# Homework 2 Solution

```
options(digits=2)
library("knitr")
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
```

```
library(boot)
library(ggplot2)
opts_chunk$set(tidy=TRUE)
set.seed(108)
```

## I. Matrix Representation of Multiple Linear Regression

Use the following 5 observations to estimate the models below.

$X: 1.0, 3.0, 5.0, 7.0, 9.0$ $Y: -0.1, 2.9, 6.2, 7.3, 10.7$

### 1.1 Write the simple linear regression model in matrix terms.

The linear regression model can be represented as follows

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

where $\mathbf{Y} = \begin{bmatrix} -0.1 \\ 2.9 \\ 6.2 \\ 7.3 \\ 10.7 \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 3.0 \\ 1.0 & 5.0 \\ 1.0 & 7.0 \\ 1.0 & 9.0 \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$,

and $I$ is a $5 \times 5$ identity matrix.

Based on the least squares calculations in matrix notiation, the following results are:

$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 25 \\ 25 & 165 \end{bmatrix}$, $(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{165 \times 5 - 25^2} \begin{bmatrix} 165 & -25 \\ -25 & 5 \end{bmatrix}$, $\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 27 \\ 187 \end{bmatrix}$

And the estimated coefficients are:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \begin{bmatrix} -1.1 \\ 1.3 \end{bmatrix}$$

**1.2 Write an R function that takes the vector Y and matrix X as input then calculates and returns each the following components:**

- the least squares estimates of the regression coefficients;
- the variance-covariance matrix of the least squares estimates;
- the correlation between the two regression coefficients;
- the vector of predicted values $X(X'X)^{-1}X'Y = HY$;
- the vector of residuals $(I - X(X'X)^{-1}X')Y = (I - H)Y$.

```
regression1.2 = function(Y, X){
  n = length(Y)
  # Create design matrix
  Xmat = cbind(rep(1,n),X)
  # Solve for betas
  beta_hat = solve(t(Xmat) %*% Xmat) %*% t(Xmat) %*% Y
  # Compute the Hat matrix
  H = Xmat %*% solve(t(Xmat)%*%Xmat) %*%t(Xmat)
  # Get the predicted values
  Y.pred = H%*%Y
  # Compute the residuals
  res = (diag(n)-H)%*%Y
  # Estimate sigma2
  sigma2 = crossprod(res)/(n-ncol(Xmat))
  # Compute the variance matrix for beta-hat
  vcov.beta = as.numeric(sigma2)*solve(crossprod(Xmat))
  # Compute the correlation for beta-hat
  corr.beta = vcov.beta[1,2]/sqrt(vcov.beta[1,1]*vcov.beta[2,2])

  return(list(
    beta.hat = beta_hat,
    var_cov = vcov.beta,
    corr.beta = corr.beta,
    Y.pred = Y.pred,
    resid = res
  ))
}
```

**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.**

```
X=c(1,3,5,7,9)
Y=c(-0.1, 2.9, 6.2, 7.3, 10.7)
regression1.2(Y, X)
```

```
## $beta.hat
##   [,1]
##   -1.1
## X  1.3
##
## $var_cov
##               X
##    0.341 -0.052
## X -0.052  0.010
##
```

```
## $corr.beta
## [1] -0.87
##
## $Y.pred
##      [,1]
## [1,]  0.2
## [2,]  2.8
## [3,]  5.4
## [4,]  8.0
## [5,] 10.6
##
## $resid
##      [,1]
## [1,] -0.3
## [2,]  0.1
## [3,]  0.8
## [4,] -0.7
## [5,]  0.1
```

The 95% confidence interval for the slope is

$$\hat{\beta}_1 \pm t_{0.975,n-2}\hat{se}(\hat{\beta}_1) = (1.3 \pm 3.182 * \sqrt{0.010}) = (0.976, 1.618)$$

**1.4 Suppose you have conducted a randomized controlled trial of an intervention (TRT = 1) vs. placebo (TRT = 0), where n1 and n0 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:**

**a**

$$Y_i = B_0 + B_1 X_i + \epsilon_i,\ \epsilon_i\ i.i.d. N(0, \sigma^2), X_i = 1 \text{ if TRT} = 1,\ X_i = 0 \text{ if TRT} = 0$$

Write out the model above using matrix notation and then using matrix calculations solve for B_0 and B_1. HINT: The estimate of the intercept should be the sample mean in the placebo arm and estimate of the slope should be the difference in the sample means comparing the intervention and control groups. I.E. you will show that the model above is the same as conducting a two-sample t-test, assuming the same variance in each group.

Let

$$\mathbf{Y} = \mathbf{X}B + \epsilon$$

where $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \dots \\ Y_{n_0} \\ Y_{n_0+1} \\ \dots \\ Y_{n_1+n_0} \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & \dots \\ 1 & 0 \\ 1 & 1 \\ 1 & \dots \\ 1 & 1 \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$.

The least squares estimator of the coefficient is

$$\hat{\beta} = \begin{bmatrix} \hat{B}_0 \\ \hat{B}_1 \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \begin{bmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n_0} y_i^0 + \sum_{i=1}^{n_1} y_i^1 \\ \sum_{i=1}^{n_1} y_i^1 \end{bmatrix} = \begin{bmatrix} \bar{y}^0 \\ \bar{y}^1 - \bar{y}^0 \end{bmatrix}$$

Which leads to the estimation of $\hat{B}_1$ is $\bar{y}^1 - \bar{y}^0$.

Next we compute the covariance matrix:

$$\begin{bmatrix} \mathrm{Var}(\hat{B}_0) & \mathrm{Cov}(\hat{B}_0, \hat{B}_1) \\ \mathrm{Cov}(\hat{B}_0, \hat{B}_1) & \mathrm{Var}(\hat{B}_1) \end{bmatrix} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{n_0 n_1}\begin{bmatrix} n_1 & -n_1 \\ -n_1 & n_0+n_1 \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{n_0} & -\frac{\sigma^2}{n_0} \\ -\frac{\sigma^2}{n_0} & \sigma^2(\frac{1}{n_0}+\frac{1}{n_1}) \end{bmatrix}$$

Thus the variance of $\hat{B}_1$ is $\sigma^2(1/n_0 + 1/n_1)$.

$\sigma^2$ Can be estimated in the following formula: $\hat{\sigma^2} = \frac{1}{n-\rho-1}\sum_{i=1}^{n}(\mathbf{Y_i} - \mathbf{X_{i*}}\hat{\beta})^2$. Here $X_{i*}$ is the $i$th row of $\mathbf{X}$. $n$ is number of observations ($n_0 + n_1$ here), $\rho$ is number of covariates (1 here).

If $X_{i*} = \begin{bmatrix} 1 & 0 \end{bmatrix}$, then $X_{i*}\hat{\beta} = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} \bar{y}^0 \\ \bar{y}^1 - \bar{y}^0 \end{bmatrix} = \bar{y}^0$, we have $n_0$ such $X_{i*}$.

Otherwise, if $X_{i*} = \begin{bmatrix} 1 & 1 \end{bmatrix}$, then $X_{i*}\hat{\beta} = \begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} \bar{y}^0 \\ \bar{y}^1 - \bar{y}^0 \end{bmatrix} = \bar{y}^1$, we have $n_1$ such $X_{i*}$.

Thus $\hat{\sigma^2} = \frac{1}{n-\rho-1}\sum_{i=1}^{n}(\mathbf{Y_i} - \mathbf{X_{i*}}\hat{\beta})^2 = \frac{1}{n_1+n_2-2}(\sum_{i=1}^{n_0}(y_i - \bar{y}^0)^2 + \sum_{i=n_0+1}^{n_0+n_1}(y_i - \bar{y}^1)^2)$.

To test the following hypothesis

$$H_0 : B_1 = 0,\ H_1 : B_1 \neq 0$$

The test statistic is

$$T = \frac{\bar{y}^1 - \bar{y}^0}{\sqrt{\hat{\sigma}^2(1/n_0 + 1/n_1)}}$$

where

$$\hat{\sigma}^2 = [\sum_{i=1}^{n_0}(y_i^0 - \bar{y}^0)^2 + \sum_{i=1}^{n_1}(y_i^1 - \bar{y}^1)^2]/(n_0 + n_1 - 2)$$

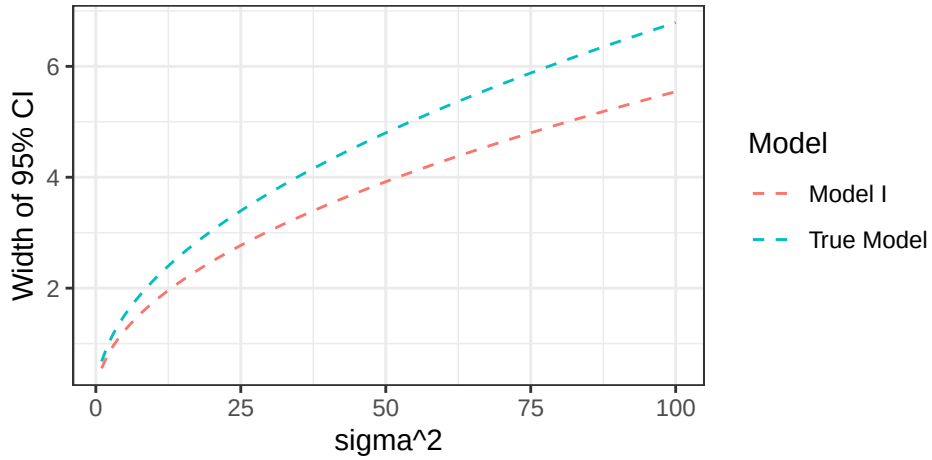
This is exactly the same as two-sample t-test under the assumption of the same variance in each group.

**b** The length for 95% confidence interval of $B_1$ based on model I will be $2\, z_{0.975}\sqrt{var(\hat{B}_1)} = 2\, z_{0.975}\sqrt{\sigma^2(1/n_0 + 1/n_1)}$. If we substitute $var(\hat{B}_1)$ with its true variance, the width is $2\, z_{0.975}\sqrt{\sigma^2(2/n_0 + 1/n_1)}$.

For demonstration purpose, here we assume $n_0 = n_1 = 100$. We plot the CI length based on Model I, or true model, versus different $\sigma^2$.

```
n0 = 100
n1 = 100
sigma2 = 1:100
CI.length = data.frame(sigma2 = sigma2,
                       l = 2*qnorm(0.975)*sqrt(sigma2 * (1/n0+1/n1)),
                       Model = rep('Model I',length(sigma2))
                       )

CI.length = rbind( CI.length,
                   data.frame(sigma2 = sigma2,
                              l = 2*qnorm(0.975)*sqrt(sigma2 * (2/n0+1/n1)),
                              Model = rep('True Model',length(sigma2))
                              )
                   )

ggplot(CI.length) +
```

```
    geom_line(aes(x = sigma2,y = l, col = as.factor(Model)),lty=2) +
    labs(y = "Width of 95% CI", x = "sigma^2") +
    theme_bw() +
    scale_colour_discrete(name="Model")
```




This means that if we were to incorrectly assume that the variance was the same in each group, we would be underestimating the width of the 95% CI for $B_1$. The result would be to more frequently than expected reject the null hypothesis that $B_1 = 0$ (i.e. we could find more frequently than expected that 0 is outside the bounds of the 95% confidence interval).

**1.5 Under the Gaussian multiple linear regression framework, write the log likelihood function for the regression coefficients and residual variance in matrix terms and derive the mle's for the regression coefficients. Derive their joint distribution as well as that of the predicted values and residuals.**

Under the Gaussian multiple linear regression framework, we have $\mathbf{Y} - \mathbf{X}\beta \sim N(0, \sigma^2 I)$. Then the log-likelihood is

$$l(\beta, \sigma^2) = -\frac{5}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)/\sigma^2$$

Set its first derivative equal zero, we have

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = 0$$

Then the mle of regression coefficient is

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

Since $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, then

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta, ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2\mathbf{I}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')') = N((\beta, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2).$$

Then

$$\hat{Y} = \mathbf{X}\beta \sim N(\mathbf{X}\beta, \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$$
$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'))$$

## II. Advanced Inferences for Linear Regression

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

```
load("nmes.rdata")
data.2=nmes[nmes$lastage>=65,]
```

**2.1 Fit a MLR of expenditures on:**

**age-65 + age_sp1 = (age- 75)+ + age_sp2=(age-85)+ + female (1-female; 0 - male) + female*( age-65 + age_sp1 + age_sp2). Write a short, scientific interpretation of the estimate (with confidence interval) for each of the coefficients in the model.**

```
data.2$agem65 = data.2$lastage - 65
data.2$age_sp1 = ifelse(data.2$lastage>75, data.2$lastage-75, 0)
data.2$age_sp2 = ifelse(data.2$lastage>85, data.2$lastage-85, 0)
data.2$female = 1-data.2$male

fit_2.1 = lm(totalexp ~ (agem65 + age_sp1 + age_sp2) * female, data=data.2)
summary(fit_2.1)
```

```
##
## Call:
## lm(formula = totalexp ~ (agem65 + age_sp1 + age_sp2) * female,
##     data = data.2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8573  -3997  -3180   -982 170901
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4159.4      473.4    8.79   <2e-16 ***
## agem65             11.9       75.8    0.16     0.88
## age_sp1           132.8      153.9    0.86     0.39
## age_sp2          -254.6      327.1   -0.78     0.44
## female           -974.5      614.7   -1.59     0.11
## agem65:female     117.2       98.2    1.19     0.23
## age_sp1:female   -154.6      197.0   -0.78     0.43
## age_sp2:female    512.5      406.3    1.26     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10600 on 5686 degrees of freedom
## Multiple R-squared:  0.00582,    Adjusted R-squared:  0.00459
## F-statistic: 4.75 on 7 and 5686 DF,  p-value: 2.44e-05
```

## Interpretations of coefficients

(Intercept): The average medical expenditure for 65-year old males is estimated to be 4159.39 (confident interval (3231.33, 5087.44)) dollars.

agem65: The average difference in medical expenditures comparing two males who differ in age by 1 year but are 65 to 75 years of age is is estimated to be 11.86 (confident interval (-136.8, 160.51)) dollars.

agesp1: The difference in the average annual increase in medical expenditures comparing males 75 to 85 years of age to males under 75 years of age is estimated to be 132.81 (confident interval (-168.84, 434.46)) dollars.

agesp2: The difference in the average annual increase in medical expenditures comparing males over 85 years

of age to males 75 to 85 years of age is estimated to be -254.6 (confident interval (-895.81, 386.61)) dollars.

female: The difference in average medical expenditures for a 65-year old female compared to a 65-year old male is -974.52 (confident interval (-2179.57, 230.53)) dollars.

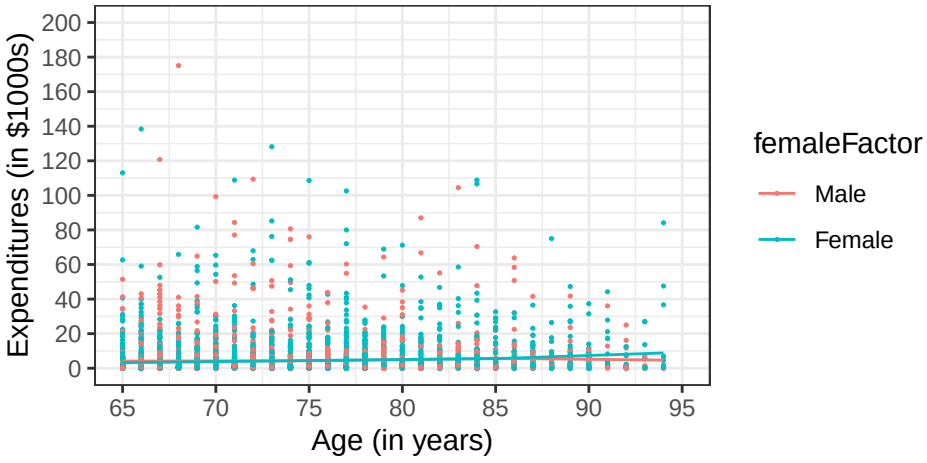
age65:female: The difference in average annual increase in medical expenditures comparing females to males younger than 75 years of age is estimated to be 117.22 (confident interval (-75.25, 309.7)) dollars.

agesp1:female: The difference between the female and male **additional** average annual increase in medical expenditures comparing persons 75 to 85 years of age compared to persons younger than 75 is estimated to be -154.6 (confident interval (-540.82, 231.61)) dollars.

agesp2:female: The difference between the female and male **additional** average annual increase in medical expenditures comparing persons over 85 years of age compared to persons 75 to 85 years of age is estimated to be 512.46 (confident interval (-284.04, 1308.96)) dollars.

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

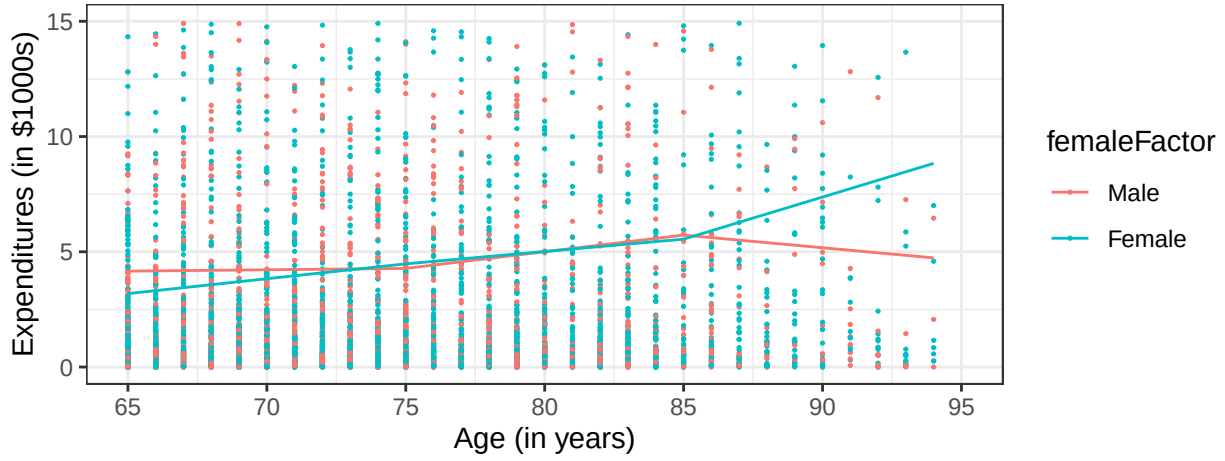
```
data.2$observed_y = data.2$totalexp / 1000
data.2$predicted_y = fit_2.1$fitted.values / 1000
data.2$femaleFactor = factor(data.2$female,levels=c(0,1),labels=c("Male","Female"))
max_y = max(data.2$observed_y, data.2$predicted_y)
min_y = min(data.2$observed_y, data.2$predicted_y)

ggplot(data.2, aes(x = lastage, y = observed_y,color=femaleFactor)) +
    geom_point(size=0.25) +
    geom_line(aes(x = lastage, y = predicted_y)) +
    theme_bw() +
    scale_y_continuous(breaks=seq(0,200,20),limits=c(0,200)) +
    scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95)) +
    labs(y = "Expenditures (in $1000s)", x = "Age (in years)")
```




We can also zoom in on the plot to show the average medical expenditures for females and males.

```
ggplot(data.2, aes(x = lastage, y = observed_y,color=femaleFactor)) +
    geom_point(size=0.25) +
    geom_line(aes(x = lastage, y = predicted_y)) +
    theme_bw() +
    scale_y_continuous(breaks=seq(0,15,5),limits=c(0,15)) +
```

```
    scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95)) +
    labs(y = "Expenditures (in $1000s)", x = "Age (in years)")
```




**2.3 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 this difference is a simple function of the estimated coefficients from the model.**

Based on this problem, the model we will fit is

$$Y = B_0 + B_1(age - 65) + B_2(age - 75)^+ + B_3(age - 85)^+ + B_4 female +$$

$$B_5 female(age - 65) + B_6 female(age - 75)^+ + B_7 female(age - 85)^+ + \epsilon$$

If we substitute $female = 0$ for male, and $female = 1$ for female, then:

$$E[Y|female = 1] = B_0 + B_1(age - 65) + B_2(age - 75)^+ + B_3(age - 85)^+ +$$

$$B_4 + B_5(age - 65) + B_6(age - 75)^+ + B_7(age - 85)^+ + \epsilon$$

$$E[Y|female = 0] = B_0 + B_1(age - 65) + B_2(age - 75)^+ + B_3(age - 85)^+ + \epsilon$$

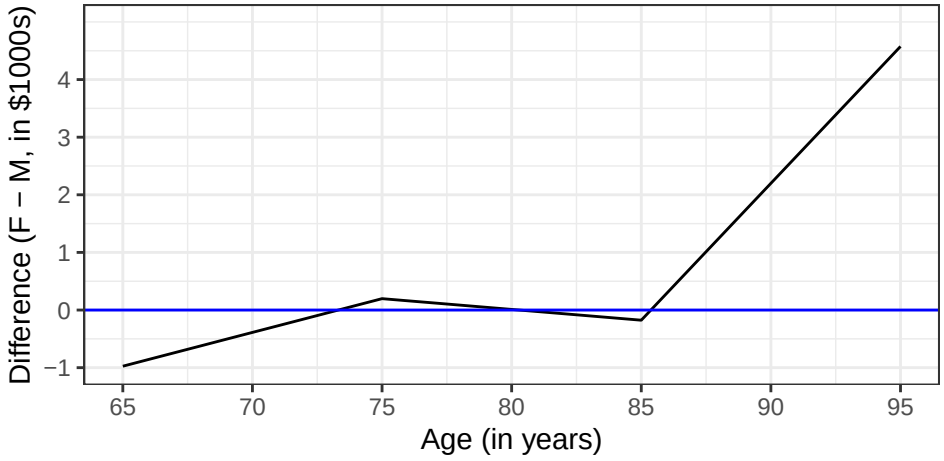
Then, the difference in medical expenditures comparing females to males is:

$$E[Y|female = 1] - E[Y|female = 0] = B_4 + B_5(age - 65) + B_6(age - 75)^+ + B_7(age - 85)^+$$

```
age = seq(65,95)
agesp1 = ifelse(age>=75,age-75,0)
agesp2 = ifelse(age>=85,age-85,0)
age65 = age-65
AGE = rbind(rep(1,31),age65,agesp1,agesp2)
expectDiff=t(AGE)%*%coefficients(fit_2.1)[5:8]
V = vcov(fit_2.1)[5:8,5:8]
V.big = t(AGE)%*%V%*%AGE
var = diag(V.big)
diffage = data.frame(age,
          expectDiff = expectDiff/1000,
```

```
          std = sqrt(var),
          min = (expectDiff-1.96*sqrt(var))/1000,
          max =(expectDiff+1.96*sqrt(var))/1000)

ggplot(diffage,aes(x = age,y = expectDiff)) +
    geom_line(data = diffage) +
    geom_hline(yintercept = 0,col = 'blue',data= diffage) +
    theme_bw() +
    scale_y_continuous(breaks=seq(-1,4,1),limits=c(-1,5)) +
    scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95)) +
    labs(y = "Difference (F - M, in $1000s)", x = "Age (in years)")
```

```
## Warning: geom_hline(): Ignoring `data` because `yintercept` was provided.
```

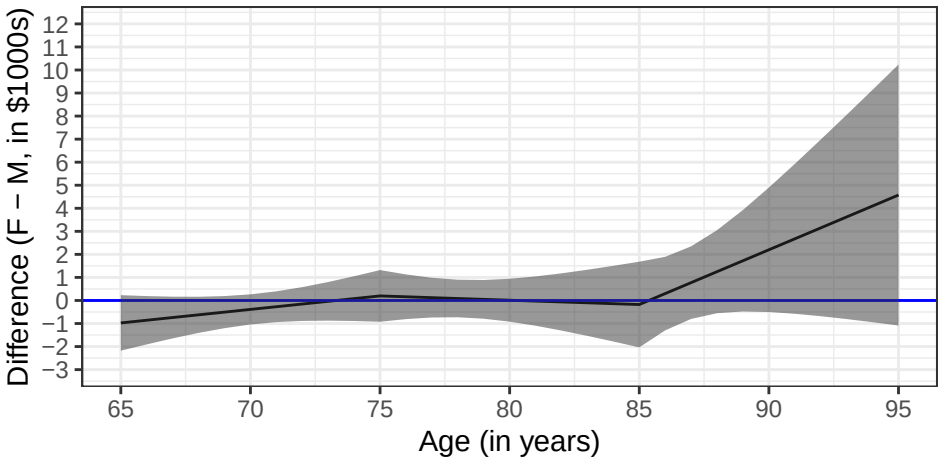



```
# +geom_ribbon(aes(ymin=min,ymax =max),alpha =0.5,data= diffage)
#conf =confint(fit_2.1)
```

You were not asked to do this but: in addition to the difference, it is also possible to add the 95% confidence interval for the curve, by adding the code `geom_ribbon(aes(ymin=min,ymax =max),alpha =0.5,data= diffage)`

```
ggplot(diffage,aes(x = age,y = expectDiff)) +
    geom_line(data = diffage) +
    geom_hline(yintercept = 0,col = 'blue',data= diffage) +
    geom_ribbon(aes(ymin=min,ymax =max),alpha =0.5,data= diffage) +
    theme_bw() +
    scale_y_continuous(breaks=seq(-3,12,1),limits=c(-3,12)) +
    scale_x_continuous(breaks=seq(65,95,5),limits=c(65,95)) +
    labs(y = "Difference (F - M, in $1000s)", x = "Age (in years)")
```

```
## Warning: geom_hline(): Ignoring `data` because `yintercept` was provided.
```




**2.4 Use the appropriate linear combination of regression coefficients to calculate the estimated difference between women and men in average expenditures and its standard error at 65, 75 and 85 years of age. Complete the table below. (Hint: Start out by first expressing the average expenditure for males and females at 65, 75 and 85 in terms of the regression model, and determine what function of the regression coefficients gives you the difference at each age) .**

From last question (Let us use $\beta_i$ to represent $B_i$ to be consistent with lecture notes)

$$E[Y|female = 1] - E[Y|female = 0] = \beta_4 + \beta_5(age - 65) + \beta_6(age - 75)^+ + \beta_7(age - 85)^+$$

Plug in age=65, we can obtain that the difference is $\beta_4$.

Plug in age=75, we can obtain that the difference is $\beta_4 + 10\beta_5$.

Plug in age=85, we can obtain that the difference is $\beta_4 + 20\beta_5 + 10\beta_6$.

To obtain the standard error of each linear combination, recall that $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$.

```
#age 65
diff65 = coefficients(fit_2.1)[5]/1000
se65 = sqrt(vcov(fit_2.1)[5,5])/1000
CI65 = diff65 + c(-1,1)*1.96*se65

#age 75
A = c(0,0,0,0,1,10,0,0)
diff75 = as.numeric(t(A)%*% coefficients(fit_2.1))/1000
se75 = sqrt(as.numeric(t(A)%*%vcov(fit_2.1)%*%A))/1000
CI75 = diff75 + c(-1,1)*1.96*se75

#age 85
A = c(0,0,0,0,1,20,10,0)
diff85 = as.numeric(t(A)%*%coefficients(fit_2.1))/1000
se85 = sqrt(as.numeric(t(A)%*%vcov(fit_2.1)%*%A))/1000
CI85 = diff85 + c(-1,1)*1.96*se85
```

The table below provides the estimated differences in $1000s comparing females to males at age 65, 75 and 85.

| Age | Estimated Difference Women-Men | Least Squares Std Error | Least Squares 95% CI | Bootstrap Std Error | Bootstrap 95% CI |
|---|---|---|---|---|---|
| 65 | −0.97 | 0.61 | [−2.18, 0.23] | | |
| 75 | 0.2 | 0.57 | [−0.92, 1.32] | | |
| 85 | −0.18 | 0.95 | [−2.03, 1.68] | | |

**2.5 Now estimate the ratio of 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.**

```
g.prime = matrix(
  c(-coefficients(fit_2.1)[1]^(-2)*coefficients(fit_2.1)[5],
    coefficients(fit_2.1)[1]^(-1)),nrow=2)
se_ratio=sqrt(t(g.prime)%*%vcov(fit_2.1)[c(1,5),c(1,5)]%*%g.prime)
```

At age 65, the expected medical expenditures is $E[Y|65, \text{male}] = \beta_0$ for males and $E[Y|65, \text{female}] = \beta_0 + \beta_4$ for females, and we would like to know the standard error and 95% CI of quantity $(\beta_0 + \beta_4)/\beta_0$. Let function

$$f(p, q) = (p + q)/p,$$

then the standard error of $f(\hat{\beta}_0, \hat{\beta}_4)$ could be written as

$$\text{SE}_{\text{ratio}} = \sqrt{\nabla f(\hat{\beta}_0, \hat{\beta}_4)^T Cov(\hat{\beta}_0, \hat{\beta}_4) \nabla f(\hat{\beta}_0, \hat{\beta}_4)},$$

where $\nabla f(\hat{\beta}_0, \hat{\beta}_4)^T = (-\hat{\beta}_4\hat{\beta}_0^{-2} \quad \hat{\beta}_0^{-1})$, $Cov(\hat{\beta}_0, \hat{\beta}_4)$ is the variance-covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_4$. Using R, we can know that $\text{SE}_{\text{ratio}} = 0.13$. And the 95% confidence interval of $(\beta_0 + \beta_4)/\beta_0$ is $[(\hat{\beta}_0 + \hat{\beta}_4)/\hat{\beta}_0 - 1.96 * \text{SE}_{\text{ratio}}, (\hat{\beta}_0 + \hat{\beta}_4)/\hat{\beta}_0 + 1.96 * \text{SE}_{\text{ratio}}] = [0.51, 1.02]$.

**2.6 The data used in this regression are highly skewed and heteroscedastic (look up this term if it is the first time you have seen it). Hence, the assumptions of the linear regression are not consistent with the data. As you will learn shortly, the estimates are still unbiased, but the standard errors and confidence intervals are not. Hence, your inferences may be incorrect. To check, use the bootstrap to estimate the standard errors and confidence intervals for the differences in the table in part 5 and for ratio in part 6. Compare the results.**

Based on what we have learnt during the lab, it is convenient to use the `boot` function

```
bt.est <- function(data, id){
  dt <- data[id, ]
  fit = lm(totalexp ~ (agem65 + age_sp1 + age_sp2) * female, dt)
  cc = coefficients(fit)
  c1 = cc[5]
  c2 = cc[5] + 10 * cc[6]
  c3 = cc[5] + 20 * cc[6] + 10 * cc[7]
  c4 = (cc[1]+cc[5])/cc[1]
  c(c1, c2, c3, c4)
}

result = boot(data.2, bt.est, 1000)
```

We can also use the help from `boot.ci` function with basic bootstrap

```
boot.perc.ci <- sapply(1:4, function(x) boot.ci(result, index = x,type = "basic")$basic[4:5])
boot.result <- data.frame(rbind(result$t0,boot.perc.ci,sqrt(apply(result$t,2,var)))/1000)

rownames(boot.result) <- c("Est","Lower","Upper", "Standard Errors")
colnames(boot.result) <- c('se65_bt','se75_bt','se85_bt', 'se65_ratio_bt')
boot.result
```

The table below provides the estimated differences in $1000s comparing females to males at age 65, 75 and 85. The confidence intervals are based on both the least squares solution and the bootstrap.

We can learn from the table that the bootstrap standard errors and bootstrap 95% CI are close to their ordinary least squares counterparts when age = 65 or 75, but are larger and wider for age 85.

For the ratio $(\beta_0+\beta_4)/\beta_0$, the standard error and 95% CI derived using the ordinary least squares assumption and the delta method are 0.13 and [0.51, 1.02], respectively. The bootstrap standard error is 0.12 and the bootstrap 95% CI is $[0.48, 0.97]$, which are more precise than their ordinary least squares counterparts.

| Age | Estimated Difference Women-Men | Least Squares Std Error | Least Squares 95% CI | Bootstrap Std Error | Bootstrap 95% CI |
|---|---|---|---|---|---|
| 65 | $-0.97$ | 0.61 | $[-2.18, 0.23]$ | 0.58 | [-2.14, 0.11] |
| 75 | 0.2 | 0.57 | $[-0.92, 1.32]$ | 0.57 | [-0.89, 1.36] |
| 85 | $-0.18$ | 0.95 | $[-2.03, 1.68]$ | 1.09 | [-2.27, 1.99] |

If you feel that it is not comfortable to use `boot` function, another choice is to write the bootstrap functions by yourself.

```
#bootstrap
par_bootstrap = array(0,c(1000,length(coefficients(fit_2.1))))
colnames(par_bootstrap) = names(coefficients(fit_2.1))
for (k in 1:1000)
{
  idx_rd=sample(1:nrow(data.2),size=nrow(data.2),replace=TRUE)
  data.2.4=data.2[idx_rd,]
  data.2.4$agem65 = data.2.4$lastage - 65
  data.2.4$age_sp1 = ifelse(data.2.4$lastage>75, data.2.4$lastage-75, 0)
  data.2.4$age_sp2 = ifelse(data.2.4$lastage>85, data.2.4$lastage-85, 0)
  data.2.4$female = 1-data.2.4$male
  fit_2.4 = lm(totalexp ~ (agem65 + age_sp1 + age_sp2) * female, data=data.2.4)
  par_bootstrap[k,]=coefficients(fit_2.4)
}

#age 65
se65_bt = sd(par_bootstrap[,5])/1000
CI65_bt = quantile(par_bootstrap[,5],probs=c(0.025,0.975))/1000

#age 75
se75_bt = sd(par_bootstrap[,5]+10*par_bootstrap[,6])/1000
CI75_bt = quantile(par_bootstrap[,5]+10*par_bootstrap[,6],probs=c(0.025,0.975))/1000

#age 85
se85_bt = sd(par_bootstrap[,5]+20*par_bootstrap[,6]+10*par_bootstrap[,7])/1000
CI85_bt = quantile(par_bootstrap[,5]+20*par_bootstrap[,6]+10*par_bootstrap[,7],probs=c(0.025,0.975))/100
```

```
#age 65 ratio
se65_ratio_bt = sd((par_bootstrap[,1]+par_bootstrap[,5])/par_bootstrap[,1])/1000
CI65_ratio_bt = quantile((par_bootstrap[,1]+par_bootstrap[,5])/par_bootstrap[,1],probs=c(0.025,0.975))/
```

Update the table in 2.5.

| Age | Estimated Difference Women-Men | Least Squares Std Error | Least Squares 95% CI | Bootstrap Std Error | Bootstrap 95% CI |
|---|---|---|---|---|---|
| 65 | −0.97 | 0.61 | [−2.18, 0.23] | 0.57 | [−2.05, 0.15] |
| 75 | 0.2 | 0.57 | [−0.92, 1.32] | 0.57 | [−0.91, 1.29] |
| 85 | −0.18 | 0.95 | [−2.03, 1.68] | 1.1 | [−2.31, 2.03] |

**2.7 Test the null hypothesis that mean expenditure is the same function of age for men and women. Use a likelihood ratio test performed by fitting a null and extended model and comparing the change in −2*log likelihood to the appropriate chi-square statistic. Now perform an F-test for the same null hypothesis. Write a sentence or two that summarizes what you learned about the relationship of medical expenditures to age from this test and the similarity/difference of the two tests.**

**Likelihood ratio test** The null hypothesis is that the mean expenditure is the same function fo age for men and women. This null hypothesis translates to a null model that includes only the main effects for the age variables. Let $L_0$ be the likelihood of the null model and $L_1$ be the likelihood of the full model, then

$$2\log L_1 - 2\log L_0 \sim \chi^2_{df_1 - df_0},$$

where $df_1$ is the number of explanatory variables (including intercept) in the full model and $df_0$ is the number of explanatory variables in the null model, where the models are nested (all variables in the null model also appear in the full model).

**F test** Use $RSS_0$ and $RSS_1$ to denote the residual sum of squares of the null model and alternative model, respectively. Then the $F$ statistic is

$$F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/(n - df_1)}.$$

```
fit_2.2_null=lm(totalexp ~ agem65 + age_sp1 + age_sp2, data=data.2)
logLik.null=logLik(fit_2.2_null)
logLik.null
```

```
## 'log Lik.' -60831 (df=5)
```

```
logLik.full=logLik(fit_2.1)
logLik.full
```

```
## 'log Lik.' -60828 (df=9)
```

```
#compute test statistic
Dev_2.2=as.numeric(2*logLik.full - 2*logLik.null)

#compute Pr(X>D),
# where X is Chi-squared with 4 df
pchisq(Dev_2.2, df=4, lower.tail=FALSE)
```

```
## [1] 0.27
```

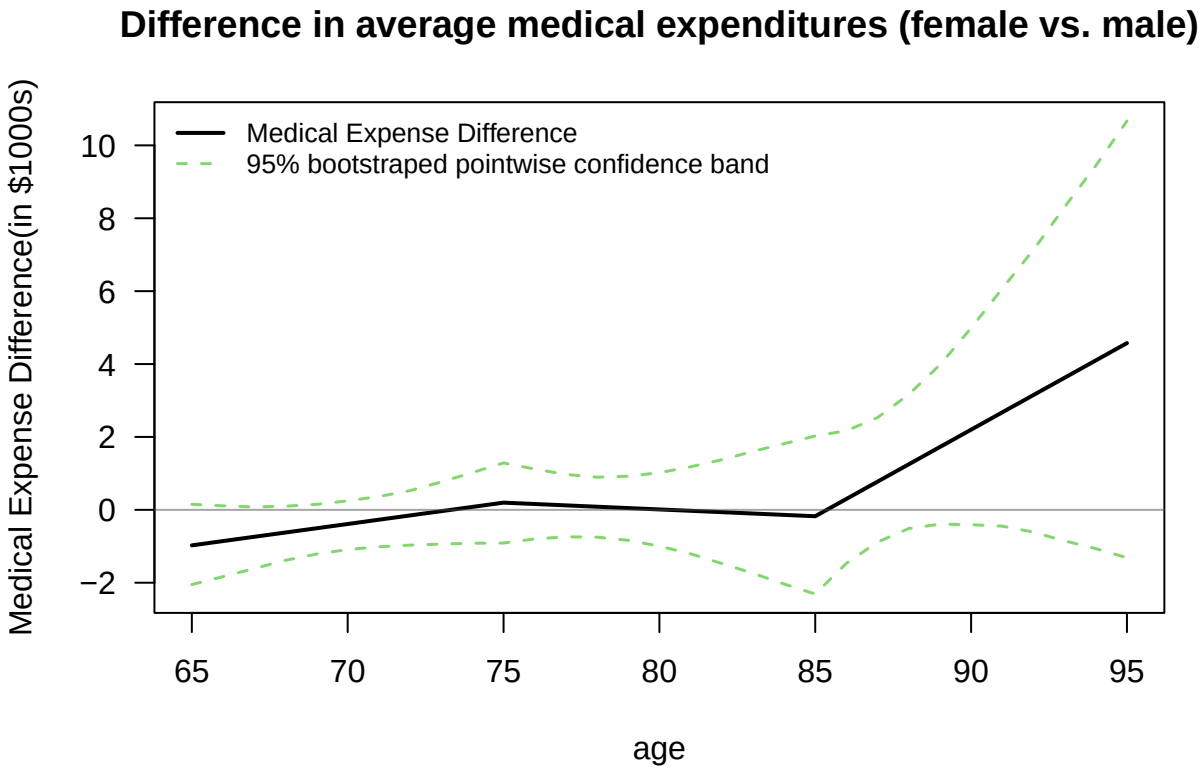
```
#Compute residual sum of squares
RSS0 = sum(residuals(fit_2.2_null)^2)
RSS1 = sum(residuals(fit_2.1)^2)
df_diff = fit_2.2_null$df.residual-fit_2.1$df.residual
df_alt = fit_2.1$df.residual
Fstat = ((RSS0-RSS1)/df_diff)/(RSS1/df_alt)
pf(Fstat, df_diff, df_alt, lower.tail = FALSE)
```

```
## [1] 0.27
```

```
anova(fit_2.2_null,fit_2.1)
```

According to the likelihood ratio test and the F test, we do not have sufficient evidence to reject the null hypothesis that males and females have the same relationship between age and medical expenses, with $p_{\chi^2} = 0.27$ and $p_F = 0.27$. Two tests give similar results.

**2.8 Using the results of 1-7, write a brief report with sections: objective, data, methods, results, summary as if for a health services journal.**




Objective: To determine whether older, i.e. at least 65 years of age, men and women of the same age use roughly the same quantity of medical services measured by their annual medical expenditures.

Data: Annual medical expenditures for men and women 65 to 95 years of age were obtained from the 1987 National Medical Expenditure Survey.

Methods: The average annual medical expenditures were modeled as a non-linear function of age (via a linear spline with knots at 75 and 85 years of age) allowed to be distinct for each sex (via statistical interaction terms). Using the fit of the linear regression model, the difference in the average annual medical expenditures

comparing females to males ages 65 to 95 was computed. Due to the positive skew in the distribution of annual medical expenditures, 95% confidence intervals for the differences were derived using the bootstrap method where 1000 bootstrap samples were drawn with replacement from the original sample and the percentile bootstrap method was used to compute the confidence interval.

Results: Figure 1 displays the estimated average difference in annual medical expenditures comparing females to males 65 to 95 years of age. Among 65 year olds, females have lower estimated average annual medical expenditures compared to males (estimated difference: 0.57 dollars, 95% bootstrap confidence interval: -2.05 to 0.15). Among 75 year olds, the estimated difference in annual medical expenditures comparing females to males is 0.57 (95% bootstrap confidence interval: [-0.91 to 1.29). After age 75, females are estimated to have higher average annual medical expenditures compared to men of the same age; for example, among 85 year olds, the estimated difference is 1.1 dollars (95% bootstrap confidence interval: -2.31 to 2.03) comparing females to males. However, the observed differences at each age were not found to be not statistically significant (p-value 0.27 based on likelihood ratio test comparing the model described above to a model that assumed the average annual medical expenditures could change with age, via linear spline with knots at 75 and 85 years of age, but were not different among males and females of the same age).

Summary: Based on data from the National Medical Expenditure Survey, we estimated that females have increasingly higher average annual medical expenditures as they age; however this difference did not reach statistical significance.