HW2

Zixuan_Yu

I.Matrix Representation of Multiple Linear Regression

1.Use the following 5 observations and write the simple linear regression model in matrix terms. Then using the least squares calculations in matrix notation, compute estimates for the simple linear regression intercept and slope.

```
y <- c(-0.1, 2.9, 6.2, 7.3, 10.7)
x <- matrix(c(1,1,1,1,1,1,3,5,7,9),nrow=5,ncol=2) ## The design matrix
x; y
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    3
## [3,]    1    5
## [4,]    1    7
## [5,]    1    9
```

```
## [1] -0.1  2.9  6.2  7.3 10.7
```

$$Y = \begin{bmatrix} -0.1 \\ 2.9 \\ 6.2 \\ 7.3 \\ 10.7 \end{bmatrix}, X = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 9 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

Model: $Y = \beta X + \epsilon$
Where ϵ_i $N(0, \sigma^2)$ and $Cov(\epsilon_i, \epsilon_j) = 0$ $\hat{\beta} = (X^T X)^{-1} X^T Y$ So,

$$\hat{\beta} = (\begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 9 \end{bmatrix}^T * \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 9 \end{bmatrix})^{-1} * \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 9 \end{bmatrix}^T * \begin{bmatrix} -0.1 \\ 2.9 \\ 6.2 \\ 7.3 \\ 10.7 \end{bmatrix} = \begin{bmatrix} -1.1 \\ 1.3 \end{bmatrix}$$

```
solve(t(x) %*% x) %*% t(x) %*% y
```

```
##      [,1]
## [1,] -1.1
## [2,]  1.3
```

```
y <- c(-0.1, 2.9, 6.2, 7.3, 10.7)
x <- matrix(c(1,1,1,1,1,1,3,5,7,9),nrow=5,ncol=2) ## The design matrix
x
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    3
## [3,]    1    5
## [4,]    1    7
## [5,]    1    9
```

```
y
```

```
## [1] -0.1  2.9  6.2  7.3 10.7
```

2.Write an R function that takes the vector Y and matrix X as input then calculates and returns each the following components:
a.the least squares estimates of the regression coefficients

```
lse <- function(x,y){
    matrix(solve(t(x) %*% x) %*% t(x) %*% y, 2, 1)
}
lse(x,y)
```

```
##      [,1]
## [1,] -1.1
## [2,]  1.3
```

b.the variance-covariance matrix of the least squares estimates

```
vcm <- function(x,y){
    H = x %*% solve(t(x) %*% x) %*% t(x)
    yhat = H  %*% y
    yvar = sum((y-yhat)^2)/(length(y)-2)
    vcmatrix = yvar * solve(t(x) %*% x)
vcmatrix
}
vcm(x,y)
```

```
##             [,1]        [,2]
## [1,]  0.34100000 -0.05166667
## [2,] -0.05166667  0.01033333
```

c.the correlation between the two regression coefficients

```
corr_beta <- function(x,y){
vcmatrix = vcm(x,y)
vcmatrix[1,2]/(sqrt(vcmatrix[1,1])*sqrt(vcmatrix[2,2]))
}
corr_beta(x,y)
```

```
## [1] -0.8703883
```

d.the vector of predicted values X(X'X)-1X'Y = HY

```
cpred <- function(x,y){
    H = x %*% solve(t(x) %*% x) %*% t(x)
    H  %*% y
}
cpred(x,y)
```

```
##      [,1]
## [1,]  0.2
## [2,]  2.8
## [3,]  5.4
## [4,]  8.0
## [5,] 10.6
```

e.the vector of residuals (I- X(X'X)-1X')Y = (I-H)Y.

```
vres <- function(x,y){
    H = x %*% solve(t(x) %*% x) %*% t(x)
    (diag(nrow(H))-H) %*% y
}
vres(x,y)
```

```
##      [,1]
## [1,] -0.3
## [2,]  0.1
## [3,]  0.8
## [4,] -0.7
## [5,]  0.1
```

3. Using the R function from Question 2, verify your estimates of the simple linear regression intercept and slope computed in Question 1. Using the standard error estimate for the simple linear regression model slope, construct a 95% confidence interval for the true slope.

```
lse(x,y) ## same as what I got in Question1
```

```
##      [,1]
## [1,] -1.1
## [2,]  1.3
```

```
var_beta = vcm(x,y)[2,2]
beta_1 = lse(x,y)[2]
beta_1 + c(-1,1)*qt(0.975,df=3)*sqrt(var_beta)
```

```
## [1] 0.9764948 1.6235052
```

4. Suppose you have conducted a randomized controlled trial of an intervention (TRT = 1) vs. placebo (TRT = 0), where n_1 and n_0 patients received the intervention and placebo, respectively. For each patient, you have measured a continuous outcome Y with the goal of comparing E(Y|TRT=1) to E(Y|TRT=0). I ask that you fit the following linear regression model:

$$Y_i = B_0 + B_1 X_i + \varepsilon_i, \varepsilon_i \; iid \; N(0, \sigma^2), X_i = 1 \; if \; TRT = 1, 0 \; if \; TRT = 0$$

   a. Write out the model above using matrix notation and then using matrix calculations solve for the least squares estimates of B_0 and B_1 and $Var(\hat{B}_1)$. 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 $Var(\hat{B}_1) = \sigma^2/n_0 + \sigma^2/n_1$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \underset{\sim}{Y} = X\underset{\sim}{\beta} + E = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

the score equations are: (and setting ... to 0)

$$U_\beta(\beta) = \frac{\partial (Y-X\beta)'(Y-X\beta)}{\partial \beta} = X'(Y-X\beta) = 0$$

$$X'X\beta = X'Y$$

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

then plug in X, Y.

$$\hat{\beta} = \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_n \end{bmatrix} \right)^{-1} \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}} \underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}$$

Let the number of people in the treatment group be n_1, control group be n_0, and $n = n_1 + n_2$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$= \begin{bmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \quad 2\times 1$$

$$\begin{bmatrix} n_0+n_1 & n_1 \\ n_1 & n_1 \end{bmatrix}^{-1} = \frac{1}{(n_0+n_1)n_1 - n_1^2}\begin{bmatrix} n_1 & -n_1 \\ -n_1 & n_1+n_0 \end{bmatrix} = \begin{bmatrix} \frac{1}{n_0} & -\frac{n_1}{n_0} \\ -\frac{n_1}{n_0} & \frac{n_1+n_0}{n_0 n_1} \end{bmatrix}$$

2×2

$$\rightarrow = \begin{bmatrix} \sum_{i=1}^{n} \frac{(1-x_i)y_i}{n_0} \\ \sum_{i=1}^{n} \frac{x_i y_i (n_1+n_0)}{n_0 n_1} - \sum_{i=1}^{n} \frac{y_i}{n_0} \end{bmatrix}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \frac{(1-x_i)y_i}{n_0} = \sum_{i=1}^{n} \frac{y_i}{n_0} - \sum_{i=1}^{n} \frac{y_i x_i}{n_0}$$

$$= \frac{\sum_{i=1}^{n} y_i}{n_0} - \frac{\sum_{i=n_1+1}^{n} x_i}{n_0} = \frac{\sum_{i=1}^{n_0} x_i}{n_0} = \bar{y}_0$$

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

$$= \frac{\sum_{i=1}^{n_1} y_i (n_1+n_0)}{n_0 n_1} - \frac{\sum_{i=1}^{n_1} y_i}{n_0} - \frac{\sum_{i=n_1+1}^{n} y_i}{n_0}$$

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

$$= \frac{\sum_{i=1}^{n_1} y_i}{n_1} - \bar{y}_0$$

$$= \bar{y}_1 - \bar{y}_0$$

$\bar{y}_0$ is the average expenditure in placebo arm, $\bar{y}_1$ is the average expenditure in the intervention arm.

$$Var(\hat{B}_1) = Var(\bar{y}_1 - \bar{y}_0) = Var(\bar{y}_1) + Var(\bar{y}_0)$$

$$= \frac{Var(y_0)}{n_0} + \frac{Var(y_1)}{n_1}$$

$$= \frac{\sigma^2}{n_0} + \frac{\sigma^2}{n_1}$$

b. Now suppose that the true model is:

$$Y_i = B_0 + B_1 X_i + \varepsilon_i$$

where $X_i = 1\ if\ TRT = 1, 0\ if\ TRT = 0$ and $\varepsilon_i \sim N(0, \sigma^2(X_i))$, where $\sigma^2(X_i) = \sigma^2$ if $X_i = 1$ and $\sigma^2(X_i) = 2\sigma^2$ if $X_i = 0$ and $Cov(\varepsilon_i, \varepsilon_j)$ for all i and j. Under the true model, $Var(\hat{B}_1) = \sigma^2/n_1 + 2\sigma^2/n_0$.

Make a figure to compare the width of the 95% confidence interval for B_1 based on the model I asked you to fit (part a) and the true model. Set values for n_1 and n_0 and allow σ^2 to range from 1 to 100. Describe the impact of fitting the model I asked you to fit (i.e. a model that assumes the same variance in each group) vs. a more flexible model that would allow the variance of the residuals to depend on the assigned treatment group.

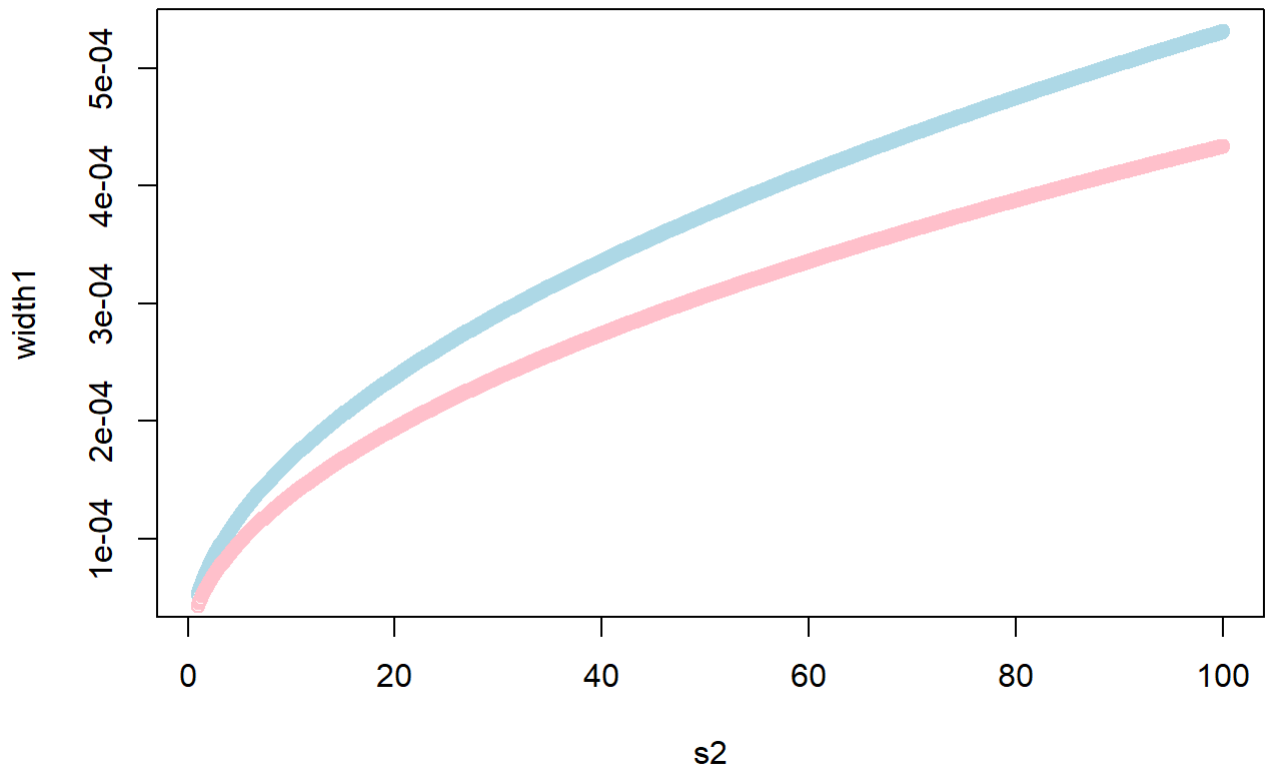
hw2_1_figure

Zixuan Yu

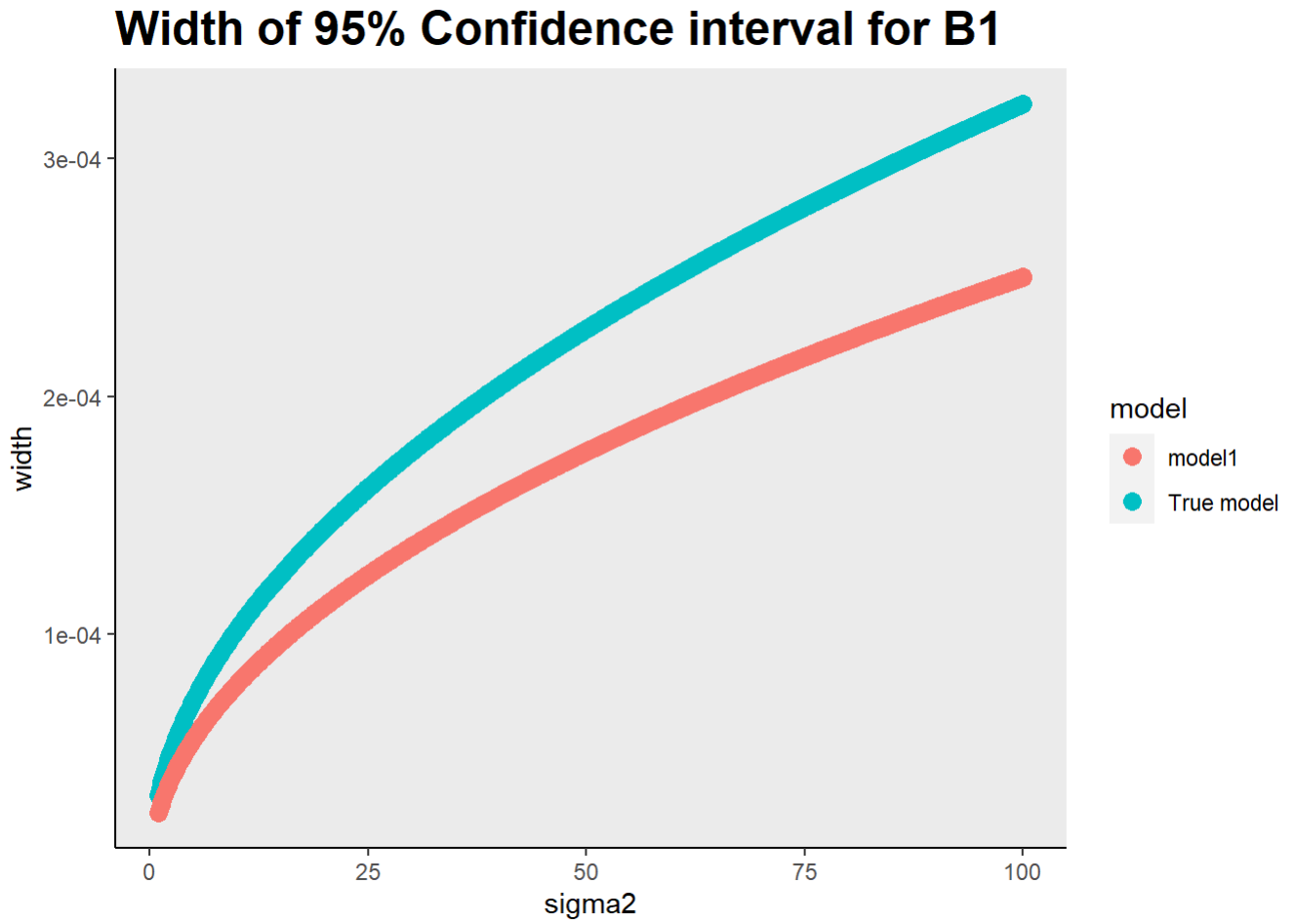
2/25/2022

b. plot width

```
## Equal Sample Size
s2 = seq(1,100, length=1000)
n1 = 1600
n2 = 1600
var_beta1 = s2/n1 + 2*s2/n2
var_beta2 = s2/n1 + 1*s2/n2
df = n1+n2-2
width1 = 2*qt(0.975, df = df )*sqrt(var_beta1)/df
width2 = 2*qt(0.975, df = df )*sqrt(var_beta2)/df
plot(s2, width1, col='lightblue')
points(s2, width2, col='pink')
```



```
## Unequal Sample Size
s2 = seq(1,100, length=1000)
n1 = 3200
n2 = 1600
var_beta1 = s2/n1 + 2*s2/n2
var_beta2 = s2/n1 + 1*s2/n2
df = n1+n2-2
width1 = 2*qt(0.975, df = df )*sqrt(var_beta1)/df
width2 = 2*qt(0.975, df = df )*sqrt(var_beta2)/df
df1 <- data.frame(sigma2 = s2, width = c(width1,width2), model = c(rep("True model",1000), rep(
"model1",1000)) )
library(ggplot2)
ggplot(data=df1, aes(x=sigma2, y=width, color=model))+geom_point(size=3) +
  labs(title = "Width of 95% Confidence interval for B1") +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        plot.title = element_text(size = 18, face = "bold"))
```



Whether sample size are equal or not does not seems to influence the shape of the curve. The true model has larger with, and the relationship between σ^2 and width is not linear. By assuming same variance across groups, we will bias standard error done, thus getting a smaller width of the confidence interval. A more flexible model that would allow the variance of the residuals to depend on the treatment group will more accurately reflec the reality.

II. Advanced Inferences for Linear Regression

Use the NMES data set on persons 65 years of age and above to address the question of whether older men and women of the same age use roughly the same quantity of medical services. That is, estimate the difference in average medical expenditures between men and women as a function of age.

```
load("C:/Users/19092/OneDrive - Johns Hopkins/Term3/Biostat653/Data/nmes.rdata")
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
```

```
d <- nmes %>% filter(lastage>=65) %>%
                mutate(agem65 = lastage-65,
                age_sp1 = ifelse(lastage>75,lastage-75,0),
                age_sp2 = ifelse(lastage>85,lastage-85,0),
                female = ifelse(male==1,0,1))
```

Fit a MLR of expenditures on age and gender and spline terms:

```
model1 <- lm(totalexp~ ((agem65 + age_sp1 + age_sp2) + female + female*(agem65 + age_sp1 + age_sp2)), data = d)
summary(model1)
```

```
##
## Call:
## lm(formula = totalexp ~ ((agem65 + age_sp1 + age_sp2) + female +
##     female * (agem65 + age_sp1 + age_sp2)), data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8573  -3997  -3180   -982 170901
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4159.39     473.41   8.786   <2e-16 ***
## agem65            11.86      75.83   0.156    0.876
## age_sp1          132.81     153.87   0.863    0.388
## age_sp2         -254.60     327.08  -0.778    0.436
## female          -974.52     614.70  -1.585    0.113
## agem65:female    117.22      98.18   1.194    0.233
## age_sp1:female  -154.60     197.01  -0.785    0.433
## age_sp2:female   512.46     406.30   1.261    0.207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10560 on 5686 degrees of freedom
## Multiple R-squared:  0.005818,   Adjusted R-squared:  0.004594
## F-statistic: 4.753 on 7 and 5686 DF,  p-value: 2.44e-05
```

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

$\beta_0 = 4159.39$, it is the average expenditure for 65-year-old male. Its 95% confidence interval is $[3231.33, 5087.44]$, indicating we have 95% chance that the CI will overlap the true mean.

$\beta_1 = 11.86$, it is the average expenditure change per year growth of age for male aged between 65 and 75. The 95% confidence interval $[-136.80, 160.51]$ overlaps 0, suggesting no significant difference in average expenditure for male aged between 65 and 75 with each year growth of age at $\alpha = 0.05$ level.

$\beta_2 = -132.81$, it is the estimated difference in average expenditure growth rate per year of age between male aged 75~85 and male aged 65~75. The 95% confidence interval is $[-168.84, 434.46]$, suggesting no significant difference in average expenditure growth rate.

$\beta_3 = -254.60$, it is the estimated difference in average expenditure growth rate per year of age between male aged >=85 and male aged 75~85. The 95% confidence interval $[-895.81, 386.61]$ overlaps 0, suggesting no significant in average expenditure growth rate between male aged >=85 and male aged 65~75 at $\alpha = 0.05$ level.

$\beta_4 = -974.52$, it is the difference in average expenditure between 65-year-old female and 65-year-old male. Its 95% confidence interval $[-2179.56620230.5313]$ overlaps 0, indicating no significant difference in average expenditure between 65-year-old female and 65-year-old male at $\alpha = 0.05$ level.

$\beta_5 = -117.22$, it is the estimated difference in average expenditure growth rate per year of age between male aged 65~75 and female aged 65~75 . Its 95% confidence interval $[-75.25, 309.70]$ overlaps 0, indicating no significant difference in average expenditure growth rate per year of age between male aged 65~75 and female aged 65~75 at $\alpha = 0.05$ level.

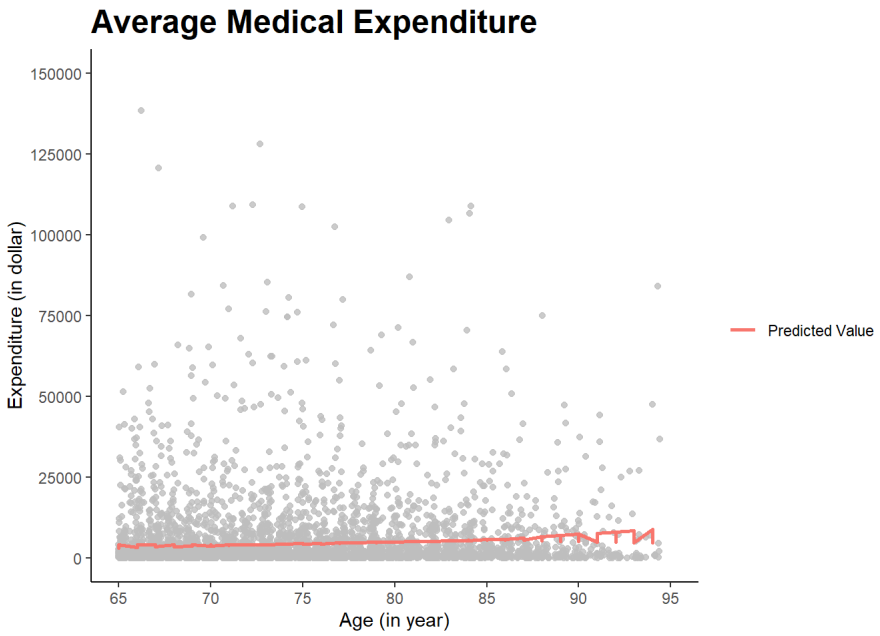
$\beta_6 = -154.60$, it is the **difference** in average expenditure growth rate change per year of age between male aged 75~85 and male aged 65~75 **when compared to** average expenditure growth rate change per year of age between female aged 75~85 and female aged 65~75. Its 95% confidence interval $[-540.82, 231.61]$ overlaps 0, indicating no significant difference at $\alpha = 0.05$ level.

$\beta_7 = 512.46$, it is the **difference** in average expenditure growth rate change per year of age between male aged >=85 and male aged 75~85 **when compared to** average expenditure growth rate change per year of age between female aged >=85 and female aged 75~85. Its 95% confidence interval $[-284.04, 1308.96]$ overlaps 0, indicating no significant difference at $\alpha = 0.05$ level.

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

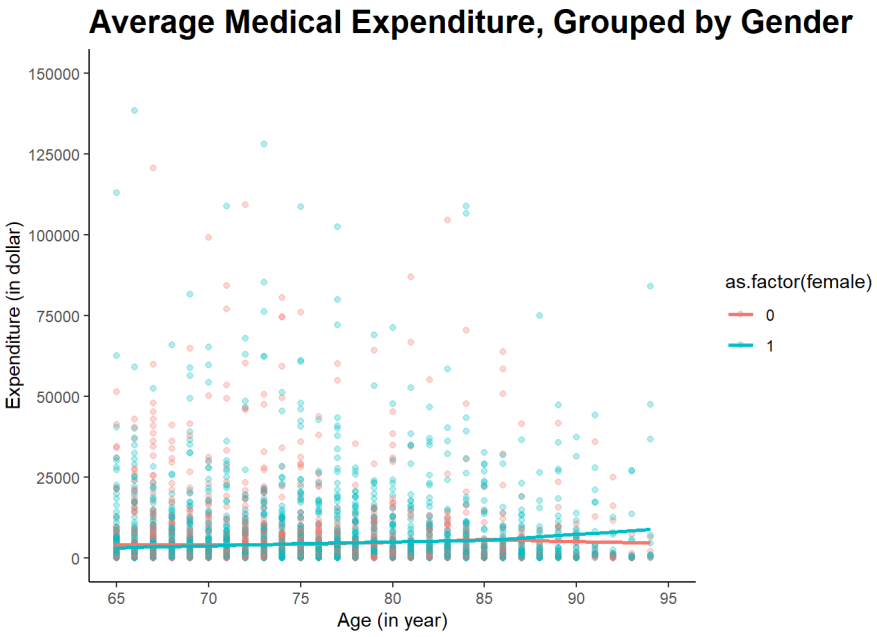
```
library(ggplot2)
ggplot(d, aes(lastage, totalexp)) +
  geom_jitter(alpha = 0.8, color = 'grey') +
  theme_bw() +
  geom_line(aes(lastage, model1$fitted.values, color = 'Predicted Value'), size = 1) +
  scale_y_continuous(name="Expenditure (in dollar)",limits=c(0,150000),breaks=seq(0,150000,25000)) +
  scale_x_continuous(name="Age (in year)",limits=c(65,95),breaks=seq(65,95,5)) +
  #scale_colour_manual(name = 'Predicted value', values = ('c1' = 'pink'),labels = ('MLR')) +
  labs(title = "Average Medical Expenditure",
       color = '') +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        plot.title = element_text(size = 18, face = "bold"))
```

```
## Warning: Removed 354 rows containing missing values (geom_point).
```




```
library(ggplot2)
ggplot(d, aes(lastage, totalexp, col = as.factor(female))) +
  geom_point(alpha = 0.3) +
  theme_bw() +
  geom_line(aes(lastage, model1$fitted.values), size = 1) +
  scale_y_continuous(name="Expenditure (in dollar)",limits=c(0,150000),breaks=seq(0,150000,25000)) +
  scale_x_continuous(name="Age (in year)",limits=c(65,95),breaks=seq(65,95,5)) +
  #scale_colour_manual(name = 'Gender', values = c("0","1"),labels = c("Male","Female")) +
  labs(title = "Average Medical Expenditure, Grouped by Gender") +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        plot.title = element_text(size = 18, face = "bold"))
```

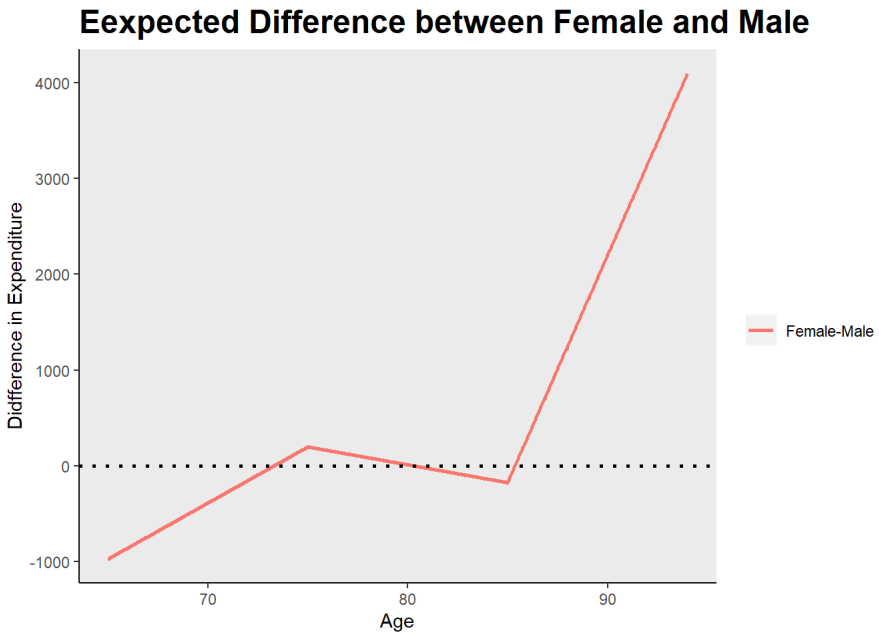
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



3. Using the model fit in Step 1 above, make a plot of the expected difference between women and men in expenditures as a function of age. Add a horizontal line at 0. Note that this difference is a simple function of the estimated coefficients from the model. (Hint: Start by writing out the regression model for females and males, both will be a function of age and the regression coefficients. Then take the difference and plug in the estimated regression coefficients and allow age to range from 65 to 94.)

```
## The function calculates the difference as male-female
## Later I use -diff_val(age) to obtain the value of female-male
betas = coef(model1)
diff_val <- function(age){
    sapply(age, function(age){
        ydiff=numeric()
    if (age<75){
        ydiff = -betas['female'] -betas['agem65:female']*(age-65)
    }else if (age<85){
        age75 = ifelse(age-75>0, age-75,0)
        ydiff = -betas['female']-betas['agem65:female']*(age-65) -betas['age_sp1:female']*(age75)
    }else if(age>=85){
        age75 = ifelse(age-75>0, age-75,0)
        age85 = ifelse(age-85>0, age-85,0)
        ydiff = -betas['female']-betas['agem65:female']*(age-65) -betas['age_sp1:female']*(age75) - betas['age_sp2:female']*
(age85)
    }
    return(as.numeric(ydiff))
})
}
```

```
xdiff = seq(65,94, length.out = 1000)
dat <- data.frame(xdfif = xdiff)
dat$ydiff= diff_val(xdiff)
ggplot(data = dat) +
geom_line(aes(x = xdiff, y = -ydiff, color = "Female-Male"), size = 1)+
geom_hline(yintercept=0, color = "black", size=1, linetype="dotted")+
ggtitle('Eexpected Difference between Female and Male')+
ylab('Didfference in Expenditure')+
xlab('Age')+
labs(color = "")+
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        plot.title = element_text(size = 18, face = "bold"))
```



4. Use the appropriate linear combination of regression coefficients to calculate the estimated difference between females and males in average expenditures and its standard error at ages 65, 75 and 85 years. Complete the table below.

```
## age at 65
LMSE65 = 614.70
coef(model1)['female'];confint(model1)['female',]
```

```
##    female
## -974.5175
```

```
##      2.5 %     97.5 %
## -2179.5662   230.5313
```

```
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
## age at 75
df = summary(model1)$df[2] ## df = 5686
summary(glht(model1, linfct = "`female`+ 10*`agem65:female`=0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = totalexp ~ ((agem65 + age_sp1 + age_sp2) + female +
##     female * (agem65 + age_sp1 + age_sp2)), data = d)
##
## Linear Hypotheses:
##                                   Estimate Std. Error t value Pr(>|t|)
## female + 10 * `agem65:female` == 0    197.7      572.0   0.346     0.73
## (Adjusted p values reported -- single-step method)
```

```
LMSE75 = 572.0
CI75 =  197.7 + c(-1,1)*(qt(0.975,df))*572.0
CI75
```

```
## [1] -923.6381 1319.0381
```

```
## age at 85
df = summary(model1)$df[2] ## df = 5686
summary(glht(model1, linfct = "`female`+ 20*`agem65:female` + 10*`age_sp1:female`=0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = totalexp ~ ((agem65 + age_sp1 + age_sp2) + female +
##     female * (agem65 + age_sp1 + age_sp2)), data = d)
##
## Linear Hypotheses:
##                                                        Estimate Std. Error
## female + 20 * `agem65:female` + 10 * `age_sp1:female` == 0   -176.1      948.0
##                                                        t value Pr(>|t|)
## female + 20 * `agem65:female` + 10 * `age_sp1:female` == 0  -0.186    0.853
## (Adjusted p values reported -- single-step method)
```

```
LMSE85 = 948.0
CI85 = -176.1 + c(-1,1)*(qt(0.975,df))*948.0
CI85
```

```
## [1] -2034.541  1682.341
```

Bootstrap Standard Error and 95% Confidence Interval

```
library(boot) ## Bootstrap using data set "d", model1
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:survival':
##
##     aml
```

```
exp.est <- function(data, id){
    dt <- data[id,] # allows boot to select sample
    fit1 <- lm(formula = totalexp ~ ((agem65 + age_sp1 + age_sp2) + female + female * (agem65 + age_sp1 + age_sp2)), data =
 dt)
    diff65 <- coef(fit1)[5]
    diff75 <- coef(fit1)[5] + 10*coef(fit1)[6]
    diff85 <- coef(fit1)[5] + 20*coef(fit1)[6] + 10*coef(fit1)[7]
    ratio <- 1 + coef(fit1)[5]/coef(fit1)[1]
    c(diff65,diff75,diff85,ratio)
}

set.seed(520)
exp.result <- boot(d, exp.est, R=11000)
exp.result
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = d, statistic = exp.est, R = 11000)
##
##
## Bootstrap Statistics :
##         original      bias     std. error
## t1* -974.5174711  3.09035155  587.1540521
## t2*  197.7029449 -5.25251182  577.4572612
## t3* -176.1208926 12.43827593 1086.9514538
## t4*    0.7657065  0.01023329    0.1247678
```

```
# The element $t contains a total number of R values (i.e. number of bootstrap samples) for each statistic generated by the
 bootstrap procedure:
head(exp.result$t)
```

```
##           [,1]      [,2]       [,3]      [,4]
## [1,]  -810.6234 1075.1126  -289.1875 0.8004511
## [2,] -1263.8750 1183.7965  -663.2746 0.6978618
## [3,] -1085.0706 -315.7258  1436.3726 0.7452094
## [4,]  -475.3900  659.8262  -665.6100 0.8855904
## [5,] -1611.8564  960.3684 -2357.9717 0.6787147
## [6,] -2422.9755  828.6640   325.6920 0.5365114
```

```
# Using bca is too slow and can not be finished within 1hr, so I choose type = "norm"
exp.bootci <- sapply(1:4,function(x) boot.ci(exp.result,index = x,type = "bca")$bca[4:5])
exp.boot.result <- data.frame(rbind(exp.result$t0,exp.bootci))
rownames(exp.boot.result) <- c("Est","Lower","Upper")
colnames(exp.boot.result) <- c("diff65","diff75","diff85","ratio")
round(exp.boot.result, 3)
```

```
##          diff65   diff75    diff85 ratio
## Est    -974.517  197.703  -176.121 0.766
## Lower -2181.420 -956.731 -2438.179 0.561
## Upper   143.920 1313.500  1855.661 1.050
```

Age	Est. Diff	LM Std.Error	LM 95% CI	BS Diff	BS Std.Err	BS 95% CI
65	-974.5174711	614.7	$[-2179.57, 230.53]$	-974.517	587.15	$[-2181.42, 143.92]$
75	197.7029449	572.0	$[-923.64, 1319.04]$	197.703	577.46	$[-956.73, 1313.50]$
85	-176.1208926	948.0	$[-2034.54, 1682.34]$	-176.121	1086.95	$[-2438.179, 1855.66]$

	Estimation	LM Std.Error	LM 95% CI	BS ratio	BS Std.Err	BS 95% CI
ratio	0.766	0.128	$[0.514, 1.017]$	0.766	0.125	$[0.561, 1.050]$

5. Now estimate the ratio of the average expenditures comparing women to men 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.

for female aged 65, estimated average expenditure = $\beta_0 + \beta_4$

for male aged 65, estimated average expenditure = β_0

$$g(\vec{\beta}) = \frac{\beta_0 + \beta_4}{\beta_0} = 1 + \frac{\beta_4}{\beta_0}$$

$$\nabla = \begin{bmatrix} -\frac{\beta_4}{\beta_0^2} \\ 0 \\ 0 \\ 0 \\ \frac{1}{\beta_0} \\ \cdots \\ 0 \end{bmatrix}$$

https://www.physicsread.com/latex-gradient-operator/ (https://www.physicsread.com/latex-gradient-operator/)

```
reg.vc = vcov(model1)
# Compute the estimate of g(beta)
g.est = 1 + coef(model1)[5]/coef(model1)[1]
as.numeric(g.est)
```

```
## [1] 0.7657065
```

```
# Define the vector of the derivative of g(beta) wrt beta
g.prime <- matrix(c(-betas['female']/betas[1]^2,0,0,0,1/betas[1],0,0,0),ncol = 1)
g.prime
```

```
##              [,1]
## [1,] 5.632887e-05
## [2,] 0.000000e+00
## [3,] 0.000000e+00
## [4,] 0.000000e+00
## [5,] 2.404201e-04
## [6,] 0.000000e+00
## [7,] 0.000000e+00
## [8,] 0.000000e+00
```

```
# Compute the variance of g(beta.hat)
g.var =  t(g.prime) %*% reg.vc %*% g.prime
# Compute the standard error
sqrt(g.var)
```

```
##           [,1]
## [1,] 0.1283814
```

```
# Compute the 95% CI for g(beta)
g.est + c(-1,1)*qt(0.975,df=summary(model1)$df[2]) * sqrt(g.var)
```

```
## Warning in c(-1, 1) * qt(0.975, df = summary(model1)$df[2]) * sqrt(g.var): Recycling array of length 1 in vector-array ar
ithmetic is deprecated.
##   Use c() or as.vector() instead.
```

```
## [1] 0.5140299 1.0173830
```

The point estimate of the ratio of the average expenditures comparing women to men at age 65 is 0.766, The standard error is 0.128, and the 95% CI is $[0.51, 1.02]$.

6. The data used in this regression are highly skewed and heteroscedastic (unequal variances across observations). Hence, the assumptions of the linear regression are not consistent with patterns in the data. As you will learn shortly, the estimates are still unbiased, but the standard errors and confidence intervals are likely biased. Hence, your inferences (tests and CIs) that depend on both the mean and variance estimates may be incorrect.

To check, use the bootstrap procedure to estimate the standard errors and confidence intervals for the differences in the table in Question 4 and for the ratio in Question 5. Compare the results obtained directly from the linear regression with those obtained using bootstrapping.

The bootstrap steps, results and table are all listed in *Question 4*.
Compare linear combination of coefficients: When R is large (i.e. >= 11,000), the point estimate from bootstrapping and linear regression are almost identical, but the 95% confidence intervals are somewhat different.For example, when age = 85, the 95% CI for the differences $[-2034.54, 1682.34]$ from linear regression, and $[-2438.179, 1855.66]$ from bootstrapping. When age =85, bootstrapping seems to be more accurately reflect the original expenditure distribution which is highly right-skewed. When age =65/76, the 95% CI are very similar. For value obtained from linear regression, since the assumptions of the linear regression are not consistent with patterns in the data, the point estimates is unbiased, but the standard errors and confidence intervals are likely biased.

For more detailed of comparisons, please refer to the table in *Question4*.

Ratio:
From bootstrap: The point estimate of the ratio of the average expenditures comparing women to men at age 65 is 0.766, The standard error is 0.125, and the 95% CI is $[0.561, 1.050]$.
From linea regression: The point estimate of the ratio of the average expenditures comparing women to men at age 65 is 0.766, The standard error is 0.125, and the 95% CI is $[0.514, 1.017]$.
The point estimate and 95% CI from the two methods are almost identical.

6. The data used in this regression are highly skewed and heteroscedastic (unequal variances across observations). Hence, the assumptions of the linear regression are not consistent with patterns in the data. As you will learn shortly, the estimates are still unbiased, but the standard errors and confidence intervals are likely biased. Hence, your inferences (tests and CIs) that depend on both the mean and variance estimates may be incorrect.

To check, use the bootstrap procedure to estimate the standard errors and confidence intervals for the differences in the table in Question 4 and for the ratio in Question 5. Compare the results obtained directly from the linear regression with those obtained using bootstrapping.

The bootstrap steps, results and table are all listed in *Question 4*.
Compare linear combination of coefficients: When R is large (i.e. >= 11,000), the point estimate from bootstrapping and linear regression are almost identical, but the 95% confidence intervals are somewhat different.For example, when age = 85, the 95% CI for the differences $[-2034.54, 1682.34]$ from linear regression, and $[-2438.179, 1855.66]$ from bootstrapping. When age =85, bootstrapping seems to be more accurately reflect the original expenditure distribution which is highly right-skewed. When age =65/76, the 95% CI are very similar. For value obtained from linear regression, since the assumptions of the linear regression are not consistent with patterns in the data, the point estimates is unbiased, but the standard errors and confidence intervals are likely biased.

For more detailed of comparisons, please refer to the table in *Question4*.

Ratio:
From bootstrap: The point estimate of the ratio of the average expenditures comparing women to men at age 65 is 0.766, The standard error is 0.125, and the 95% CI is $[0.561, 1.050]$.
From linea regression: The point estimate of the ratio of the average expenditures comparing women to men at age 65 is 0.766, The standard error is 0.125, and the 95% CI is $[0.514, 1.017]$.
The point estimate and 95% CI from the two methods are almost identical.

Test the null hypothesis that on average, males and females use the same quantity of medical services; i.e. are the mean expenditures at any age the same for males and females? Use a likelihood ratio test performed by fitting a null and extended model and comparing the change in –2log likelihood to the appropriate chi-square statistic. In addition, perform an F-test for the same null hypothesis. Write a sentence or two that summarizes what you learned about the medical expenditures and age from this test and the similarity/difference of the two tests.

Null model: $E[Y] = \beta_0 + \beta_1(age - 65) + \beta_2(age - 75)^+ + \beta_3(age - 85)^+ + \epsilon$
Extended model (Model1):
$E[Y] = \beta_0 + \beta_1(age - 65) + \beta_2(age - 75)^+ + \beta_3(age - 85)^+ + \beta_4 * female + \beta_5(age - 65) * female + \beta_6(age - 75)^+ * female + \beta_7(age - 85)^+ * female$
Another Extended model (Model2): $E[Y] = \beta_0 + \beta_1(age - 65) + \beta_2(age - 75)^+ + \beta_3(age - 85)^+ + \beta_4 * female + \epsilon$

```
# I have alreaday fitted the extended model previously and named it model1
model0 = lm(totalexp~ ((agem65 + age_sp1 + age_sp2)), data = d)
summary(model0)
```

```
##
## Call:
## lm(formula = totalexp ~ ((agem65 + age_sp1 + age_sp2)), data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7485  -3909  -3215  -1026 171268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3582.73     301.98  11.864   <2e-16 ***
## agem65         81.69      48.17   1.696   0.0899 .
## age_sp1        40.25      95.99   0.419   0.6750
## age_sp2        85.35     193.85   0.440   0.6597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10560 on 5690 degrees of freedom
## Multiple R-squared:  0.004917,   Adjusted R-squared:  0.004392
## F-statistic: 9.371 on 3 and 5690 DF,  p-value: 3.551e-06
```

```
model2 <- lm(totalexp~ ((agem65 + age_sp1 + age_sp2) + female) , data = d)
summary(model2)
```

```
##
## Call:
## lm(formula = totalexp ~ ((agem65 + age_sp1 + age_sp2) + female),
##     data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7612  -3899  -3203  -1015 171161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3690.80     345.46  10.684   <2e-16 ***
## agem65         81.48      48.17   1.691   0.0908 .
## age_sp1        42.13      96.04   0.439   0.6610
## age_sp2        84.25     193.87   0.435   0.6639
## female       -184.17     285.88  -0.644   0.5195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10560 on 5689 degrees of freedom
## Multiple R-squared:  0.004989,   Adjusted R-squared:  0.00429
## F-statistic: 7.131 on 4 and 5689 DF,  p-value: 1.006e-05
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
#Likelihood ratio test
lrtest(model2, model1) ## adding the female*age interaction terms
```

```
## Likelihood ratio test
##
## Model 1: totalexp ~ ((agem65 + age_sp1 + age_sp2) + female)
## Model 2: totalexp ~ ((agem65 + age_sp1 + age_sp2) + female + female *
##     (agem65 + age_sp1 + age_sp2))
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 -60831
## 2   9 -60828  3 4.7426     0.1916
```

```
library(lmtest)
#Likelihood ratio test
lrtest(model0, model1)
```

```
## Likelihood ratio test
##
## Model 1: totalexp ~ ((agem65 + age_sp1 + age_sp2))
## Model 2: totalexp ~ ((agem65 + age_sp1 + age_sp2) + female + female *
##     (agem65 + age_sp1 + age_sp2))
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   5 -60831
## 2   9 -60828  4 5.158     0.2715
```

```
#F test
anova(model2, model1)## adding the female*age interaction terms
```

```
## Analysis of Variance Table
##
## Model 1: totalexp ~ ((agem65 + age_sp1 + age_sp2) + female)
## Model 2: totalexp ~ ((agem65 + age_sp1 + age_sp2) + female + female *
##     (agem65 + age_sp1 + age_sp2))
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   5689 6.3434e+11
## 2   5686 6.3381e+11  3 528132711 1.5793 0.1921
```

```
#F test
anova(model0, model1)
```

```
## Analysis of Variance Table
##
## Model 1: totalexp ~ ((agem65 + age_sp1 + age_sp2))
## Model 2: totalexp ~ ((agem65 + age_sp1 + age_sp2) + female + female *
##     (agem65 + age_sp1 + age_sp2))
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   5690 6.3439e+11
## 2   5686 6.3381e+11  4 574407318 1.2883 0.2721
```

Take the comparison between model1 to model0 for an example, the results of Log likelihood test and F test are almost the same. P-value of Log likelihood test and of F test are all greater than 0.05. All other p-value of comparisons are > 0.05. According these p-values, we would fail to reject the null hypothesis that the interaction terms and $\beta_4(female)$ are not needed. We have no evidence that the extended model are more superior than the null model, suggesting same amount expenditure between female and males.

8. Using the results of Questions 1-7, write a brief report with sections: Background and objective, data, methods, results, discussion as if for a health services journal. NOTE: The data section should briefly (in a sentence or two) describe the data source (e.g. 1987 NMES). Recall the question: Do older males and females of the same age use roughly the same quantity of medical services.

1987 NMES is the data from the 1987 U.S. National Medical Expenditure Survey, and its key variables in this data set are annual medical expenditures, disease status, current smoker or not, race, etc. For this problem set, we are focusing on age, gender, and expenditure. I analyze persons aged 65 years and above to address the question of whether older men and women of the same age use roughly the same quantity of medical services.

First, multiple linear regression is used to address the question. Spline terms and interaction terms are used.
$E[Y] = \beta_0 + \beta_1(age - 65) + \beta_2(age - 75)^+ + \beta_3(age - 85)^+ + \beta_4 * female + \beta_5(age - 65) * female + \beta_6(age - 75)^+ * female + \beta_7(age -$
We find that the 95% confidence interval of β_4, β_5 and β_7 all overlap 0, suggesting no significant difference in female and male at certain age, and no statistical significant interactions between female gender and age categories. I plot the original data and the fitted value, no apparent difference between female expenditure and male expenditure at a given age could be seen.

Second, we estimate the average expenditure difference between female and male at the age of 65, 75 and 85, using both linear regression method and bootstrap. The results obtained from these two methods are almost identical: No statistical significant difference in health expenditure between female and male at 65, 75 or 85. We also calculate ratio of the average expenditures between female and male at age 65, using both linear regression method and bootstrap. When calculate the ratio from linear regression, delta method is applied. Both two methods show that the ratio is not significantly different form 1.

Third, we fit the data to a simpler model and named it 'the Null Model (Model0)'
$E[Y] = \beta_0 + \beta_1(age - 65) + \beta_2(age - 75)^+ + \beta_3(age - 85)^+ + \epsilon$. The original model is named 'the Extended Model (Model1)'. We perform Log likelihood test and F test to these two models. The p-value of Log likelihood test is 0.2715 and p-value of F test is 0.2721. According these p-values, we fail to reject the null hypothesis that β_4, β_5 and β_7 are all equal to 0. Another extended model is fit to the data (Model2):
$E[Y] = \beta_0 + \beta_1(age - 65) + \beta_2(age - 75)^+ + \beta_3(age - 85)^+ + \beta_4 * female + \epsilon$. Then I compare these models using Log likelihood test and F test, the p-values are greater than 0.05. In conclusion, there is no evidence that the extended model are more superior than the null model (the interaction terms and $\beta_4(female)$ are not needed), adding evidence to the conclusion that female expenditure are not significantly different from male expenditure at a given age.

Conclusion and Discussion: Males and females of the same age use roughly the same quantity of medical services (same expenditure). For detailed discussion please refer to the previous paragraph.