ProblemSet2_Siyu

Siyu Zou

2024-02-20

I.Analysis of Variance

1.Link between two-sample t-test, linear regression and ANOVA.

a a.Write out the model above using matrix notation and then using matrix calculations solve for the least squares estimates of , and . What is the estimate for HINT: You will show that the model above is the same as conducting a two-sample t-test, assuming the same variance in the intervention and placebo groups. The estimate of the intercept should be the sample mean in the placebo arm, the estimate of the slope should be the difference in the sample means comparing the intervention and control groups and the.

b b.Fill in the ANOVA table for the two-sample t-test. Write the expressions for SS(Total), SS(Model) and SS(Error) using the correct combinations of and . Show that the F-statistic = (t-statistic)2

c c.Using data from the NMES, perform an analysis comparing the mean total expenditures for 65 year old ever vs. never smokers using three methods: two-sample t-test, analysis of variance and a simple linear regression model.

```
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
```

```
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
```

$$Y_i = B_0 + B_1 X_i + \varepsilon_i , \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad \begin{cases} x_i = 1 & \text{if } TRT=1 \\ x_i = 0 & \text{if } TRT=0 \end{cases}$$

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} , \quad \varepsilon \sim MVN(\underset{\sim}{0}, \sigma^2 I)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot (\beta_0, \beta_1) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1} X' Y$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ \underbrace{0 \; 0 \; 0}_{n_0} & & & \cdots & \underbrace{1 \; 1}_{n_1} \end{bmatrix}_{2\times n} \cdot \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix}_{n\times 2} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}_{2\times 2} = \begin{bmatrix} n_0+n_1 & n_1 \\ n_1 & n_1 \end{bmatrix}_{2\times 2}$$

n_0: numbe of $X_i=0$ $\qquad$ n_1: number of $X_i=1$

$$n_0 = 1 - \sum_{i=1}^n x_i \qquad n_1 = \sum_{i=1}^n x_i$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}_{2\times 2} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}_{2\times 1}$$

$$(X'X)^{-1} = \frac{1}{(n_0+n_1)n_1 - n_1^2} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n_0+n_1 \end{pmatrix}_{2\times 2}$$

$$\hat{\beta} = (X'X)^{-1} X'Y = \frac{1}{(n_0+n_1)n_1 - n_1^2} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n_0+n_1 \end{pmatrix}_{2\times 2} \begin{bmatrix} \sum_{i=1}^{n_0+n_1} y_i \\ \sum_{i=n_0+1}^{n_0+n_1} y_i \end{bmatrix}_{2\times 1}$$

$$= \frac{1}{n_0 n_1} \begin{bmatrix} n_1 \sum_{i=1}^{n_0+n_1} y_i - n_1 \sum_{i=n_0+1}^{n_0+n_1} y_i \\ -n_1 \sum_{i=1}^{n_0+n_1} y_i + (n_0+n_1) \sum_{i=n_0+1}^{n_0+n_1} y_i \end{bmatrix}_{2\times 1}$$

$$= \frac{1}{n_0 n_1} \begin{bmatrix} n_1 \sum_{i=n_1}^{n_0+n_1} y_i \\ n_0 \sum_{i=n_0+1}^{n_0+n_1} y_i - n_1 \sum_{i=1}^{n_0} y_i \end{bmatrix}_{2\times 1}$$

$$= \begin{bmatrix} \frac{1}{n_0} \sum_{i=1}^{n_0} y_i \\ \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} y_i - \frac{1}{n_0} \sum_{i=1}^{n_0} y_i \end{bmatrix}_{2\times 1}$$

Figure 1: partA-1

So $\hat{\beta}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} y_i$

$\hat{\beta}_1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} y_i - \frac{1}{n_0} \sum_{i=1}^{n_0} y_i$

$\hat{\beta} = \underbrace{(X'X)^{-1} X'}_{A} Y$

$$
\begin{aligned}
Var(\hat{\beta}) &= Var(Ay) \\
&= A \, Var(y) A' \\
&= (X'X)^{-1} X' (\sigma^2 I) X (X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1} \\
&= \sigma^2 \frac{1}{n_0 n_1} \begin{bmatrix} n_1 & -n_1 \\ -n_1 & n_0+n_1 \end{bmatrix}_{2\times 2} \\
&= \sigma^2 \begin{bmatrix} \frac{1}{n_0} & -\frac{1}{n_0} \\ -\frac{1}{n_0} & \frac{n_0+n_1}{n_0 n_1} \end{bmatrix}_{2\times 2}
\end{aligned}
$$

So $Var(\hat{\beta}_1) = \frac{n_0+n_1}{n_0 n_1} \sigma^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_0}$, $Var(\hat{\beta}_0) = \frac{\sigma^2}{n_0}$

$\sigma^2 = Var(\varepsilon_i)$

$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \mu_i(\hat{\beta}, X_i))^2 / n-(p+1)$

in this model, $p=1$. so $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}{n-2}$

Figure 2: partA-2

Source	Sums of Squares	Df	Mean Squares	F-stat
Model	$SSM = \frac{n_0 n_1}{n_0+n_1}\hat{\beta}_1^2$	$k-1=1$	$MS(Model) = SSM/(k-1) = \frac{n_0 n_1}{n_0+n_1}\hat{\beta}_1^2$	$\frac{n_0 n_1}{n_0+n_1}\hat{\beta}_1^2 / \sigma^2$
Error	$SSE = (n-2)\sigma^2$	$n-k=n-2$	$MSE = SSE/(n-k) = \sigma^2$	
Total	$\frac{n_0 n_1}{n_0+n_1}\hat{\beta}_1^2 + (n-2)\sigma^2$	$n-1$		

$$\bar{y} = \frac{\bar{y}_1 n_1 + \bar{y}_0 n_0}{n_1+n_0} = \frac{(\hat{\beta}_0+\hat{\beta}_1)n_1 + \hat{\beta}_0 n_0}{n_1+n_0}$$

$$SSM = \Sigma_{i=1}^{n_0}(\bar{y}_0-\bar{y})^2 + \Sigma_{i=1}^{n_1}(\bar{y}_1-\bar{y})^2$$

$$= \Sigma_{i=1}^{n_0}\left(\hat{\beta}_0 - \frac{(\hat{\beta}_0+\hat{\beta}_1)n_1+\hat{\beta}_0 n_0}{n_1+n_0}\right)^2 + \Sigma_{i=n_0+1}^{n_0+n_1}\left(\hat{\beta}_0+\hat{\beta}_1 - \frac{(\hat{\beta}_0+\hat{\beta}_1)n_1+\hat{\beta}_0 n_0}{n_1+n_0}\right)^2$$

$$= \Sigma_{i=1}^{n_0}\left(\frac{-\hat{\beta}_1 n_1}{n_1+n_0}\right)^2 + \Sigma_{i=n_0+1}^{n_0+n_1}\left(\frac{\hat{\beta}_1 n_0}{n_1+n_0}\right)^2$$

$$= n_0\frac{(\hat{\beta}_1 n_1)^2}{(n_1+n_0)^2} + n_1\frac{(\hat{\beta}_1 n_0)^2}{(n_1+n_0)^2}$$

$$= \frac{n_0 n_1}{n_0+n_1}\hat{\beta}_1^2$$

Figure 3: partB-1

```
##
##  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
```

```r
# 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|)
```

$$SSE = (n-2)\,MSE = (n-2)\sum_{i=1}^{n}(y_i-\bar{y})^2 = (n-2)\,\sigma^2$$

$$SSTotal = SSM + SSE = \frac{n_0 n_1}{n_0+n_1}\hat{\beta}_1 + (n-2)\,\sigma^2$$

$$F\text{-stat} = \frac{MS(Model)}{MSE} = \frac{SSM/\overset{k=2}{(k-1)}}{SSE/(n-2)} = \frac{\hat{\beta}_1^2\,\frac{n_0 n_1}{(n_1+n_0)}}{\sigma^2}$$

$$= \frac{\hat{\beta}_1^2\,\frac{1}{\frac{1}{n_0}+\frac{1}{n_1}}}{\sigma^2}$$

$$= \frac{\hat{\beta}_1^2}{\sigma^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}$$

$$\because\ \hat{\beta}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} y_i - \frac{1}{n_0}\sum_{i=1}^{n_0} y_i$$

$$E(\hat{\beta}_1) = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} E(y_i) - \frac{1}{n_0}\sum_{i=1}^{n_0} E(y_i)$$

$$= 0$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n_0} + \frac{\sigma^2}{n_1}$$

$$F\text{-stat} = \frac{\hat{\beta}_1^2}{\sigma^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)} = \left(\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\frac{\sigma^2}{n_0}+\frac{\sigma^2}{n_1}}}\right)^2 = t^2.$$

Figure 4: partB-2

```
## (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
```

iv. Compute the square root of the “mean squared error” from the analysis of variance table and compare this to the “residual standard error” output from the lm function. Are these the same or different?

```r
# square root of the "mean squared error"
sqrt(86972246)
```

```
## [1] 9325.891
```

```r
sq_mse_aov <- sqrt(aov_summary[[1]]$'Mean Sq'[2])
rse_slm = summary(slm)$sigma
sq_mse_aov
```

```
## [1] 9325.891
```

```r
rse_slm
```

```
## [1] 9325.891
```

iv. the square root of the “mean squared error” from the analysis of variance table is 9325.891. “residual standard error” output from the lm function is 9325.891. These two measures are the same. They both quantify the dispersion of the observed values around the fitted values.

v. Compare the (t-statistic)2 and F-statistics with corresponding p-values. Are these the same or different?

```r
t_stats <- t_test$statistic
t_squared <- t_stats*t_stats
print(paste("t-statistic2 from t-test:", t_squared))
```

```
## [1] "t-statistic2 from t-test: 4.38358741856612"
```

```r
f_statistic_lm <- summary(slm)$fstatistic["value"]
print(paste("F-statistic from lm:", f_statistic_lm))
```

```
## [1] "F-statistic from lm: 4.38358741856612"
```

```r
p_value_ttest <- t_test$p.value
p_value_lm <- summary(slm)$coefficients[2,4] # p-value for the slope (ever)
print(paste("p-value from t-test:", p_value_ttest))
```

```
## [1] "p-value from t-test: 0.0371174755536529"
```

```r
print(paste("p-value from lm:", p_value_lm))
```

```
## [1] "p-value from lm: 0.0371174755536532"
```

(t-statistic)2 from the two-sample t-test is 4.38358. The F-statistic from the lm is 4.38358. These two results are the same.

P-value form the two-sample t-test is 0.03711, p-value from lm: 0.03711, these two results are the same.

2.Extend the ideas above to compare the mean total expenditures for 65 year old current, former and never smokers using two methods: analysis of variance and linear regression model.

```
# reate a new variable X that is 0 = never smoker, 1 = former smoker, 2 = current smoker
d$X = ifelse(d$current=="1",2,
    ifelse(d$former=="1",1,
    ifelse(d$current=="." & d$former==".",NA,0)))
d = d[!is.na(d$X),]

# analysis of variance
summary(aov(totalexp~as.factor(X),data=d))
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## as.factor(X)    2 4.161e+08 208031361   2.351  0.097 .
## Residuals     297 2.628e+10  88474011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# linear regression
summary(lm(totalexp~as.factor(X),data=d))
```

```
##
## Call:
## lm(formula = totalexp ~ as.factor(X), data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4452  -3663  -1881   -722 108605
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2092.8      789.3   2.651  0.00845 **
## as.factor(X)1   2358.2     1210.6   1.948  0.05237 .
## as.factor(X)2   2359.6     1514.1   1.558  0.12018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9406 on 297 degrees of freedom
## Multiple R-squared:  0.01559,    Adjusted R-squared:  0.008958
## F-statistic: 2.351 on 2 and 297 DF,  p-value: 0.09701
```

iii. The linear regression model has an intercept and two slopes: β_0,β_1,β_2. Write out the definition of β_0,β_1,β_2 with respect to the group means μ_{never},μ_{former},$\mu_{current}$ Show that the null hypothesis: H__0:(μ_{never})=μ_{former}= $\mu_{current}$ is equivalent to H__0:β_1=0 and β_2=0. β_0 is the intercept, representing the mean total expenditure for the reference group, which in this case is the never smokers (μ_{never}).
β_1 represents the difference in mean total expenditure between former smokers and never smokers (μ_{former} - μ_{never}).
β_2 represents the difference in mean total expenditure between current smokers and never smokers ($\mu_{current}$ - μ_{never}).

The equivalence of the two hypotheses: if β_1=0 and β_2=0, we could get μ_{former} - μ_{never} $= 0$ and $\mu_{current}$ - μ_{never} $= 0$; it implies that there is no difference in mean total expenditures between never smokers and the other two groups (former and current smokers), which equals to the ANOVA null hypothesis of equal mfeans across groups ($\mu_{never} = \mu_{former} = \mu_{current}$).

iv. Using the F-tests, what do you conclude regarding differences in the mean total expenditures for 65 year old current, former and never smokers? The F-test shows F-statistic is 2.351 on 2 and 297 DF, and p-value equals to 0.09701. We don't reject the null hypothesis, $\mu_{never} = \mu_{former} = \mu_{current}$. So the mean total expenditures for 65 year old current, former and never smokers are the same.

II.Advanced Inferences for Linear Regression

```
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
        )

reg_1 <- lm(data = data1, totalexp~agem65 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp1 + age_sp2)
summary(reg_1)
```

```
##
## Call:
## lm(formula = totalexp ~ agem65 + age_sp1 + age_sp2 + ever + ever *
##     (agem65 + age_sp1 + age_sp2), data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9846  -3739  -2838   -882 171074
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2445.18     469.31   5.210 1.97e-07 ***
## agem65          161.72      73.38   2.204   0.0276 *
## age_sp1        -102.24     140.94  -0.725   0.4682
## age_sp2         546.81     257.02   2.127   0.0334 *
## ever1          1513.54     624.50   2.424   0.0154 *
## agem65:ever1   -140.64     100.12  -1.405   0.1602
## age_sp1:ever1   261.66     206.39   1.268   0.2049
## age_sp2:ever1  -964.30     463.21  -2.082   0.0374 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10040 on 4720 degrees of freedom
## Multiple R-squared:  0.008323,   Adjusted R-squared:  0.006852
## F-statistic: 5.659 on 7 and 4720 DF,  p-value: 1.591e-06
```

```
conf_intervals <- confint(reg_1, level=0.95)
print(conf_intervals)
```

```
##                      2.5 %     97.5 %
## (Intercept)    1525.10262 3365.25200
## agem65           17.86463  305.56559
## age_sp1        -378.53673  174.06160
## age_sp2          42.93004 1050.68155
## ever1           289.22976 2737.85040
## agem65:ever1   -336.91405   55.63338
## age_sp1:ever1  -142.96473  666.28276
## age_sp2:ever1 -1872.40197  -56.19044
```

1.Write a short, scientific interpretation of each coefficient in the model; use the estimated coefficient with corresponding confidence interval.

Intercept β_0: The estimated total medical expenditure for a adult at the age of 65 and never smoker is 2445.18 units with 95% confidence interval (1525.10, 3365.25).
β_1 (agem65): For never-smokers adults aged 65 to 75 years, the estimated expenditure increases by 161.72 units (95%CI 17.86, 305.57) for every year increases.
β_2 (age_sp1): For never smoker aged 75 to 85 years, the estimated expenditure decreases by 102.24 units (95%CI -378.54, 174.06) compared to never smoker aged 65 to 75 years.

β_3 (age_sp2): For never smoker aged 85 years and above, the estimated expenditure increases by 546.81 units (95%CI 42.93, 1050.68) compared to never smoker aged 75 to 85 years.

β_4 (ever1): Ever smokers have higher total expenditures by 1513.54 units (95%CI 289.23, 2737.85) compared to never smokers.

β_5 (agem65:ever1): The interaction term agem65:ever1 has a coefficient of -140.64 (95%CI 289.23, 55.63), which is not statistically significant ($p > 0.05$). This means we do not have enough evidence to suggest that the effect of being an ever smoker on total expenditures is different for those aged 65 or older compared to younger individuals.

β_6 (age_sp1:ever1): The coefficient for the interaction age_sp1:ever1 is 261.66 (95%CI -142.96, 666.28), and it's not significant ($p > 0.05$), suggesting no clear effect modification by ever smoking status in the aged 75 to 85 years group.

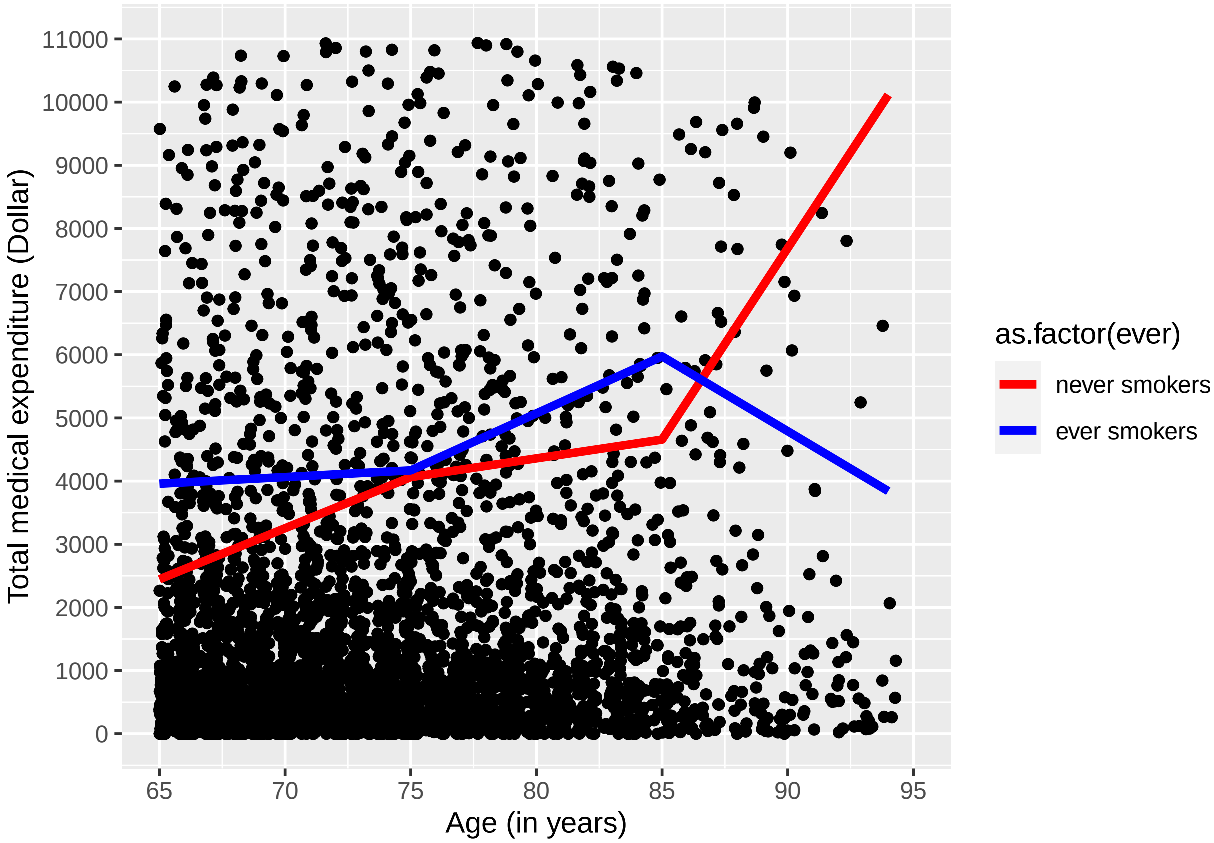
β_7 (age_sp2:ever1): The interaction suggests a decrease on total expenditures by 964.30 units (95%CI -1872.40, -56.19) for being an over 85 years ever smoker compared to the reference group (adults aged under 85 years and never-smoker).

2.Create a figure that displays the data and the predicted values from the fit of the MLR model from Question1.

```
data1 <- data1 |>
  mutate(model1_pred = predict(reg_1))

plot1 <- data1 |>
  ggplot( aes(x = age, y = totalexp) ) +
  geom_jitter() +
  geom_line(data = data1, aes( y = model1_pred,
                              x = age, color = as.factor(ever)),  linewidth = 1.5) +
  labs(y = "Total medical expenditure (Dollar)", x = "Age (in years)") +
  scale_y_continuous(breaks = seq(0, 15000, 1000), limits = c(0, 11000)) +
  scale_x_continuous(breaks = seq(65, 95, 5), limits = c(65, 95))  +
  scale_color_manual(breaks=c("0","1"),
                     values=c("red","blue"),
                     labels = c("never smokers", "ever smokers"))
```

```
plot1
```

```
## Warning: Removed 715 rows containing missing values (`geom_point()`).
```

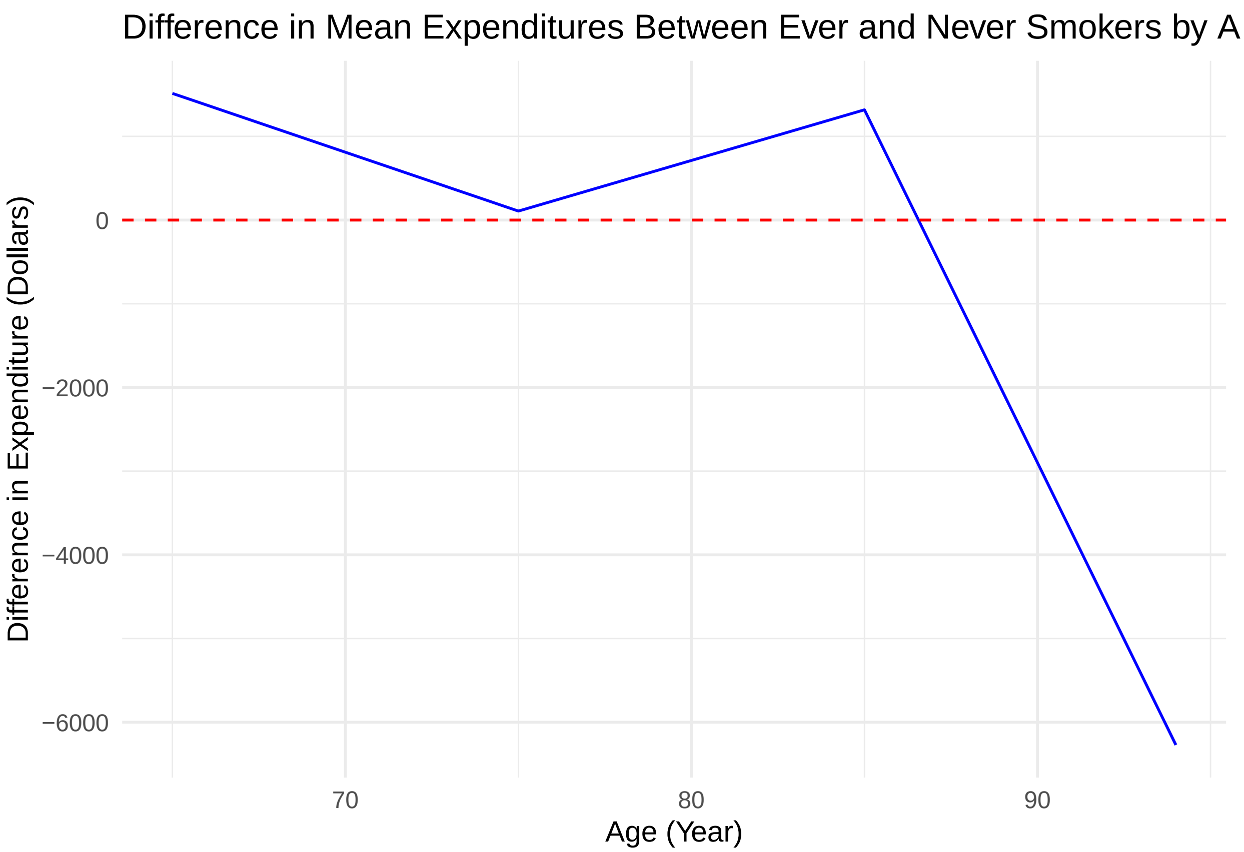


3.Using the model fit in Step 1 above, make a plot of the difference in mean expenditures between ever and never smokers as a function of age.

Difference <- coef[“ever1”] + coef[“agem65:ever1”]*agem65 + coef[“age_sp1:ever1”]*age_sp1 + coef[“age_sp2:ever1”]*age_sp2

```
coef <- reg_1$coefficients
# to calculate the difference in expenditures between ever and never smokers
expenditure_difference <- function(age) {
  agem65 <- age - 65
  age_sp1 <- ifelse(age >= 75, age - 75, 0)
  age_sp2 <- ifelse(age >= 85, age - 85, 0)
  return(coef["ever1"] + coef["agem65:ever1"] * agem65 + coef["age_sp1:ever1"]  * age_sp1 + coef["age_sp
}

# Create an age range from 65 to 94
age_range <- 65:94

# Calculate the difference for each age
differences <- sapply(age_range, expenditure_difference)

# Create a data frame for plotting
data_plot <- data.frame(Age = age_range, Difference = differences)
```

```r
# Plotting using ggplot2
ggplot(data_plot, aes(x = Age, y = Difference)) +
  geom_line(color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Difference in Mean Expenditures Between Ever and Never Smokers by Age",
       x = "Age (Year)",
       y = "Difference in Expenditure (Dollars)") +
  theme_minimal()
```



Comment on why you think the average expenditures for ever smokers are less than the average expenditures for never smokers among persons over 85 years of age.

I think one potential reason is the survival bias. Old adults who ever smoking and still lived to over 85 years are more likely to be those who have healthy body, both genetic and physical level. So they use less medical expenditures compared to never smokers.

4.Use the appropriate linear combination of regression coefficients to calculate the estimated difference between ever and never smokers in average expenditures and its standard error at ages 65, 75, and 90 years. Complete the table below.

```r
expenditure_difference(65)
```

```
##   ever1
## 1513.54
```

```r
expenditure_difference(75)
```

```
##    ever1
```

Age	Estimated difference in expenditures Ever vs. Never Smokers	Linear Model Std Error	Linear Model 95% CI	Bootstrap Std Error	Boostrap 95% CI
65	1514	624	(289,2738)	562.9	(409.6, 2678.5)
75	107	587	(-1044, 1259)	599.7	(-1009.6, 1282.4)
90	-2899	1673	(-6179, 381)	2198.9	(-7126.4, 1652.6)

Figure 5: part4

```
## 107.1367
```

```r
expenditure_difference(90)
```

```
##     ever1
## -2899.064
```

```r
reg1.vc = vcov(reg_1)

coef(reg_1)
```

```
##   (Intercept)        agem65       age_sp1       age_sp2         ever1
##     2445.1773      161.7151     -102.2376      546.8058     1513.5401
##  agem65:ever1 age_sp1:ever1 age_sp2:ever1
##     -140.6403      261.6590     -964.2962
```

```r
# linear combination of betas
##### Age = 65
A = matrix(c(0,0,0,0,1,0,0,0), nrow = 1, ncol = 8)

A %*% coef
```

```
##         [,1]
## [1,] 1513.54
```

```r
A %*% reg1.vc %*% t(A)
```

```
##          [,1]
## [1,] 389999.5
```

```r
# standard error
sqrt(A %*% reg1.vc %*% t(A))
```

```
##          [,1]
## [1,] 624.4994
```

```r
# 95% CI for beta
A %*% coef - qt(0.975, df=summary(reg_1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##          [,1]
## [1,] 289.2298
```

```r
A %*% coef + qt(0.975, df=summary(reg_1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##         [,1]
## [1,] 2737.85
```

```r
#### Age = 75
A = matrix(c(0,0,0,0,1,10,0,0), nrow = 1, ncol = 8)

A %*% coef
```

```
##          [,1]
## [1,] 107.1367
```

```r
A %*% reg1.vc %*% t(A)
```

```
##          [,1]
## [1,] 344932.5
```

```r
# standard error
sqrt(A %*% reg1.vc %*% t(A))
```

```
##          [,1]
## [1,] 587.3095
```

```r
# 95% CI for beta
A %*% coef - qt(0.975, df=summary(reg_1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##           [,1]
## [1,] -1044.264
```

```r
A %*% coef + qt(0.975, df=summary(reg_1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##          [,1]
## [1,] 1258.537
```

```r
#### Age = 90
A = matrix(c(0,0,0,0,1,25,15,5), nrow = 1, ncol = 8)

A %*% coef
```

```
##           [,1]
## [1,] -2899.064
```

```r
A %*% reg1.vc %*% t(A)
```

```
##         [,1]
## [1,] 2798905
```

```r
# standard error
sqrt(A %*% reg1.vc %*% t(A))
```

```
##          [,1]
## [1,] 1672.993
```

```r
# 95% CI for beta
A %*% coef - qt(0.975, df=summary(reg_1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##           [,1]
## [1,] -6178.911
```

```r
A %*% coef + qt(0.975, df=summary(reg_1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##          [,1]
## [1,] 380.7829
```

5.Now estimate the ratio of the average expenditures comparing ever to never smokers at age 65. This is a non-linear function of the regression coefficients from step 1. Use the delta method to estimate the standard error of this statistic and make a 95% confidence interval for the true value given the model.

```r
# # ratio of the average expenditures comparing ever to never smokers
# expenditure_ratio <- function(age) {
#   agem65 <- age - 65
#   age_sp1 <- ifelse(age >= 75, age - 75, 0)
#   age_sp2 <- ifelse(age >= 85, age - 85, 0)
#   return(1 + (coef["ever1"] + coef["agem65:ever1"] * agem65 + coef["age_sp1:ever1"]  * age_sp1 + coef
# }
#
# # Create an age range from 65 to 94
# age_range <- 65:94
# ratio <- sapply(age_range, expenditure_ratio)
#
# ratio_at_65 <- expenditure_ratio(65)
# ratio_at_65
# # estimate the standard error
# reg1.vc = vcov(reg_1)
#
# library(numDeriv)
# # Calculate the gradient at the coefficients
# grad <- grad(expenditure_ratio, coef)
#
# # Calculate the variance using the delta method
# var_ratio <- t(grad) %*% reg1.vc %*% grad
#
# se_ratio <- sqrt(var_ratio)
# se_ratio
# # 95% confidence interval
# ratio_at_65 - qt(0.975, df=summary(reg_1)$df[2]) * sqrt(var_ratio)
#
# ratio_at_65 + qt(0.975, df=summary(reg_1)$df[2]) * sqrt(var_ratio)


model_21 = lm(totalexp ~ agem65 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp1 + age_sp2), data1)
ratio_pt = 1 + coef(model_21)[5]/coef(model_21)[1]
coef_21 = coef(model_21)
model_21_vcov = vcov(model_21)
ratio_gprime <-
  matrix(c(-coef_21[5] / coef_21[1]^2, 0, 0, 0, 1 / coef_21[1], 0, 0, 0),
         nrow = 1,
         ncol = 8)

ratio_se <- sqrt(ratio_gprime %*% model_21_vcov %*% t(ratio_gprime))
ratio_ci <- ratio_pt + c(-1,1)*qt(0.975,df=summary(model_21)$df[2])*ratio_se
```

```
## Warning in c(-1, 1) * qt(0.975, df = summary(model_21)$df[2]) * ratio_se: Recycling array of length 
##   Use c() or as.vector() instead.
```

```r
ratio_ci <- ratio_pt + c(-1,1)*qnorm(0.975)*ratio_se
```

```
## Warning in c(-1, 1) * qnorm(0.975) * ratio_se: Recycling array of length 1 in vector-array arithmeti
```

```
##   Use c() or as.vector() instead.
```

```r
## Generate a 95% CI for the ratio
print(paste(c(round(ratio_pt,3), ' 95%CI:', round(ratio_ci,3)), collapse = ' '))
```

```
## [1] "1.619  95%CI: 0.926 2.312"
```

Use the delta method, the ratio of the average expenditures comparing ever to never smokers at age 65 is 1.619 with 95% CI (0.926 2.312). The Standard error of this statistic is 0.3534826.

6.use the bootstrap procedure to estimate the standard errors and confidence intervals for the difference in Question 4.

```r
# 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 <- data[indices, ]

  # Calculate the age terms for the specified age
  agem65 <- age - 65
  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_sp1 + age_sp2 + ever + ever*(agem65 + age_sp1 + age_sp2), data = res

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

  return(difference)
}
```

```r
# Perform the bootstrap for each age
results <- lapply(c(65, 75, 90), 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")
```

```r
row.names(bootstrap_results_df) <- c("Age 65", "Age 75", "Age 90")

# Print the results
print(bootstrap_results_df)
```

Bootstrap Std Error

```
##           Estimate        SE   CI_lower CI_upper
## Age 65  1548.8305  562.9474   409.6451 2678.510
## Age 75   105.1697  599.7071 -1009.6343 1282.429
## Age 90 -3077.4616 2198.9148 -7126.4454 1652.596
```

The bootstrapped std error and 95% CI of estimated difference between ever and never smokers in average expenditures for people aged 65 is similar to the model-based std error and 95% CI. For people aged 75 and 90, the estimated difference between ever and never smokers have bigger bootstrapped std error and wider bootstrapped 95% CI than the model-based one.

6b.use the bootstrap procedure to estimate the standard errors and confidence intervals for the ratio in Question 5.

```r
set.seed(653)

ratio_boot <- 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_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_sp1 + age_sp2 + ever + ever*(agem65 + age_sp1 + age_sp2), data = res

  # Calculate the difference using the model coefficients
  coef <- coef(fit)
  ratio <- 1 + (coef["ever1"] + coef["agem65:ever1"] * agem65 + coef["age_sp1:ever1"]  * age_sp1 + coef
  return(ratio)
}
```

```r
# Perform the bootstrap for each age
results_2 <- lapply(c(65, 75, 90), function(age) {
  boot(data1, ratio_boot, R = 1000, age = age)
})


# Extract the bootstrap standard errors and confidence intervals
bootstrap_results <- sapply(results_2, 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")
```

```r
row.names(bootstrap_results_df) <- c("Age 65", "Age 75", "Age 90")

# Print the results
print(bootstrap_results_df)
```

```
##         Estimate        SE  CI_lower CI_upper
## Age 65 1.6597419 0.2910671 1.1431385 2.323650
## Age 75 1.0381855 0.1530871 0.7810630 1.367224
## Age 90 0.6160127 0.2674605 0.2374612 1.248163
```

Using bootstrapping, the ratio of the average expenditures comparing ever to never smokers at age 65 is 1.659 (95% CI 1.143, 2.323), the standard errors for the ratio is 0.291. Compared with results obtained directly from the linear regression, the bootstrapped 95% CI is wider than the model-based one, the bootstrapped standard error is larger than the model-based one.

7.Test the null hypothesis that on average, ever and never smokers use the same quantity of medical services; i.e. are the mean expenditures at any age the same for ever and never smokers?

`Likelihood Ratio Test`

```r
# install.packages("lmtest")
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
null_model <- lm(totalexp ~ agem65 + age_sp1 + age_sp2, data = data1) # null model (without ever or its
extended_model <- lm(totalexp ~ agem65 + age_sp1 + age_sp2 + ever + ever*(agem65 + age_sp1 + age_sp2), d

# Likelihood ratio test
lr.test.stat = as.numeric(2*logLik(extended_model)-2*logLik(null_model))
pchisq(lr.test.stat, df=4, lower.tail = FALSE) # 0.0152
```

```
## [1] 0.01519584
```

```r
lrtest(null_model, extended_model) # 0.0152
```

```
## Likelihood ratio test
##
## Model 1: totalexp ~ agem65 + age_sp1 + age_sp2
## Model 2: totalexp ~ agem65 + age_sp1 + age_sp2 + ever + ever * (agem65 +
##     age_sp1 + age_sp2)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   5 -50278
## 2   9 -50272  4 12.309     0.0152 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# F-test
anova(null_model, extended_model) # F = 3.076, P-value =  0.01532
```

```
## Analysis of Variance Table
```

```
##
## Model 1: totalexp ~ agem65 + age_sp1 + age_sp2
## Model 2: totalexp ~ agem65 + age_sp1 + age_sp2 + ever + ever * (agem65 +
##     age_sp1 + age_sp2)
##   Res.Df        RSS Df  Sum of Sq     F  Pr(>F)
## 1   4724 4.7747e+11
## 2   4720 4.7622e+11  4 1241422709 3.076 0.01532 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: $\beta_4 = 0$, $\beta_5 = 0$, $\beta_6 = 0$, $\beta_7 = 0$, $\beta_8 = 0$ H1: at least one not equal to 0 The likelihood ratio test shows the p-value = 0.0152, which is less than 0.05. We reject the null hypothesis, indicating that the mean expenditures are different for ever and never smokers at any age. The F test also shows the p-value = 0.0153, F stat = 3.076. The result indicates the full model is better than the null model. It suggests that ever and never smokers use the different quantity of medical services, and varies with age. The results of likelihood ratio test and F test are similar, suggests that smoking status significantly affects medical expenditures, and this effect varies with age.

8.Using the results of Questions 1-7, write a brief report with sections: Objective, data, methods, results, and discussion as if for a health services journal.

`Objective`: To explore whether older adults aged 65 above ever and never smokers of the same age use roughly the same quantity of medical services.
`Data`: The study used 1987 National Medical Expenditure Survey (NMES) dataset, including 13648 study participants.
`Methods`: We conducted multiple linear regression (MLR) allowing the total medical expenditures to change as a function of age (linear spline with knot at 75 and 85 years), separately for ever and never smokers. To address data skewness and heteroscedasticity, we applied the delta method, bootstrapping for robust standard error estimation, and conducted likelihood ratio and F-tests to test whether mean expenditures differ between ever and never smokers.
`Results`: The regression analysis indicates that at age 65, never smokers have an estimated mean total medical expenditure of 2,445.18 dollars, while ever smokers have higher medical expenditure of 3,958.72 dollars. The difference in costs between ever and never smokers at age 65 is 1,513.54 dollars with 95% confidence interval (289.23 to 2,737.85). For ages 75 and 90, the differences in projected costs between the two groups are 107.14 and -2,899.06, respectively, but these are not statistically significant as their confidence intervals include zero. The expenditure ratio for ever versus never smokers at age 65 is approximately 1.62, with a 95% confidence interval of 0.926 to 2.312, confirming higher costs for ever smokers. Statistical tests yield a p-value of 0.015, indicating a significant difference in medical expenditures between the groups across all ages analyzed. `Discussions`: Our findings suggest that age has association with total medical expenditures, with this effect varying by somking status The use of advanced statistical methods, including bootstrapping, provided a more nuanced understanding of the expenditure patterns, accounting for the data's non-normal distribution and heteroscedasticity. The study underscores the economic impact of smoking on healthcare costs and highlights the importance to understand these effects accurately.

III. Estimating rates of change from smooth functions

1.Describe in words what as a function of age looks like for a linear spline with knots at 75 and 85 years.

The slope(age) as a function of age for a linear spline with knots at 75 and 85 years equals to three constant at different age range, and changes at these knot points. From ages 65 to 75, the slope(age) is constant β_1; from ages 75 to 85, the slope(age) is constant $\beta_1+\beta_2$; from ages 85 to 95, the slope(age) is constant $\beta_1+\beta_2+\beta_3$.

2.Apply the procedure described above to estimate for a cubic spline model.

a.Subset the data you used in Part II to include only the ever smokers. Fit a cubic spline model with knots at 75 and 85 years of age to the data for the ever smokers. Save the estimated regression coefficients and variance matrix for the estimated regression coefficients.

```
data_ever <- data1 |>
  filter(ever == 1) |>
  mutate(
          age2 = (age-65) * (age-65) ,
          age3 = (age-65) * (age-65) * (age-65) ,
          age_csp1 = ifelse(age-75>0, (age-75)^3,0),
          age_csp2 = ifelse(age-85>0, (age-85)^3,0),
          )

reg_cubic <- lm(data = data_ever, totalexp ~ agem65 + age2 + age3 + age_csp1 + age_csp2)
summary(reg_cubic)
```

```
##
## Call:
## lm(formula = totalexp ~ agem65 + age2 + age3 + age_csp1 + age_csp2,
##     data = data_ever)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7135  -3789  -3150  -1028 171322
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4647.7895   654.2924   7.104 1.59e-12 ***
## agem65      -452.8363   440.1793  -1.029    0.304
## age2          58.7578    75.6385   0.777    0.437
## age3          -1.6373     3.5571  -0.460    0.645
## age_csp1      -0.6856     6.2595  -0.110    0.913
## age_csp2      14.8693    16.2933   0.913    0.362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10490 on 2416 degrees of freedom
## Multiple R-squared:  0.003386,   Adjusted R-squared:  0.001323
## F-statistic: 1.642 on 5 and 2416 DF,  p-value: 0.1456
```

```
coef_cubic <- reg_cubic$coefficients
var_cubic <- vcov(reg_cubic)
```

b.For ages, in years, 65 to 94, derive and create and compute and slope(age) and var(slope(age)).

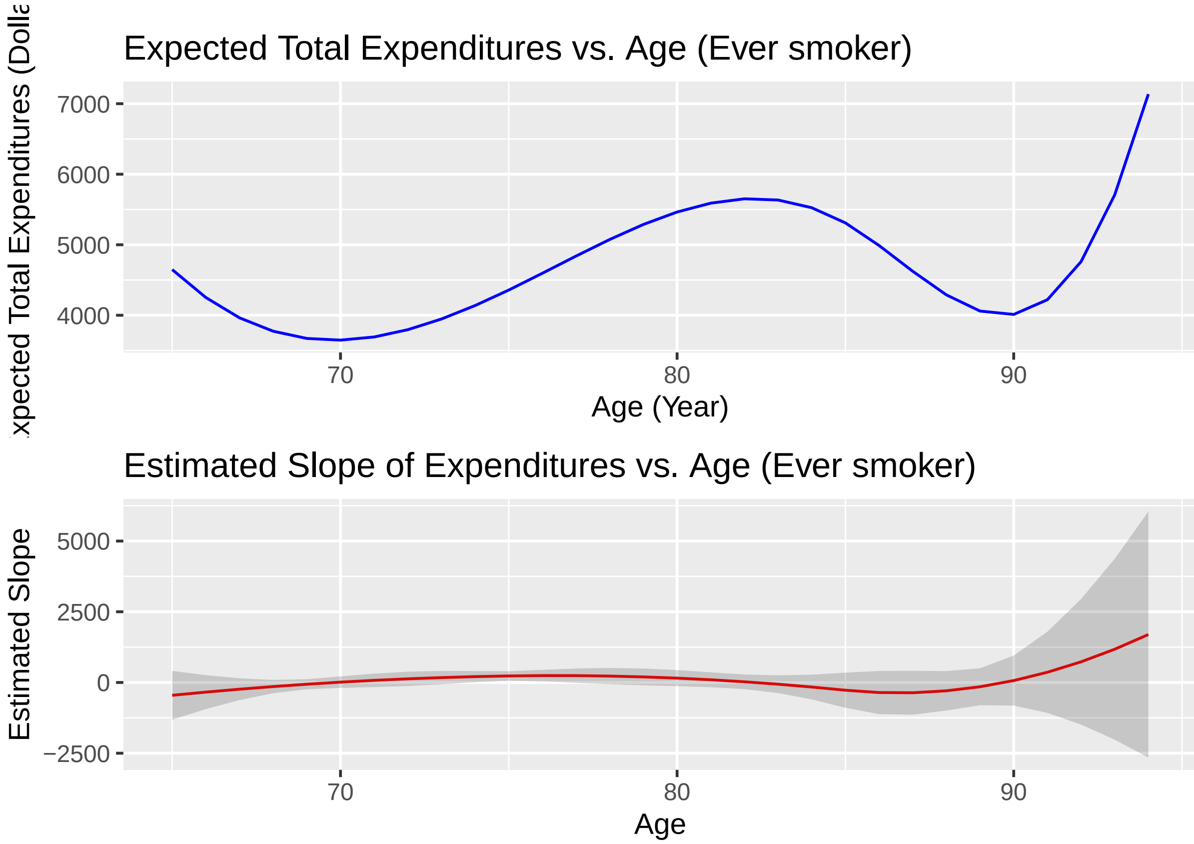
```
# dXage <- matrix(c(0,1,2*(age-65),3*(age-65)^2, 3*(age-75)^2, 3*(age-85)^2), )
ages <- 65:94
dX <- data.frame(
  Intercept = rep(0, length(ages)),
  Linear = rep(1, length(ages)),
  Quadratic = 2 * (ages - 65),
  Cubic = 3 * (ages - 65)^2,
  Spline1 = 3 * ifelse(ages > 75, (ages - 75)^2, 0),
  Spline2 = 3 * ifelse(ages > 85, (ages - 85)^2, 0)
```

```r
)

# Compute estimated slopes
slope_hat <- as.matrix(dX) %*% coef_cubic

# Compute variance of the slope estimates
var_slope_hat <- diag(as.matrix(dX) %*% var_cubic %*% t(as.matrix(dX)))

# Standard error of the slopes
se_slope_hat <- sqrt(var_slope_hat)

# 95% CI for the slopes
ci_lower <- slope_hat - 1.96 * se_slope_hat
ci_upper <- slope_hat + 1.96 * se_slope_hat

# Results
slopes_df <- data.frame(Age = ages, Slope = slope_hat, SE = se_slope_hat, CI_Lower = ci_lower, CI_Upper

slopes_df
```

```
##    Age       Slope        SE     CI_Lower   CI_Upper
## 1   65  -452.83625  440.17935 -1315.587776  409.91527
## 2   66  -340.23275  305.81645  -939.632999  259.16750
## 3   67  -237.45329  196.91447  -623.405646  148.49906
## 4   68  -144.49788  120.30140  -380.288629   91.29286
## 5   69   -61.36652   91.14742  -240.015460  117.28242
## 6   70    11.94080  103.47656  -190.873255  214.75486
## 7   71    75.42407  121.77813  -163.261051  314.10920
## 8   72   129.08330  128.71390  -123.195938  381.36254
## 9   73   172.91848  120.60004   -63.457599  409.29457
## 10  74   206.92962  100.59400     9.765374  404.09387
## 11  75   231.11671   85.92194    62.709704  399.52372
## 12  76   243.42309  103.37001    40.817871  446.02831
## 13  77   241.79208  129.11894   -11.281051  494.86521
## 14  78   226.22368  146.27363   -60.472636  512.92001
## 15  79   196.71791  151.25161   -99.735241  493.17105
## 16  80   153.27474  145.25249  -131.420143  437.96963
## 17  81    95.89419  134.19225  -167.122610  358.91100
## 18  82    24.57626  132.35269  -234.835008  283.98753
## 19  83   -60.67906  159.70028  -373.691610  252.33350
## 20  84  -159.87176  223.18102  -597.306553  277.56304
## 21  85  -273.00185  316.60680  -893.551165  347.54747
## 22  86  -355.46128  389.22865 -1118.349426  407.42687
## 23  87  -362.64201  396.12810 -1139.053082  413.76906
## 24  88  -294.54405  355.65013  -991.618313  402.53021
## 25  89  -151.16739  333.51199  -804.850894  502.51611
## 26  90    67.48796  450.77812  -816.037156  951.01308
## 27  91   361.42201  731.76763 -1072.842546 1795.68657
## 28  92   730.63476 1132.34052 -1488.752656 2950.02218
## 29  93  1175.12620 1630.48505 -2020.624495 4370.87690
## 30  94  1694.89634 2218.22482 -2652.824302 6042.61699
```

c.Make a two panel figure displaying E(total expenditures) vs. age and slope(age) vs. age (with corresponding 95% confidence intervals).

```r
ages <- 65:94
new_data <- data.frame(age = ages)
# Add columns for the cubic spline terms
new_data$agem65 <- new_data$age - 65
new_data$age2 <- (new_data$age - 65)^2
new_data$age3 <- (new_data$age - 65)^3
new_data$age_csp1 <- ifelse(new_data$age - 75 > 0, (new_data$age - 75)^3, 0)
new_data$age_csp2 <- ifelse(new_data$age - 85 > 0, (new_data$age - 85)^3, 0)

# Predict E(total expenditures) using the cubic spline model
new_data$predicted_totalexp <- predict(reg_cubic, newdata = new_data)

# data_ever <- data_ever |>
#   mutate(predicted_totalexp = predict(reg_cubic, newdata = data_ever))

p1 <- new_data |>
  ggplot(aes(x=age, y=predicted_totalexp)) +
  geom_line(color = "blue") +
  labs(title = "Expected Total Expenditures vs. Age (Ever smoker)", x = "Age (Year)", y = "Expected Tota

p2 <- slopes_df |>
  ggplot(aes(x=Age, y=Slope)) +
  geom_line(color = "red") +
  geom_ribbon(aes(ymin = CI_Lower, ymax = CI_Upper), alpha = 0.2) +
  labs(title = "Estimated Slope of Expenditures vs. Age (Ever smoker)", x = "Age", y = "Estimated Slope

# Use the 'patchwork' library to combine plots
library(patchwork)
combined_plot <- p1 + p2 + plot_layout(ncol = 1)
print(combined_plot)
```



3.Next, we will utilize the splines2 package in R to help us generate to estimate slope(age) for a natural cubic spline model with knots at 75 and 85 years. Install the splines2 package if not already done so and include library(splines2) in your R code.

```r
# install.packages("splines2")
library(splines2)
```

a.Fit the natural cubic spline model with knots at 75 and 85 years and save the coefficients and variance of the estimated coefficients.

fit = lm(totalexp~-1 + nsp(lastage,knots=c(75,85),intercept=TRUE),data=data.ever) fit.coeff = fit$coefficients V.coeff = vcov(fit)

```r
fit = lm(totalexp~-1 + nsp(lastage,knots=c(75,85),intercept=TRUE),data=data_ever)
fit.coeff = fit$coefficients
V.coeff = vcov(fit)
```

b.Generate the design matrix X and derivative of the design matrix, , evaluated at ages 65 through 94, and the estimated E(total expenditures) and for ages 65 through 94.

```r
# First create the design matrix and derivative of the design matrix
X = naturalSpline(seq(65,94),knots=c(75,85),intercept=TRUE)
dXage = naturalSpline(seq(65,94),knots=c(75,85),intercept=TRUE,derivs=1)

# Estimate E(total expenditures) and slope(age) for ages 65 to 94
mean.Y.age = predict(X,coef=fit.coeff)
slope.age = predict(dXage,coef=fit.coeff)
```

c.Compute the and create a two panel figure displaying E(total expenditures) and for ages 65 to 94 for ever smokers.

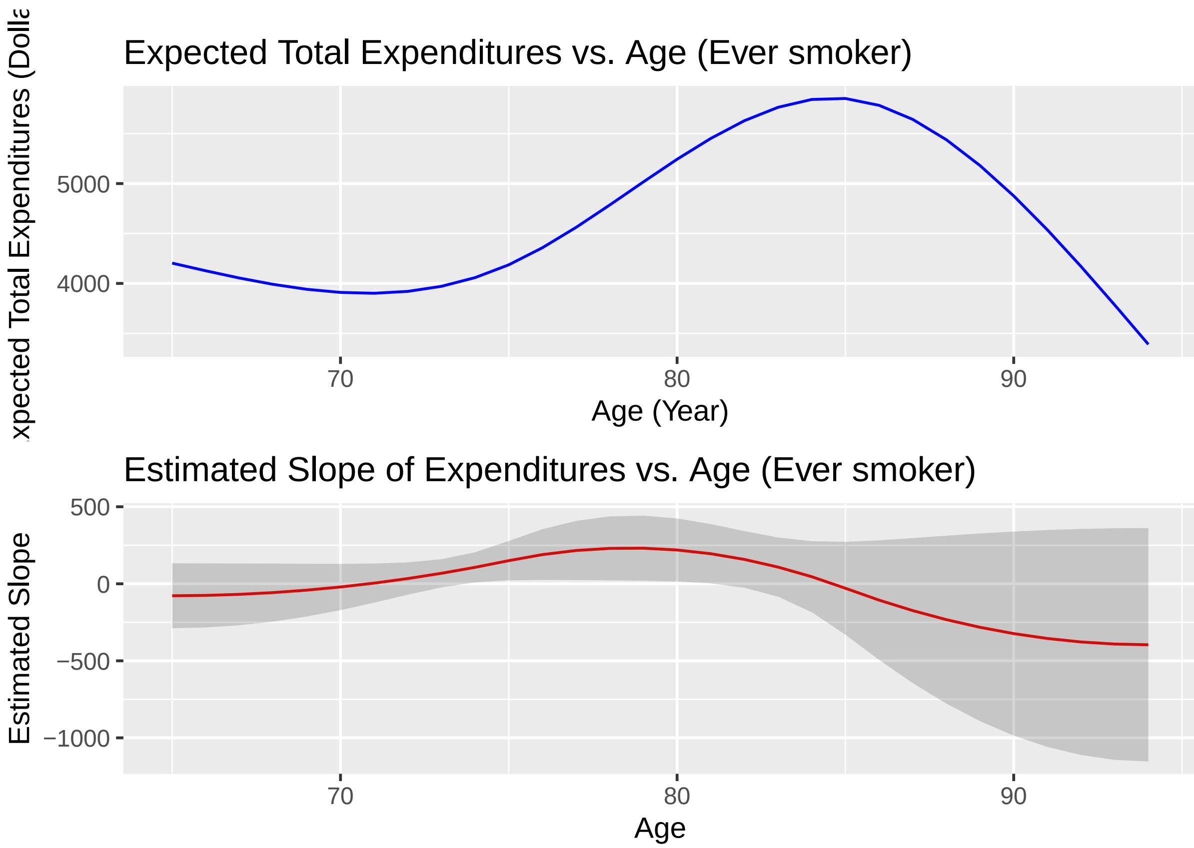
NOTE: the variance is: diag(dXage %% *V.coeff* %% t(dXage))

```
# Compute variance of the slope estimates
var_slope.age <- diag(dXage %*% V.coeff %*% t(dXage))

# Standard error of the slopes
se_slope_age <- sqrt(var_slope.age)

# 95% CI for the slopes
ci_lower <- slope.age - 1.96 * se_slope_age
ci_upper <- slope.age + 1.96 * se_slope_age

# Results
slopes_ns <- data.frame(Age = ages, Estimate = mean.Y.age, Slope = slope.age, SE = se_slope_age, CI_Lowe

slopes_ns
```

```
##    Age Estimate       Slope        SE     CI_Lower CI_Upper
## 1   65 4203.177  -77.646694 107.41659  -288.183217 132.8898
## 2   66 4126.290  -75.367099 106.12829  -283.378541 132.6443
## 3   67 4053.962  -68.528315 102.28584  -269.008560 131.9519
## 4   68 3990.753  -57.130342  95.96564  -245.223003 130.9623
## 5   69 3941.222  -41.173179  87.33165  -212.343212 129.9969
## 6   70 3909.927  -20.656827  76.72174  -171.031438 129.7178
## 7   71 3901.428    4.418714  64.87105  -122.728546 131.5660
## 8   72 3920.284   34.053445  53.50281   -70.812066 138.9190
## 9   73 3971.054   68.247365  46.55960   -23.009447 159.5042
## 10  74 4058.298  107.000475  49.91687     9.163402 204.8375
## 11  75 4186.575  150.312773  65.19952    22.521709 278.1038
## 12  76 4357.566  189.551567  84.06721    24.779843 354.3233
## 13  77 4561.443  216.084159  98.04557    23.914850 408.2535
## 14  78 4785.499  229.910552 105.91048    22.326005 437.4951
## 15  79 5017.028  231.030744 107.74181    19.856794 442.2047
## 16  80 5243.325  219.444736 104.47496    14.673811 424.2157
## 17  81 5451.682  195.152527  98.31095     2.463064 387.8420
## 18  82 5629.395  158.154119  93.73259   -25.561767 341.8700
## 19  83 5763.755  108.449509  98.00145   -83.633323 300.5323
## 20  84 5842.058   46.038700 117.67354  -184.601446 276.6788
## 21  85 5851.597  -29.078310 153.75395  -330.436059 272.2794
## 22  86 5783.293 -106.022303 197.90517  -493.916436 281.8718
## 23  87 5642.570 -173.914062 240.14549  -644.599229 296.7711
## 24  88 5438.482 -232.753586 278.11697  -777.862837 312.3557
## 25  89 5180.080 -282.540875 310.87421  -891.854329 326.7726
## 26  90 4876.418 -323.275930 337.97928  -985.715310 339.1635
## 27  91 4536.546 -354.958751 359.20723 -1059.004912 349.0874
## 28  92 4169.517 -377.589337 374.43507 -1111.482080 356.3034
## 29  93 3784.385 -391.167689 383.59470 -1143.013294 360.6779
## 30  94 3390.199 -395.693806 386.65145 -1153.530653 362.1430
```

```
p3 <- slopes_ns |>
  ggplot(aes(x=Age, y=Estimate)) +
  geom_line(color = "blue") +
  labs(title = "Expected Total Expenditures vs. Age (Ever smoker)", x = "Age (Year)", y = "Expected Tota
```

```r
p4 <- slopes_ns |>
  ggplot(aes(x=Age, y=Slope)) +
  geom_line(color = "red") +
  geom_ribbon(aes(ymin = CI_Lower, ymax = CI_Upper), alpha = 0.2) +
  labs(title = "Estimated Slope of Expenditures vs. Age (Ever smoker)", x = "Age", y = "Estimated Slope

combined_plot2 <- p3 + p4 + plot_layout(ncol = 1)
print(combined_plot2)
```



d.Repeat the process above for never smokers and create a final two panel figure with E(total expenditures) and for ages 65 to 94 separately for ever and never smokers. Include confidence intervals for .

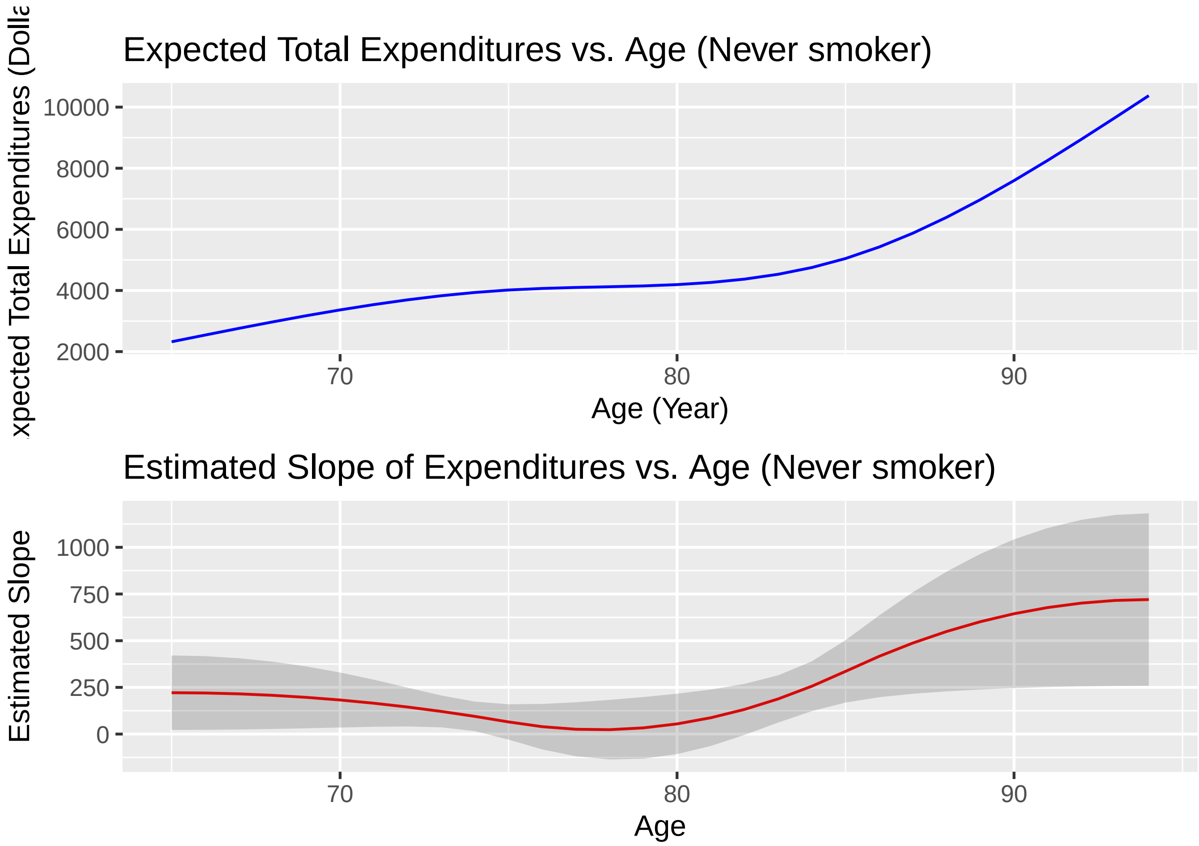
```r
data_never <- data1 |>
  filter(ever == 0) |>
  mutate(
          age2 = (age-65) * (age-65) ,
          age3 = (age-65) * (age-65) * (age-65) ,
          age_csp1 = ifelse(age-75>0, (age-75)^3,0),
          age_csp2 = ifelse(age-85>0, (age-85)^3,0),
          )
fit = lm(totalexp~-1 + nsp(lastage,knots=c(75,85),intercept=TRUE),data=data_never)
fit.coeff = fit$coefficients
V.coeff = vcov(fit)

# First create the design matrix and derivative of the design matrix
X = naturalSpline(seq(65,94),knots=c(75,85),intercept=TRUE)
dXage = naturalSpline(seq(65,94),knots=c(75,85),intercept=TRUE,derivs=1)
```

```r
# Estimate E(total expenditures) and slope(age) for ages 65 to 94
mean.Y.age = predict(X,coef=fit.coeff)
slope.age = predict(dXage,coef=fit.coeff)

# Compute variance of the slope estimates
var_slope.age <- diag(dXage %*% V.coeff %*% t(dXage))

# Standard error of the slopes
se_slope_age <- sqrt(var_slope.age)

# 95% CI for the slopes
ci_lower <- slope.age - 1.96 * se_slope_age
ci_upper <- slope.age + 1.96 * se_slope_age

# Results
slopes_ns2 <- data.frame(Age = ages, Estimate = mean.Y.age, Slope = slope.age, SE = se_slope_age, CI_Lo

slopes_ns2
```

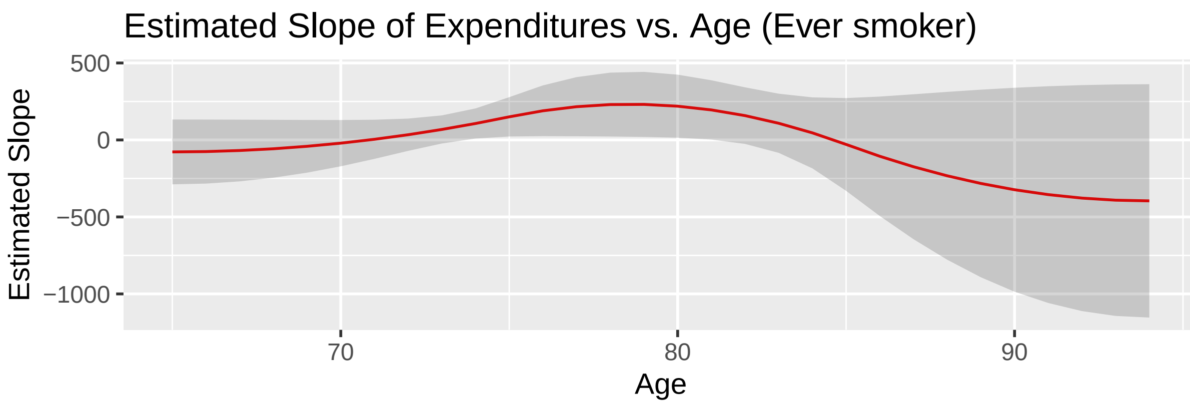
```
##    Age  Estimate     Slope        SE    CI_Lower  CI_Upper
## 1   65  2322.343 221.51487 101.77431   22.037224  420.9925
## 2   66  2543.336 219.94950 100.67134   22.633665  417.2653
## 3   67  2761.198 215.25339  97.37651   24.395430  406.1113
## 4   68  2972.799 207.42653  91.93701   27.229990  387.6231
## 5   69  3175.008 196.46893  84.45171   30.943591  361.9943
## 6   70  3364.693 182.38060  75.11704   35.151199  329.6100
## 7   71  3538.725 165.16151  64.34189   39.051409  291.2716
## 8   72  3693.973 144.81169  53.05393   40.825982  248.7974
## 9   73  3827.305 121.33113  43.50249   36.066254  206.5960
## 10  74  3935.591  94.71982  40.41244   15.511440  173.9282
## 11  75  4015.701  64.97777  48.28599  -29.662761  159.6183
## 12  76  4066.971  39.50883  62.12312  -82.252482  161.2701
## 13  77  4098.611  25.71684  73.79691 -118.925115  170.3588
## 14  78  4122.297  23.60180  81.29862 -135.743496  182.9471
## 15  79  4149.707  33.16372  84.15012 -131.770518  198.0980
## 16  80  4192.517  54.40260  82.50171 -107.300746  216.1059
## 17  81  4262.405  87.31843  77.09892  -63.795447  238.4323
## 18  82  4371.046 131.91121  69.76379   -4.825815  268.6482
## 19  83  4530.119 188.18095  64.50634   61.748514  314.6134
## 20  84  4751.301 256.12764  68.01833  122.811718  389.4436
## 21  85  5046.267 335.75129  85.30510  168.553285  502.9493
## 22  86  5423.167 416.46552 111.74623  197.442903  635.4881
## 23  87  5876.033 487.68395 138.75528  215.723602  759.6443
## 24  88  6395.369 549.40660 163.58847  228.773188  870.0400
## 25  89  6971.681 601.63345 185.22131  238.599680  964.6672
## 26  90  7595.471 644.36451 203.20906  246.074761 1042.6543
## 27  91  8257.244 677.59978 217.33438  251.624403 1103.5752
## 28  92  8947.505 701.33926 227.48269  255.473182 1147.2053
## 29  93  9656.758 715.58295 233.59214  257.742352 1173.4235
## 30  94 10375.506 720.33084 235.63177  258.492579 1182.1691
```

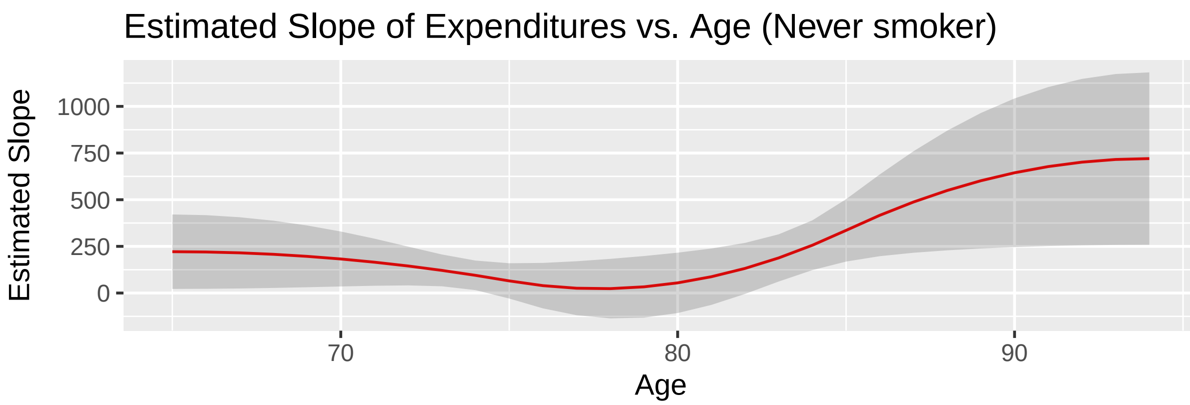
```r
p5 <- slopes_ns2 |>
  ggplot(aes(x=Age, y=Estimate)) +
  geom_line(color = "blue") +
```

```r
  labs(title = "Expected Total Expenditures vs. Age (Never smoker)", x = "Age (Year)", y = "Expected To

p6 <- slopes_ns2 |>
  ggplot(aes(x=Age, y=Slope)) +
  geom_line(color = "red") +
  geom_ribbon(aes(ymin = CI_Lower, ymax = CI_Upper), alpha = 0.2) +
  labs(title = "Estimated Slope of Expenditures vs. Age (Never smoker)", x = "Age", y = "Estimated Slope

combined_plot2 <- p5 + p6 + plot_layout(ncol = 1)
print(combined_plot2)
```




```r
combined_plot3 <- p4 + p6 + plot_layout(ncol = 1)
print(combined_plot3)
```





e.In two or three sentences, describe the different patterns you observe in how E(total expenditures) change with age, i.e. , for both ever and never smokers.

For ever smoker, the estimated total expenditures decreases with age when adults aged 65-75 years, then the estimated total expenditures increases with age when adults aged 75-85 years, and then the estimated total expenditures decreases with age when adults aged 85-95 years.

For never smokers, the estimated total expenditures consistently increases with age, implying that expenditures continue to grow as age increased.

This indicates a different pattern where the relationship between age and medical expenditures is more consistent and linear for never smokers, while for ever smokers, it's non-linear with a peak in the rate of change at 85 ages.