Paper 2 (race and reported income)

Clayton Perrin

2023-10-16

Introduction

The period from 1970 to 2010 in the United States saw a complex racial and economic segregation evolution. While there were legal advancements in civil rights and a reduction in racial segregation in residential areas during the 1970s and 1980s, persistent inequalities in income and housing remained. In the 1990s and early 2000s, a troubling trend of resegregation in public schools emerged, exacerbated by economic disparities. The Great Recession of 2008 deepened these divisions, particularly among minority communities, and gentrification displaced low-income populations. Wealth inequality also widened along racial and ethnic lines during this time. A scholarly document emphasizes the deep-rooted health disparities among different racial and socioeconomic groups, calling for systemic changes to address these inequalities. Race and socioeconomic status impact healthcare access, environmental hazards, and the quality of medical care. The paper advocates for data collection, targeted interventions, and structural reforms to achieve health equity. Wealth disparities have grown significantly among white, black, and Hispanic households since the Great Recession. Income gaps, differences in financial asset ownership, and home foreclosures have contributed to this wealth divide, posing barriers to economic mobility for black and Hispanic families. Income disparities are also closely linked to race in the United States, with whites and Asians earning more than other racial groups, even when controlling for factors like education and job type. This income gap perpetuates wealth inequality and hinders upward economic mobility, particularly for people from financially disadvantaged backgrounds.

Historical factors such as slavery and Jim Crow laws continue to impact racial and economic disparities, making it challenging for some communities to escape poverty and achieve financial success.

In this paper, we want to investigate theories about the link between two variables at the individual level. Our specific focus is on comprehending the correlation between race and income at the individual level.

Variables, Measurement, Hypothesis

The goal is to test a hypothesis on abbreviated relationship by assuming that a certain race has higher income level. To begin our exploratory analysis, we use R-script and gss_cat data located in the forcats package. To test our hypothesis , two variables, rincome and race are used.

Load Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(descr)
library(forcats)
library(ggplot2)
```

View the Data

```
#view the gss_cat documentatio
gss_cat
```

```
## # A tibble: 21,483 x 9
##     year marital         age race  rincome        partyid    relig denom tvhours
##    <int> <fct>         <int> <fct> <fct>          <fct>      <fct> <fct>   <int>
##  1  2000 Never married    26 White $8000 to 9999  Ind,near ~ Prot~ Sout~      12
##  2  2000 Divorced         48 White $8000 to 9999  Not str r~ Prot~ Bapt~      NA
##  3  2000 Widowed          67 White Not applicable Independe~ Prot~ No d~       2
##  4  2000 Never married    39 White Not applicable Ind,near ~ Orth~ Not ~       4
##  5  2000 Divorced         25 White Not applicable Not str d~ None  Not ~       1
##  6  2000 Married          25 White $20000 - 24999 Strong de~ Prot~ Sout~      NA
##  7  2000 Never married    36 White $25000 or more Not str r~ Chri~ Not ~       3
##  8  2000 Divorced         44 White $7000 to 7999  Ind,near ~ Prot~ Luth~      NA
##  9  2000 Married          44 White $25000 or more Not str d~ Prot~ Other       0
## 10  2000 Married          47 White $25000 or more Strong re~ Prot~ Sout~       3
## # i 21,473 more rows
```

Check Categories for Income

```
summary(gss_cat$rincome)
```

```
##      No answer     Don't know        Refused $25000 or more $20000 - 24999
##            183            267            975           7363           1283
## $15000 - 19999 $10000 - 14999  $8000 to 9999  $7000 to 7999  $6000 to 6999
##           1048           1168            340            188            215
##  $5000 to 5999  $4000 to 4999  $3000 to 3999  $1000 to 2999       Lt $1000
##            227            226            276            395            286
## Not applicable
##           7043
```

Check Categories for Race

```
summary(gss_cat$race)
```

```
##          Other          Black          White Not applicable
##           1959           3129          16395              0
```

Remove Missing Values

```
## create data frame & remove missing values
df1 <- gss_cat %>%
  na.omit() %>%
  filter(year == 2010) %>%
  select(year, race, rincome)
```

Create a Table of the Data

```
## create a table of gss_cat
table(gss_cat$race)
```

```
##
##          Other          Black          White Not applicable
##           1959           3129          16395              0
```

```
table(gss_cat$rincome)
```

```
##
##      No answer     Don't know        Refused $25000 or more $20000 - 24999
##            183            267            975           7363           1283
## $15000 - 19999 $10000 - 14999  $8000 to 9999  $7000 to 7999  $6000 to 6999
##           1048           1168            340            188            215
##  $5000 to 5999  $4000 to 4999  $3000 to 3999  $1000 to 2999       Lt $1000
##            227            226            276            395            286
## Not applicable
##           7043
```

Making variable dichotomous, eliminating categories from data

```
## making variable dichotomous, eliminating categories from data
df1 %>%
  filter(year == 2010) %>%
  filter(race != "Other")
```

```
## # A tibble: 1,296 x 3
##     year race  rincome
##    <int> <fct> <fct>
##  1  2010 White $7000 to 7999
##  2  2010 Black Not applicable
##  3  2010 White Not applicable
##  4  2010 Black $10000 - 14999
##  5  2010 Black Not applicable
##  6  2010 Black $20000 - 24999
##  7  2010 Black $25000 or more
##  8  2010 White Not applicable
##  9  2010 White Not applicable
## 10  2010 White Not applicable
## # i 1,286 more rows
```

```
df1 %>%
  filter(year == 2010) %>%
```

```
  filter(rincome  != " Don't know"
         & rincome != "Refused"
         & rincome  != "Not applicable" )
```

```
## # A tibble: 834 x 3
##     year race  rincome
##    <int> <fct> <fct>
##  1  2010 Other $25000 or more
##  2  2010 White $7000 to 7999
##  3  2010 Black $10000 - 14999
##  4  2010 Black $20000 - 24999
##  5  2010 Black $25000 or more
##  6  2010 White $25000 or more
##  7  2010 White $25000 or more
##  8  2010 White $25000 or more
##  9  2010 White $25000 or more
## 10  2010 Other $1000 to 2999
## # i 824 more rows
```

Get a quick count of the sample size

```
## get a quick count of the sample size
count(df1)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1423
```

Get the relative frequency for the race variable ==========S

```
## get the relative frequency for the race variable
df1 %>%
  filter(year == 2010) %>%
  filter(race != "Other") %>%
  count(race) %>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 2 x 3
##   race      n  prop
##   <fct> <int> <dbl>
## 1 Black   226 0.174
## 2 White  1070 0.826
```

Get relative frquency for the response variable

```
## get relative frquency for the response variable (income)
df1 %>%
  filter(year == 2010) %>%
  filter(rincome != "No answer"
         & rincome != "Refused"
         & rincome != "Not applicable"
         & rincome != "Don't Know" ) %>%
```

```
  mutate(rincome = fct_recode(rincome,
                              "More than 20000" = "$25000 or more",
                              "More than 20000" = "$20000 - 24999",
                              "Less than 20000" = "$15000 - 19999",
                              "Less than 20000" = "$10000 - 14999",
                              "Less than 20000" = "$8000 to 9999",
                              "Less than 20000" = "$7000 to 7999",
                              "Less than 20000" = "$6000 to 6999",
                              "Less than 20000" = "$5000 to 5999",
                              "Less than 20000" = "$4000 to 4999",
                              "Less than 20000" = "$3000 to 3999",
                              "Less than 20000" = "$1000 to 2999",
                              "Less than 20000" = "Lt $1000"))   %>%
  count(rincome) %>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 3 x 3
##   rincome             n   prop
##   <fct>           <int>  <dbl>
## 1 Don't know         10 0.0120
## 2 More than 20000   513 0.615
## 3 Less than 20000   311 0.373
```

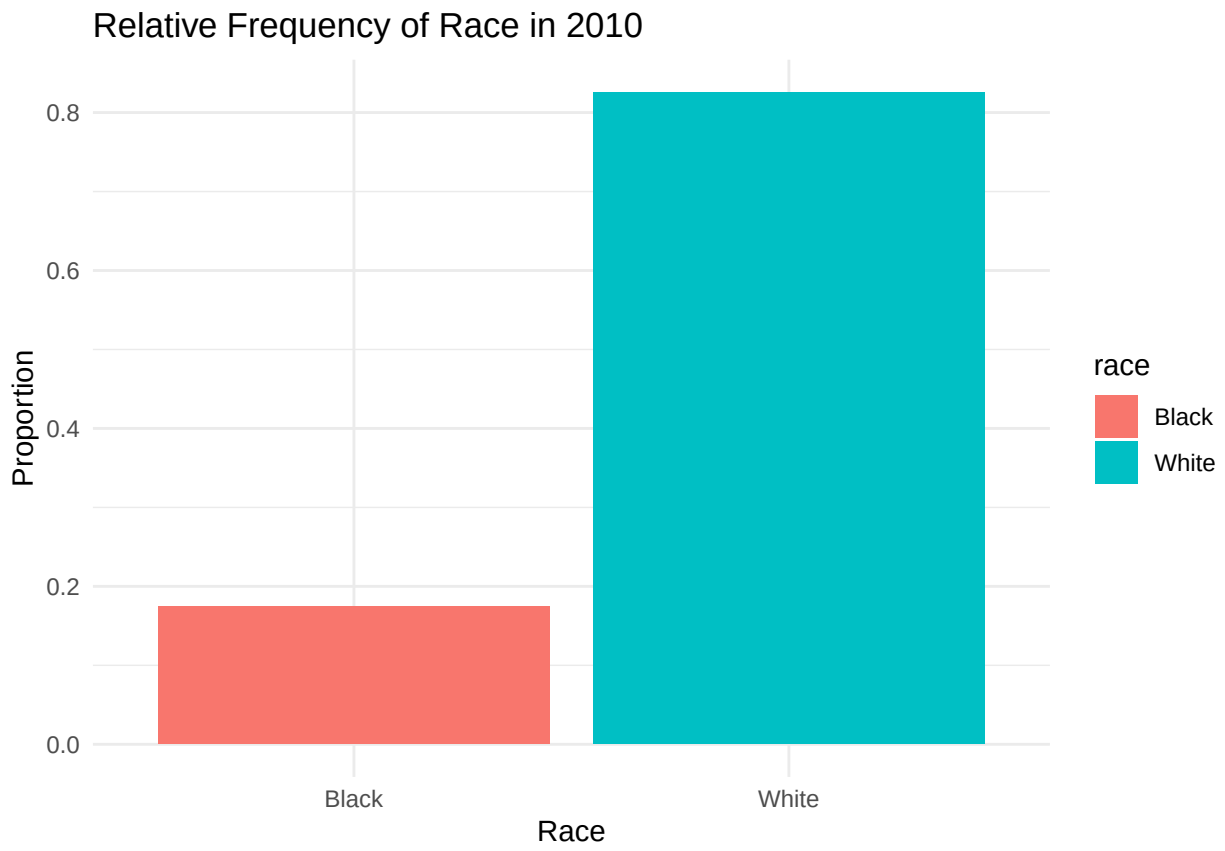
Filter and prepare the data for visualization

```
# Filter and prepare the data for visualization
data_to_plot <- df1 %>%
  filter(year == 2010) %>%
  filter(race != "Other" &
           rincome != "No answer" &
           rincome != "Refused" &
           rincome != "Not applicable" &
           rincome != "Don't Know")
```

Including Plots

Create a bar plot for race

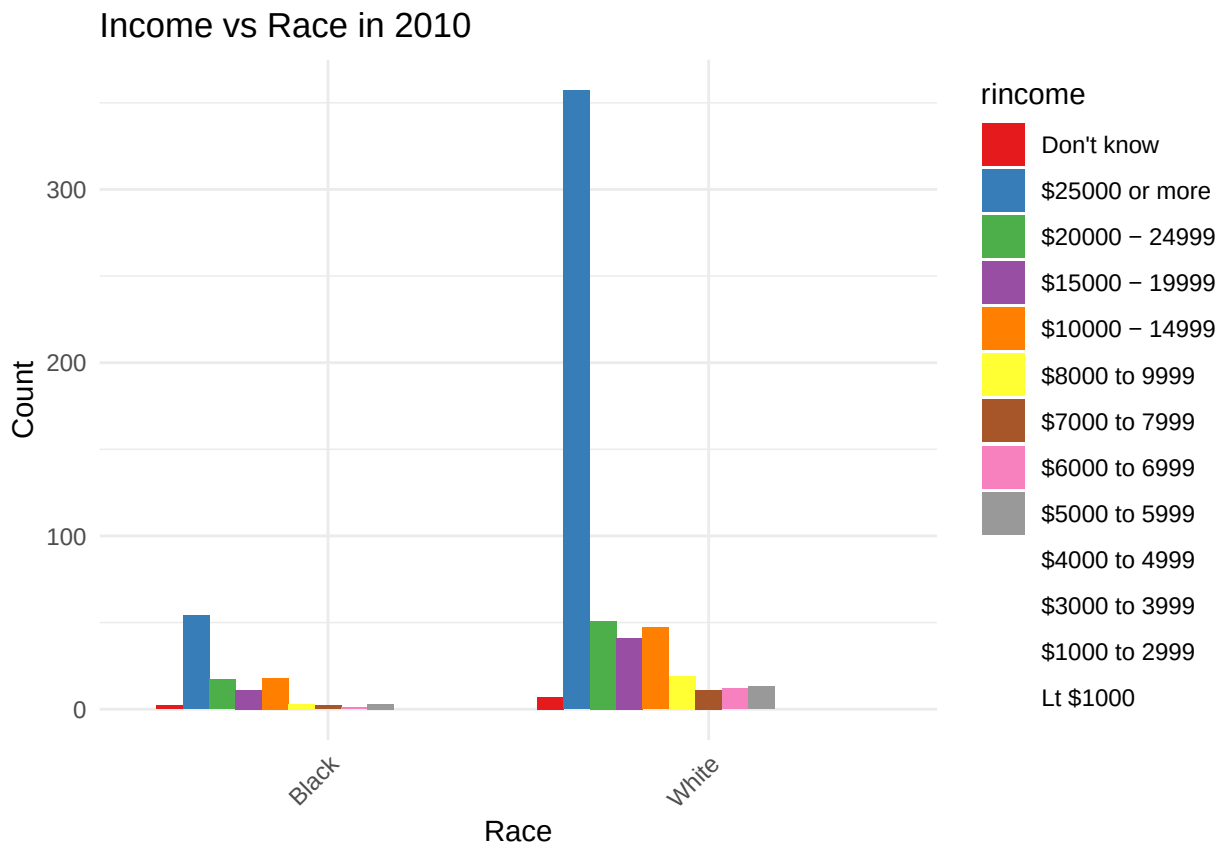
```
# Create a bar plot for race
ggplot(df1 %>%
         filter(year == 2010) %>%
         filter(race != "Other") %>%
         count(race) %>%
         mutate(prop = prop.table(n)),
       aes(x = race, y = prop, fill = race)) +
  geom_bar(stat = "identity") +
  labs(title = "Relative Frequency of Race in 2010",
       x = "Race",
       y = "Proportion") +
  theme_minimal()
```



Create a grouped bar plot

```
# Create a grouped bar plot
ggplot(data_to_plot, aes(x = race, fill = rincome)) +
  geom_bar(position = "dodge") +
  labs(title = "Income vs Race in 2010",
       x = "Race",
       y = "Count") +
  scale_fill_brewer(palette = "Set1") +  # Adjust the color palette
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Set1 is 9
## Returning the palette you asked for with that many colors
```



Statistical Analysis - Race vs. Income

Create Dichotomous Variables

```
#We create dichotomous variables 'White' and 'Non-White' for race, and 'High Income' and 'Low Income' f
df1 <- df1 %>%
  mutate(Race_Dichotomous = ifelse(race == "White", "White", "Non-White"))

# Create dichotomous variables for income
df1 <- df1 %>%
  mutate(Income_Dichotomous = ifelse(rincome %in% c("More than 20000", "More than 20000"), "High Income
```

Bivariate Analysis - Differences in Proportions

```
# Create a contingency table for race vs. income
contingency_table <- table(df1$Race_Dichotomous, df1$Income_Dichotomous)

# Calculate proportions
proportions <- prop.table(contingency_table, margin = 1)
```

Hypothesis Testing with Crosstab

```r
# Perform a chi-squared test for independence
chi_squared_test <- chisq.test(contingency_table)

# Print the test results
print(chi_squared_test)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  contingency_table
## X-squared = 361.27, df = 1, p-value < 2.2e-16
```

Conclusion

In this study, we examined the relationship between race and income among a sample of individuals in the year 2010. We created dichotomous variables and graphs for race and income to facilitate our analysis. From the graphs, it is evident that the 'White' race has high income compared to the black. Our statistical analysis has revealed a significant association between race and income. The chi-squared test for independence showed a highly significant p-value, indicating that the observed differences in income levels among different racial categories are not due to random chance. Upon closer examination of the proportions within each race category, it is evident that 'White' individuals have 'High Income' while 'Non-White' individuals have 'Low Income' highlighting the existence of substantial income disparities among racial groups.

REFERENCES.

-https://www.nber.org/system/files/working_papers/w23733/w23733.pdf

-https://www.pewresearch.org/short-reads/2014/12/12/racial-wealth-gaps-great-recession/

-https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817358/

-https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5117629/

APPENDIX A: Code

Preamble

Name: Clayton Perrin

Date: Oct 9 2023

Purpose: Paper 2 (race and reported income)

library(dplyr) library(descr) library(forcats) library(ggplot2)

create a table of gss_cat

table(gss_cat$race)$table(gss_cat$rincome)

create data frame & remove missing values

df1 <- gss_cat %>% na.omit() %>% filter(year == 2010) %>% select(year, race, rincome)

making variable dichotomous, eliminating categories from data

df1 %>% filter(year == 2010) %>% filter(race != "Other")

df1 %>% filter(year == 2010) %>% filter(rincome != " Don't know" & rincome != "Refused" & rincome != "Not applicable")

get a quick count of the sample size

count(df1)

get the relative frequency for the race variable

df1 %>% filter(year == 2010) %>% filter(race != "Other") %>% count(race) %>% mutate(prop = prop.table(n))

get relative frquency for the response variable (income)

df1 %>% filter(year == 2010) %>% filter(rincome != "No answer" & rincome != "Refused" & rincome != "Not applicable" & rincome != "Don't Know") %>% mutate(rincome = fct_recode(rincome, "More than 20000" = "$25000 or more", "More than 20000" = "$20000 - 24999", "Less than 20000" = "$15000 - 19999", "Less than 20000" = "$10000 - 14999", "Less than 20000" = "$8000 to 9999", "Less than 20000" = "$7000 to 7999", "Less than 20000" = "$6000 to 6999", "Less than 20000" = "$5000 to 5999", "Less than 20000" = "$4000 to 4999", "Less than 20000" = "$3000 to 3999", "Less than 20000" = "$1000 to 2999", "Less than 20000" = "Lt $1000")) %>% count(rincome) %>% mutate(prop = prop.table(n))

Filter and prepare the data for visualization

data_to_plot <- df1 %>% filter(year == 2010) %>% filter(race != "Other" & rincome != "No answer" & rincome != "Refused" & rincome != "Not applicable" & rincome != "Don't Know")

Create a bar plot for race

ggplot(df1 %>% filter(year == 2010) %>% filter(race != "Other") %>% count(race) %>% mutate(prop = prop.table(n)), aes(x = race, y = prop, fill = race)) + geom_bar(stat = "identity") + labs(title = "Relative Frequency of Race in 2010", x = "Race", y = "Proportion") + theme_minimal()

Create a grouped bar plot

ggplot(data_to_plot, aes(x = race, fill = rincome)) + geom_bar(position = “dodge”) + labs(title = “Income vs Race in 2010”, x = “Race”, y = “Count”) + scale_fill_brewer(palette = “Set1”) + # Adjust the color palette theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels for better readability

Create a bar plot for reported income

ggplot(df1 %>% filter(year == 2010) %>% filter(rincome != “No answer” & rincome != “Refused” & rincome != “Not applicable” & rincome != “Don’t Know”), aes(x = rincome, fill = rincome)) + geom_bar() + labs(title = “Relative Frequency of Reported Income in 2010”, x = “Reported Income”, y = “Count”) + theme_minimal()