Problem Set 4

Siyu Zou

2024-03-13

```
library(medicaldata)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(splines)
load("nmes.rdata")

d <- nmes |>
  filter(lastage == 65) |>
  filter(!is.na(lastage) & !is.na(totalexp) & !is.na(eversmk)) |>
  filter(eversmk != ".") |>
  arrange(lastage) |>
  mutate(ever = eversmk)

# two-sample t-test
t_test <- t.test(totalexp~ever,data=d,var.equal=TRUE)
t_test
```

```
##
##  Two Sample t-test
##
## data:  totalexp by ever
## t = -2.0937, df = 303, p-value = 0.03712
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4348.020  -134.757
## sample estimates:
## mean in group 0 mean in group 1
##        2092.803        4334.192
```

```
# analysis of variance
aov_summary <- summary(aov(totalexp ~ ever, data = d))
# simple linear regression
slm <- lm(totalexp ~ ever, data = d)
```

```
summary(slm)
```

```
##
## Call:
## lm(formula = totalexp ~ ever, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4334  -3629  -1885   -723 108723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2092.8      782.6   2.674   0.0079 **
## ever1         2241.4     1070.5   2.094   0.0371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9326 on 303 degrees of freedom
## Multiple R-squared:  0.01426,    Adjusted R-squared:  0.01101
## F-statistic: 4.384 on 1 and 303 DF,  p-value: 0.03712
```

```
data1 <- nmes |>
  filter(lastage >= 65) |>
  filter(!is.na(lastage) & !is.na(totalexp) & !is.na(eversmk)) |>
  filter(eversmk != ".") |>
  arrange(lastage)
```

Fit a MLR of expenditures on age and smoking status as:

```
data1 <- data1 |>
  mutate(
    age = lastage,
    agem65 = age - 65,
    age_sp1 = ifelse(age>=75, age -75, 0),
    age_sp2 = ifelse(age>=85, age-85, 0),
    ever = eversmk
         )

# Number of patients by ever smoker
data1 %>%
 summarise(num_smoker = n_distinct(pidx),
          mean_age = mean(age),
          sd_age = sd(age))
```

```
##   num_smoker mean_age   sd_age
## 1       4728 73.42259 6.427373
```

```
data1 %>%
 group_by(ever) %>%
 summarise(num_smoker = n_distinct(pidx),
          mean_age = mean(age),
          sd_age = sd(age))
```

```
## # A tibble: 2 x 4
##   ever  num_smoker mean_age sd_age
##   <chr>      <int>    <dbl>  <dbl>
```

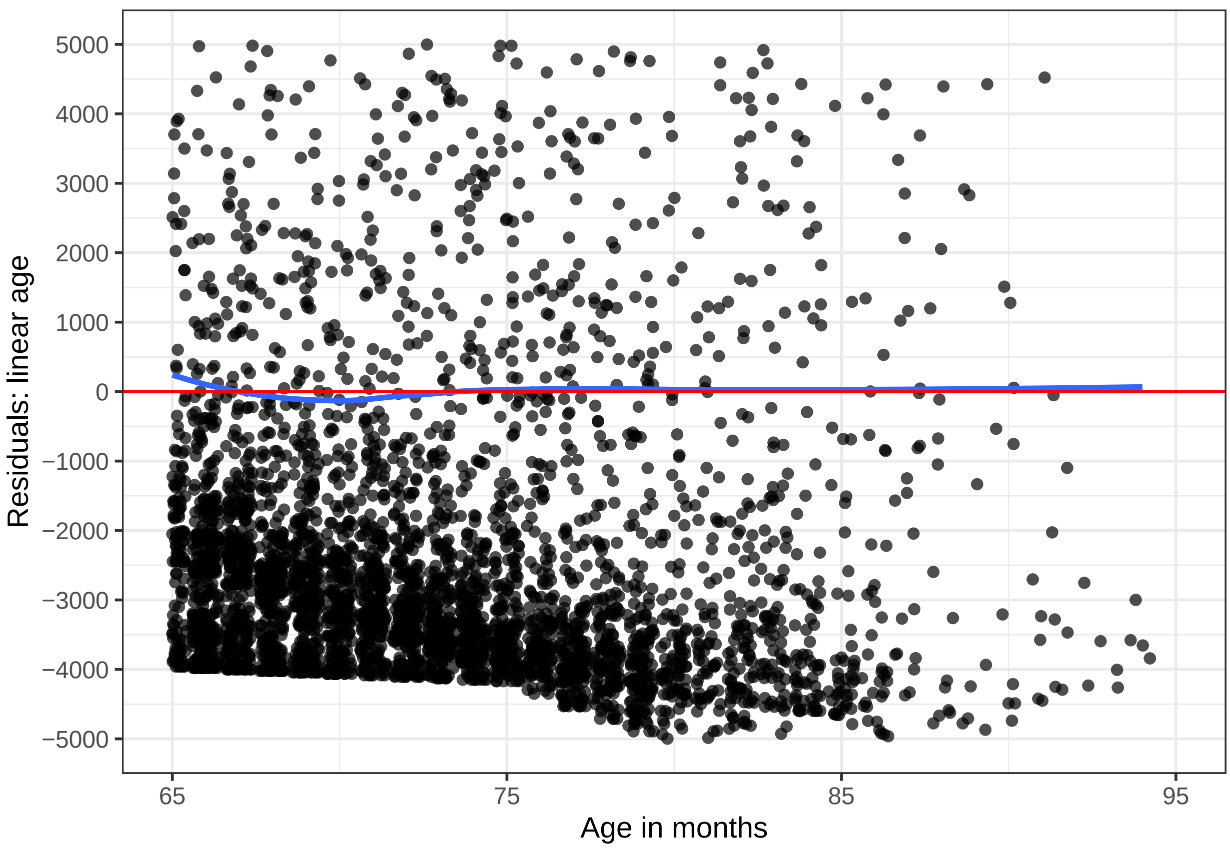
```
## 1 0            2306     74.5   6.93
## 2 1            2422     72.4   5.71
```

Q1 check the assumption

```
fit = lm(data = data1, totalexp~agem65 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp1 + age_sp2))

res.fit = lm(fit$residual~ ns(data1$age,4))
data1$residuals = residuals(fit)

#data1$residuals = residuals(model1)
ggplot(data1,aes(x=age, y= residuals)) +
    geom_jitter(alpha = 0.7) +
    theme_bw() +
    geom_smooth(aes(x = data1$age, y = res.fit$fitted.values), method = 'loess') +
    geom_hline(yintercept=0,color="red") +
    labs(y="Residuals: linear age",x="Age in months") +
    scale_y_continuous(breaks=seq(-5000,5000,1000),limits=c(-5000,5000)) +
  scale_x_continuous(breaks=seq(65,95,10),limits=c(65,95))
```

assumption E(Y |X) = X β

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 807 rows containing missing values (`geom_point()`).
```

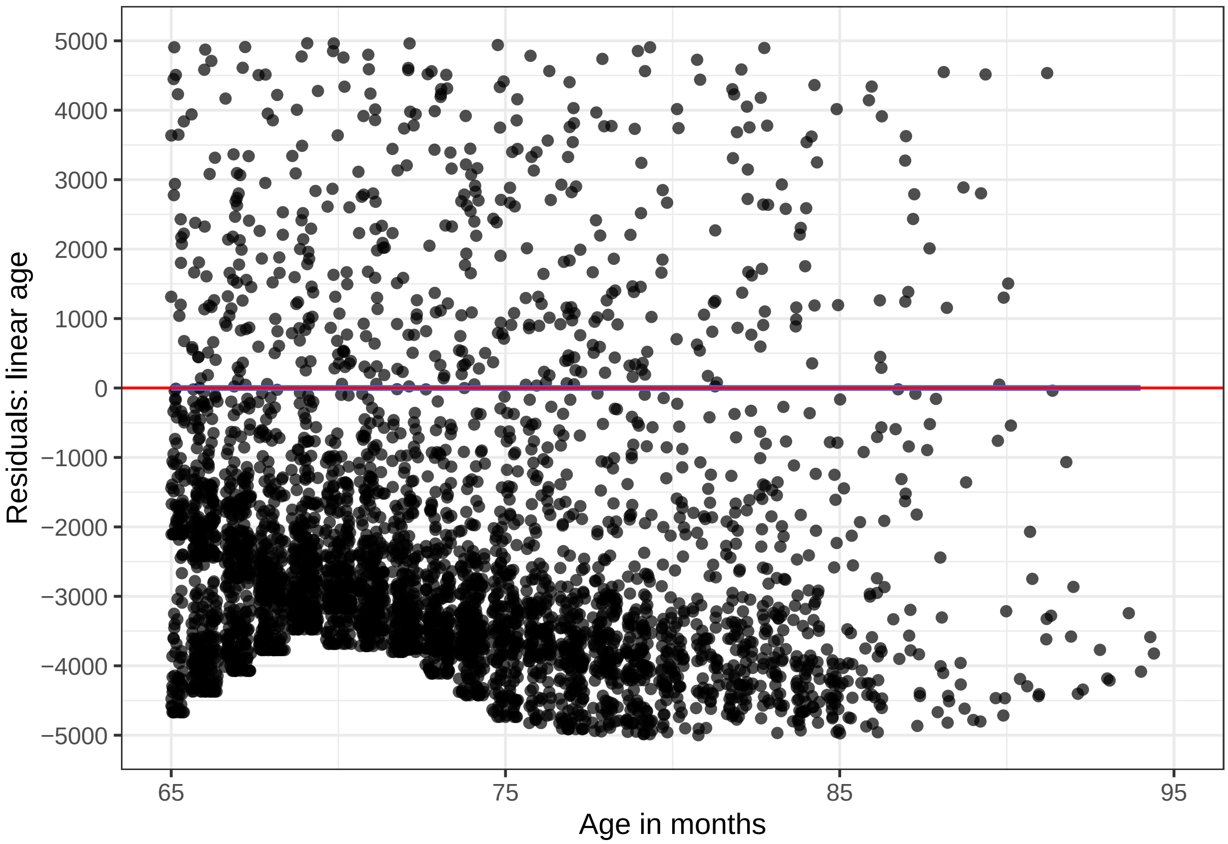



```
data1 <- data1 |>
  mutate( age_sp0 = ifelse(age>=70, age -70, 0) )
```

The residual not equal to 0 at age 65 to 70, so I try to add another knots at age 70 years in the new model.

```r
model2 <- lm(data = data1, totalexp~ agem65 + age_sp0 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_s

# check the mean model
res.fit2 = lm(model2$residual~ ns(data1$age,knots = 4))
data1$residuals2 = residuals(model2)

ggplot(data1,aes(x=age, y= residuals2)) +
    geom_jitter(alpha = 0.7) +
    theme_bw() +
    geom_smooth(aes(x = data1$age, y = res.fit2$fitted.values), method = 'loess') +
    geom_hline(yintercept=0,color="red") +
    labs(y="Residuals: linear age",x="Age in months") +
    scale_y_continuous(breaks=seq(-5000,5000,1000),limits=c(-5000,5000)) +
  scale_x_continuous(breaks=seq(65,95,10),limits=c(65,95))
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 807 rows containing missing values (`geom_point()`).
```



We could see the alternative model is more suitable with most residuals equal to 0.

Q2 Potential confounder

male: 1 – male, 0 – female RACE3: 1 – white, 2 – black, 3 – other educate: Education: 1 – college grad, 2 – some college, 3 – hs grad, 4 – other marital: 1 – married, 2 – widowed, 3 – divorced, 4 – separated, 5 – never married povstalb: Poverty status: 1 – poor, 2 – near poor, 3 – low income, 4 – middle income, 5 – high income

```r
model_adjusted <- lm(data = data1, totalexp~ agem65 + age_sp0 + age_sp1 + age_sp2 + ever + ever*(agem65

summary(model_adjusted)
```

```
##
## Call:
## lm(formula = totalexp ~ agem65 + age_sp0 + age_sp1 + age_sp2 +
##     ever + ever * (agem65 + age_sp0 + age_sp1 + age_sp2) + male +
##     RACE3 + educate + marital + povstalb, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22016  -3597  -2695   -725 170206
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    31699.70    5360.96   5.913 3.60e-09 ***
## agem65           312.32     183.85   1.699 0.089436 .
## age_sp0         -272.65     311.08  -0.876 0.380813
## age_sp1           35.75     226.34   0.158 0.874502
## age_sp2          411.07     262.12   1.568 0.116896
## ever1           2389.28     798.32   2.993 0.002778 **
## male             183.35     340.92   0.538 0.590735
## RACE3            149.52     332.47   0.450 0.652933
## educate           32.18     173.97   0.185 0.853243
## marital1       -9903.96    1490.12  -6.646 3.34e-11 ***
## marital2      -10079.98    1485.55  -6.785 1.30e-11 ***
## marital3       -8705.07    1613.00  -5.397 7.12e-08 ***
## marital4      -11374.67    1924.07  -5.912 3.62e-09 ***
## marital5      -10868.94    1631.92  -6.660 3.05e-11 ***
## povstalb1     -20313.25    5019.60  -4.047 5.28e-05 ***
## povstalb2     -18958.91    5031.25  -3.768 0.000166 ***
## povstalb3     -20020.14    5012.64  -3.994 6.60e-05 ***
## povstalb4     -20513.64    5007.99  -4.096 4.27e-05 ***
## povstalb5     -20297.86    5009.23  -4.052 5.16e-05 ***
## agem65:ever1    -596.03     244.07  -2.442 0.014642 *
## age_sp0:ever1    852.32     417.42   2.042 0.041217 *
## age_sp1:ever1   -254.46     322.72  -0.788 0.430454
## age_sp2:ever1   -650.34     470.28  -1.383 0.166761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9974 on 4705 degrees of freedom
## Multiple R-squared:  0.0253, Adjusted R-squared:  0.02074
## F-statistic:  5.55 on 22 and 4705 DF,  p-value: 1.243e-15
```

Q3 unadjusted and adjusted difference

We choose the bootstrap procedure to estimate the unadjusted and adjusted differences in average medical expenditures between ever and never smokers as a function of age, with correspnding standard errors and confidence intervals. But we also calculate the model-based standard errors.

model-based standard error

```r
model2 <- lm(data = data1, totalexp~ agem65 + age_sp0 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp

summary(model2)
```

```
##
## Call:
## lm(formula = totalexp ~ agem65 + age_sp0 + age_sp1 + age_sp2 +
##     ever + ever * (agem65 + age_sp0 + age_sp1 + age_sp2), data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9777  -3697  -2844   -872 171323
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2106.76     599.44   3.515 0.000445 ***
## agem65          315.50     184.76   1.708 0.087763 .
## age_sp0        -283.84     312.97  -0.907 0.364499
## age_sp1          59.71     227.45   0.263 0.792944
## age_sp2         496.27     262.89   1.888 0.059126 .
## ever1          2562.05     796.71   3.216 0.001310 **
## agem65:ever1   -613.99     245.36  -2.502 0.012367 *
## age_sp0:ever1   894.17     419.66   2.131 0.033167 *
## age_sp1:ever1  -283.11     324.35  -0.873 0.382795
## age_sp2:ever1  -755.30     471.99  -1.600 0.109613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10040 on 4718 degrees of freedom
## Multiple R-squared:  0.009496,   Adjusted R-squared:  0.007606
## F-statistic: 5.026 on 9 and 4718 DF,  p-value: 9.036e-07
```

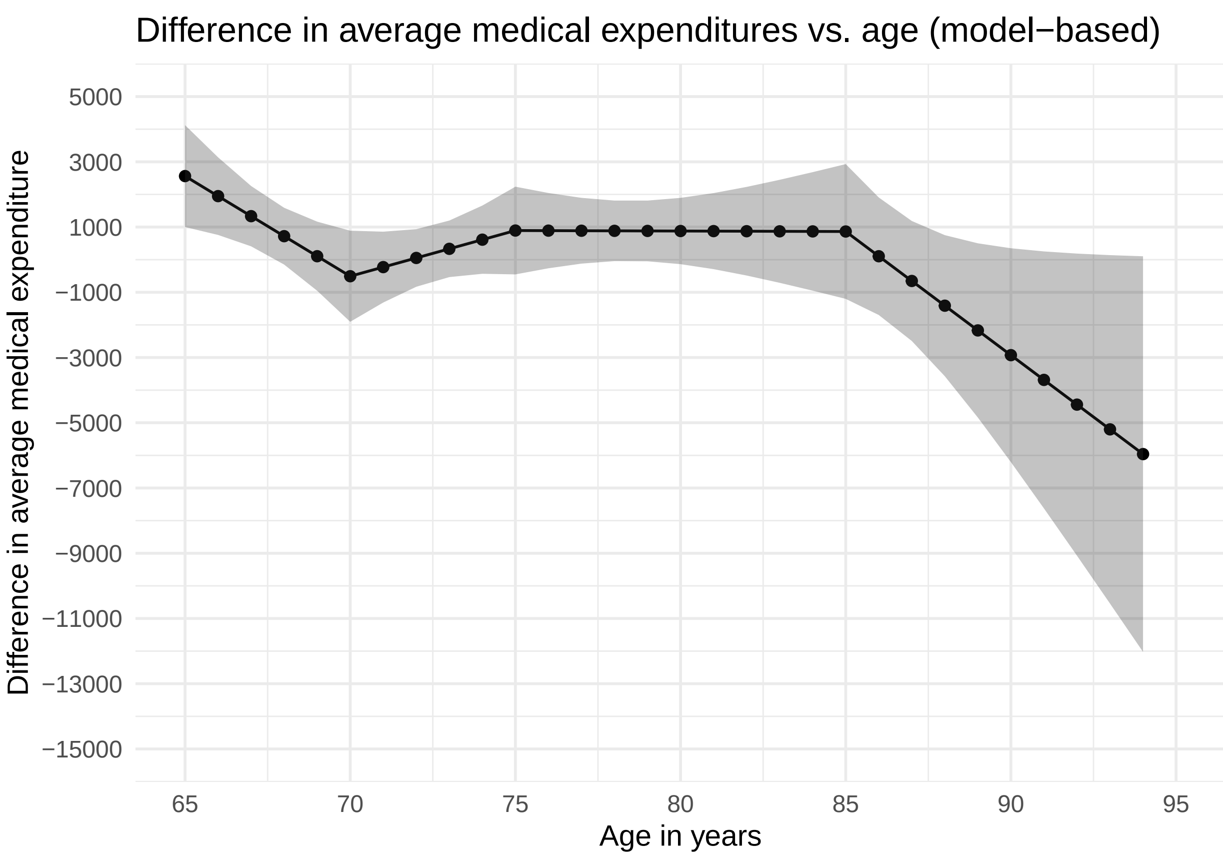
```r
coef_unadjusted <- model2$coefficients

# Function to calculate the difference in expenditures between ever and never smokers
expenditure_difference <- function(age) {
  agem65 <- age - 65
  age_sp0 <- ifelse(age >= 70, age - 70, 0)
  age_sp1 <- ifelse(age >= 75, age - 75, 0)
  age_sp2 <- ifelse(age >= 85, age - 85, 0)
  coef_unadjusted["ever1"] + coef_unadjusted["agem65:ever1"] * agem65 +
  coef_unadjusted["age_sp0:ever1"] * age_sp0 + coef_unadjusted["age_sp1:ever1"] * age_sp1 +
  coef_unadjusted["age_sp2:ever1"] * age_sp2
}


linear_combination <- function(age) {
  # Adjustments for age splines
  agem65 <- age - 65
  age_sp0 <- ifelse(age >= 70, age - 70, 0)
  age_sp1 <- ifelse(age >= 75, age - 75, 0)
  age_sp2 <- ifelse(age >= 85, age - 85, 0)
```

```
  lc <- c(0, 0, 0, 0, 0, 1, agem65, age_sp0, age_sp1,age_sp2 )
  matrix(lc, nrow = length(lc), ncol = 1)
}

ages <- 65:94
differences <- numeric(length(ages))
standard_errors <- numeric(length(ages))
lower_bounds <- numeric(length(ages))
upper_bounds <- numeric(length(ages))

reg1.vc <- vcov(model2)

for (i in seq_along(ages)) {
  differences[i] <- expenditure_difference(ages[i])
  lc <- linear_combination(ages[i])
  var <- t(lc) %*% reg1.vc %*% lc
  standard_errors[i] <- sqrt(diag(var)[1]) # The diagonal contains variances for each coefficient, we t
  lower_bounds[i] <- differences[i] - 1.96 * standard_errors[i]
  upper_bounds[i] <- differences[i] + 1.96 * standard_errors[i]
}

# Combine the ages, differences, and standard errors into a data frame for easy viewing
results <- data.frame(
  Age = ages,
  ExpenditureDifference = differences,
  StandardError = standard_errors,
  Lower95CI = lower_bounds,
  Upper95CI = upper_bounds
)

# View the results
print(results)
```

```
##    Age ExpenditureDifference StandardError    Lower95CI Upper95CI
## 1   65            2562.04767      796.7133  1000.48966 4123.6057
## 2   66            1948.05395      607.0472   758.24150 3137.8664
## 3   67            1334.06023      471.8692   409.19652 2258.9239
## 4   68             720.06650      444.0889  -150.34781 1590.4808
## 5   69             106.07278      540.5270  -953.36022 1165.5058
## 6   70            -507.92094      712.4072 -1904.23913  888.3972
## 7   71            -227.74894      553.9545 -1313.49978  858.0019
## 8   72              52.42306      450.3827  -830.32707  935.1732
## 9   73             332.59507      442.1098  -533.94012 1199.1303
## 10  74             612.76707      533.5869  -433.06331 1658.5974
## 11  75             892.93907      686.0076  -451.63582 2237.5140
## 12  76             890.00407      588.8314  -264.10542 2044.1136
## 13  77             887.06906      514.3853  -121.12613 1895.2643
## 14  78             884.13406      473.5145   -43.95437 1812.2225
## 15  79             881.19906      474.9670   -49.73621 1812.1343
## 16  80             878.26405      518.3871  -137.77461 1894.3027
## 17  81             875.32905      594.6516  -290.18818 2040.8463
## 18  82             872.39405      693.0006  -485.88722 2230.6753
## 19  83             869.45904      805.3838  -709.09325 2448.0113
## 20  84             866.52404      926.7093  -949.82627 2682.8743
```

```
## 21  85             863.58904     1053.8934  -1202.04193 2929.2200
## 22  86             105.35872      919.2048  -1696.28271 1907.0002
## 23  87            -652.87160      938.4479  -2492.22950 1186.4863
## 24  88           -1411.10191     1103.5996  -3574.15721  751.9534
## 25  89           -2169.33223     1362.6007  -4840.02953  501.3651
## 26  90           -2927.56255     1672.4024  -6205.47117  350.3461
## 27  91           -3685.79286     2009.6466  -7624.70019  253.1145
## 28  92           -4444.02318     2362.6108  -9074.74041  186.6940
## 29  93           -5202.25350     2725.1938 -10543.63330  139.1263
## 30  94           -5960.48381     3094.0157 -12024.75455  103.7869
```

```r
plot_0 <- ggplot(results, aes(x = Age, y = ExpenditureDifference)) +
  geom_point() +
  geom_line(aes(y = ExpenditureDifference)) + # Plot the fitted line
  geom_ribbon(aes(ymin = Lower95CI, ymax = Upper95CI), alpha = 0.3) +
  labs(title = "Difference in average medical expenditures vs. age (model-based)",
       x = "Age in years",
       y = "Difference in average medical expenditure") +
  theme_minimal() +
  scale_y_continuous(breaks=seq(-15000,5000,2000),limits=c(-15000,5000)) +
  scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95))

plot_0
```



bootstrap standard error

Unadjusted difference:

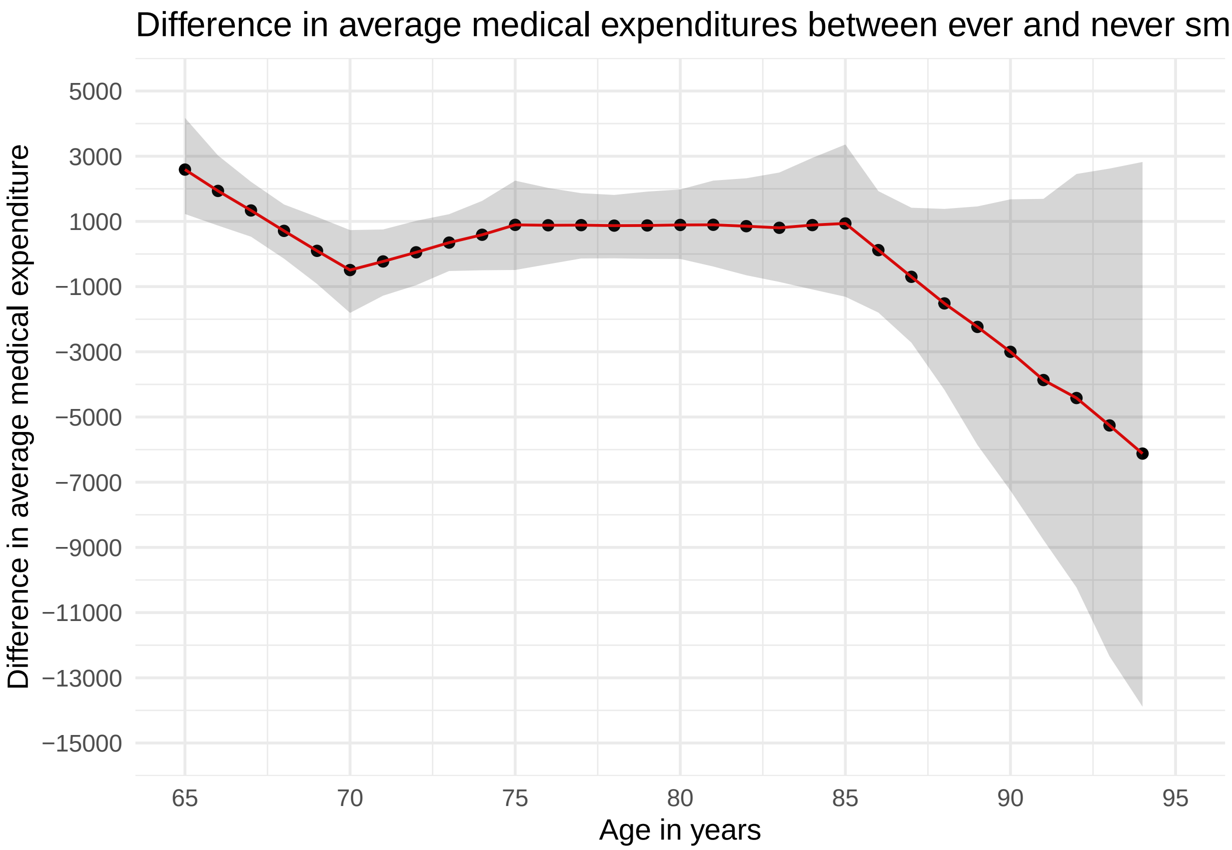
```
# Set seed
set.seed(653)
library(boot)

# Define a function to calculate the difference in expenditures
difference_calc <- function(data, indices, age) {
  # Ensure the data is correctly sampled
  resample <- data1[indices, ]

  # Calculate the age terms for the specified age
  agem65 <- age - 65
  age_sp0 <- ifelse(age >= 70, age - 70, 0)
  age_sp1 <- ifelse(age >= 75, age - 75, 0)
  age_sp2 <- ifelse(age >= 85, age - 85, 0)

  # Fit the model on the sampled data
  fit <- lm(totalexp ~ agem65 + age_sp0 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp0 + age_sp1 +

  # Calculate the difference using the model coefficients
  coef_fit <- coef(fit)
  difference <- coef_fit["ever1"] +
                coef_fit["agem65:ever1"] * agem65 +
                coef_fit["age_sp0:ever1"] * age_sp0 +
                coef_fit["age_sp1:ever1"] * age_sp1 +
                coef_fit["age_sp2:ever1"] * age_sp2

  return(difference)
}
```

```
# Perform the bootstrap for each age
results <- lapply(65:94, function(age) {
  boot(data1, difference_calc, R = 1000, age = age)
})

# Extract the bootstrap standard errors and confidence intervals
bootstrap_results <- sapply(results, function(b) {
  se <- boot.ci(b, type = "perc")
  return(c(Estimate = mean(b$t), SE = sd(b$t), CI_lower = se$percent[4], CI_upper = se$percent[5]))
})

# Combine the results into a data frame
bootstrap_results_df <- as.data.frame(t(bootstrap_results))
names(bootstrap_results_df) <- c("Estimate", "SE", "CI_lower", "CI_upper")
row.names(bootstrap_results_df) <- paste("Age in years", 65:94)

# Print the results
bootstrap_results_df$age = c(65:94)
print(bootstrap_results_df)
```

```
##                    Estimate        SE    CI_lower  CI_upper age
## Age in years 65  2589.61850  730.0634   1227.2971 4176.0328  65
```

```
## Age in years 66  1935.15497  551.6222    872.7999 3022.2633  66
## Age in years 67  1335.49412  422.0613    531.0154 2215.3409  67
## Age in years 68   709.55643  410.3522   -142.9970 1518.6287  68
## Age in years 69    97.24267  527.6316   -918.9574 1135.9623  69
## Age in years 70  -492.24470  669.9154  -1805.9768  733.4786  70
## Age in years 71  -227.96737  521.5171  -1278.7091  752.2582  71
## Age in years 72    51.51105  462.2794   -956.8276 1016.4556  72
## Age in years 73   349.63834  454.8331   -522.1453 1221.0862  73
## Age in years 74   589.85105  530.1283   -499.2269 1629.0927  74
## Age in years 75   893.76067  702.7589   -490.4021 2249.5130  75
## Age in years 76   881.63762  609.0494   -313.5264 2026.7851  76
## Age in years 77   885.57347  515.3403   -137.5138 1865.4776  77
## Age in years 78   870.58176  489.3207   -132.2836 1811.9214  78
## Age in years 79   875.44033  514.2259   -148.3853 1913.8509  79
## Age in years 80   891.43686  545.1850   -151.5206 1979.4165  80
## Age in years 81   895.92793  659.7519   -384.0267 2251.2258  81
## Age in years 82   853.31928  755.5422   -652.0150 2323.5445  82
## Age in years 83   803.56371  860.7304   -856.6012 2499.1006  83
## Age in years 84   887.35063 1018.4518  -1086.0398 2947.8056  84
## Age in years 85   937.30843 1205.9243  -1311.5194 3358.8116  85
## Age in years 86   119.33005  990.6222  -1797.1145 1928.5821  86
## Age in years 87  -700.14902 1049.4144  -2719.6537 1418.9740  87
## Age in years 88 -1515.67482 1407.1260  -4166.6433 1383.6281  88
## Age in years 89 -2236.58261 1800.6345  -5859.4930 1459.8533  89
## Age in years 90 -3000.82175 2272.4840  -7258.6979 1676.8289  90
## Age in years 91 -3866.38158 2645.0330  -8773.5173 1691.3379  91
## Age in years 92 -4417.64977 3200.4014 -10229.9130 2454.9698  92
## Age in years 93 -5259.03824 3864.4680 -12342.0470 2621.2244  93
## Age in years 94 -6122.33819 4259.7935 -13887.4795 2822.9851  94
```

```r
plot_1 <- ggplot(bootstrap_results_df, aes(x = age, y = Estimate)) +
  geom_point() +
  geom_line(color="red") +
  geom_ribbon(aes(ymin = CI_lower, ymax = CI_upper), alpha = 0.2) +
  labs(title = "Difference in average medical expenditures between ever and never smokers across differe
       x = "Age in years",
       y = "Difference in average medical expenditure") +
  theme_minimal() +
  scale_y_continuous(breaks=seq(-15000,5000,2000),limits=c(-15000,5000)) +
  scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95))

plot_1
```



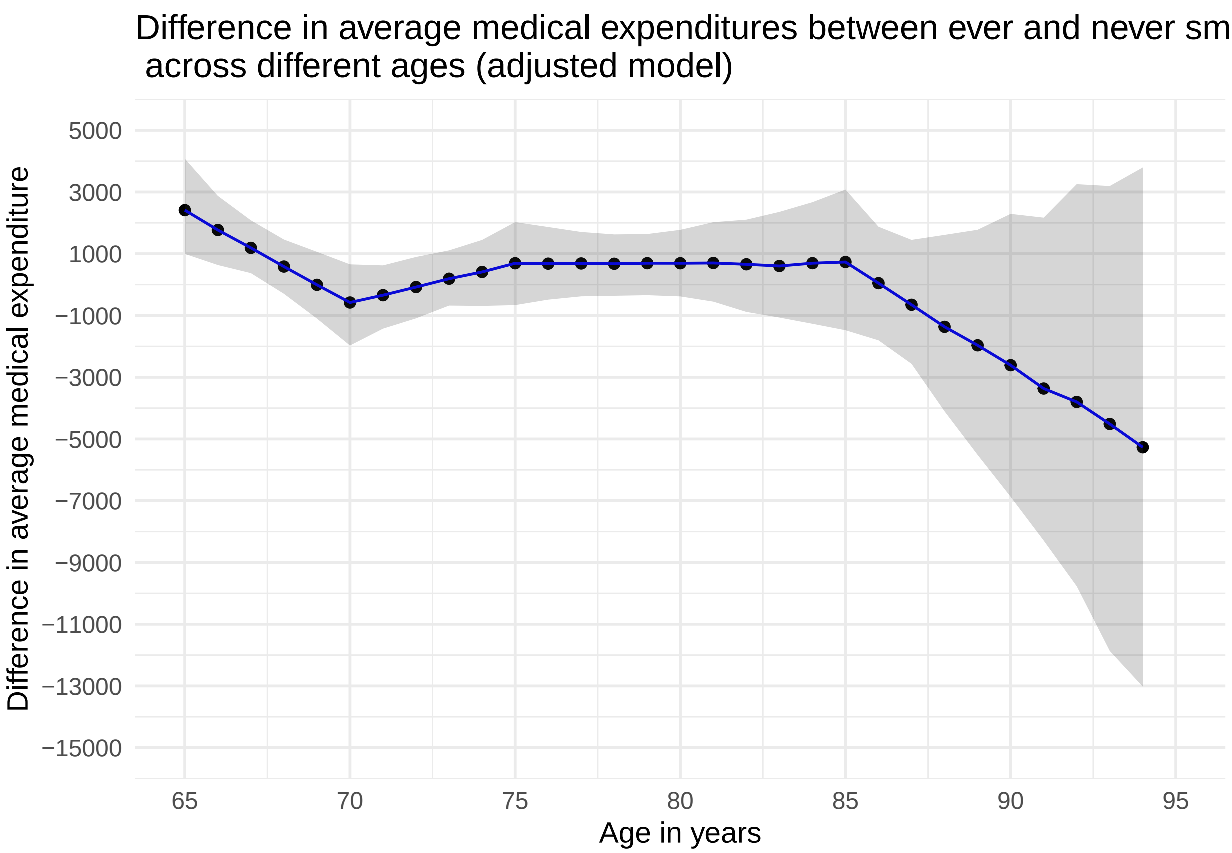
Adjusted difference:

```
# Set seed for reproducibility
set.seed(653)
# Adjusted function to calculate the difference in expenditures
difference_calc_adjusted <- function(data, indices, age) {
  # Ensure the data is correctly sampled
  resample <- data[indices, ]

  # Calculate the age terms for the specified age
  agem65 <- age - 65
  age_sp0 <- ifelse(age >= 70, age - 70, 0)
  age_sp1 <- ifelse(age >= 75, age - 75, 0)
  age_sp2 <- ifelse(age >= 85, age - 85, 0)

  # Fit the adjusted model on the sampled data
  fit_adjusted <- lm(totalexp ~ agem65 + age_sp0 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp0 + a

  # Calculate the difference using the model coefficients
  coef_fit_adjusted <- coef(fit_adjusted)
  difference <- coef_fit_adjusted["ever1"] +
                coef_fit_adjusted["agem65:ever1"] * agem65 +
                coef_fit_adjusted["age_sp0:ever1"] * age_sp0 +
                coef_fit_adjusted["age_sp1:ever1"] * age_sp1 +
                coef_fit_adjusted["age_sp2:ever1"] * age_sp2
```

```r
  return(difference)
}

# Perform the bootstrap for each age from 65 to 94
results_adjusted <- lapply(65:94, function(age) {
  boot(data1, difference_calc_adjusted, R = 1000, age = age)
})

# Extract the bootstrap standard errors and confidence intervals
bootstrap_results_adjusted <- sapply(results_adjusted, function(b) {
  se <- boot.ci(b, type = "perc")
  return(c(Estimate = mean(b$t), SE = sd(b$t), CI_lower = se$percent[4], CI_upper = se$percent[5]))
})

# Combine the results into a data frame
bootstrap_results_df_adjusted <- as.data.frame(t(bootstrap_results_adjusted))
names(bootstrap_results_df_adjusted) <- c("Estimate", "SE", "CI_lower", "CI_upper")
row.names(bootstrap_results_df_adjusted) <- paste("Age", 65:94)

# Print the results
bootstrap_results_df_adjusted$age = c(65:94)
print(bootstrap_results_df_adjusted)
```

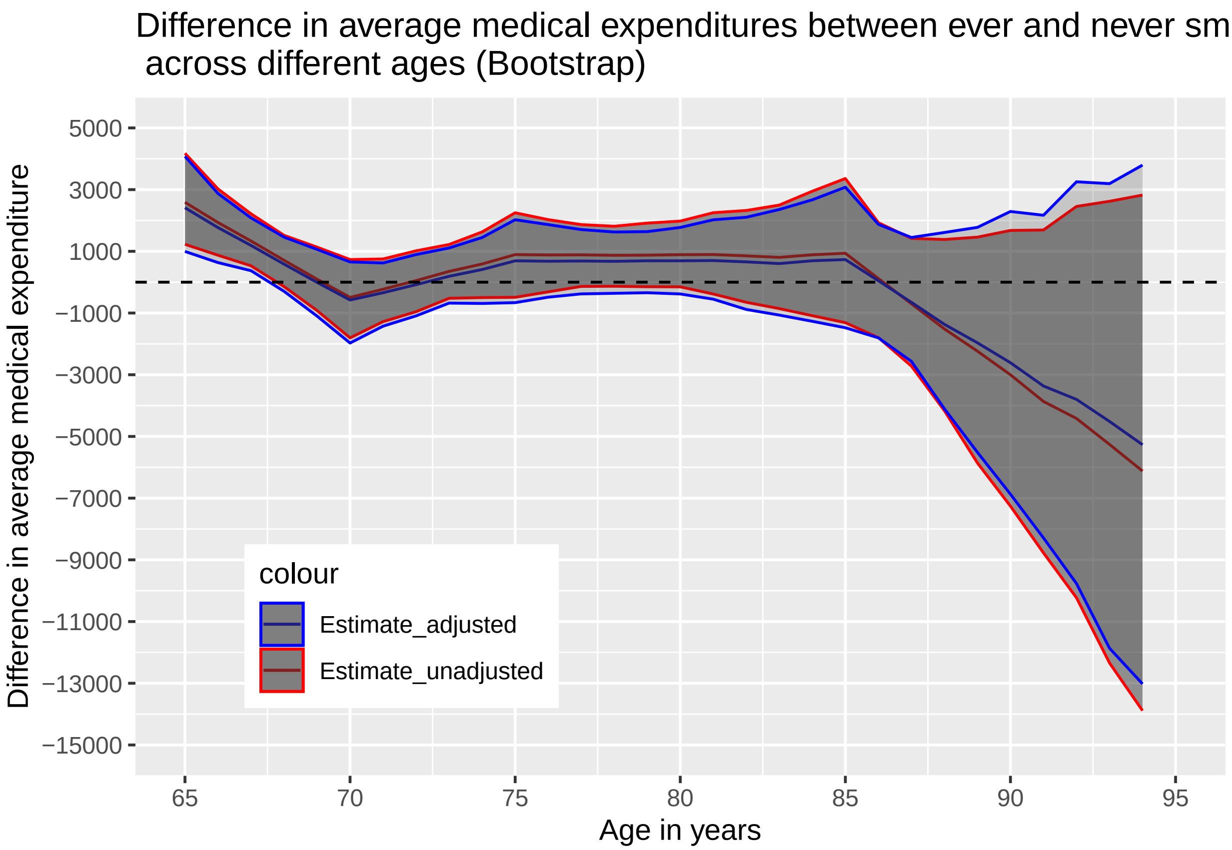
```
##            Estimate        SE    CI_lower  CI_upper age
## Age 65  2412.171108  745.2357    996.1473 4078.0425  65
## Age 66  1769.270787  565.3784    634.5817 2874.4490  66
## Age 67  1190.943336  436.2521    371.9999 2083.2098  67
## Age 68   584.802867  436.0324   -296.8850 1460.0753  68
## Age 69    -6.304707  549.9490  -1098.9717 1062.2998  69
## Age 70  -579.062353  679.1375  -1971.9542  656.0496  70
## Age 71  -342.529118  540.8773  -1426.4175  622.3253  71
## Age 72   -79.595146  475.2125  -1091.8301  895.3986  72
## Age 73   190.888276  469.8312   -678.1433 1108.8563  73
## Age 74   409.970030  539.3407   -689.4278 1447.4074  74
## Age 75   691.915964  705.4342   -663.5371 2024.3769  75
## Age 76   678.912142  611.4756   -486.4151 1865.5300  76
## Age 77   685.218181  535.9641   -379.4540 1705.0355  77
## Age 78   675.294965  496.0946   -361.3919 1626.1205  78
## Age 79   693.489479  511.9139   -342.0235 1638.0918  79
## Age 80   692.514030  543.0493   -381.7510 1773.6846  80
## Age 81   700.718374  649.0251   -551.5334 2021.6334  81
## Age 82   658.244073  732.6592   -884.3259 2103.8958  82
## Age 83   602.922779  853.9984  -1066.8919 2355.0233  83
## Age 84   694.418624  976.0765  -1268.6112 2670.0709  84
## Age 85   732.431617 1165.8978  -1477.2989 3078.0594  85
## Age 86    46.771242  964.2744  -1804.9586 1875.0212  86
## Age 87  -652.441579 1042.1836  -2566.1364 1448.9216  87
## Age 88 -1366.975100 1435.4486  -4105.4812 1610.5467  88
## Age 89 -1965.854858 1819.3244  -5514.8644 1777.0628  89
## Age 90 -2610.419639 2292.6501  -6871.8009 2291.8198  90
## Age 91 -3365.826984 2681.1817  -8289.4612 2168.4071  91
## Age 92 -3800.316593 3231.8182  -9768.8755 3253.8374  92
## Age 93 -4513.852852 3888.8295 -11864.5478 3193.7449  93
## Age 94 -5269.003802 4317.6484 -13021.0750 3794.5718  94
```

```r
plot_2 <- ggplot(bootstrap_results_df_adjusted, aes(x = age, y = Estimate)) +
  geom_point() +
  geom_line(color="blue") + # Plot the fitted line
  geom_ribbon(aes(ymin = CI_lower, ymax = CI_upper), alpha = 0.2) +
  labs(title = "Difference in average medical expenditures between ever and never smokers \n across dif
       x = "Age in years",
       y = "Difference in average medical expenditure") +
  theme_minimal() +
  scale_y_continuous(breaks=seq(-15000,5000,2000),limits=c(-15000,5000)) +
  scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95))

plot_2
```




```r
Plot_combine <- ggplot(bootstrap_results_df_adjusted, aes(x = age, y = Estimate)) +
  geom_line(data = bootstrap_results_df, aes(x = age, y = Estimate, color="Estimate_unadjusted")) +
  geom_line(data= bootstrap_results_df_adjusted, aes(x = age, y=Estimate,color="Estimate_adjusted")) +
  labs(title = "Difference in average medical expenditures between ever smokers and never smokers",
       y = "Average difference of expenditures",
       x = "Age in years") +
  geom_ribbon(data = bootstrap_results_df, aes(ymin = CI_lower, ymax = CI_upper, color="Estimate_unadju
  geom_ribbon(data= bootstrap_results_df_adjusted, aes(ymin = CI_lower, ymax = CI_upper, color="Estimate
    labs(title = "Difference in average medical expenditures between ever and never smokers \n across di
       x = "Age in years",
       y = "Difference in average medical expenditure") +
   scale_color_manual(values = c("Estimate_unadjusted" = "red","Estimate_adjusted"="blue")) +
   scale_y_continuous(breaks=seq(-15000,5000,2000),limits=c(-15000,5000)) +
   scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95)) +
```

```r
  geom_hline(yintercept = 0, color = "black", linetype = "dashed" ) +
  theme(legend.position = c(0.1, 0.1), legend.justification = c(0, 0))

Plot_combine
```



Q4 Findings

Write up the findings with sections: objective, data, methods, results, summary as if for a health services journal.

Objective: This analysis aims to explore if the difference in average medical expenditures of patients comparing ever and never smokers changes with age.

Data: We use the 1987 National Medical Expenditure Survey (NMES) Dataset extract from Johns Hopkins Biostatistics Center. This data contains detailed information on health expenditures through the use of several component surveys.

Methods: First, we fit a multiple linear regression model for total expenditure as a linear spline function of age (knots at 70, 75 and 85 years of age), smoking status (ever smoker vs. never smoker) and the interaction of age terms and smoking status. We conduct the analysis to check the appropriateness of the mean model. Then we implemented an adjusted model included sex (male vs. female), race (white, black, other), education (college grad, some college, hs grad, other), marital status (married, widowed, divorced, separated, never married), poverty status (poor, near poor, low income, middle income, high income) as the covariates. The standard error and 95% confidence intervals (CIs) for unadjusted and adjusted difference in average medical expenditure between ever and never smokers as a function of age, were constructed via the percentile bootstrap procedure using 1000 bootstrap samples of participants. Analyses were performed in R, version 4.3.1 (R Foundation).

Results: There are 4728 participants aged 65 to 94 included in the analysis, with mean age of 73.42 (SD = 6.43) years and half were ever smoker (51.2%) . The unadjusted regression analysis indicates that at age 65, never smokers have an estimated mean total medical expenditure of 2106.76 dollars, while ever smokers have higher medical expenditure of 4668.81 dollars. The interaction of age and smoking status are significant when adults aged 65-70 years and 70-75 years. Using percentile bootstrap procedure, the difference in expenditure between ever and never smokers at age 65 is 2589.62 dollars (95% CI 1227.30 to 4146.03). This difference declines with age when adults aged 65 to 70, as seen in the figure 1. For adults aged 70, the differences in medical expenditure between ever and never smokers are -492.24 (95%CI -1805.98 to 733.48). Then the difference increases with age when adults aged 70 to 75. For adults aged 75, the differences in medical expenditure between ever and never smokers are 893.76 (95%CI -490.40 to 2249.51). Then the difference declines with age when adults aged 75 to 94. For adults aged 85 and 94, the differences in medical expenditure between ever and never smokers are 937.31 (95%CI -1311.52 to 3358.81) and -5269.00 (95% CI -13021.07 to 3794.57) respectively. But the difference when aged 70 to 94 are not statistically significant as their confidence intervals include zero (Figure 1). In the adjusted model, the estimated average medical expenditure for ever smokers was significantly higher than never smokers by $2389.28 (SE = 798.32, p = 0.003), adjusting for sex, race, education, marital status, and poverty status. The interaction of age and smoking status remained significant when adults aged 65-70 years and 70-75 years. Using percentile bootstrap procedure, the difference in expenditure between ever and never smokers at age 65 is 2412.17 dollars (95% CI 996.15 to 4078.04). For ages 70, 75 and 85, the differences in medical expenditure between ever and never smokers are -579.06 (95%CI -1971.95 to 656.05), 691.92 (95%CI -663.54 to 2024.38), and 732.43 (95%CI -1477.30 to 3078.06) respectively, but these are not statistically significant as their confidence intervals include zero (Figure 1).

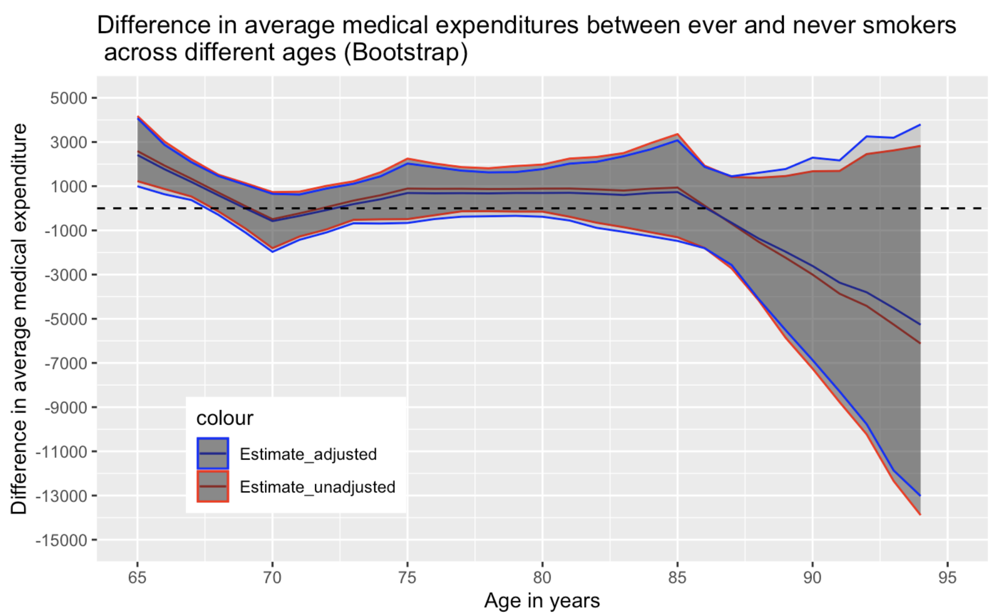


Figure 1: Difference in average medical expenditures between ever and never smokers across different ages (Bootstrap Model)

Discussion: Our analysis shows that ever smokers have higher medical expenditures than never smokers, and this difference is affected by age. The estimated difference in expenditure between ever and never smokers decreases with age when adults aged 65-70 years, increases with age when adults aged 70-75 years, and then continue decreases with age for adults aged above 75 years. After adjusting for demographic and socioeconomic factors, the difference remains significant for adults aged 65-75 years. This suggests that the impact of smoking on healthcare costs persists across different age groups. Overall, the findings underscore the health economic impact of smoking and the influence of age and sociodemographic variables on medical expenditures. The bootstrap estimates confirm our initial findings that while ever smokers tend to have higher medical expenditures at earlier ages, this difference diminishes and becomes nonsignificant as individuals age, particularly after the age of 70. The observed variability in the trend at age 70 may suggest a complex relationship between aging and healthcare costs influenced by smoking status, potentially indicating a survival bias. The bootstrap results, particularly the wide confidence intervals at older ages, highlight the increasing uncertainty about expenditure differences in this older population. Despite adjusting for some demographic and socioeconomic factors, there could be other confounders that influence both smoking status and medical expenditures that were not available in the NMES dataset. These could include environmental exposures, lifestyle factors (like diet and physical activity), and access to healthcare services. Without accounting for these factors, the estimated effect of smoking on medical expenditures might be biased.