Subject: Simpson's Paradox

About the Analysis:

In 1972-1974, in Whickham, a town in the north-east of England, located approximately 6.5 kilometers south-west of Newcastle upon Tyne, a survey of one-sixth of the electorate was conducted in order to inform work on thyroid and heart disease (Tunbridge et al. 1977). A continuation of this study was carried out twenty years later. (Vanderpump et al. 1995). Some of the results were related to smoking and whether individuals were still alive at the time of the second study. For the purpose of simplicity, we will restrict the data to women and among those to the 1314 who were categorized as "smoking currently" or "never smoked". There were relatively few women in the initial survey who smoked but have since quit (162) and very few for whom information was not available (18). Survival at 20 years was determined for all women of the first survey.

All these data are available in Subject6_smoking csv file . You will find on each line if the person smokes or not, whether alive or dead at the time of the second study, and his age at the time of the first survey.

Question 1:

Tabulate the total number of women alive and dead over the period according to their smoking habits. Calculate in each group (smoking/non-smoking) the mortality rate (the ratio of the number of women who died in a group to the total number of women in that group). You can graph these data and calculate confidence intervals if you wish. Why is this result surprising?

Answer 1:

Let's read the data and examine the mortality rates in each group, without considering age. This overview will provide us with a broad understanding of the mortality patterns across different groups.

```r
# Load necessary 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
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```r
library(ggplot2)

#The dataset used in my analysis comprises three columns: Age, Status, and Smoker,

data <- read.csv("./data_smoker_women/Subject6_smoking.csv")

summary(data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.30   44.80   47.36   60.60   89.90
```

```r
summary(data$Status)
```

```
##    Length     Class      Mode
##      1314 character character
```

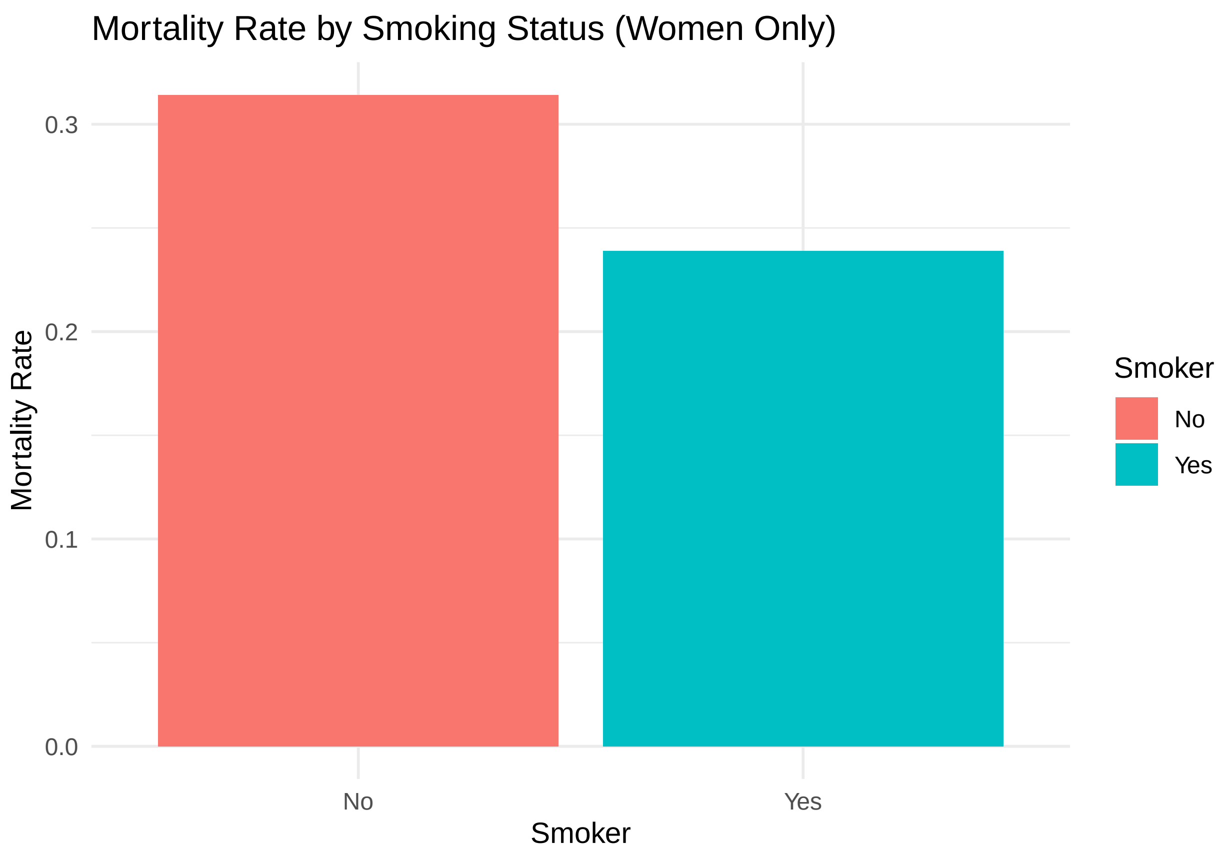
```r
summary(data$Smoker)
```

```
##    Length     Class      Mode
##      1314 character character
```

```r
# Create a contingency table
table <- data %>%
  group_by(Smoker, Status) %>%
  summarize(Count = n(), .groups = "drop") %>%
  pivot_wider(names_from = Status, values_from = Count)

# Calculate mortality rate
table <- table %>%
  mutate(Mortality_Rate = Dead / (Dead + Alive))

# Print table and mortality rates
print(table)
```

```
## # A tibble: 2 x 4
##   Smoker Alive  Dead Mortality_Rate
##   <chr>  <int> <int>          <dbl>
## 1 No       502   230          0.314
## 2 Yes      443   139          0.239
```

```r
# Create a bar plot
ggplot(data = table, aes(x = Smoker, y = Mortality_Rate, fill = Smoker)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Mortality Rate by Smoking Status (Women Only)",
       x = "Smoker",
       y = "Mortality Rate") +
  theme_minimal()
```



The table and the bar chart show the counts of women categorized by smoking habits ('No' or 'Yes'), along with the number of women alive, dead, and the calculated mortality rate in each group. The mortality rate is the ratio of the number of women who died to the total number of women in each group.

Although I observe a higher mortality rate in the non-smoker group compared to the smoker group, it's essential to consider the potential influence of age on these results. I will investigate the distribution of women in each age group for both smokers and non-smokers. This will help us assess if there is any bias in the data due to variations in age, ensuring a more comprehensive understanding of the observed mortality rates.

Perhaps, some of these women might have started smoking recently also mbe some of them dead in mid ages like 40, and this factor could influence the results. Additionally, there are a few women for whom we lack information , also as I said before we take the moratality rate for all ages,that not accurate because there is bias , some ages have already alot of dead people without smoking habits consideration ,Also the number of people in each age group have different counts.

Question 2:

Go back to question 1 (numbers and mortality rates) and add a new category related to the age group. For example, the following classes will be considered: 18-34 years, 34-54 years, 55-64 years, over 65 years.

Why is this result surprising? Can you explain this paradox? Similarly, you may wish to provide a graphical representation of the data to support your explanations.

Answer 2:

Begin by eliminating the missing values and categorizing individuals into age groups.

```
# Remove rows with missing values
data <- na.omit(data)
# Create age groups
data$Age_Group <- cut(data$Age, breaks = c(18, 34, 54, 64, 90),
                      labels = c("18-34", "35-54", "55-64", "65-90"))
# Count the number of women in each age group based on smoking status
count_by_age_group <- data %>%
  group_by(Smoker, Age_Group) %>%
  summarize(Count = n(), .groups = "drop")
print(count_by_age_group)
```

```
## # A tibble: 10 x 3
##    Smoker Age_Group Count
##    <chr>  <fct>     <int>
##  1 No     18-34       218
##  2 No     35-54       199
##  3 No     55-64       121
##  4 No     65-90       193
##  5 No     <NA>          1
##  6 Yes    18-34       177
##  7 Yes    35-54       237
##  8 Yes    55-64       115
##  9 Yes    65-90        49
## 10 Yes    <NA>          4
```

The table provides the count of women in each age group, allowing us to observe the distribution across different age categories.

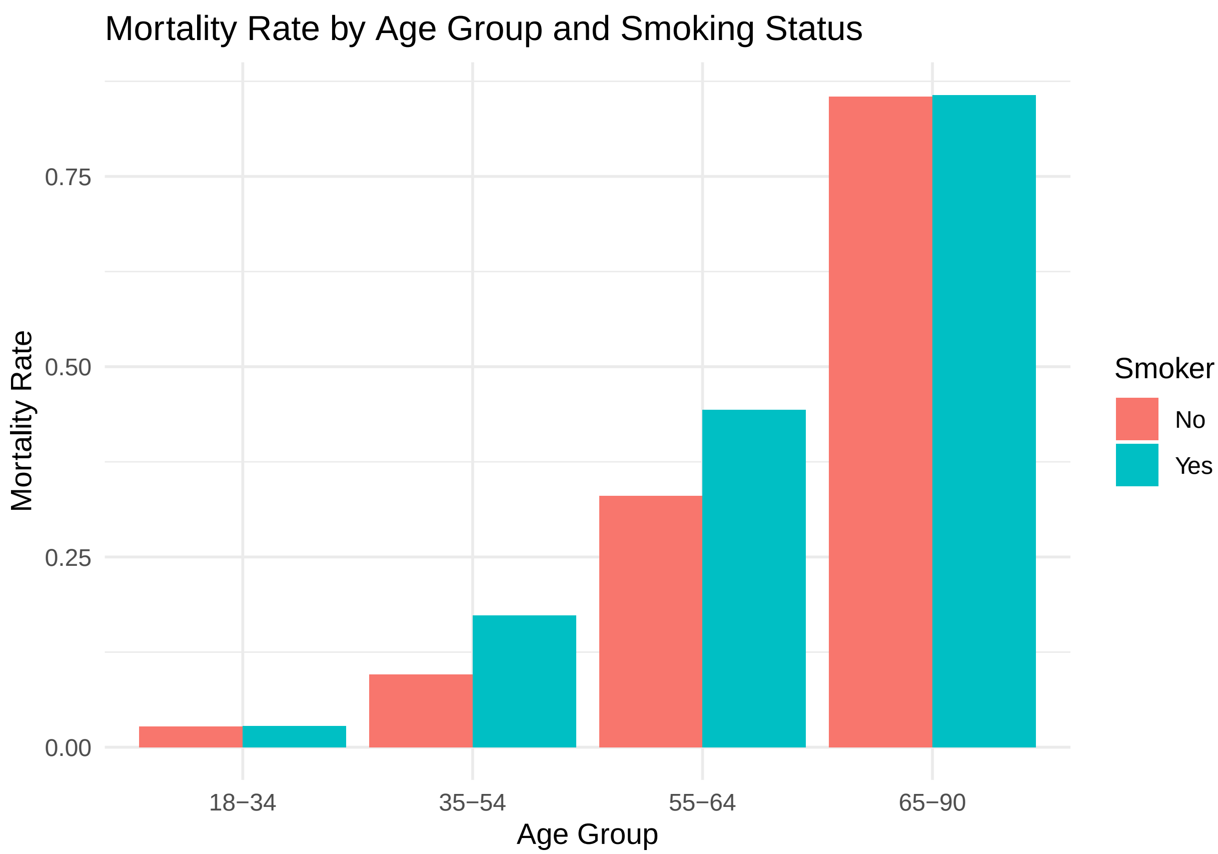
```
# Remove rows with missing values in the 'Age' variable
data <- na.omit(data)

# Create age groups
data$Age_Group <- cut(data$Age, breaks = c(18, 34, 54, 64, 90),
                      labels = c("18-34", "35-54", "55-64", "65-90"))

# Count the number of women in each age group based on smoking status
count_by_age_group <- data %>%
  group_by(Smoker, Age_Group) %>%
  summarize(
    Count = n(),
    Dead = sum(Status == "Dead"),
    .groups = "drop"
  )

# Calculate the mortality rate in each age group
mortality_by_age_group <- data %>%
  group_by(Smoker, Age_Group) %>%
```

```r
  summarize(
    Count = n(),
    Dead = sum(Status == "Dead"),
    Mortality_Rate = Dead / n(),
    .groups = "drop"
  ) %>%
  ungroup()

print(mortality_by_age_group)
```

```
## # A tibble: 8 x 5
##   Smoker Age_Group Count  Dead Mortality_Rate
##   <chr>  <fct>     <int> <int>          <dbl>
## 1 No     18-34       218     6         0.0275
## 2 No     35-54       199    19         0.0955
## 3 No     55-64       121    40         0.331
## 4 No     65-90       193   165         0.855
## 5 Yes    18-34       177     5         0.0282
## 6 Yes    35-54       237    41         0.173
## 7 Yes    55-64       115    51         0.443
## 8 Yes    65-90        49    42         0.857
```

Now, examining the Mortality Rate in each group, I observe higher rates in certain age ranges, such as 55-64, within the smokers' group. However, it's essential to note that the number of measurements is limited, especially in the 65-90 range, where there are only 49 women who smoke compared to 193 non-smokers ,

```r
library(ggplot2)


# Create a bar plot
ggplot(mortality_by_age_group, aes(x = Age_Group, y = Mortality_Rate, fill = Smoker)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Mortality Rate by Age Group and Smoking Status",
       x = "Age Group",
       y = "Mortality Rate",
       fill = "Smoker") +
  theme_minimal()
```



The bar chart compares the mortality Rate for each group , we can see how the smoking ipact the mortality rate in the two groups , 35-54 and 55-64 , but fro younger people not that much , for older whether they are smokers or not they have a high morailty rate.

Questions 3:

In order to avoid a bias induced by arbitrary and non-regular age groupings, it is possible to try to perform a logistic regression. If we introduce a Deathvariable of 1or 0to indicate whether the individual died during the 20-year period, we can study the Death ~ Agemodel to study the probability of death as a function of age according to whether one considers the group of smokers or non-smokers. Do these regressions allow you to conclude or not on the harmfulness of smoking? You will be able to propose a graphical representation of these regressions (without omitting the regions of confidence).

Answer 3:

For the answer we will start by creating two datasets one for smokers and one for nonsmokers, and creat the column death depend when the status is alive we put 0 and when the status is dead we put 1 in the column Death.

```
# we will see  how the probability of death changes with age.

data$Death <- as.numeric(data$Status == "Dead")  # Convert 'Status' to a binary 'Death' variable
```

```r
data$Smoker <- as.factor(data$Smoker)  # Ensure 'Smoker' is a factor

# spliit the data between smokers and not smokers.
smokers_data <- subset(data, Smoker == "Yes")

nonsmokers_data <- subset(data, Smoker == "No")
```

```r
# Logistic regression for smokers
model_smokers <- glm(Death ~ Age, family = binomial, data = smokers_data)
summary(model_smokers)
```

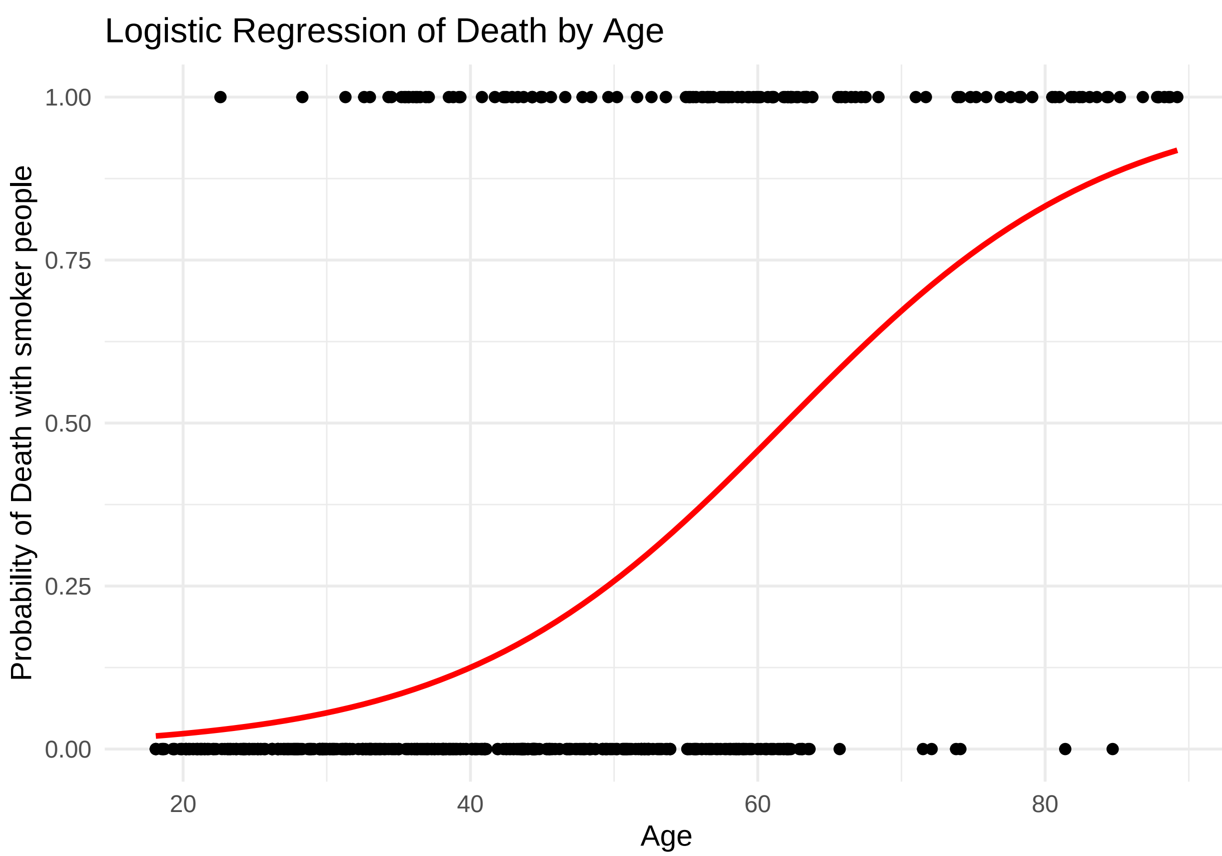
```
##
## Call:
## glm(formula = Death ~ Age, family = binomial, data = smokers_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.496493   0.467120  -11.77   <2e-16 ***
## Age          0.088772   0.008735   10.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 637.69  on 577  degrees of freedom
## Residual deviance: 480.25  on 576  degrees of freedom
## AIC: 484.25
##
## Number of Fisher Scoring iterations: 5
```

```r
# Logistic regression for non-smokers
model_nonsmokers <- glm(Death ~ Age, family = binomial, data = nonsmokers_data)
summary(model_nonsmokers)
```

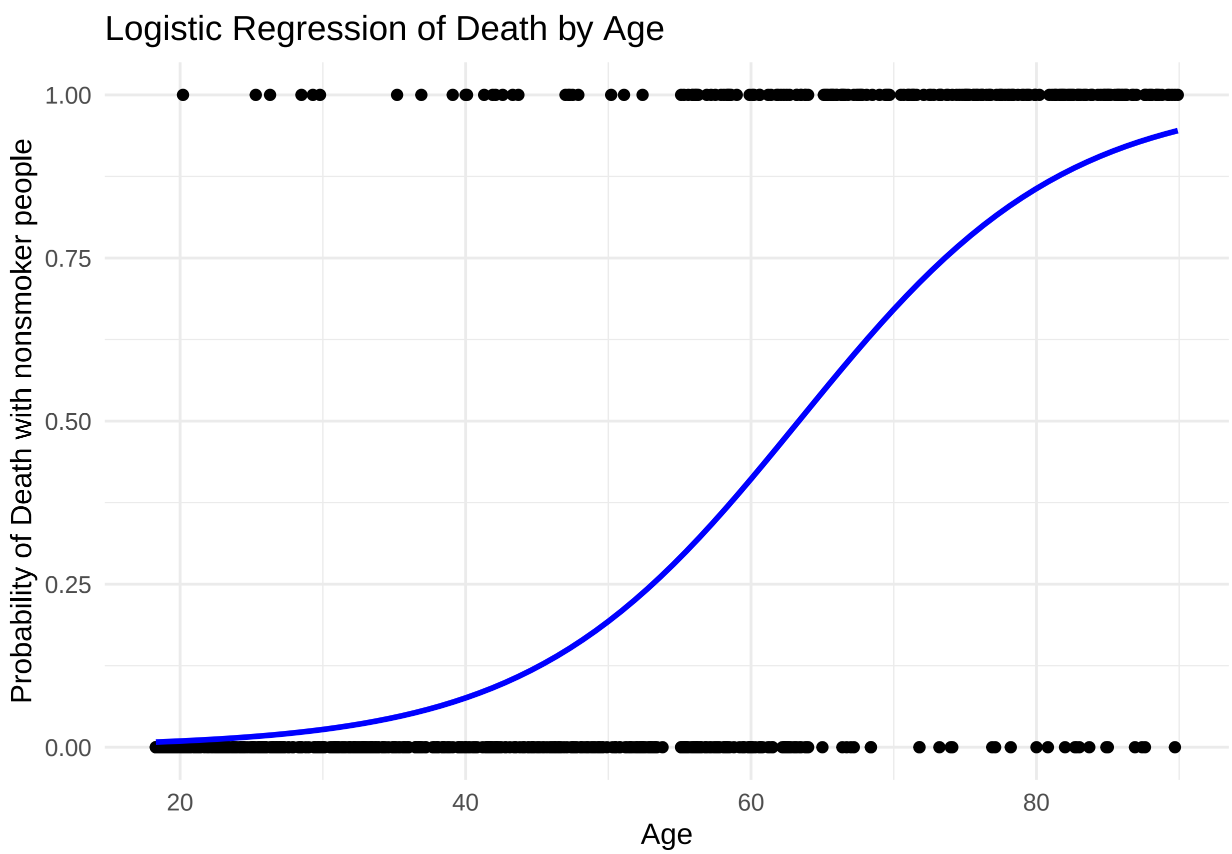
```
##
## Call:
## glm(formula = Death ~ Age, family = binomial, data = nonsmokers_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.794248   0.479545  -14.17   <2e-16 ***
## Age          0.107256   0.007808   13.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 910.48  on 730  degrees of freedom
## Residual deviance: 519.06  on 729  degrees of freedom
## AIC: 523.06
##
## Number of Fisher Scoring iterations: 6
```

we can see how the age has impact on death variable from the output above.

```
ggplot(smokers_data, aes(x = Age, y = Death)) +
  geom_point() +  # Plot the raw data points
  stat_smooth(method = "glm", method.args = list(family = "binomial"),
              formula = y ~ x,    # Logistic regression formula
              geom = "smooth",    # Add a smoothed line
              se = FALSE,         # Optionally, set to FALSE to not display the confidence interval
              color = "red") +  # Color of the smoothed line
  labs(x = "Age", y = "Probability of Death with smoker people",
       title = "Logistic Regression of Death by Age") +
  theme_minimal()
```



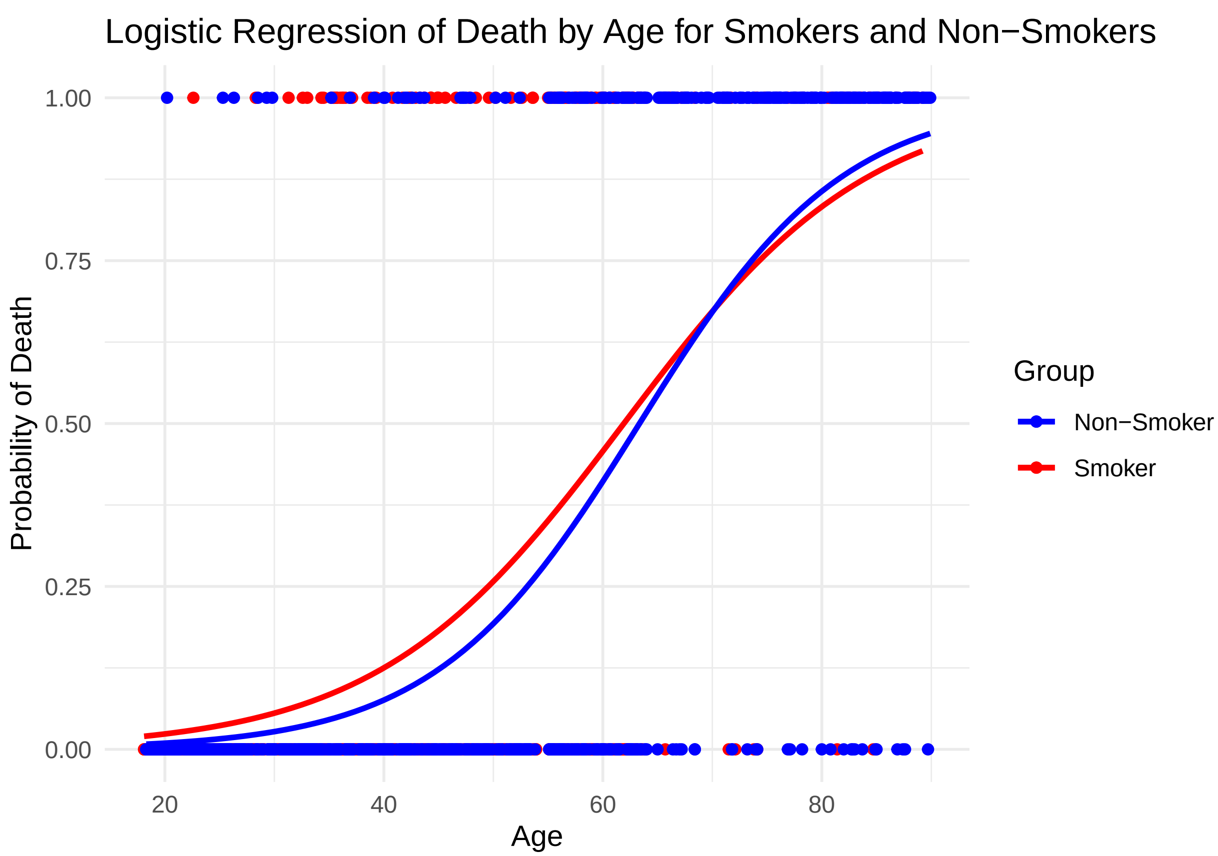

```
ggplot(nonsmokers_data, aes(x = Age, y = Death)) +
  geom_point() +  # Plot the raw data points
  stat_smooth(method = "glm", method.args = list(family = "binomial"),
              formula = y ~ x,    # Logistic regression formula
              geom = "smooth",    # Add a smoothed line
              se = FALSE,         # Optionally, set to FALSE to not display the confidence interval
              color = "blue") +  # Color of the smoothed line
  labs(x = "Age", y = "Probability of Death with nonsmoker people",
       title = "Logistic Regression of Death by Age") +
  theme_minimal()
```



now both of the models from above in the same graph , so we can compare.

```
smokers_data$Group <- 'Smoker'
nonsmokers_data$Group <- 'Non-Smoker'

library(ggplot2)

# Start with an empty ggplot object specifying the aesthetics common to both datasets
p <- ggplot() +
  labs(x = "Age", y = "Probability of Death",
       title = "Logistic Regression of Death by Age for Smokers and Non-Smokers") +
  theme_minimal()

# Add points and a smooth line for smokers, mapping 'Group' to color within aes()
p <- p + geom_point(data = smokers_data, aes(x = Age, y = Death, color = Group)) +
  stat_smooth(data = smokers_data, aes(x = Age, y = Death, color = Group),
              method = "glm", method.args = list(family = "binomial"),
              formula = y ~ x, geom = "smooth", se = FALSE)

# Add points and a smooth line for non-smokers, mapping 'Group' to color within aes()
p <- p + geom_point(data = nonsmokers_data, aes(x = Age, y = Death, color = Group)) +
  stat_smooth(data = nonsmokers_data, aes(x = Age, y = Death, color = Group),
              method = "glm", method.args = list(family = "binomial"),
              formula = y ~ x, geom = "smooth", se = FALSE)
```

```r
# Use scale_color_manual to customize the legend and colors
p <- p + scale_color_manual(name = "Group",
                            values = c("Smoker" = "red", "Non-Smoker" = "blue"),
                            labels = c("Smoker" = "Smoker", "Non-Smoker" = "Non-Smoker"))

# Print the plot
print(p)
```



we can see from model above how the age between 35-65 has a higher probabilty compare to nonsmoker people in the same age.

also we can see in general people with age 65 and more have a higher propabilty to die in smokers and nonsmokers people , we should not forget the number of people with age more than 75 is small when the status is smokers so we can see a real impact from smoking (maybe because most of them dead in earlier).

I finished my evaluation of my peers' work and they were good. I put my notes in a MOOC and now I am waiting for others to evaluate my work.