Project Part 1

Your name - Blank for Part 1

Data collection

The data set that will be covered is titled 'NCHS - Death rates and life expectancy at birth' . This data set was collected by the researcher from data.gov under the health section. The metadata was updated in August 20, 2018. This dataset from the U.S Department of Health & Human Services highlights the difference in age-adjusted death rates and life expectancy at birth by race and sex from the 1900 until 2015.

The data was collected from the U.S Department of Health & Human services, which collected their samples from 1900 to 2015. The columns are separated into 5 columns of year, race, sex, Average Life Expectancy (Years), and Age-adjusted Death Rate. The years are represented as a numeric for example class(2015) is an integer or numeric. The race is a string value that takes either the values of White, Black, or All Races. The Sex column is a string value that takes either the values of Male, Female, Both Sexes. The Average Life Expectancy is shown as a numeric number that represents how many years a person born in a certain year will have on average. The Age-adjusted represents the deaths per 100,000 which is calculated based on the 2000 U.S. standard population.

The data for age-adjusted death rate was collected for the population by postcensal estimates based on the 2010 census, estimated as of July 1, 2010. Rates for non-census years between 2000 and 2010 are revised using updated intercensal population estimates. This also may differ from rates previously published. Date on age-adjusted death rates prior to 1999 are taken from historical data from various references published.

In this data set Life expectancy data are available up to 2014 with 2015 being a year we do not focus on due to the missing information, which is why we focused on data from 1900-2014

Possible problems

Although this dataset has many valuable information it is subject to some problems as with any dataset. The one problem that is clearly visible is its representation of the population. Especially in America were poverty is affecting 12.3 percent of the population (Reference 2), it is possible that some people do not fill out or report in the consensus. Especially in

population/counties that are in poverty it is sometimes common to not report a birth or a death. From the data collection we can see that not everyone was included. Including everyone in itself is a hard method to do in collecting data so this study/project was focusing on the information provided by a reliable source. Also, another big problem that could have occurred was the reporting of infant deaths in the earlier years. Since in the older years it was hard conditions to live in due to the lack of technology many infants died or weren't even reported until the age of 1 in some states. This data points would have heavily skewed the data points in comparison to the modern time when infant mortality is not as common, which means the overall data is not as much affected.

Data summary

A summary of the data is provided both numerically and graphically. The table (tablemeans) shows the mean of Average_Life_Expectancy, mean of Age_Adjusted_Death_Rate, standard deviation of Average_Life_Expectancy, and standard deviation of Age_Adjusted_Death_Rate. This was done by race (All Races, Black, and White) which will be further discussed in the Data Summary section. We can clearly see that there is a big difference between Black vs White or All Races. We can also see the distribution of the data. As shown by the bar graph we can see that the data is skewed left. This is due to the death rate constantly increasing. This is further explained in the graphical summary of the linearized model in the Data Summary. It was also evident that life epxectancy and age-adjusted death rate had a strong linear relationship with years, which is why we decide to look at this problem in further detail (tablecorr).

The numerical represented is shown by the table named tablerace19002010. This table represents the average life expectancy and age-adjusted death rate of the years 2010 and 1900. Looking at the data collected in the year 2010 the Average Life Expectancy was 78.7, 75.1, 78.9 for All races, black, and white respectively. This shows that there isn't a great difference in Life Expectancy in comparison of each race. However, the statistics for 2010 were drastically different than 1900s. In the year 1900 the Average Life Expectancy was 47.3, 33.0, and 47.6 for all races, black, and white respectively. We can see that in comparison to other races the blacks had a lower life expectancy. This may be because of the lack of resources in the older days as well as because of segregation laws restricting black people to receive the resources they need. We can see that as time progressed the life expectancy became similar to other races. The Age-adjusted death rate for 2010 was 747.0, 898.2, and 741.8 for All races, Black, and white respectively. It seems that All races and White is similar in Age-adjusted

death rate for 2010 with Black having a higher death rate in comparison there can be many other factors associated with this with one being poverty. Even until this day we see that more black children are living in poverty compared to other races. (Reference 3). To look at this data in comparison we need to look at the data for 1900. The Age-adjusted death rate for 1900 was 2518.0, 3423.3, and 2501.2 for All races, Black, and white respectively. We can see this trend of black people with a higher age-adjusted death rate is higher in comparison to all the other races, which makes logical sense in our history of segregation. It seems to be difference has decreased over time as well as Age-adjusted death rate in general has decreased over time.

The graphical representation is a graph of life expectancy of all races over time. As the graph shows the x-axis is the years, which is divided into 10 years. However, there are points for every year, and for every year represented there are three class divisions of All Races, Black, and White as shown as the legend. The scatter plot (the points) are set to a transparency of 0.3 so that it doesn't distract the reader from the linear regression line. The linear regression lines also have a confidence interval, but due to the high number of observations it is very faint and narrow. We can infer from the linear regression model that as time increased the Average_Life_Expectancy for all races increased. It is also visible that the difference between life expectancy of races decreased by a large factor over time, which may have factors associated with it such as health care and modern technologies.

The conclusion that is common for both of the studies is that it is true that the average life expectancy increases from 1900 to 2014. Another interesting discovery was that the difference in the average life expectancy decreased throughout the years as the average life expectancy increased. This could be related to social norms and practices being changed, but that was not the focus of this study. An improvement or further research that could be done will be to compare average life expectancy by state instead of the country itself if data permits such research. This will be able to provide more information about specific locations and their effect on the average life expectancy.

Data manipulation/Representation

```
library(ggplot2)
setwd("/Users/maxryoo/Documents/Fall 2018/STAT3080/Project1")
dataset <- read.csv("deathrateslifeexpect.csv")
## Years represented excluding 2015 due to its lack of life expectancy
yrs <- unique(dataset$Year)
```

```
race.list <- lapply(1:length(yrs), function(x)
  dataset[dataset$Year == yrs[x] & dataset$Year != 2015,])
## Average life expectancy and Age adjusted death rate (every 5 years all races)
yrs5 <- yrs[yrs%%5 == 0]
raceboth<- lapply(2:length(yrs5), function(x)
  dataset[dataset$Sex == "Both Sexes" & dataset$Year == yrs5[x] ,])
for (i in 1:length(raceboth)) {
  names(raceboth[[i]]) <-
    c("Year","Race","Sex","Average_Life_Expectancy", "Age_Adjusted_Death_Rate")
}
## 1900 vs 2014
bothsex2010lifeex <- raceboth[[1]]$Average_Life_Expectancy
bothsex2010death <- raceboth[[1]][5]$Age_Adjusted_Death_Rate
bothsex1900lifeex <-
  raceboth[[length(raceboth)]]$Average_Life_Expectancy
bothsex1900death <- raceboth[[length(raceboth)]]$Age_Adjusted_Death_Rate
tablerace19002010 <-
  cbind(a = bothsex2010lifeex, b=bothsex1900lifeex,
        c=bothsex2010death, d=bothsex1900death)
rownames(tablerace19002010) <- raceboth[[1]]$Race
colnames(tablerace19002010) <-
  c("LE 2010", "LE 1900", "AADR 2010", "AADR 1900")
## Numerical Summary
diffrace <- unique(dataset$Race)
byrace <- lapply(1:length(diffrace), function(x)
  dataset[dataset$Race == diffrace[x] & dataset$Year != 2015,])
for (i in 1:length(byrace)) {
  names(byrace[[i]]) <-
    c("Year","Race","Sex","Average_Life_Expectancy", "Age_Adjusted_Death_Rate")
}
avgbyracelife <- sapply(byrace, function(x) mean(x$Average_Life_Expectancy))
avgbyracedeath <- sapply(byrace, function(x) mean(x$Age_Adjusted_Death_Rate))
sdbyracelife <- sapply(byrace, function(x) sd(x$Average_Life_Expectancy))
sdbyracedeath <- sapply(byrace, function(x) sd(x$Age_Adjusted_Death_Rate))
tablemeans <- rbind(a = avgbyracelife, b=avgbyracedeath,
                    c=sdbyracelife ,d=sdbyracedeath)
```

```
colnames(tablemeans) <- diffrace
rownames(tablemeans) <- c("Average_Life_Expectancy", "Age_Adjusted_Death_Rate",
                          "SD of Average_Life_Expectancy",
                          "SD of Age_Adjusted_Death_Rate")
corbyracelife <- sapply(byrace, function(x) cor(x$Average_Life_Expectancy,x$Year))
corbyracedeath <- sapply(byrace, function(x) cor(x$Age_Adjusted_Death_Rate,x$Year))
corbyracelifesqr <- corbyracelife * corbyracelife
corbyracedeath <- corbyracedeath*corbyracedeath
tablecorr <- rbind(a = corbyracelife, b= corbyracedeath,
                   c=corbyracelifesqr, d=corbyracedeath)
colnames(tablecorr) <- diffrace
rownames(tablecorr) <- c("r (Years vs Average Life Expectancy)",
                         "r (Years vs Age Adjusted Death)",
                         "r^2 (Years vs Average Life Expectancy)",
                         "r^2 (Years vs Age Adjusted Death)")
#LE = Life Expectancy AADR = Age-Adjusted Death Rate
print(tablerace19002010)
```

```
##           LE 2010 LE 1900 AADR 2010 AADR 1900
## All Races    78.7    47.3     747.0    2518.0
## Black        75.1    33.0     898.2    3423.3
## White        78.9    47.6     741.8    2501.2
```

```
print(tablemeans)
```

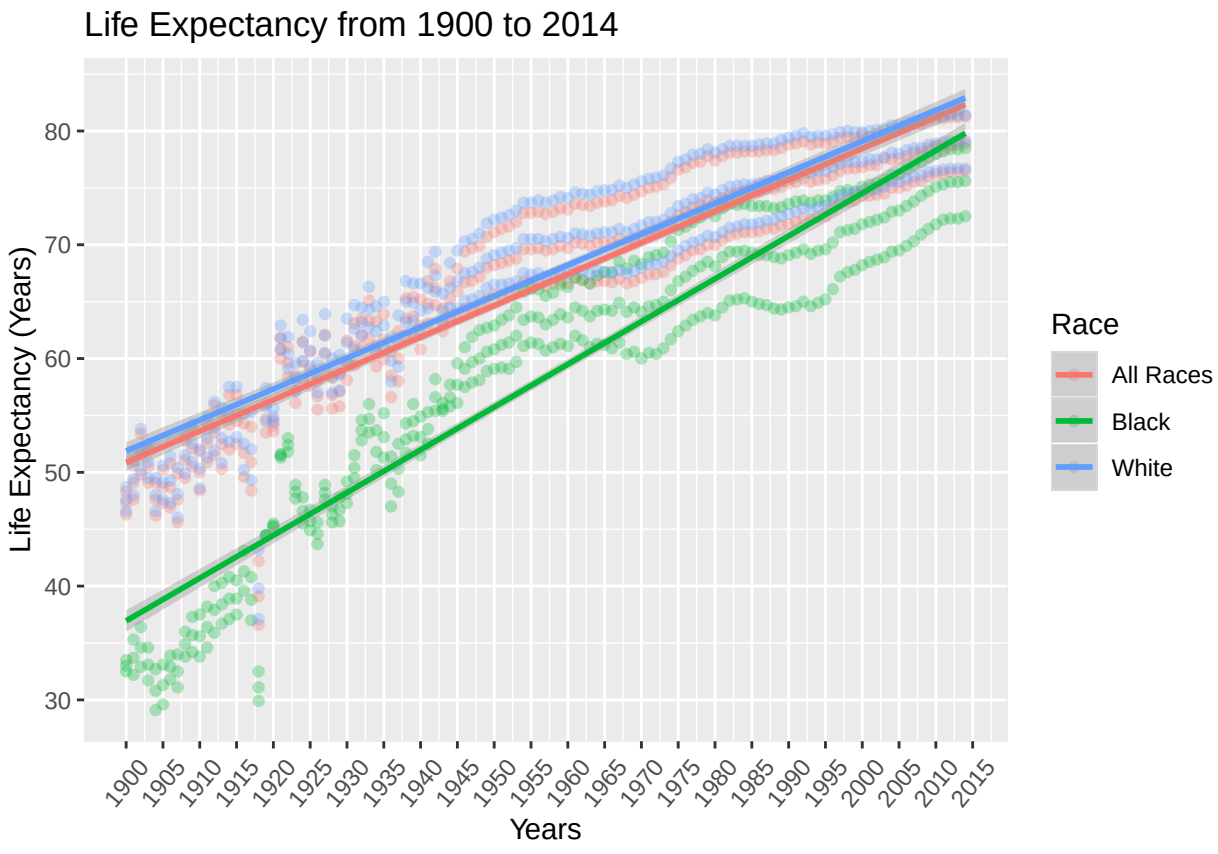
```
##                                 All Races      Black       White
## Average_Life_Expectancy         66.587246   58.36870   67.395942
## Age_Adjusted_Death_Rate       1495.082029 1905.05797 1463.769565
## SD of Average_Life_Expectancy    9.823418   13.25659    9.761832
## SD of Age_Adjusted_Death_Rate  566.209390  788.50431  557.865899
```

```
print(tablecorr)
```

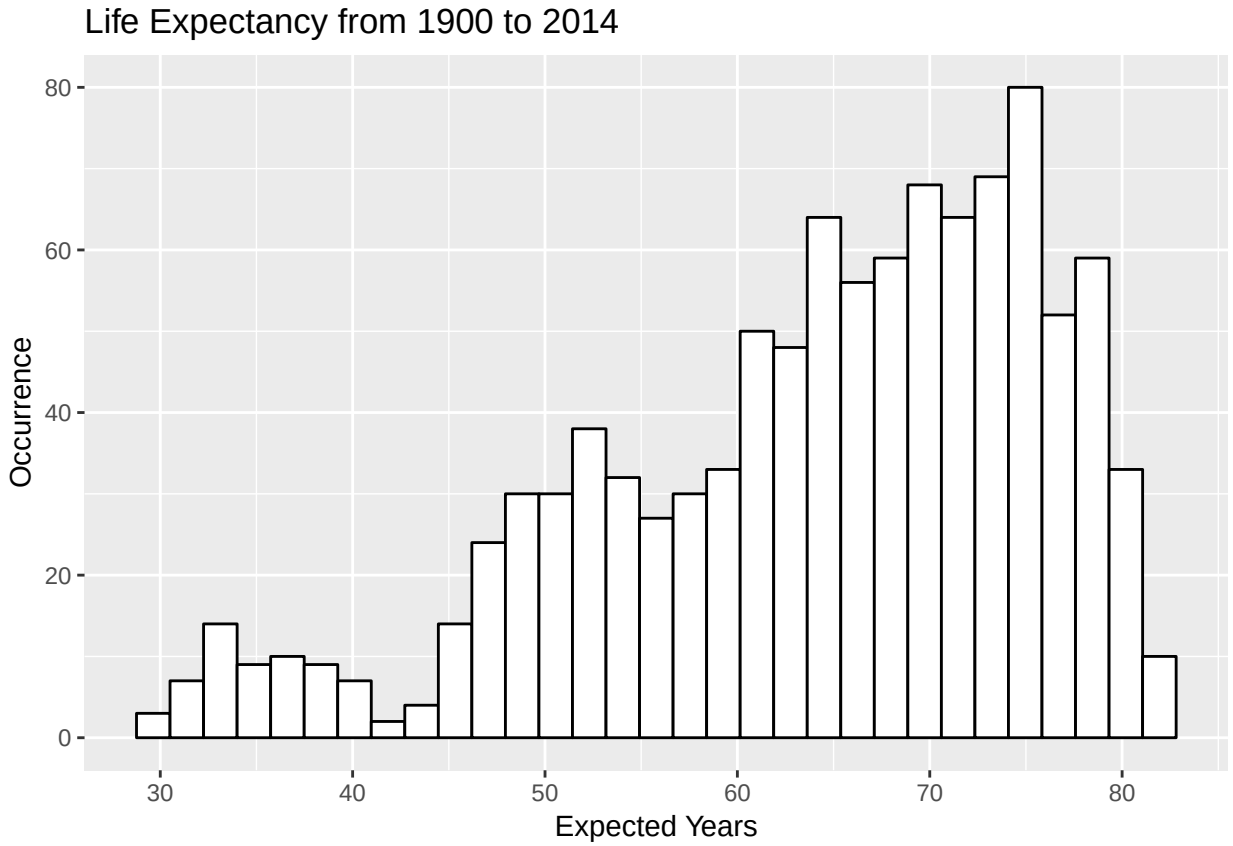
```
##                                          All Races     Black     White
## r (Years vs Average Life Expectancy)   0.9337486 0.9426009 0.9272683
## r (Years vs Age Adjusted Death)        0.8868192 0.8662649 0.8844759
## r^2 (Years vs Average Life Expectancy) 0.8718865 0.8884965 0.8598265
## r^2 (Years vs Age Adjusted Death)      0.8868192 0.8662649 0.8844759
```

```
exclude.2015 <- dataset[dataset$Year != 2015,]
```

```
colnames(exclude.2015) <- c("Year", "Race",
                            "Sex", "Average_Life_Expectancy",
                            "Age_Adjusted_Death_Rate")
lm.plot <- ggplot(exclude.2015, aes(x=Year,
                                    y=Average_Life_Expectancy,colour=Race)) +
  geom_point(alpha=0.3) + geom_smooth(method=lm)
lm.plot.names <- lm.plot+labs(title="Life Expectancy from 1900 to 2014",
                              y="Life Expectancy (Years)", x="Years") +
  scale_x_continuous(breaks=yrs5) +
  theme(axis.text.x=element_text(angle=50,vjust=0.5))
hist1 <- ggplot(exclude.2015, aes(x=exclude.2015$Average_Life_Expectancy))
binsize <- diff(range(exclude.2015$Average_Life_Expectancy))/30
bargraph <- hist1 + geom_histogram(binwidth=binsize, fill="white", colour="black") +
  labs(title="Life Expectancy from 1900 to 2014", x="Expected Years", y="Occurrence")
print(lm.plot.names)
```




```
print(bargraph)
```



References

1. <https://catalog.data.gov/dataset/age-adjusted-death-rates-and -life-expectancy-at-birth-all-races-both-sexes-united-sta-1900>
2. https://www.usnews.com/news/healthiest-communities/articles/2018-09-12/poverty-in-america-new-census-data-paint-an-unpleasant-picture
3. http://www.nccp.org/media/releases/release_34.html