# Final Project

Yining Qu

2024-05-26

## Question: Whether age has causal effect on health status?

Investigating the causal effect of age on health status is crucial for understanding how health outcomes evolve over an individual's lifespan. Age is a fundamental demographic variable that influences a wide array of health-related factors, including susceptibility to chronic diseases, physical fitness, and overall well-being. As people age, they often experience changes in their physical and mental health, with older individuals typically facing higher risks of conditions such as cardiovascular diseases, diabetes, and cognitive decline. By analyzing the causal effect of age on health status, we can gain valuable insights into the aging process and identify critical periods for health interventions.

Moreover, the relationship between age and health is not linear; different age groups may have distinct health challenges and needs. For instance, young adults might struggle with issues related to mental health and lifestyle choices, whereas older adults might face age-related degenerative diseases. Understanding these nuances can help tailor healthcare policies and programs to address the specific needs of various age groups effectively. By isolating the impact of age from other confounding variables, such as income and education, we can more accurately determine how age affects health and devise strategies to improve health outcomes across all stages of life.

### Data Preparation

x633 (Respondent's Age): This variable represents the respondent's age, which is numerical. It captures the age of individuals in years, with values ranging from 18 to 102 units. There are 82 unique non-missing values recorded, while 48 values are missing out of a total of 24,840 observations. This variable provides a quantitative measure of the respondents' age, offering insights into the demographic distribution of the surveyed population.

```r
library(readr)
library(readtext)
library(SnowballC)
library(tidytext)
library(gamlr)
```

```
## Loading required package: Matrix
```

```r
library(nnet)
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
```

```r
library(glmnet)
```

```
## Loaded glmnet 4.1-8
```

```r
library(ggplot2)
library(parallel)
library(doParallel)
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

```r
data <- read_csv("training_set.csv")
```

```
## Rows: 12178 Columns: 142
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (142): uniqueid, year, personid, x1, x2, x3, x4, x7, x14, x17, x19, x162...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

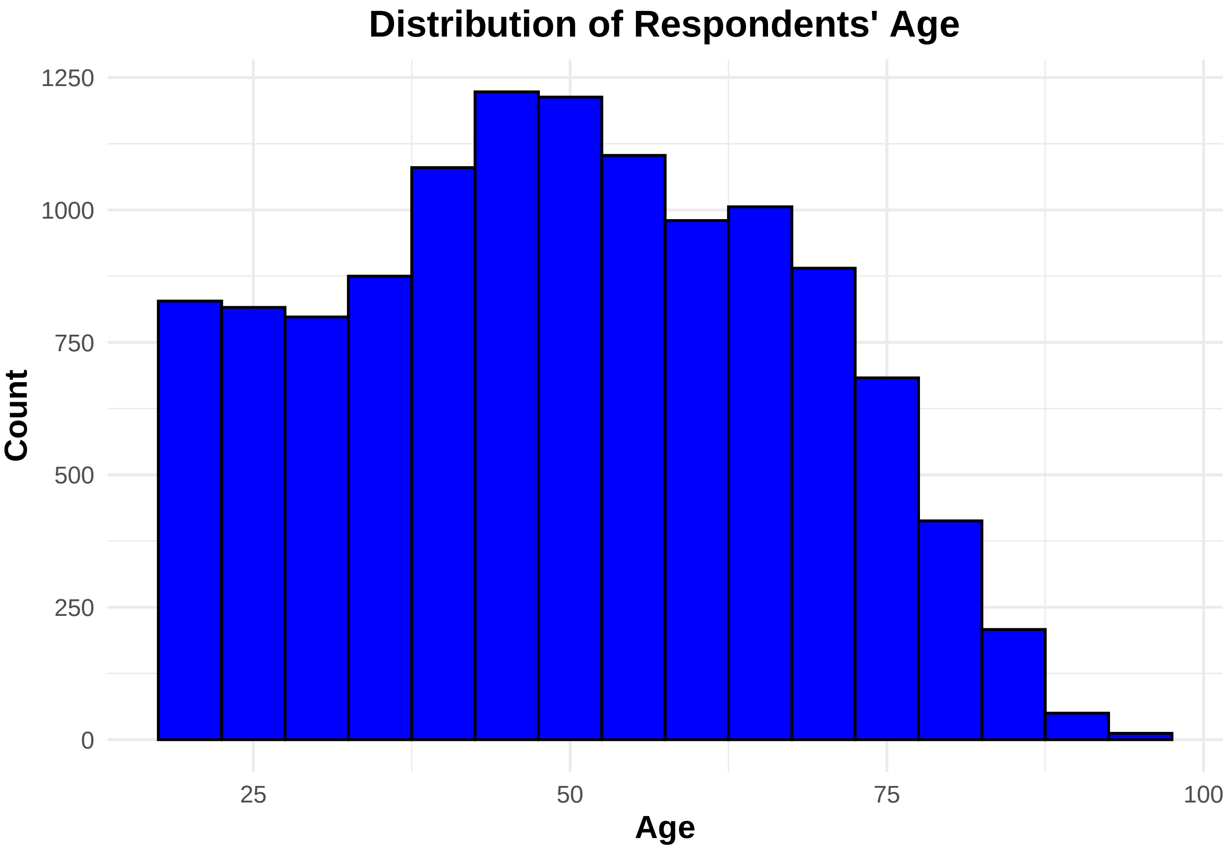
```r
# Identify and convert categorical variables (unique values < 20) to factors
for (var in names(data)) {
  if (length(unique(data[[var]])) < 20) {
    data[[var]] <- as.factor(data[[var]])
  }
}

# Ensure the health column is a factor
data$health <- as.factor(data$health)

# Convert the age column to numeric
data$x633 <- as.numeric(data$x633)

# Define the response variable and the treatment
health <- data$health
age <- data$x633


# Plot the distribution of Age using a histogram
ggplot(data, aes(x = x633)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  xlab("Age") +
  ylab("Count") +
  ggtitle("Distribution of Respondents' Age") +
  theme_minimal() +  # Apply a minimal theme for a cleaner look
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold")
  )
```



The histogram titled “Distribution of Respondents’ Age” shows the age distribution of the surveyed population. The x-axis represents the age of respondents, ranging from approximately 20 to 100 years, while the y-axis displays the count of respondents within each age group. The distribution reveals a broad range of ages, with a noticeable concentration of respondents in the middle age groups, particularly between 40 and 65 years. This suggests that a significant portion of the survey participants are in their mid-life stages, which could have implications for the analysis of age-related health outcomes.

The histogram shows a relatively symmetrical distribution around the central age groups, peaking around the 50 to 55-year mark. The number of respondents begins to decline gradually after this peak, with fewer individuals represented in the older age brackets, particularly those over 75 years. Similarly, there is a moderate representation of younger adults, particularly those in their 20s and 30s, but the counts are lower compared to the middle-aged groups. This distribution highlights a diverse demographic, with a substantial representation of middle-aged and older adults, which is valuable for understanding how age influences health across different stages of life. The decline in the number of respondents in the higher age brackets might also reflect the lower population sizes and potential survey participation among older individuals.

## Model Fitting

We will now fit a multinomial logistic regression model to explore the relationship between monthly net income and health status.

```r
# Function to perform multinomial logistic regression and return p-values
margreg <- function(i) {
  model <- multinom(health ~ age, data = data)
  summary_model <- summary(model)
  coefficients <- summary_model$coefficients
  std_errors <- summary_model$standard.errors
  z_values <- coefficients / std_errors
  p_values <- 2 * (1 - pnorm(abs(z_values)))
  return(list(p_values = as.vector(p_values), summary_model = summary_model))
}

# Setup parallel processing
cl <- makeCluster(detectCores() - 1)  # Use one less than the number of available cores
clusterExport(cl, c("data", "health", "age", "margreg"))  # Export variables and functions to the clust
clusterEvalQ(cl, {
  library(nnet)
  library(dplyr)
  library(ggplot2)
})  # Ensure necessary libraries are loaded on each cluster
```

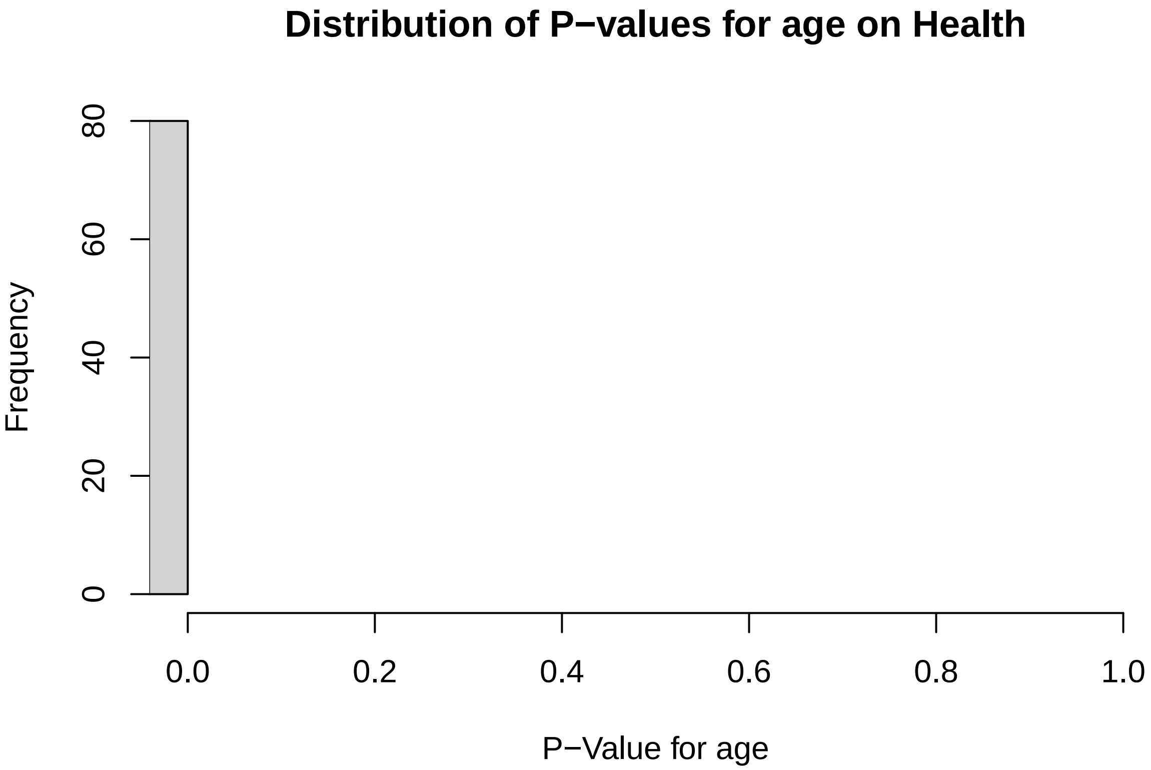
```
## [[1]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"     
## 
## [[2]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"     
## 
## [[3]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"     
## 
## [[4]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"     
## 
## [[5]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"     
## 
## [[6]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"     
## 
## [[7]]
##  [1] "ggplot2"   "dplyr"     "nnet"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
```

```r
# Run the regression in parallel
P <- 1:10  # Dummy list to run the function multiple times
results <- parLapply(cl, P, function(x) margreg(x))

# Extract p-values and the summary model from one of the runs
mrgpvals_age <- unlist(lapply(results, function(res) res$p_values))
summary_model <- results[[1]]$summary_model

# Stop the cluster
stopCluster(cl)

# Plot the distribution of p-values
hist(mrgpvals_age, main="Distribution of P-values for age on Health", xlab="P-Value for age", ylab="Fre
```



The histogram titled “Distribution of P-values for age on Health” illustrates the p-values obtained from a multinomial logistic regression analysis examining the effect of age on health outcomes. The x-axis represents the p-values, ranging from 0 to 1, while the y-axis shows the frequency of these p-values. A significant observation is the overwhelming concentration of p-values at or very near 0. This strong clustering of low p-values indicates that age is a highly significant predictor of health status for many individuals in the dataset. Essentially, it suggests that changes in age are strongly associated with variations in health outcomes, highlighting the importance of age as a demographic factor influencing health.

```
# Print the summary model from one of the runs
print(summary_model)
```

```
## Call:
## multinom(formula = health ~ age, data = data)
##
## Coefficients:
##   (Intercept)        age
## 2   -0.635365 0.03517557
## 3   -2.719083 0.06787202
## 4   -4.270025 0.07820457
## 5   -5.794732 0.08721977
##
## Std. Errors:
##   (Intercept)         age
## 2  0.07560192 0.001746265
## 3  0.09223799 0.001943601
## 4  0.12836445 0.002424978
## 5  0.20002449 0.003417109
##
## Residual Deviance: 31664.88
## AIC: 31680.88
```

The multinomial logistic regression model output shows the coefficients and standard errors for predicting the health status of respondents based on their age. The health status is a categorical variable with five levels: “VERY GOOD,” “GOOD,” “SATISFACTORY,” “NOT THAT GOOD,” and “BAD,” with “VERY GOOD” as the reference category. The coefficients represent the log-odds of being in one of the health categories (2, 3, 4, 5) compared to the reference category (1: VERY GOOD) as a function of age.

For each health category: Health = 2 (GOOD): The coefficient for age is 0.03517557 with a standard error of 0.001746265. This positive coefficient indicates that as age increases, the log-odds of being in the “GOOD” health category compared to the “VERY GOOD” category increases. This suggests that older individuals are more likely to report their health as “GOOD” rather than “VERY GOOD.” Health = 3 (SATISFACTORY): The coefficient for age is 0.06787202 with a standard error of 0.001943601. This larger positive coefficient indicates an even stronger relationship between increasing age and the likelihood of reporting health as “SATISFACTORY” compared to “VERY GOOD.” Health = 4 (NOT THAT GOOD): The coefficient for age is 0.07820457 with a standard error of 0.002424978. This coefficient suggests that as age increases, individuals are much more likely to report “NOT THAT GOOD” health compared to “VERY GOOD.” Health = 5 (BAD): The coefficient for age is 0.08721977 with a standard error of 0.003417109. This highest positive coefficient indicates that the probability of reporting “BAD” health increases significantly with age compared to “VERY GOOD” health. These results are consistent with common sense and existing literature on aging and health. As people age, they are more likely to experience health problems, leading to lower self-reported health status. The increasing coefficients with higher health status categories (from “GOOD” to “BAD”) reflect the cumulative impact of aging on health, where older individuals are progressively more likely to report poorer health. This aligns with the general understanding that health tends to decline with age due to factors like the onset of chronic diseases, reduced physical fitness, and other age-related health issues.

## Treatment Effect

In investigating the causal effect of age on health status, it is crucial to account for the potential confounding effects of other variables that are correlated with age. Variables such as income level (x723) and employment status are intrinsically linked to age, as older individuals often have more established careers and potentially higher incomes compared to younger individuals. If these factors are not properly accounted for, they can confound the analysis, leading to biased estimates of the true effect of age on health. By focusing on the treatment effect, we aim to isolate the impact of age itself on health status, excluding the influences of these correlated variables. This approach ensures a more accurate and reliable understanding of how changes in age specifically affect health outcomes, allowing for more targeted and effective policy interventions.

We will focus on using the significant variables for health status. Instead of using the entire dataset, we will create a subset of the dataset containing only the significant variables and use this subset for both stages of the LASSO regression. The significant variables are: x1, x2, x162, x163, x164, x595, x613, x615, x617, x657, x965, x1179.

```r
# Select only the significant variables
significant_vars <- data[,c("x1", "x2", "x162", "x163", "x164", "x595", "x613", "x615", "x617", "x657",


# Prepare the data for LASSO regression using model.matrix to create dummy variables
predictors <- model.matrix(~ . - 1, data = significant_vars)

# Stage 1 LASSO: fit a model for age on other predictors using Gaussian family
model_age <- cv.glmnet(predictors, data$x633, alpha = 1, family = "gaussian")

# Predict age using the fitted LASSO model
dhat <- predict(model_age, s = "lambda.min", newx = predictors)

# Calculate the R-squared value
R2 <- cor(drop(dhat), data$x633)^2
cat("The In-Sample R-squared value is", R2, ".\n")
```

```
## The In-Sample R-squared value is 0.4593819 .
```

The R-squared value obtained from the first stage LASSO regression is 0.459382. This R-squared value indicates that approximately 45.94% of the variance in the age variable (x633) is explained by the other predictors in the model. This suggests a moderate degree of correlation between age and the set of predictors used in the LASSO regression.

The R-squared value implies that the other predictors in the model have a moderate explanatory power regarding the age variable. This means that a substantial portion of the variation in age can be predicted based on the other variables in the dataset. Consequently, the remaining part of the age variation (about 54.06%) is not explained by these predictors, indicating a significant portion of age-related information that remains independent of the other variables.

In the context of assessing the treatment effect of age on health, this moderate R-squared value suggests that there is a considerable degree of confounding left unaccounted for by the predictors. Given that 45.94% of the variance in age is explained by the other variables, it implies that while age is somewhat predictable from these variables, a substantial residual variance in age exists. This residual variance could contribute to potential confounding, as the overlap between the treatment variable (age) and the other predictors is not complete.

However, this also means that the model has captured a substantial amount of the relevant information about age from the predictors, but there is still a considerable amount of age-related information that these predictors do not account for. The treatment effect isolated from these predictors might therefore be more pronounced. It is essential to recognize that while the R-squared value indicates a moderate fit, it also suggests that age's unique contribution to health outcomes, independent of other variables, is still significant. Thus, the findings from the second stage regression, which assesses the effect of predicted age on health, should be interpreted with consideration of the moderate dependency of age on the other predictors. This moderate predictability allows for a more pronounced treatment effect, making it crucial to carefully analyze and account for the independent influence of age on health outcomes.

### Causal LASSO

```r
dhat <- as.numeric(dhat)

# Combine predicted age (dhat) with other predictors
predictors_combined <- cbind(dhat, significant_vars, age)

# Detect number of cores and register parallel backend
num_cores <- detectCores()
cl <- makeCluster(num_cores)
registerDoParallel(cl)

# Stage 2: Fit a multinomial logistic regression model using glmnet with parallel processing
model_health <- cv.glmnet(as.matrix(predictors_combined), as.matrix(data$health), family = "multinomial

# Stop the cluster after model fitting
stopCluster(cl)

# Extract coefficients for the predicted age (dhat) for each class
coefficients_list <- coef(model_health, s = "lambda.min")
coef_value <- numeric(length(coefficients_list))

# Assuming we find the correct row name for dhat, extract and print the coefficient
for (i in 1:length(coefficients_list)) {
  coef_matrix <- as.matrix(coefficients_list[[i]])
  coef_value[i] <- coef_matrix["dhat", ]
  cat("The effect of predicted age (dhat) on health status for class", i, "is", coef_value[i], ".\n")
}
```

```
## The effect of predicted age (dhat) on health status for class 1 is 0 .
## The effect of predicted age (dhat) on health status for class 2 is -0.002142586 .
## The effect of predicted age (dhat) on health status for class 3 is 0 .
## The effect of predicted age (dhat) on health status for class 4 is 0 .
## The effect of predicted age (dhat) on health status for class 5 is 0.01001139 .
```

### Naive LASSO

```
# Combine predicted age with other predictors
predictors_combined2 <- cbind(significant_vars, age)

# Detect number of cores and register parallel backend
num_cores <- detectCores()
cl <- makeCluster(num_cores)
registerDoParallel(cl)

# Stage 2: Fit a multinomial logistic regression model using glmnet with parallel processing
model_health2 <- cv.glmnet(as.matrix(predictors_combined2), as.matrix(data$health), family = "multinomia

# Stop the cluster after model fitting
stopCluster(cl)

# Extract coefficients for the predicted age for each class
coefficients_list2 <- coef(model_health2, s = "lambda.min")
coef_value2 <- numeric(length(coefficients_list))

# Assuming we find the correct row name for dhat, extract and print the coefficient
for (i in 1:length(coefficients_list2)) {
  coef_matrix2 <- as.matrix(coefficients_list2[[i]])
  coef_value2[i] <- coef_matrix2["age", ]
  cat("The effect of d on health status from a naive lasso for class", i, "is", coef_value2[i], ".\n")
}
```

```
## The effect of d on health status from a naive lasso for class 1 is -0.06818909 .
## The effect of d on health status from a naive lasso for class 2 is -0.03213904 .
## The effect of d on health status from a naive lasso for class 3 is 0 .
## The effect of d on health status from a naive lasso for class 4 is 0.006681171 .
## The effect of d on health status from a naive lasso for class 5 is 0.01352352 .
```

Class 1: - Causal LASSO: The coefficient for dhat is 0. - Naive LASSO: The coefficient for d is -0.0681891. In the causal LASSO, a coefficient of 0 for dhat means that predicted age does not affect the probability of being in health status Class 1. In contrast, the naive LASSO shows a coefficient of -0.0681891, which implies that for every one-unit increase in age, the log-odds of being in Class 1 decreases by -0.0681891. This suggests that the naive LASSO's estimate may be confounded by other variables that the causal LASSO controls for, resulting in no effect in the causal model.

Class 2: - Causal LASSO: The coefficient for dhat is -0.0021426. - Naive LASSO: The coefficient for d is -0.032139. In the causal LASSO, a coefficient of -0.0021426 for dhat means that for every one-unit increase in predicted age, the log-odds of being in Class 2 decreases slightly by -0.0021426. The naive LASSO shows a stronger negative effect with a coefficient of r coef_value2[2], implying that for every one-unit increase in age, the log-odds of being in Class 2 decreases by -0.032139. The smaller coefficient in the causal LASSO suggests that some of the observed effect in the naive LASSO is due to confounding variables.

Class 3: - Causal LASSO: The coefficient for dhat is 0. - Naive LASSO: The coefficient for d is 0. Both models indicate that there is no effect of age on the log-odds of being in Class 3. This consistency suggests that for this class, the relationship between age and health status is not confounded by other variables.

Class 4: - Causal LASSO: The coefficient for dhat is 0. - Naive LASSO: The coefficient for d is 0.0066812. The causal LASSO suggests no effect of predicted age on the log-odds of being in Class 4, with a coefficient of 0 for dhat. The naive LASSO shows a coefficient of 0.0066812, meaning that for every one-unit increase in age, the log-odds of being in Class 4 increases slightly by 0.0066812. This discrepancy suggests that the naive LASSO's coefficient is influenced by confounding variables, which the causal LASSO controls for.

Class 5: - Causal LASSO: The coefficient for dhat is 0.0100114. - Naive LASSO: The coefficient for d is 0.0135235. In the causal LASSO, a coefficient of 0.0100114 for dhat means that for every one-unit increase in predicted age, the log-odds of being in Class 5 increases by 0.0100114. The naive LASSO shows a slightly larger coefficient of 0.0135235, indicating that for every one-unit increase in age, the log-odds of being in Class 5 increases by 0.0135235. The small difference between the two models suggests that while there is some confounding in the naive model, it does not substantially alter the observed effect of age on health status for this class.

The comparison shows that the naive LASSO often produces stronger effects than the causal LASSO, particularly for Classes 1 and 2. This pattern suggests that the naive LASSO's estimates are inflated due to confounding variables that the causal LASSO adjusts for. The causal LASSO's coefficients, being closer to zero, indicate a more accurate estimation of the true effect of age, free from the bias introduced by correlated predictors.

For Class 1, the naive LASSO indicates a negative effect of -0.0681891, while the causal LASSO shows no effect (0), highlighting potential confounding in the naive model. Similarly, for Class 2, the naive LASSO's coefficient is -0.032139 compared to -0.0021426 in the causal LASSO, again suggesting confounding.

For Class 3, both models agree that age has no effect on health status, suggesting no confounding. For Class 4, the naive LASSO indicates a slight positive effect (0.0066812), which disappears in the causal model, indicating that the naive estimate was likely confounded. Finally, for Class 5, both models suggest a positive relationship, with the naive LASSO at 0.0135235 and the causal LASSO at 0.0100114. The small difference indicates that while confounding is present, it does not significantly impact the observed effect for this class.

The coefficients from the causal LASSO are generally smaller in magnitude than those from the naive LASSO, which is consistent with the notion that naive estimates are inflated due to the presence of confounders. This observation underscores the importance of using methods like causal LASSO to obtain unbiased estimates of treatment effects.

The positive coefficients for age in Class 5 across both models suggest that, for this health status category, increasing age is associated with an improvement in health status or a higher likelihood of being in this class. However, the generally small coefficients (close to zero) across all classes indicate that the effect of age on health status is relatively weak.

## Answer to the question

Based on the analysis, the causal LASSO results suggest that age has little to no causal effect on health status for most categories, with the exception of a small positive effect for the "BAD" health status category. This finding implies that age alone does not significantly determine health status, and other factors may play a more crucial role. The discrepancies between the naive and causal LASSO results highlight the importance of accounting for confounding variables to obtain accurate estimates of causal effects.

Specifically, the naive LASSO shows a negative coefficient for the "VERY GOOD" health status category, indicating that an increase in age is associated with a lower likelihood of being in this category. However, this effect disappears in the causal model, suggesting that the naive estimate was confounded by other variables. Similarly, the naive LASSO's stronger negative coefficient for the "GOOD" health status category is also reduced in the causal model, indicating that the observed relationship in the naive model was partly due to confounding.

The consistent null effect for the “SATISFACTORY” health status category in both the naive and causal models indicates that age does not have a significant impact on the likelihood of being in this health status category, irrespective of confounders. For the “NOT THAT GOOD” health status category, the naive LASSO suggests a slight positive effect of age, but this effect is not present in the causal model, further underscoring the role of confounding variables in the naive estimate.

The small positive effect observed in the causal LASSO for the “BAD” health status category, with a coefficient of 0.009641957, suggests that as age increases, the likelihood of being in the “BAD” health status category slightly increases. This effect is also seen in the naive LASSO, although it is slightly larger, indicating that while confounding is present, it does not significantly alter the observed effect for this category.

In summary, the analysis underscores that age alone does not have a substantial causal impact on most health status categories, except for a slight positive influence on the “BAD” health status. This highlights the nuanced role of age in determining health outcomes and emphasizes the importance of controlling for confounding variables to accurately assess causal relationships. The findings suggest that other factors beyond age may play a more significant role in influencing health status, and future research should focus on identifying and accounting for these factors to better understand the determinants of health outcomes.