Perception towards premarital sex in relation to age and sex

Statistical inference on General Social Survey data

Tosin Akingbemisilu

29/06/2021

Getting Started

This work is an exploratory data analysis and inferential tests on the **General Social Survey (GSS) dataset**. The important R packages and datasets are loaded below to commence this analysis. Background is also stated immediately after the loading of packages and data.

Load packages

We will start by loading the relevant packages that will be needed in this work

```
library(ggplot2)
library(dplyr)
library(statsr)
library(tidyverse)
library(tidyr)
```

Load data

We will also be loading our data from the gss.Rdata file in the same local directory. As shown below, the data has 57061 observations with 114 variables.

```
load("gss.Rdata")
dim(gss)
```

```
## [1] 57061   114
```

Part 1: Data

Background

The GSS data is a sociological survey that contains data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. It contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered

are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. The questions cover a diverse range of issues including national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior.

This project uses an extract of the data from the General Social Survey (GSS) Cumulative File 1972-2012. As stated above, the data consists of 57061 observations with 114 variables. Each variable corresponds to a specific question asked to the respondent.

Methodology

The GSS survey is an in-person interview conducted face-to-face by NORC at the University of Chicago, with a target population of adults (18+) living in United States households. Respondents are randomly sampled from a mix of urban, suburban, and rural geographic areas, and participation is strictly voluntary.

Scope of Inference

The GSS performs random sampling, for broad generalization to the US population. It is however an observational study - with no explicit random assignments to treatments and the survey adopts a combination of cluster random sampling and stratified random sampling. The Primary Sampling Units (PSUs) employed are Standard Metropolitan Statistical Areas (SMSAs) or non-metropolitan counties (i.e., clusters). These SMSAs and counties are stratified by region, age, and race before selection.

We can therefore infer that the sample data should allow us to generalize to the population of interest, since random sampling techniques have been extensively used for the survey. Also, since the study is observational and not an experiment, there is no random assignment. Thus, we can infer correlation between the explanatory and response variables, but not causation.

In addition, potential biases are associated with non-response because this is a voluntary in-person survey that takes approximately 90 minutes. Some potential respondents may choose not to participate.

Part 2: Research question

Question 1

Is there a relationship between race and premarital sex perception?

Question 2

Is there a relationship between age and premarital sex perception?

This research will try to find the relationship between race, age and perception of each as a unique element towards premarital sex. This will help to understand the attitude of either the race or age of the population towards having premarital sex. Sexual behaviors of adolescents and youth have always been categorized as one of the main health priorities of a society because of high prevalence of human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS), sexually transmitted infections (STIs), and unwanted pregnancies. Drawing inference from the GSS data, this work would help to strengthen understanding on the

race and age of the population that are more prone to these diseases, based on their perception towards premarital sex, believing it would give a better insight to justify the belief of most people and also guide adolescent and youth health interventions.

Based on this, the analysis was done using the following data elements from the GSS data:

```
age - Age of respondents
race - Race of respondents
premarsx - opinion of respondents on sex before marriage
```

To ensure we are reflecting the recent trends, data from year 2010 and above will be used. Also, all NAs (empty cells) will be removed. The new data set we will now be using will then be renamed as `gss_sub`.

```
#below selects our year range and then removes all empty cells from our selected column
gss_sub <- gss[ which(gss$year >= 2010 & !is.na(gss$age) & !is.na(gss$race) & !is.na(gss$prem
        arsx)), ]
#below loads only our relevant column into our new table gss_sub, but tidyverse must be loade
        d for the select function to work
gss_sub <- gss_sub %>% select(age, race, premarsx)
dim(gss_sub)
```

```
## [1] 2654    3
```

Part 3: Exploratory data analysis

After cleaning, the data we are now using for this analysis has 2654 observations, with the summary below.

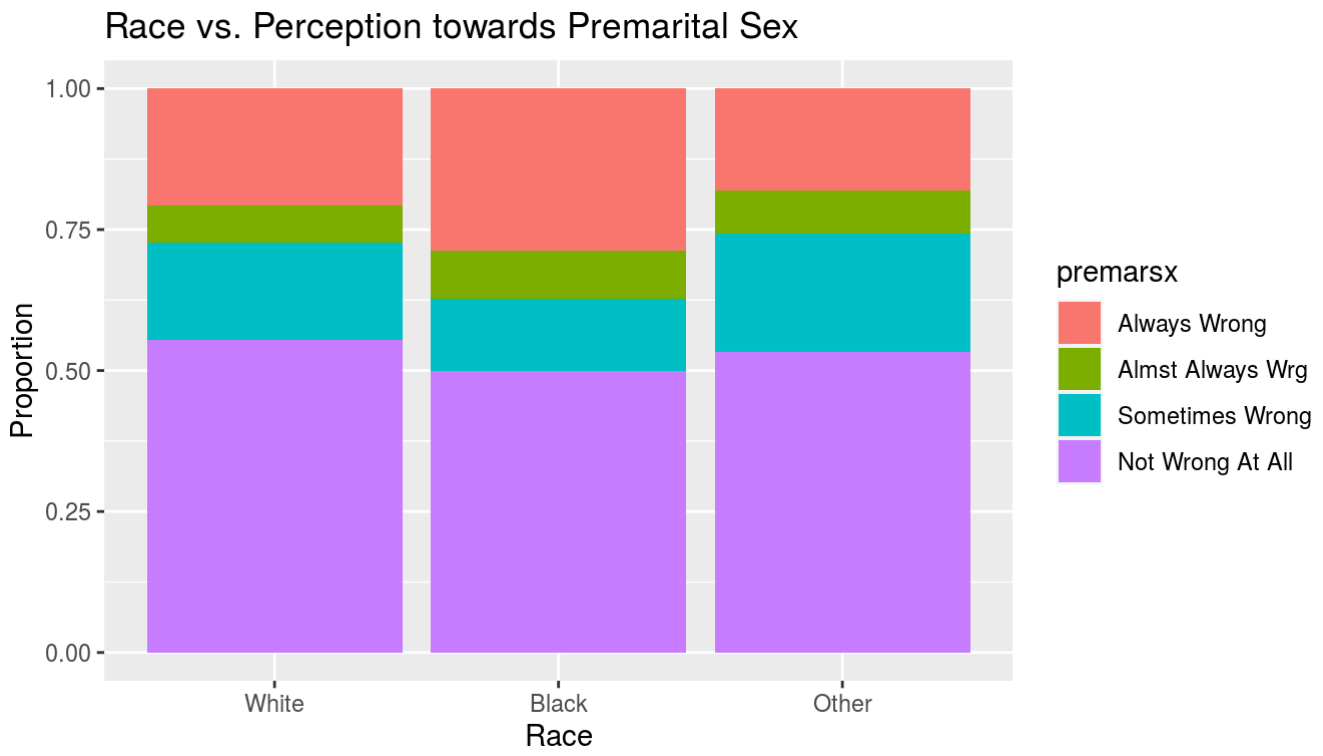
```
summary (gss_sub)
```

```
##       age           race                  premarsx
##  Min.   :18.00   White:2003   Always Wrong    : 576
##  1st Qu.:33.00   Black: 413   Almst Always Wrg: 187
##  Median :46.00   Other: 238   Sometimes Wrong : 447
##  Mean   :47.65                Not Wrong At All:1444
##  3rd Qu.:61.00                Other           :   0
##  Max.   :89.00
```

Race vs. Perception towards premarital sex

Next, let's look at the proportion of different races and their perception towards premarital sex. Since we are showing proportion, we will be using `geom_bar()` function as it would help make the height of the bar proportional to the number of our cases in each group.

```
g <- ggplot(gss_sub) + aes(x=race,fill=premarsx) + geom_bar(position = "fill") +
  labs(x="Race",y="Proportion",title="Race vs. Perception towards Premarital Sex")
g
```

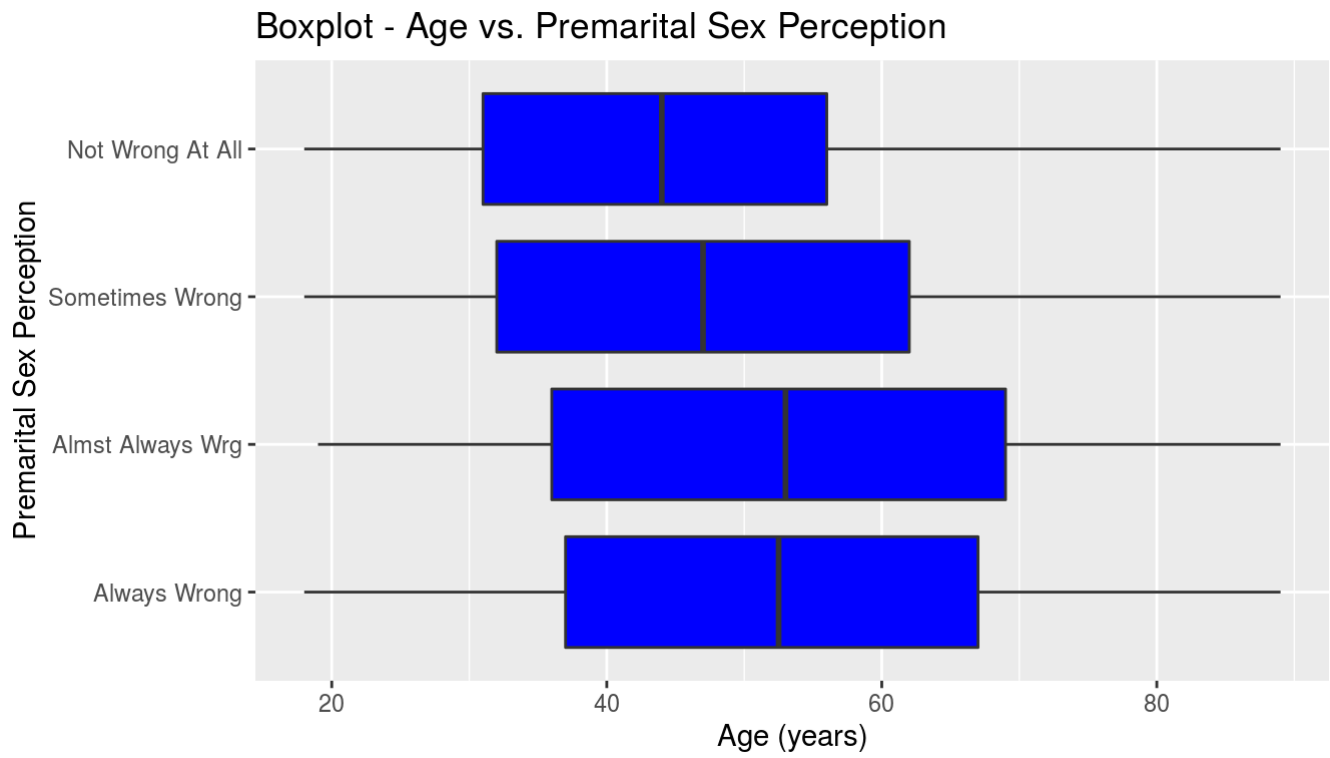


Looking at the chart above, it seems clear that race does not make much difference in the perception of respondents towards premarital sex, although there is slightly more proportion of black than white, who perceive it is always wrong. Nonetheless, majority of the respondents who are either white, black or others, believe premarital sex is not wrong at all.

Age vs. Perception towards premarital sex

Moving on, let's take a look at the opinion of respondents on premarital sex by their age. We will start by using a box plot chart, and see if we can explore or derive anything from it. For a better view, we will put Age in our x-axis and perception in our y-axis.

```
gss_sub <- gss_sub %>%
        filter(!is.na(age))

ggplot(gss_sub, aes(age, premarsx)) +
        geom_boxplot(fill = "blue") +
        labs(
                x = "Age (years)",
                y = "Premarital Sex Perception",
                title = "Boxplot - Age vs. Premarital Sex Perception"
        )
```



For more perspective into the data and better visualization insight, let's try to look at the age vs premarital sex perception using proportion bar plot.

```
g2 <- ggplot(gss_sub) +
  aes(x=age,fill=premarsx) +
  geom_bar(position = "fill") +
  labs(x="Respondent's Age",y="Proportion",title="Age vs. Perception on Premarital Sex")
g2
```

Our proportion bar chart actually gives a better insight into the age vs. premarital sex perception data. It seems glaring that age has an effect on the perception towards premarital sex. The perception that nothing is wrong at all with premarital sex reduces as age of respondents increase. We could however assume based on real life

scenario, that most of the people in the older category are already married, hence, the reason for their perception.

Part 4: Inference

Inference for Race vs Premarital Sex Perception

In this section, we perform statistical inference on the results (Race vs Perception on premarital sex). We will use Chi-Squared Test to test the independence of race and premarital sex perception. Chi-Squared Test was chosen since it is ideal for testing associated relationships between different categorical variables, which is valid in our case.

The conditions of inference will be checked for conducting a hypothesis test to compare two proportions

- **Independence:** yes, individuals are random sample and the sample also met.
- **10% condition:** yes, there are 57061 observations, which are less 10% of whole population.
- **Sample size/skew:** The boxplots we plotted in the previous section indicate a nearly normal distribution. In any case, our sample sizes are large enough to account for any skew in the distributions of the population of interest.

Hypothesis for Race vs Premarital Sex Perception defined below

We will first start by defining the hypothesis for the analysis.

- H_0 : - Race is not connected with people's perception towards premarital sex
- H_A : - Race is connected with people's perception towards premarital sex

Inference Test

Testing all observations

```
c_gss <-chisq.test(gss$race,gss$premarsx)
c_gss
```

```
##
##  Pearson's Chi-squared test
##
## data:  gss$race and gss$premarsx
## X-squared = 151, df = 6, p-value < 2.2e-16
```

Testing our cleaned data that reflects the recent trends - data from 2010 and above, with all 'NAs' removed

```
c_gsssub <-chisq.test(gss_sub$race,gss_sub$premarsx)
c_gsssub
```

```
##
##  Pearson's Chi-squared test
##
## data:  gss_sub$race and gss_sub$premarsx
## X-squared = 22.248, df = 6, p-value = 0.001092
```

The p value is nearly zero, so the null hypothesis should be rejected, and alternative hypothesis, indicating that people's Race is connected with their perception towards premarital sex should be accepted.

Inference for Age vs Premarital Sex Perception

In this section, we perform statistical inference on the results (Age vs Perception on premarital sex). Since we are comparing the averages of a numerical variable (Age) for more than two categories of a categorical variable (Premarital Sex perception), we will be using 1-way ANOVA test.

We will first start by defining the hypothesis for the analysis.

- H_0 : - There is no difference in means of people's age in relation to their perception towards premarital sex
- H_A : - There is difference in means of people's age in relation to their perception towards premarital sex

Inference Test

Testing the sub used for analysis, that reflects the recent trends, data from 2010 and above, with all NAs removed

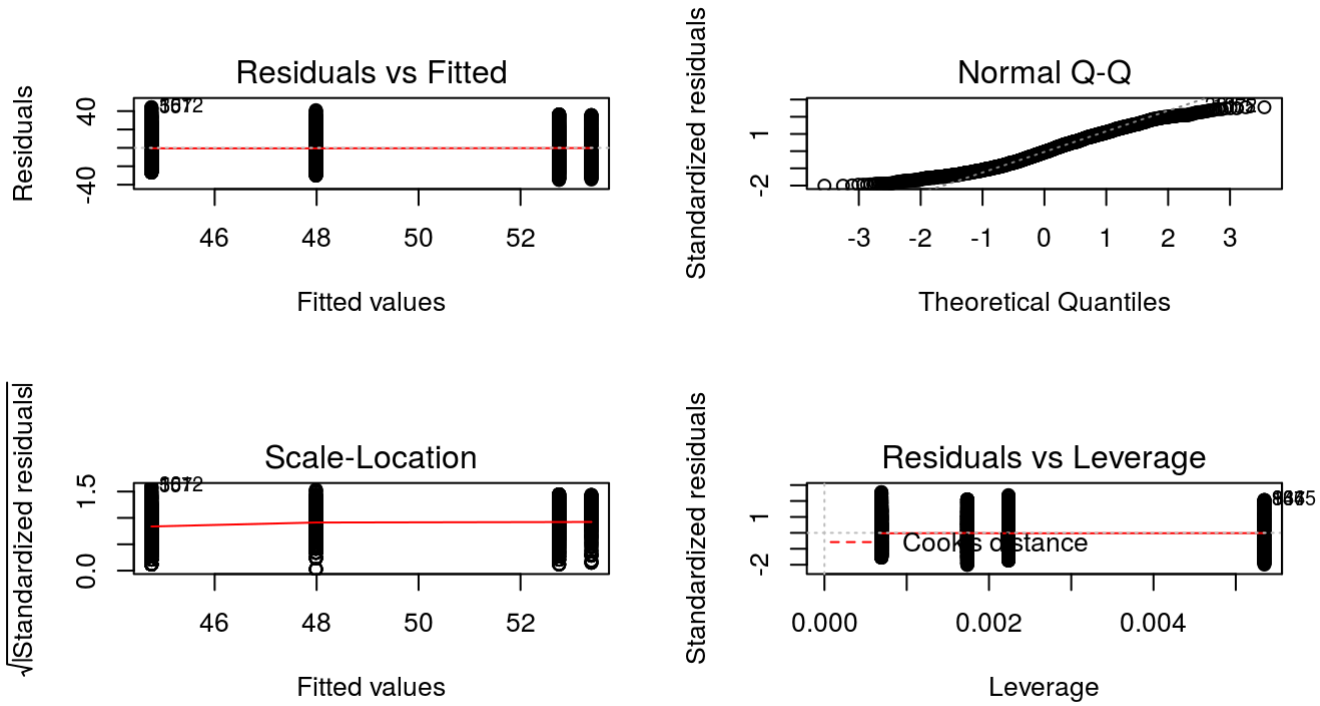
```
a_gsssub <- aov(age ~ premarsx, data = gss_sub)

summary(a_gsssub)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## premarsx       3  33298   11099   36.92 <2e-16 ***
## Residuals   2650 796574     301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Pr(>F) displays the P-value of the F-statistic. This shows how likely it is that the F-value calculated from the test would have occurred if the null hypothesis is of no difference among group means were true. In view of this, the P-value of the premarital sex perception is very low (f = 36.92, p < 0.001), therefore, the null hypothesis will be rejected and the alternative hypothesis accepted.

However, we will also go further to check whether the model fits the assumption of homoscedasticity

```
par(mfrow=c(2,2))
plot(a_gsssub)
```



Each plot above gives a specific piece of information about the model fit, but it's enough to know that the red line representing the mean of the residuals is horizontal and centered on one, in the scale-location plot, meaning that there are no large outliers that would cause bias in the model.

The normal Q-Q plot plots a regression between the theoretical residuals of a perfectly-homoscedastic model and the actual residuals of our model, so the closer to a slope of 1 this is the better. This Q-Q plot is very close, with only a bit of deviation.

From these diagnostic plots we can say that the model fits the assumption of homoscedasticity.

Overall, we can therefore say it is statistically significant that there is difference in means of people's age in relation to their perception towards premarital sex.

Conclusion

Based on our Research Questions:

1. Is there a relationship between race and premarital sex perception?

2. Is there a relationship between age and premarital sex perception?

With P-value < 0.001 in each case, we accept the alternative hypothesis that race is connected with people's perception towards premarital sex and there is statistically significance difference in people's age and their perception towards premarital sex.

Reference

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4145500/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4145500/)
- https://www.scribbr.com/statistics/anova-in-r/ (https://www.scribbr.com/statistics/anova-in-r/)