Bias in Healthcare

Setup

Load packages

```
library(tidyverse)

## -- Attaching packages ------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ---------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2)
library(ggthemes)
library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a telephone survey that collects data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. It collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

Population: Health characteristics estimated from the BRFSS pertain to the non-institutionalized adult population, aged 18 years or older, who reside in the US.

Respondent data are forwarded to CDC to be aggregated for each state, returned with standard tabulations, and published at year's end by each state. Source: https://www.cdc.gov/brfss/

In this project, I chose only the variables concerning race and its effect on the experience of seeking healthcare.The purpose of this analysis is to detect bias in healthcare delivery in the USA.

Load data

```
race_data1 <- read_csv("Data/race_data1.csv")

## Parsed with column specification:
## cols(
##   rrclass2 = col_character(),
##   rrhcare3 = col_character()
## )
```

Viewing the structure of the data

```
str(race_data1)

## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 2658 obs. of  2 v
ariables:
##  $ rrclass2: chr  "Black or African American" "White" "White" "White" ...
##  $ rrhcare3: chr  "The same as other races" "The same as other races" "The
same as other races" "The same as other races" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rrclass2 = col_character(),
##   ..   rrhcare3 = col_character()
##   .. )
```

rrclass2: How Do Other People Usually Classify You In This Country?

```
class(race_data1$rrclass2)

## [1] "character"

unique(race_data1$rrclass2)

## [1] "Black or African American" "White"
```

rrhcare3: When Seeking Health Care Past 12 Months, Was Experience Worse, Same, Better Than other races

```
class(race_data1$rrhcare3)

## [1] "character"

unique(race_data1$rrhcare3)

## [1] "The same as other races" "Better than other races"
```

Data Summarization

```
race_data1 %>%
  group_by(rrclass2) %>%
  summarize(count = n()) -> count_tab
pct <- list(c(paste0(round((168/length(race_data1$rrclass2))*100), "%"),
paste0(round((2490/length(race_data1$rrclass2))*100),"%")))
freq_tab <- cbind(count_tab, pct)
colnames(freq_tab) <- c("Race", "Frequency", " Percentage")
freq_tab

##                          Race Frequency  Percentage
## 1 Black or African American       168          6%
## 2                       White      2490         94%
```

As seen in the above frequency table, the majority of our sample is White **(94%)**.

Cross-tabulating our data

```
tab1 <- xtabs(~ rrclass2 + rrhcare3, data = race_data1)
tab1

##                            rrhcare3
## rrclass2                     Better than other races The same as other race
s
```

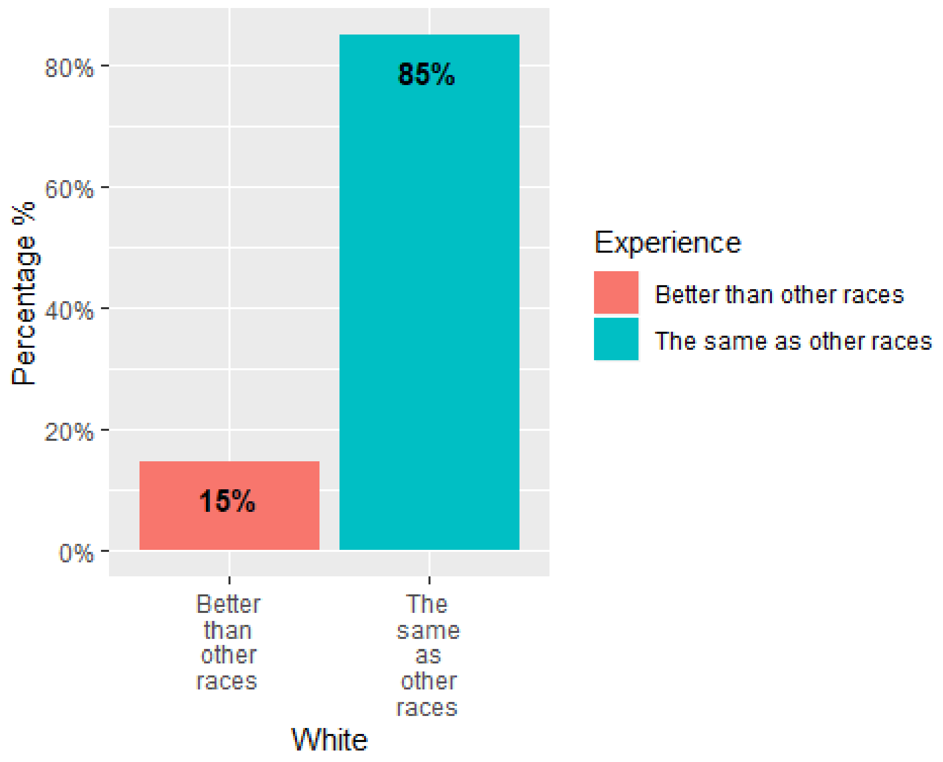
```
##    Black or African American                       17                       15
1
##    White                                          370                      212
0
```

Using visualization to find:

Which race has a higher percentage of people having a better healthcare experience than other races compared with those having the same experience as other races.

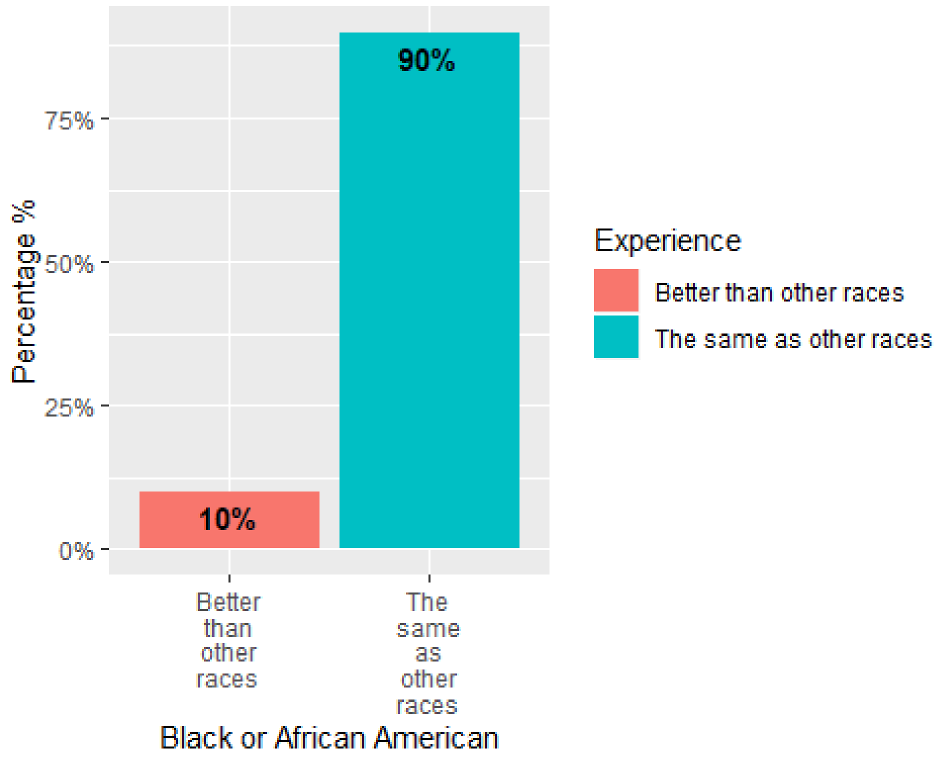
For the white race

```
race_white <- race_data1 %>%
  filter(rrclass2=="White")%>%
  group_by(rrhcare3) %>%
  summarise(counts = n())

race_white %>%
  ggplot(aes(x= rrhcare3,
             y= counts/sum(counts),
             fill= rrhcare3)) +
  geom_bar(stat = "identity") +
  labs(x= "White",
       y= "Percentage %",
       fill = "Experience") +
  scale_x_discrete(label = function(x)str_wrap(x, width = 5))+
  scale_y_continuous(labels = percent_format(accuracy = 1))+
  geom_text(
    aes(
      label= paste0(round((counts/ sum(counts))*100),"%"),
      fontface= "bold"
      ),
    vjust= 2
    ) +
  theme(axis.text.x = element_text(hjust = 0.5))
```



For the black or african american race

```
race_black <- race_data1 %>%
  filter(rrclass2=="Black or African American")%>%
  group_by(rrhcare3) %>%
  summarise(counts = n())

race_black %>%
  ggplot(aes(x= rrhcare3,
             y= counts/sum(counts),
             fill= rrhcare3)) +
  geom_bar(stat = "identity") +
  labs(x= "Black or African American",
       y= "Percentage %",
       fill = "Experience") +
  scale_x_discrete(label = function(x)str_wrap(x, width = 5))+
  scale_y_continuous(labels = percent_format(accuracy = 1))+
  geom_text(
    aes(
      label= paste0(round((counts/ sum(counts))*100),"%"),
      fontface = "bold"
      ),
    vjust = 1.5
    ) +
  theme(axis.text.x = element_text(hjust = 0.5))
```



As seen from the two graphs, the percentage of those having a better healthcare experience than other races is higher in people of **white** race**(15%)** compared to those of **black or african american** race**(10%)**, but is this difference statistically significant?

Data Analysis

I researched online on the question of using Chi-square vs logistic regression in case of 2 categorical variables with 2 levels each. There doesn’t seem to be a straight answer, but I arrived at the conclusion that both can be applied based on the question: Chi-square for describing the strength of an association, and logistic regression for modeling determinants and predicting the likelihood of an outcome.

I tested both methods here.

Performing a Chi-square test

Using Chi-square test to test the strength of association between race and healthcare experience

```
chisq.test(tab1)

## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  tab1
## X-squared = 2.4746, df = 1, p-value = 0.1157

chisq.test(tab1, correct = F)

## 
##  Pearson's Chi-squared test
## 
## data:  tab1
## X-squared = 2.8429, df = 1, p-value = 0.09178
```

According to its results, the difference is **not statistically significant**, so we fail to reject the null hypothesis of healthcare experience and race being independent from each other.

Correlation & Logistic Regression:

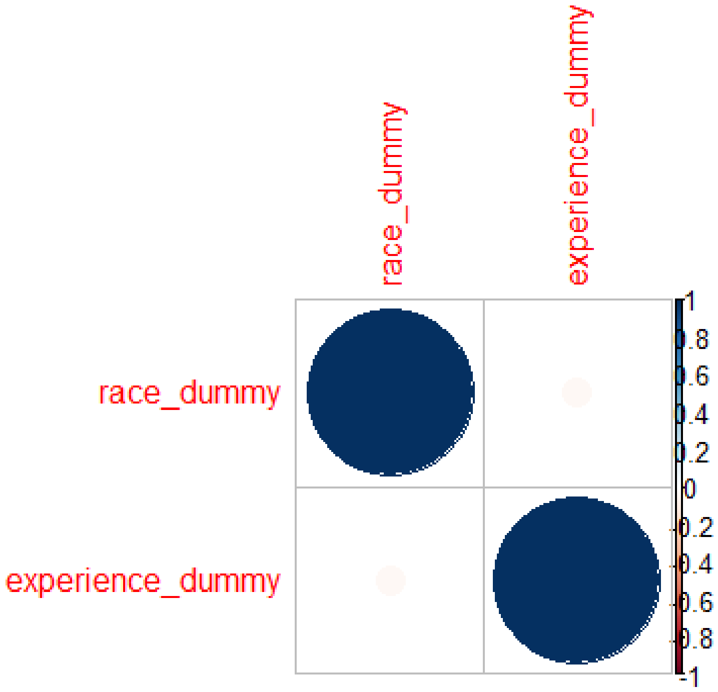
Using logistic regression to predict the likelihood of better or same healthcare experience based on race

Preparing the data

```
logis_data <- race_data1 %>%
  mutate(race_dummy= (ifelse(rrclass2=="White", 0, 1)),
         experience_dummy= (ifelse(rrhcare3=="The same as other races", 0, 1)
),
         rrclass2 = as.factor(rrclass2),
         rrhcare3 =  as.factor(rrhcare3))

logis_data <- within(logis_data, rrclass2 <- relevel(rrclass2, ref = "White")
)
```

Assessing correlation between race and the experience of seeking healthcare

```
cor(logis_data$experience_dummy, logis_data$race_dummy, method = "spearman")

## [1] -0.03270426

library(corrplot)

## corrplot 0.84 loaded

correlation <- cor(logis_data[,c(3,4)])
corrplot(correlation, method = "circle")
```



There is a **weak negative correlation** between race and the experience of seeking healthcare.

Fitting the data into a logistic regression model

```
model1 <- glm(formula = rrhcare3~ rrclass2, family = binomial, data = logis_d
ata)
summary(model1)
```

```
##
## Call:
## glm(formula = rrhcare3 ~ rrclass2, family = binomial, data = logis_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1404   0.5672   0.5672   0.5672   0.5672
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        1.74567    0.05634  30.984   <2e-16 ***
## rrclass2Black or African American  0.43840    0.26194   1.674   0.0942 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2206.1  on 2657  degrees of freedom
## Residual deviance: 2203.0  on 2656  degrees of freedom
## AIC: 2207
##
## Number of Fisher Scoring iterations: 4
```

Notice both Chi-square test and logistic regression model are reporting very similar p-values, but only when continuity correction is disabled in Chi-square.

Interpreting the model

```
exp(model1$coefficients)
```

```
##                       (Intercept) rrclass2Black or African American
##                          5.729730                          1.550222
```

The **odds** of having a better healthcare experience than other races if a person is **white** is **5.7**

The **odds** of having a better healthcare experience than other races if a person is **black or african american** is **1.6**

```r
p1 <- (exp(1.74567)/(1+exp(1.74567)))
p1

## [1] 0.8514058

p2 <- (exp(0.43840)/(1+exp(0.43840)))
p2

## [1] 0.6078777
```

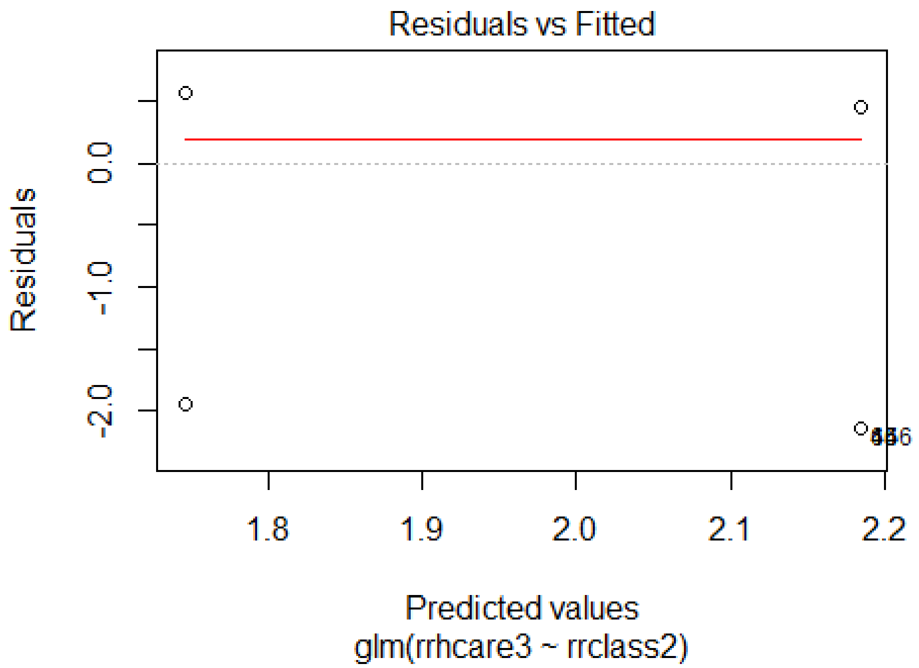
The **probability** of having a better healthcare experience than other races if a person is **white** is **85%**

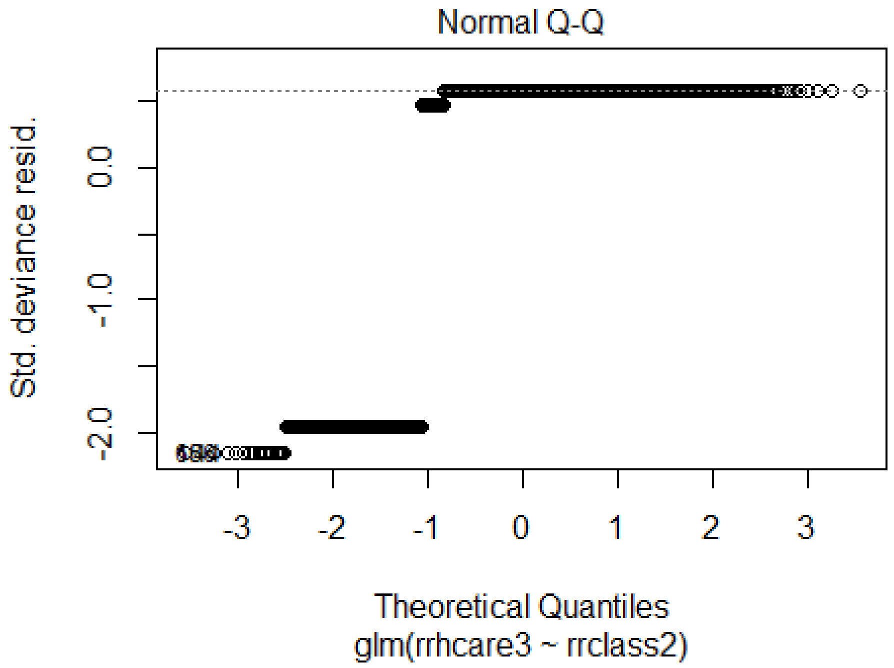
The **propability** of having a better healthcare experience than other races if a person is **black or african american** is **60%**

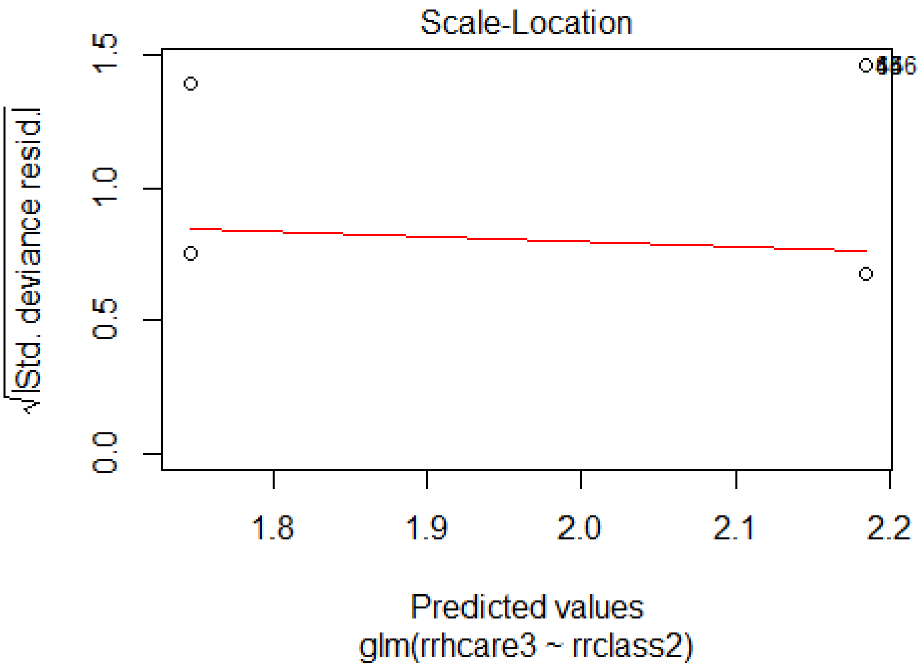
Ploting the model

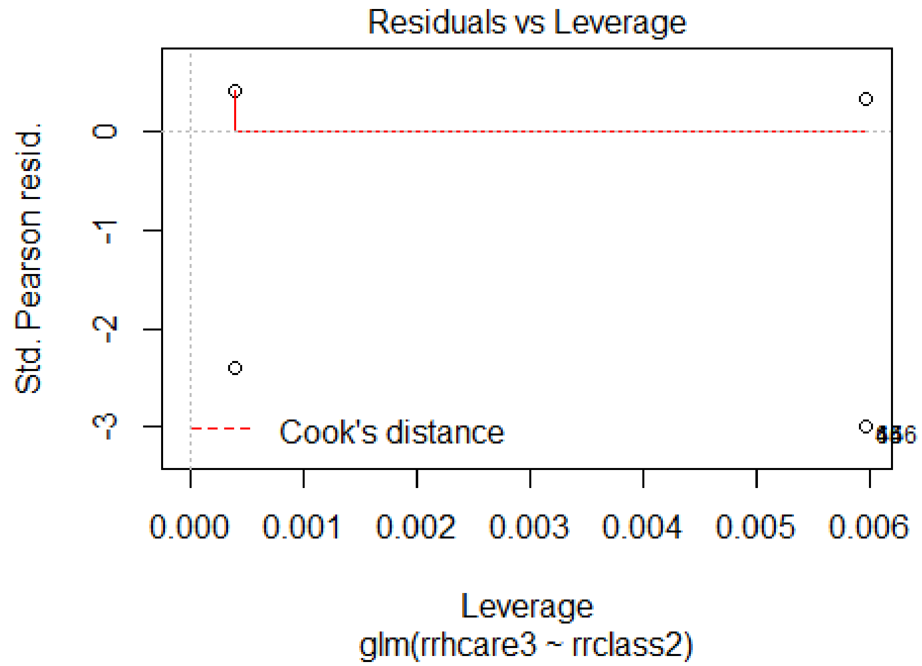
I am not sure how to interpret the plots in this type of logistic regression. I would appreciate any help.

```r
plot(model1)
```


Residuals vs Fitted
Residuals
0.0
-1.0
-2.0
1.8
1.9
2.0
2.1
2.2
Predicted values
glm(rrhcare3 ~ rrclass2)


Normal Q-Q
Std. deviance resid.
0.0
-1.0
-2.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles
glm(rrhcare3 ~ rrclass2)


Scale-Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
1.8
1.9
2.0
2.1
2.2
656
Predicted values
glm(rrhcare3 ~ rrclass2)


Residuals vs Leverage
Std. Pearson resid.
-3
-2
-1
0
Cook's distance
656
0.000
0.001
0.002
0.003
0.004
0.005
0.006
Leverage
glm(rrhcare3 ~ rrclass2)