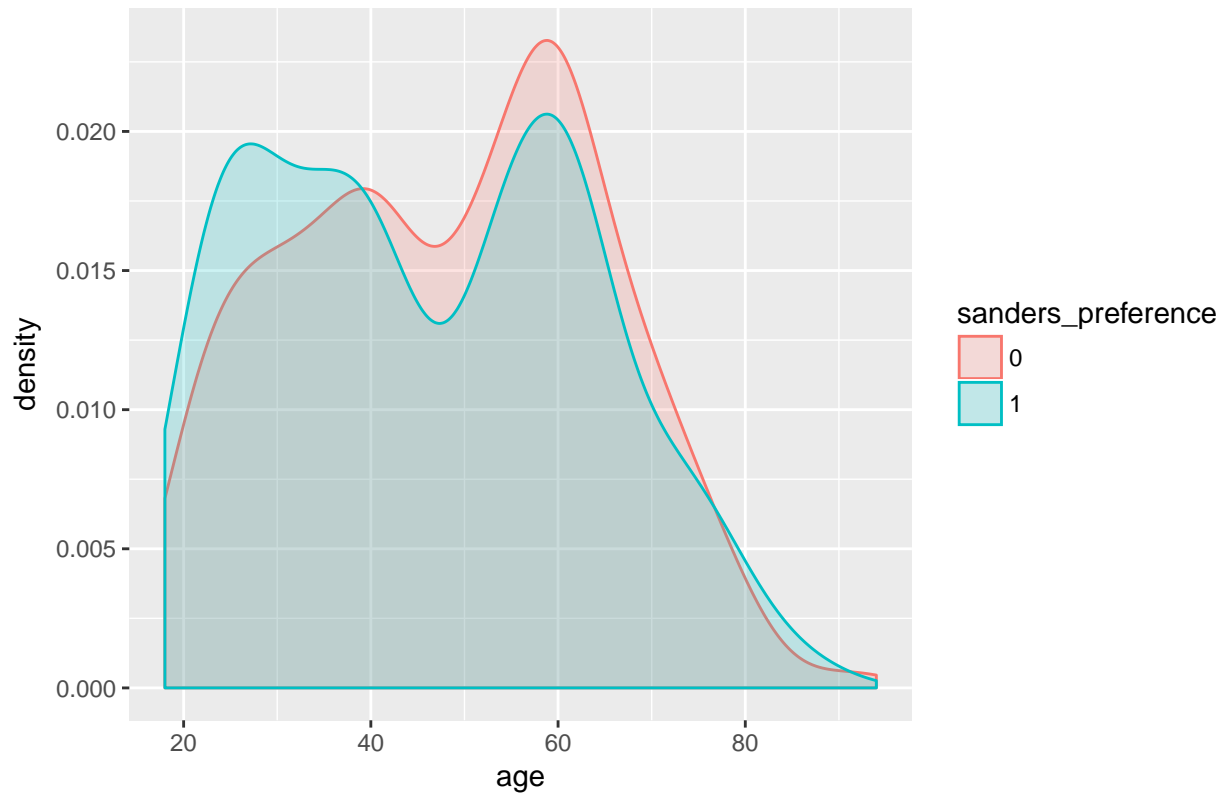
## preference vs. gender

We are going to plot a density plot for the voter age conditioned on sanders\_preference to investigate.

```
ggplot(dt, aes(age, fill = sanders_preference, color = sanders_preference)) +
  geom_density(alpha = 0.2) + ggtitle("Sanders_preference for each age") +
  theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



## Sanders\_preference for each age



```
t.test(dt$age ~ dt$sanders_preference)
```

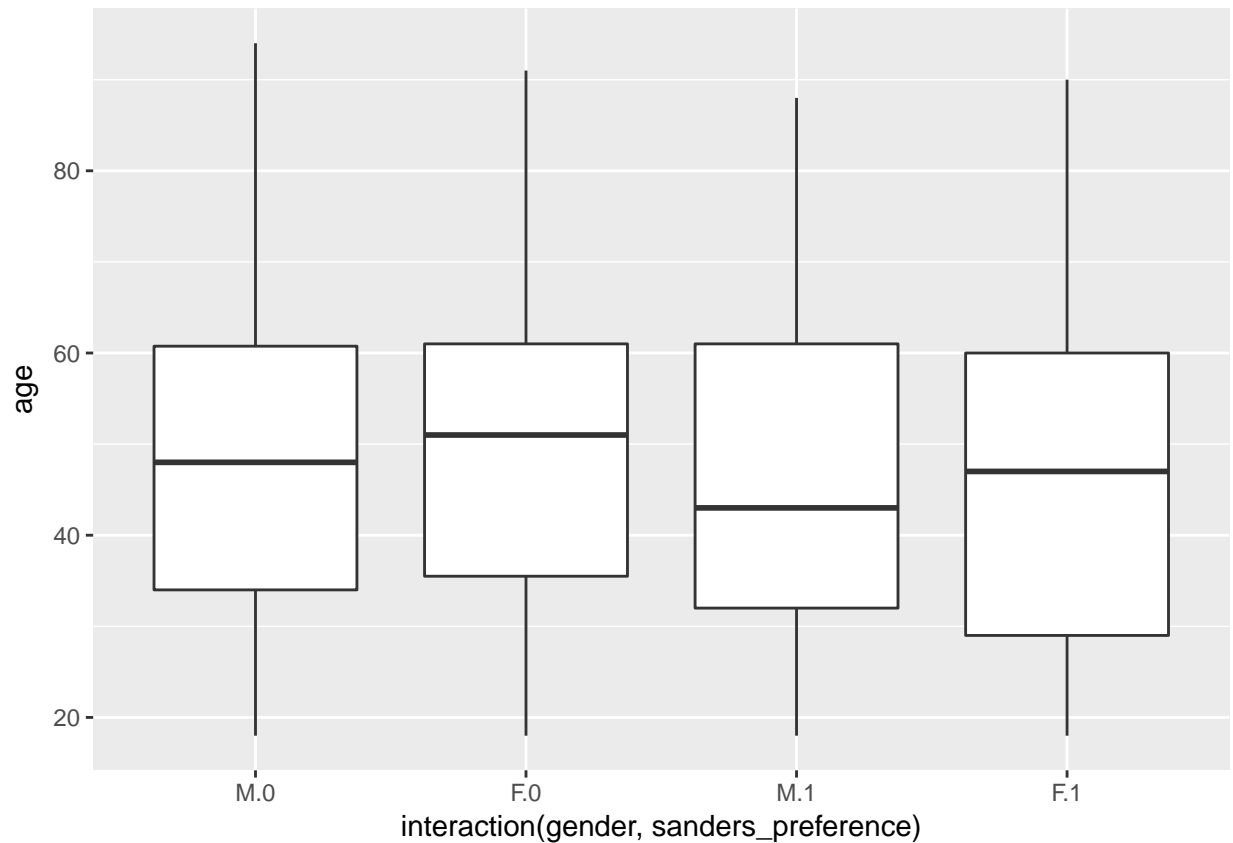
```
##
##  Welch Two Sample t-test
##
## data:  dt$age by dt$sanders_preference
## t = 2.1991, df = 1116.5, p-value = 0.02808
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2342468 4.1128247
## sample estimates:
## mean in group 0 mean in group 1
##      48.29307      46.11953
```

The density plot indicates that there is a clear difference in favor of Bernie Sanders for voters below the age of 40 and somewhat of a reverse effect for the voters above the age of 40. Also the t-test shows that there is reasonable evidence to reject to null hypothesis and accept that the mean age of voters who prefer Bernie Sanders is younger than those who prefer Hillary Clinton. **We will include age in our model.**

## Interactions

preference ~ age & gender

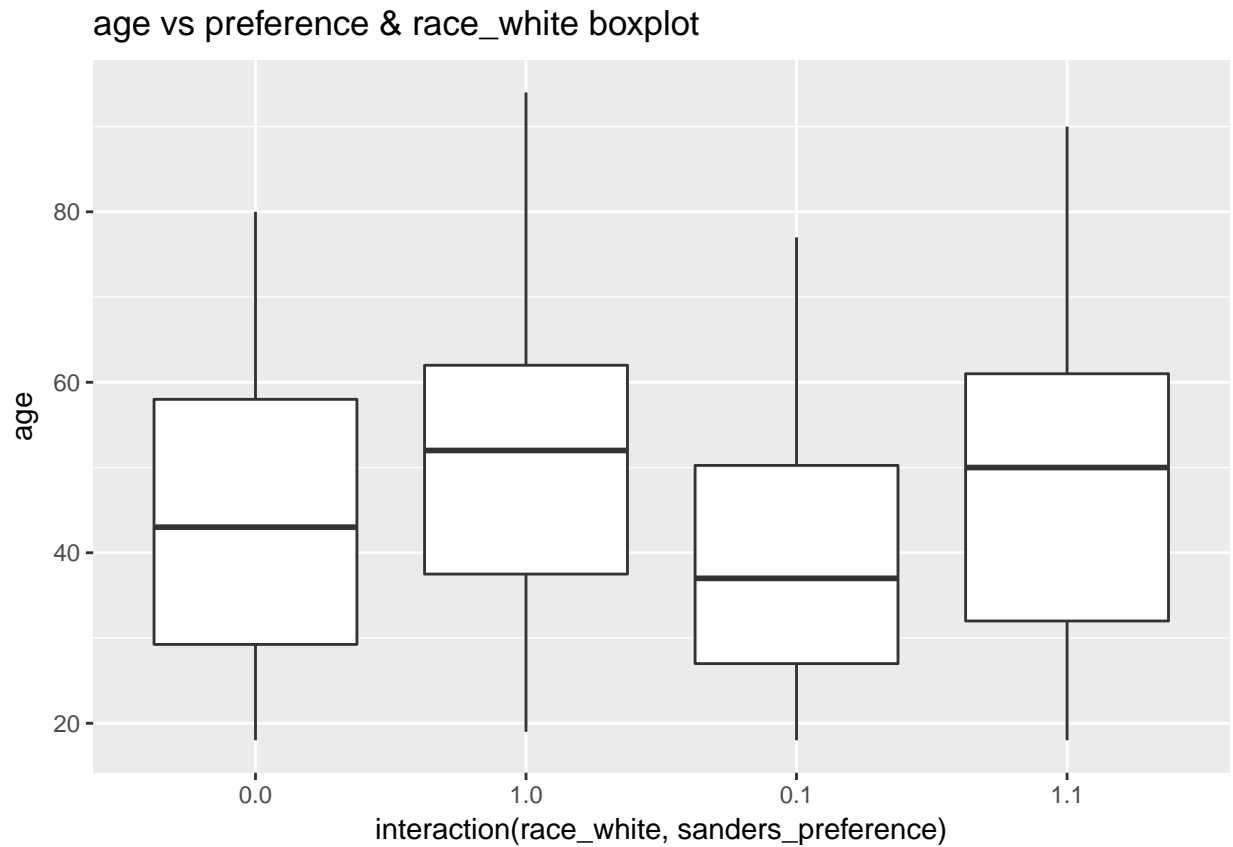
```
ggplot(dt, aes(x = interaction(gender, sanders_preference), y = age)) +  
  geom_boxplot()
```



There is a mild interaction between age, gender and sanders\_preference (males who prefer Sanders have the lowest mean age). Yet, we remain skeptical that there is evidence of an interaction between age and gender variables as an explanatory variable for the model, given that we've decided against using gender as an explanatory variable. However, we'll conduct a model test in the next section.

preference ~ age & race\_white

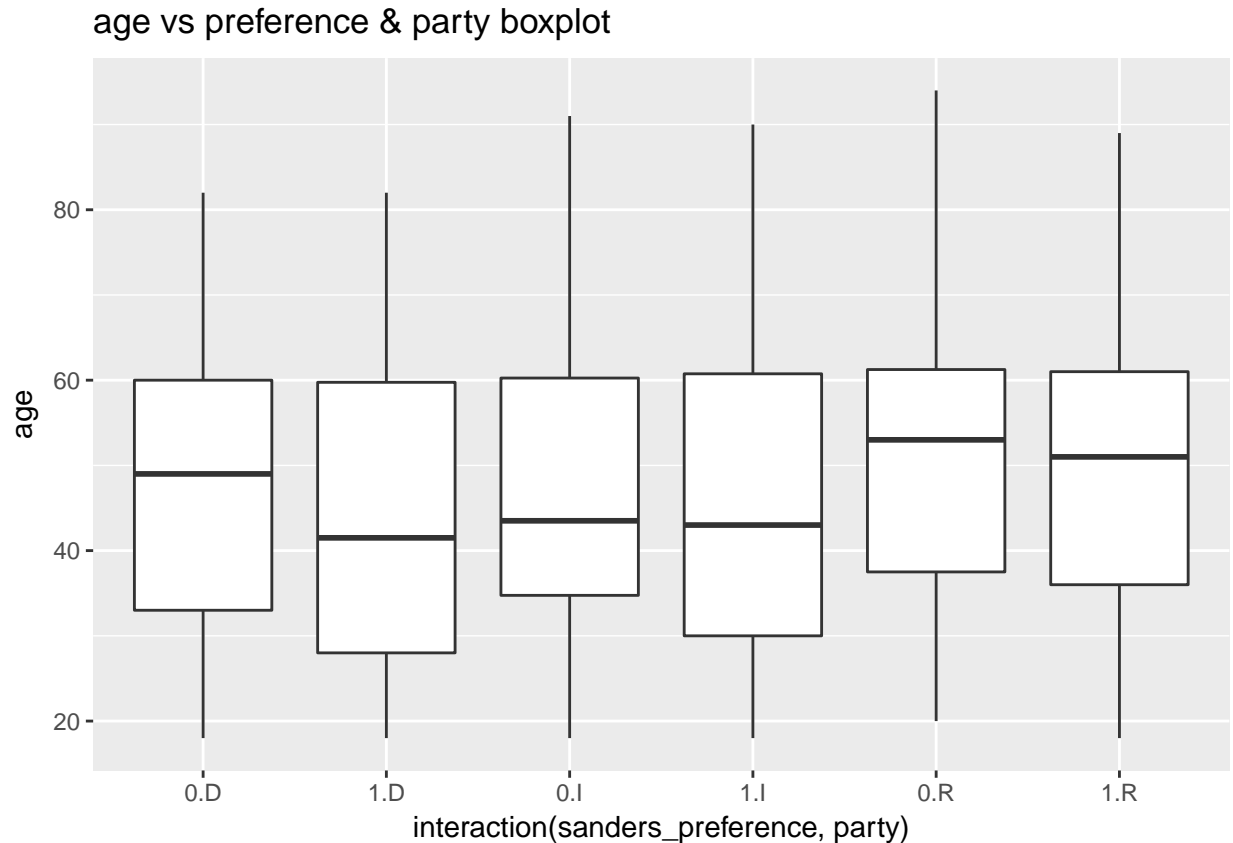
```
ggplot(dt, aes(x = interaction(race_white, sanders_preference),  
  y = age)) + geom_boxplot() + ggtitle("age vs preference & race_white boxplot")
```



Non-white voters who prefer Bernie are the youngest group (based on their 1st, 3rd quartile and mean age). We'll explore the interaction between age and race during our model exploration.

preference ~ age & party

```
ggplot(dt, aes(x = interaction(sanders_preference, party), y = age)) +  
  geom_boxplot() + ggtitle("age vs preference & party boxplot")
```



For Republicans and Independents, the mean and 3rd quartiles for age don't change that much based on sanders\_preference. But 1st quartile for Independents as well as the mean age for Democrats changes noticeably based on sanders\_preference. **We are going to explore the interaction between the age and party variables.**

### 3) Models

Based on the exploratory data analysis, our model exploration strategy is as follows:

- Investigate whether gender is a significant explanatory variable or not.
- Investigate that both **party** and **race\_white** are meaningful explanatory variables.
- Investigate additional interactions:
  - Interaction between age and race
  - Interaction between age and party
  - Interaction between party and race
  - Interaction between age and gender

#### Is gender important?

Our exploratory analysis has indicated that there is no empirical evidence to include gender as an explanatory variable. For the sake of thoroughness, we'll conduct one final analysis.

## Null Hypothesis

In our null hypothesis we are going to assume that  $\beta_{gender} = 0$

```
model1.H0 = glm(sanders_preference ~ age, dt, family = binomial(link = "logit"))
summary(model1.H0)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age, family = binomial(link = "logit"),
##      data = dt)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4074  -1.2842   0.9841   1.0620   1.1840
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.661384   0.173958   3.802 0.000144 ***
## age         -0.007522   0.003457  -2.176 0.029566 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1618.7  on 1189  degrees of freedom
## AIC: 1622.7
##
## Number of Fisher Scoring iterations: 4
```

## Alternative hypothesis

In the alternative hypothesis, we are going to assume that  $\beta_{gender} \neq 0$

```
model1.Ha = glm(sanders_preference ~ age + gender, dt, family = binomial(link = "logit"))
summary(model1.Ha)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + gender, family = binomial(link = "logit"),
##      data = dt)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4111  -1.2869   0.9838   1.0625   1.1816
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.653207   0.183064   3.568 0.000359 ***
## age         -0.007535   0.003458  -2.179 0.029346 *
## genderF       0.016857   0.117657   0.143 0.886079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1623.5 on 1190 degrees of freedom
## Residual deviance: 1618.7 on 1188 degrees of freedom
## AIC: 1624.7
##
## Number of Fisher Scoring iterations: 4
```

## Comparing H0 to Ha

```
anova(model1.H0, model1.Ha, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age
## Model 2: sanders_preference ~ age + gender
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1 1189 1618.7
## 2 1188 1618.7 1 0.020525 0.8861
```

We've confirmed once again that there is no evidence supporting the alternative hypothesis that **gender** is an explanatory variable given that **age** is in the model. We fail to reject the null hypothesis that  $\beta_{gender} = 0$ .

## Adding race and party affiliation

We are now going to include race and party. To recap:

- Null hypothesis  $H_0$ :  

$$- \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age}$$
- Alternative hypothesis  $H_a$ :  

$$- \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white}$$

```
model3.H0 = glm(sanders_preference ~ age, dt, family = binomial(link = "logit"))
model3.Ha = glm(sanders_preference ~ age + party + race_white,
  dt, family = binomial(link = "logit"))
summary(model3.Ha)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + party + race_white,
##      family = binomial(link = "logit"), data = dt)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7036  -1.1792   0.7907   0.9881   1.6662
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.139978   0.199715  -0.701 0.483374
## age         -0.012480   0.003666  -3.404 0.000664 ***
## partyI       0.713501   0.140368   5.083 3.71e-07 ***
## partyR       0.594231   0.162972   3.646 0.000266 ***
## race_white1  0.872782   0.141872   6.152 7.66e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1533.2  on 1186  degrees of freedom
## AIC: 1543.2
##
## Number of Fisher Scoring iterations: 4
# Comparing H0 and Ha
anova(model3.H0, model3.Ha, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age
## Model 2: sanders_preference ~ age + party + race_white
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1189      1618.7
## 2      1186      1533.2  3    85.511 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values for the **party(I|R)** and **race\_white1** explanatory variables are very low. In addition, the p-value for Ha (model3.Ha) is also very low. We conclude that there is strong empirical evidence for including race and party affiliation in our model.

## Exploring models with interactions

We are going to construct 4 additional models to explore the interactions:

- Between race and age
- Between party and age
- Between party and race
- Between gender and age

```
model4.H0 = glm(sanders_preference ~ age + party + race_white,
  dt, family = binomial(link = "logit")) # also = H3.HA
model4.Ha1 = glm(sanders_preference ~ age + party + race_white +
  age:race_white, dt, family = binomial(link = "logit"))
model4.Ha2 = glm(sanders_preference ~ age + party + race_white +
  age:party, dt, family = binomial(link = "logit"))
model4.Ha3 = glm(sanders_preference ~ age + party + race_white +
  party:race_white, dt, family = binomial(link = "logit"))
model4.Ha4 = glm(sanders_preference ~ age + party + race_white +
  age:gender, dt, family = binomial(link = "logit"))
stargazer(model4.H0, model4.Ha1, model4.Ha2, model4.Ha3, model4.Ha4,
  type = "text", report = ("vc*p"))
```

```
##
## =====
##                               Dependent variable:
##   -----
##                               sanders_preference
##                               (1)      (2)      (3)      (4)      (5)
```

```
## -----
## age          -0.012***   -0.021***   -0.017***   -0.012***   -0.013***
##              p = 0.001    p = 0.008    p = 0.003    p = 0.001    p = 0.001
##
## partyI       0.714***    0.709***    0.360      0.749***    0.721***
##              p = 0.00000  p = 0.00000  p = 0.377    p = 0.004    p = 0.00000
##
## partyR       0.594***    0.587***    0.151      0.574      0.597***
##              p = 0.0003   p = 0.0004   p = 0.757    p = 0.121    p = 0.0003
##
## race_white1  0.873***    0.419      0.874***    0.888***    0.874***
##              p = 0.000    p = 0.297    p = 0.000    p = 0.00001  p = 0.000
##
## age:race_white1      0.010
##                      p = 0.229
##
## age:partyI          0.008
##                    p = 0.354
##
## age:partyR          0.009
##                    p = 0.329
##
## partyI:race_white1  -0.051
##                    p = 0.869
##
## partyR:race_white1  0.019
##                    p = 0.964
##
## age:genderF          0.001
##                    p = 0.643
##
## Constant          -0.140      0.195      0.088      -0.150      -0.141
##                    p = 0.484    p = 0.569    p = 0.759    p = 0.497    p = 0.480
## -----
## Observations      1,191      1,191      1,191      1,191      1,191
## Log Likelihood    -766.601    -765.868    -765.960    -766.582    -766.493
## Akaike Inf. Crit. 1,543.201    1,543.736    1,545.920    1,547.163    1,544.986
## =====
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

We observe that none of the alternative models have any new explanatory variable with a low p-value. We'll confirm our findings with a pairwise anova table  $H_0$  vs.  $H_{a1}$ ,  $H_0$  vs.  $H_{a2}$ ,  $H_0$  vs.  $H_{a3}$  \$, and  $H_0$  vs.  $H_{a4}$

```
model4.1.anova <- anova(model4.H0, model4.Ha1, test = "Chisq")
model4.2.anova <- anova(model4.H0, model4.Ha2, test = "Chisq")
model4.3.anova <- anova(model4.H0, model4.Ha3, test = "Chisq")
model4.4.anova <- anova(model4.H0, model4.Ha4, test = "Chisq")

model4.anova.results <- data.frame(c(model4.1.anova$`Pr(>Chi)`[2],
  model4.2.anova$`Pr(>Chi)`[2], model4.3.anova$`Pr(>Chi)`[2],
  model4.4.anova$`Pr(>Chi)`[2]))
colnames(model4.anova.results) <- "Pr(>Chi)"
rownames(model4.anova.results) <- c("model4.Ha1", "model4.Ha2",
```



```
"model4.Ha3", "model4.Ha4")
model4.anova.results
```

```
##           Pr(>Chi)
## model4.Ha1 0.2260354
## model4.Ha2 0.5269330
## model4.Ha3 0.9810663
## model4.Ha4 0.6423527
```

None of the 4 interaction models yields a test statistic satisfactory to reject the null hypothesis. **There is no empirical evidence of any interaction between age and race, age and party, party and race, and age and gender.**

## Model selection summary

Based on our model study, we conclude that the most appropriate model for the specific question we want to answer (*whether it is a good idea to focus on younger voters*) given the dataset is:

### Description of the model (answer 1.a)

Our model estimates the log odds of preferring Bernie Sanders over Hillary Clinton as a linear combination of the age of the voter, party affiliation and race.

$$\begin{aligned} \text{logit}(\pi_{\text{preference\_sanders}}) &= \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white} \\ \log\left(\frac{\hat{\pi}_{\text{preference\_sanders}}}{1 - \hat{\pi}_{\text{preference\_sanders}}}\right) &= -0.1340 - 0.0125\text{age} + 0.7135\text{partyI} + 0.5942\text{partyR} + 0.8728\text{race\_white} \\ \hat{\pi}_{\text{preference\_sanders}} &= \frac{e^{-0.1340 - 0.0125\text{age} + 0.7135\text{partyI} + 0.5942\text{partyR} + 0.8728\text{race\_white}}}{1 + e^{-0.1340 - 0.0125\text{age} + 0.7135\text{partyI} + 0.5942\text{partyR} + 0.8728\text{race\_white}}} \end{aligned}$$

### Description of the variables and model's functional form (answer 1.b)

Based on the EDA (section 2), we have shown that **age, party affiliation and race** place an important role for the probability of a voter choosing Bernie Sanders over Hillary Clinton. We chose logistic regression because the dependent variable is a binary.

### Other models that have been considered (answer 1.c)

Based on the model investigation above (section 3), we have looked into the following 4 additional models:

- 1)  $H_{a1} : \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white} + \beta_5 \text{age} * \text{race\_white}$
- 2)  $H_{a2} : \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white} + \beta_5 \text{age} * \text{partyI} + \beta_6 \text{age} * \text{partyR}$
- 3)  $H_{a3} : \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white} + \beta_5 \text{partyI} * \text{race\_white} + \beta_6 \text{partyR} * \text{race\_white}$
- 4)  $H_{a4} : \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white} + \beta_5 \text{age} * \text{genderF}$

And compared them to our  $H_0 : \text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white}$   
 Our conclusion based on the section above is that we don't have enough evidence to reject the null hypothesis.  
 Our final model is  $H_0$ .

## Final model results (answer 1.d)

```
summary(model4.H0)

##
## Call:
## glm(formula = sanders_preference ~ age + party + race_white,
##      family = binomial(link = "logit"), data = dt)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7036  -1.1792   0.7907   0.9881   1.6662
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.139978   0.199715  -0.701 0.483374
## age         -0.012480   0.003666  -3.404 0.000664 ***
## partyI       0.713501   0.140368   5.083 3.71e-07 ***
## partyR       0.594231   0.162972   3.646 0.000266 ***
## race_white1  0.872782   0.141872   6.152 7.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1533.2  on 1186  degrees of freedom
## AIC: 1543.2
##
## Number of Fisher Scoring iterations: 4
```

## Relevant tests for the model (answer 1.e)

All the coefficients  $\beta_1$  through  $\beta_4$  have very low p-values. However the p-value for  $\beta_0$  (i.e, the intercept) is quite high. But that doesn't mean that the model is bad or should be thrown out. Rather, in this example, the intercept does not have a meaning (there are no 0 year old voters).

## Confidence intervals for the model

```
confint(model4.H0, level = 0.95)

## Waiting for profiling to be done...
##
##              2.5 %      97.5 %
## (Intercept) -0.53250475  0.250921131
## age         -0.01969964 -0.005319264
## partyI       0.43913512  0.989620714
```

```
## partyR      0.27611947  0.915385761
## race_white1 0.59580662  1.152289588
```

## Odds Ratios

### Odds Ratio for age

The estimated odds ratio for age is

$$\widehat{OR}_{age} = \exp(c\hat{\beta}_1)$$

With  $c = -10$ , we calculate the odds ratio and 95% intervals.

Odds ratio:

```
# Odds ratio for age
round(exp(-10 * model4.H0$coefficients[2]), 3)
```

```
## age
## 1.133
```

95% Wald interval:

```
# Wald interval for the age odds ratio
beta.ci <- confint.default(object = model4.H0, parm = "age",
  level = 0.95)
round(rev(exp(-10 * beta.ci)), 3)
```

```
## [1] 1.054 1.217
```

95% LR interval:

```
# LR interval for the age odds ratio
beta.ci <- confint(object = model4.H0, parm = "age", level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
round(rev(exp(-10 * beta.ci)), 3)
```

```
## 97.5 % 2.5 %
## 1.055 1.218
```

The Wald and LR intervals are similar.

With 95% confidence, the odds of preference for Sanders increases by an amount between 1.05 and 1.22 times for every 10-year decrease in the voter's age, holding other variables constant.

### Odds Ratio for party

#### Odds Ratio for partyI

The estimated odds ratio for partyI is

$$\widehat{OR}_{partyI} = \exp(\hat{\beta}_2)$$

Odds ratio:

```
# Odds ratio for partyI
round(exp(1 * model4.H0$coefficients[3]), 3)
```

```
## partyI
## 2.041
```

95% Wald interval:

```
# Wald interval for the partyI odds ratio
beta.ci <- confint.default(object = model4.H0, parm = "partyI",
  level = 0.95)
round(exp(1 * beta.ci), 3)
```

```
##          2.5 % 97.5 %
## partyI  1.55  2.688
```

95% LR interval:

```
# LR interval for the partyI odds ratio
beta.ci <- confint(object = model4.H0, parm = "partyI", level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
round(exp(1 * beta.ci), 3)
```

```
##  2.5 % 97.5 %
##  1.551  2.690
```

The Wald and LR intervals are similar.

With 95% confidence, the odds of a voter preferring Sanders are between 1.55 to 2.69 times as large when the voter is an Independent than when the voter is a Democrat, holding other variables constant.

### Odds Ratio for partyR

The estimated odds ratio for partyR is

$$\widehat{OR}_{partyR} = \exp(\hat{\beta}_3)$$

Odds ratio:

```
# Odds ratio for partyR
round(exp(1 * model4.H0$coefficients[4]), 3)
```

```
## partyR
##  1.812
```

95% Wald interval:

```
# Wald interval for the partyI odds ratio
beta.ci <- confint.default(object = model4.H0, parm = "partyR",
  level = 0.95)
round(exp(1 * beta.ci), 3)
```

```
##          2.5 % 97.5 %
## partyR  1.316  2.493
```

95% LR interval:

```
# LR interval for the partyI odds ratio
beta.ci <- confint(object = model4.H0, parm = "partyR", level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
round(exp(1 * beta.ci), 3)
```

```
##  2.5 % 97.5 %
##  1.318  2.498
```

The Wald and LR intervals are similar.

With 95% confidence, the odds of a voter preferring Sanders are between 1.32 to 2.49 times as large when the voter is a Republican than when the voter is a Democrat, holding other variables constant.

### Odds Ratio comparing Independent to Republican

The estimated odds ratio comparing partyI to partyR is

$$\widehat{OR}_{partyI vs. partyR} = \exp(\hat{\beta}_2 - \hat{\beta}_3)$$

Odds ratio:

```
beta.hat <- model4.H0$coefficients[-1] # matches up beta indices with [i] to help avoid mistakes
# Odds ratio comparing partyI to partyR
as.numeric(round(exp(1 * (beta.hat[2] - beta.hat[3])), 3))
```

```
## [1] 1.127
```

95% Wald interval:

```
# Wald interval for partyI vs partyR
cov.mat <- vcov(model4.H0)[2:5, 2:5]
var.I.R <- cov.mat[2, 2] + cov.mat[3, 3] - 2 * cov.mat[3, 2]
CI.betas <- beta.hat[2] - beta.hat[3] + qnorm(p = c(0.025, 0.975)) *
  sqrt(var.I.R)
round(exp(CI.betas), 3)
```

```
## [1] 0.819 1.550
```

95% LR interval:

```
# LR interval for partyI vs partyR
library(package = mcprofile)
K <- matrix(data = c(0, 0, 1, -1, 0), nrow = 1, ncol = 5, byrow = TRUE)
linear.combo <- mcprofile(object = model4.H0, CM = K)
ci.log.OR <- confint(object = linear.combo, level = 0.95, adjust = "none")
round(exp(ci.log.OR$confint), 3)
```

```
## lower upper
```

```
## 1 0.818 1.549
```

The Wald and LR intervals are similar.

With 95% confidence, the odds of a voter preferring Sanders are between 0.82 to 1.55 times as large when the voter is an Independent than when the voter is a Republican, holding other variables constant. Because 1 is inside the interval, there is insufficient evidence to indicate the preference for Sanders is different for Independent voters compared to Republican voters.

### Odds Ratio for race\_white

The estimated odds ratio for race\_white is

$$\widehat{OR}_{race\_white} = \exp(\hat{\beta}_4)$$

Odds ratio:

```

# Odds ratio for race_white
round(exp(1 * model4.H0$coefficients[5]), 3)

## race_white1
##      2.394

95% Wald interval:

# Wald interval for the race_white odds ratio
beta.ci <- confint.default(object = model4.H0, parm = "race_white1",
  level = 0.95)
round(exp(1 * beta.ci), 3)

##      2.5 % 97.5 %
## race_white1 1.813  3.161

95% LR interval:

# LR interval for the race_white odds ratio
beta.ci <- confint(object = model4.H0, parm = "race_white1",
  level = 0.95)

## Waiting for profiling to be done...
round(exp(1 * beta.ci), 3)

##  2.5 % 97.5 %
##  1.814  3.165

```

The Wald and LR intervals are similar.

With 95% confidence, the odds of a voter preferring Sanders are between 1.81 to 3.16 times as large when the voter is white than when the voter is non-white, holding other variables constant.

## Interpreting the impact of age (answer 1.f)

As our study above, the estimated odds ratio for age is:

$$\hat{O}R_{age} = \exp(c\hat{\beta}_1)$$

### 95% LR Confidence interval for

$\beta_{age}$

```

# LR interval for the age odds ratio
beta.ci <- confint(object = model4.H0, parm = "age", level = 0.95)

## Waiting for profiling to be done...
round(rev(exp(beta.ci)), 3)

## 97.5 %  2.5 %
##  0.995  0.980

```

Let's look at the impact of targeting voters that are 10 years younger (i.e  $c = -10$ ) With  $c = -10$ , we calculate the odds ratio and 95% intervals.

Odds ratio:

```

# LR interval for the age odds ratio
beta.ci <- confint(object = model4.H0, parm = "age", level = 0.95)

```

```
## Waiting for profiling to be done...
```

```
round(rev(exp(-10 * beta.ci)), 3)
```

```
## 97.5 % 2.5 %
```

```
## 1.055 1.218
```

With 95% confidence, the odds of preference for Sanders is between 1.05 and 1.22 times as large for every 10-year decrease in the voter's age, holding other (race and party) variables constant.

## Probability Plots (Answer 2)

```
#### the base model (for plotting all observations)
```

```
logit.mod.base <- glm(formula = sanders_preference ~ age, family = binomial(link = logit),  
  data = dt)
```

```
#### the full model (for plotting subgroups)
```

```
logit.mod.full <- glm(formula = sanders_preference ~ age + party +  
  race_white, family = binomial(link = logit), data = dt)
```

```
#### function for calculating C.I.
```

```
# Wald confidence interval
```

```
wald.ci.pi <- function(newdata, mod.fit.obj, alpha) {  
  linear.pred <- predict(object = mod.fit.obj, newdata = newdata,  
    type = "link", se = TRUE)  
  CI.lin.pred.lower <- linear.pred$fit - qnorm(p = 1 - alpha/2) *  
    linear.pred$se # Wald interval  
  CI.lin.pred.upper <- linear.pred$fit + qnorm(p = 1 - alpha/2) *  
    linear.pred$se # Wald interval  
  CI.pi.lower <- exp(CI.lin.pred.lower)/(1 + exp(CI.lin.pred.lower))  
  CI.pi.upper <- exp(CI.lin.pred.upper)/(1 + exp(CI.lin.pred.upper))  
  list(lower = CI.pi.lower, upper = CI.pi.upper)  
}
```

## Bubble plot of the base model for all observations

In this case, the data include all observations.

```
### aggregate the data by age for plotting observations  
w <- aggregate(formula = as.numeric(dt$sanders_preference) ~  
  1 ~ dt$age, FUN = sum) # sanders supporters at each age  
n <- aggregate(formula = as.numeric(dt$sanders_preference) ~  
  dt$age, FUN = length) # total voters at each age
```

```
names(w)[1] <- "age"  
names(w)[2] <- "preference"  
names(n)[1] <- "age"  
names(n)[2] <- "preference"
```

```
w.n <- data.frame(age = w$age, n = n$preference, w = w$preference,
  ratio = round(w$preference/n$preference, 4))
# head(w.n) # ratio = (sanders_supporters/total number of
# voters)
```

The estimated logistic model with the 95% Wald confidence intervals for the base model

$$\text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age}$$

the bubble plot also shows the observed ratio of voters who prefer Sanders at each age, with the plotting size being proportional to the number of observations at that age.

```
# Plot data points
# #####
symbols(x = w$age, y = w$preference/n$preference, circles = sqrt(n$preference),
  inches = 0.12, xlab = "age", ylab = "Estimated probability",
  xlim = c(15, 100), panel.first = grid(col = "gray", lty = "dotted"))

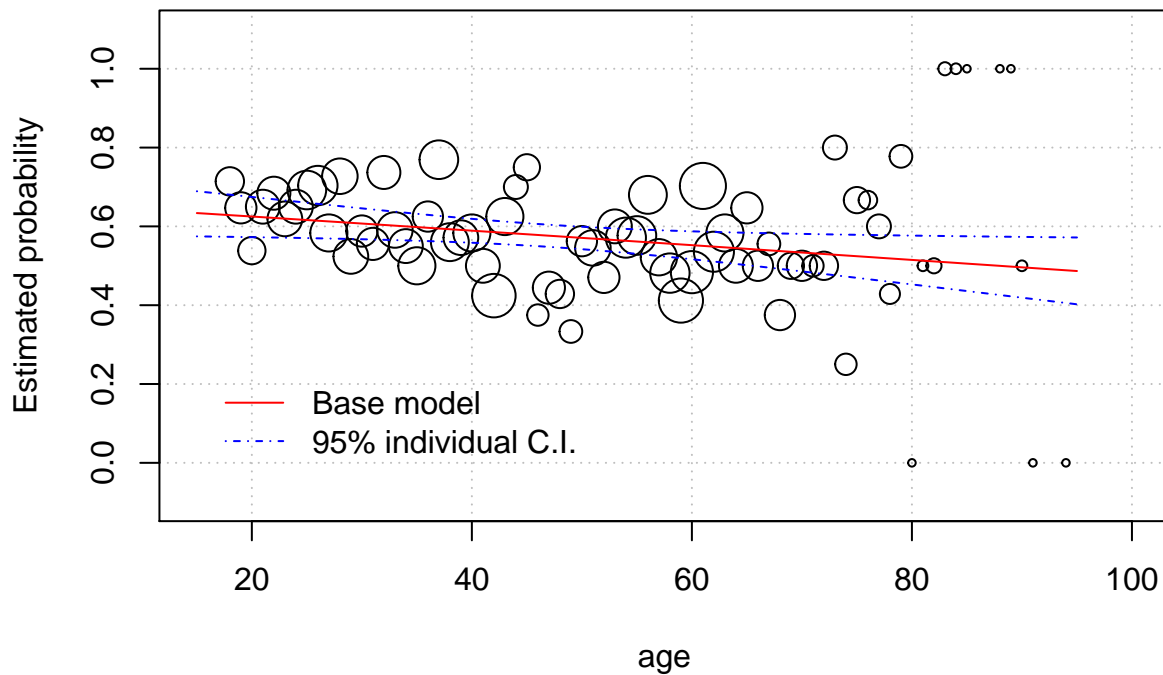
# Plot model fit
curve(expr = predict(object = logit.mod.base, newdata = data.frame(age = x),
  type = "response"), col = "red", add = TRUE, xlim = c(15,
  95))

# Plot C.I. bands
curve(expr = wald.ci.pi(newdata = data.frame(age = x), mod.fit.obj = logit.mod.base,
  alpha = 0.05)$lower, col = "blue", lty = "dotdash", add = TRUE,
  xlim = c(15, 95))

curve(expr = wald.ci.pi(newdata = data.frame(age = x), mod.fit.obj = logit.mod.base,
  alpha = 0.05)$upper, col = "blue", lty = "dotdash", add = TRUE,
  xlim = c(15, 95))

# Legend
legend(x = 15, y = 0.25, legend = c("Base model", "95% individual C.I."),
  lty = c("solid", "dotdash"), col = c("red", "blue"), bty = "n")
```





The base model shows the preference for Sanders declines as the voter age increases.

## Comparing different voter groups using the full model

For comparing different groups of voters using the full model, we will plot the observed data points for the subgroup of voters each model represents with the estimated probability and 95% confidence intervals.

Since our model is

$$\text{logit}(\pi_{\text{preference\_sanders}}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{partyI} + \beta_3 \text{partyR} + \beta_4 \text{race\_white}$$

the *gender* factor has been excluded in our model, we will only include *party* and *race\_white* in the subgroup plotting function.

```
#### Define function plotsubgroup.pi.ci() for subgroup
#### probability plot
plotsubgroup.pi.ci <- function(newdata, mod.fit.obj, Party, Race) {

  party.G = Party
  race_white.G = Race

  # Strings for plot title
  if (Race == "1") {
    raceStr = "white"
  } else if (Race == "0") {
    raceStr = "non-white"
  } else {
```

```

    raceStr = " "
  }

  if (Party == "D") {
    partyStr = "Democratic"
  } else if (Party == "I") {
    partyStr = "Independent"
  } else if (Party == "R") {
    partyStr = "Republican"
  } else {
    partyStr = " "
  }

  # Creating subgroup data frame for the specified party and
  # race
  dt.G <- newdata[newdata$race_white == race_white.G & newdata$party ==
    party.G, ]

  # Creating aggregated data for plotting observations
  # #####
  w.G <- aggregate(formula = as.numeric(dt.G$sanders_preference) ~
    1 ~ dt.G$age, FUN = sum) # sanders supporters for each age
  n.G <- aggregate(formula = as.numeric(dt.G$sanders_preference) ~
    dt.G$age, FUN = length) # total voters for each age

  # change to easy to understand column names
  names(w.G)[1] <- "age"
  names(w.G)[2] <- "preference"
  names(n.G)[1] <- "age"
  names(n.G)[2] <- "preference"

  w.n.G <- data.frame(age = w.G$age, n = n.G$preference, w = w.G$preference,
    ratio = round(w.G$preference/n.G$preference, 4))

  # PLOT
  # #####

  # Plot data points
  symbols(x = w.G$age, y = w.G$preference/n.G$preference, circles = sqrt(n.G$preference),
    inches = 0.05, xlab = "age", ylab = "Estimated probability",
    xlim = c(15, 100), panel.first = grid(col = "gray", lty = "dotted"),
    main = paste(partyStr, ":", raceStr))

  # Plot model fit
  curve(expr = predict(object = logit.mod.full, newdata = data.frame(age = x,
    party = party.G, race_white = race_white.G), type = "response"),
    col = "red", add = TRUE, xlim = c(15, 100), ylim = c(0,
    1), xlab = "age", ylab = expression(hat(pi)))

  # Plot C.I. bands
  curve(expr = wald.ci.pi(newdata = data.frame(age = x, party = party.G,
    race_white = race_white.G), mod.fit.obj = logit.mod.full,

```

```

    alpha = 0.05)$lower, col = "blue", lty = "dotdash", add = TRUE,
    xlim = c(15, 100))

curve(expr = wald.ci.pi(newdata = data.frame(age = x, party = party.G,
  race_white = race_white.G), mod.fit.obj = logit.mod.full,
  alpha = 0.05)$upper, col = "blue", lty = "dotdash", add = TRUE,
  xlim = c(15, 100))

# legend(x=55, y=1.0, legend = c('model', '95% C.I.'), lty =
# c('solid', 'dotdash'), col = c('red', 'blue'), bty = 'n')
}

```

Estimated probability of preferring sanders vs observations for different voter groups

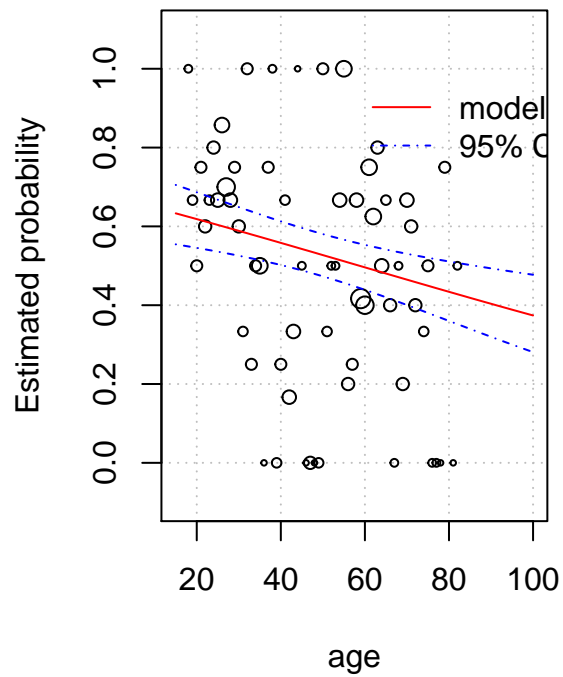
```

# layout(matrix(c(1,4,2,5,3,6), 3, 2, byrow = TRUE))
# layout(matrix(c(1,2,3,4,5,6), 2, 3, byrow = TRUE))

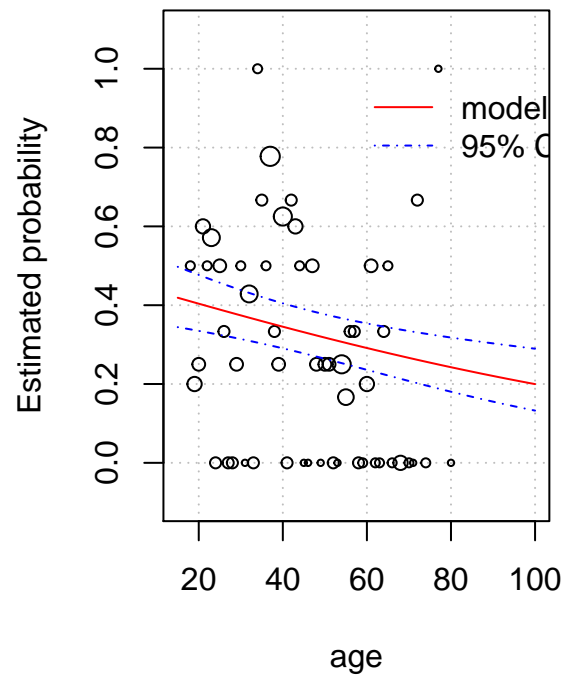
par(mfrow = c(1, 2))
# White Democratic
plotsubgroup.pi.ci(newdata = dt, mod.fit.obj = logit.mod.full,
  Party = "D", Race = "1")
# Non-white Democratic
plotsubgroup.pi.ci(newdata = dt, mod.fit.obj = logit.mod.full,
  Party = "D", Race = "0")

```

**Democratic : white**

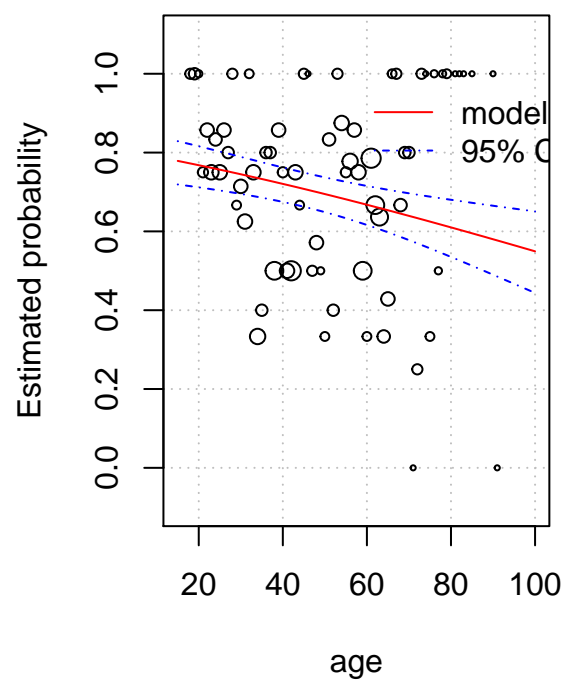


**Democratic : non-white**

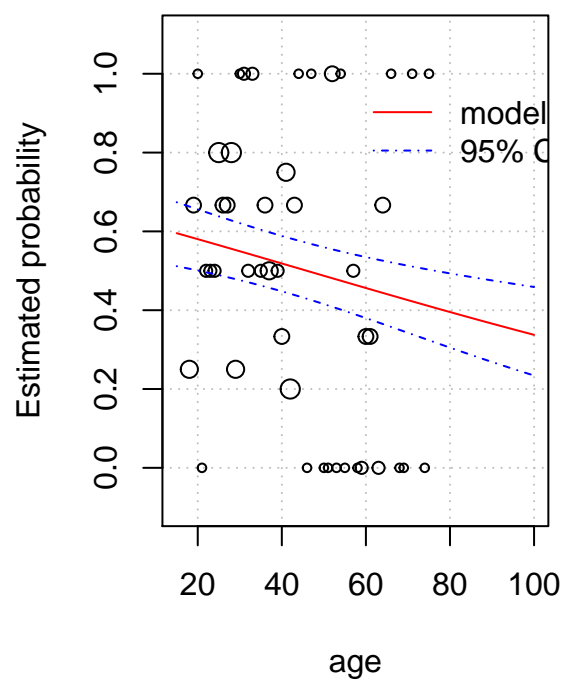


```
par(mfrow = c(1, 2))
# White Independent
plotsubgroup.pi.ci(newdata = dt, mod.fit.obj = logit.mod.full,
  Party = "I", Race = "1")
# Non-white Independent
plotsubgroup.pi.ci(newdata = dt, mod.fit.obj = logit.mod.full,
  Party = "I", Race = "0")
```

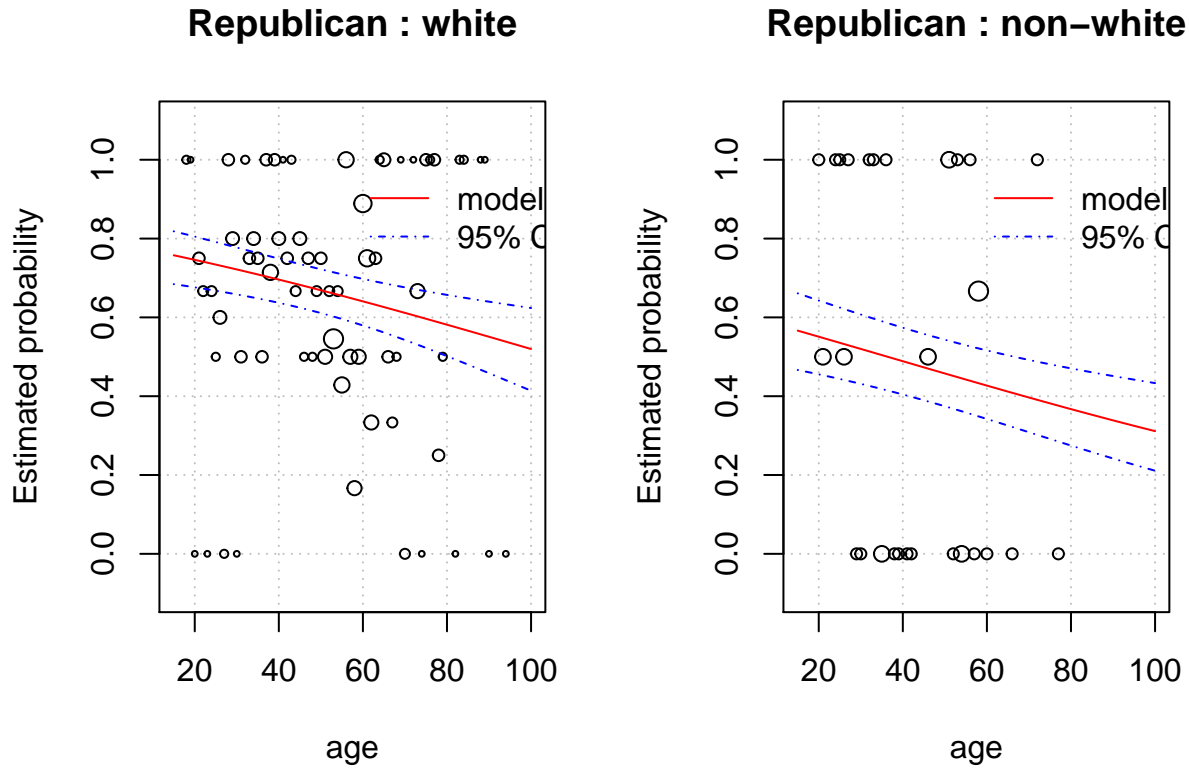
**Independent : white**



**Independent : non-white**



```
par(mfrow = c(1, 2))
# White Republican
plotsubgroup.pi.ci(newdata = dt, mod.fit.obj = logit.mod.full,
  Party = "R", Race = "1")
# Non-white Republican
plotsubgroup.pi.ci(newdata = dt, mod.fit.obj = logit.mod.full,
  Party = "R", Race = "0")
```



The estimated probability plots (the model prediction is shown in red, the 95% confidence intervals are shown by the dashed blue lines) based on our model show that the preference for Bernie Sanders was higher among the white voters than the non-white voters within each political party. In addition, Sanders enjoyed a similar level of support among the Independent and Republican voters, however, the preference for Sanders was lower among the Democratic voters compared to the Independent and Republican voters.

### Conclusion (answer 3)

Through our analysis, we have found that age is a statistically significant variable in predicting a preference for Sanders. Based on the sample, we find that 63% of voters under the age of 40 prefer Sanders, compared to 55% of voters 40 or older. In the broadest approach, this segmentation can be used for marketing purposes if the goal is to target the largest number of people.

Given we are looking to market to a specific demographic, however, it would make sense to narrow the segment of voters in order to get the most return for our investment. In addition to age, we have found that race, and party preference are also statistically significant. The odds of a voter preferring Sanders are between 1.81 and 3.16 times (95% confidence) as large when the voter is white, 1.55 to 2.69 times as large when the voter identifies as an Independent, and 1.32 to 2.49 times as large when the voter identifies as a Republican. We note that it is unlikely that a Republican voter would purchase a Democratic candidate's T-shirt, and that the preference in the data is between two Democratic candidates, thus not validating the preference for Sanders over a Republican candidate.

With this additional information, we can refine our strategy to focus on younger (<40) white voters. Since we are trying to sell a product, we would recommend still targeting Democratic voters since they will likely have the highest chance of purchasing Democratic merchandise. We would also include Independents as they

are between parties. In the same vein, we would recommend being cautious in targeting Republicans as they will be less likely to purchase Democratic merchandise.