HW 03

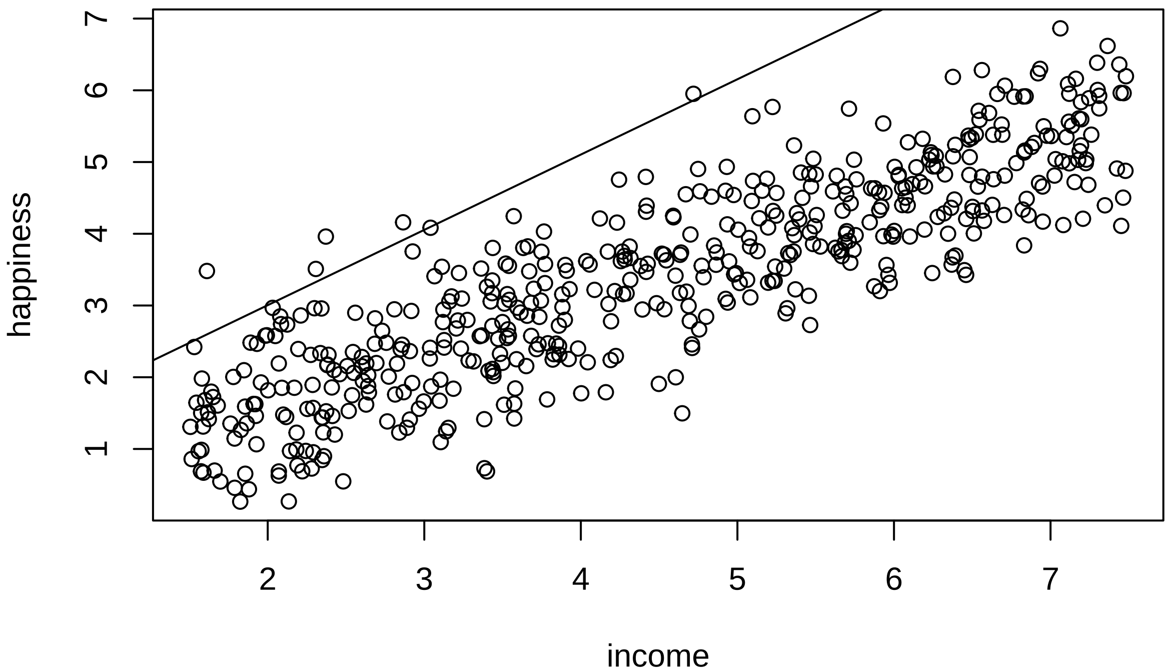
Tuan Bui

Question 01:

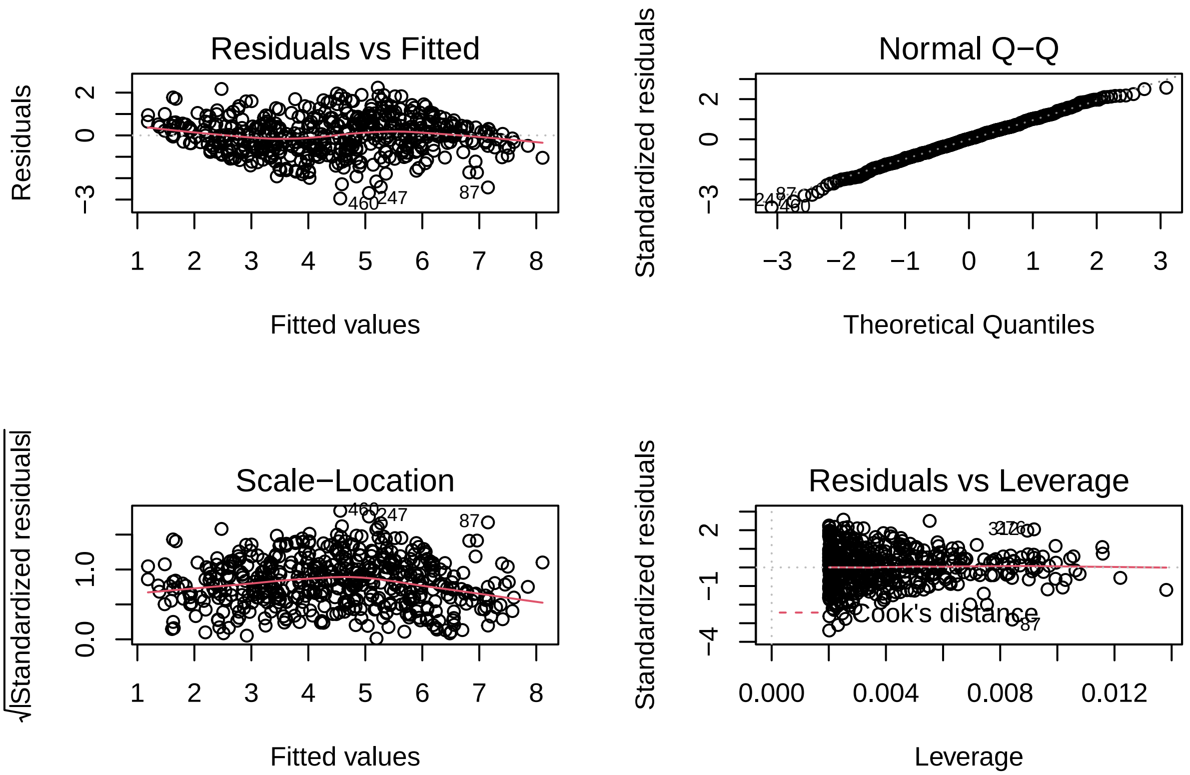
```
income_data <- read.csv('~/OneDrive - Stony Brook University/SBU/MAT + AMS/Fall 2021/AMS 380/hw/03/incom

attach(income_data)

fit <- lm(income ~ happiness)
fit
```

```
##
## Call:
## lm(formula = income ~ happiness)
##
## Coefficients:
## (Intercept)    happiness
##      0.9053       1.0497
```

```
# a. The least square regression line equation: income = 0.9053 + 1.0497 * happiness

# b. Plot the points and regression line in the same figure
plot(income, happiness)
abline(fit)
```




```
# c. Check assumptions:
par(mfrow = c(2,2))
plot(fit)
```




```
## 1. Linearity: it is satisfied because the residuals are symmetrically distributed around the 0-line

## 2. Homoscedasticity: it is satisfied because the square root of standardized residuals is symmetrica

## 3. Independence: assume it is satisfied

## 4. Normality:
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.99682, p-value = 0.4377
```

```
### p-value is 0.4377 greater than the significance level 0.05, so residuals is normal distributed, nor

# d. Sample correlation coefficient between the 2 variables:
cor(income, happiness)
```

```
## [1] 0.8656337
```

```
## Sample correlation coefficient is 0.8656337
```

```r
## The corresponding population correlation test:
cor.test(income, happiness)
```

```
##
##  Pearson's product-moment correlation
##
## data:  income and happiness
## t = 38.505, df = 496, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8417942 0.8861031
## sample estimates:
##       cor
## 0.8656337
```

```r
### p-value is less than 2.2e-16, which is less than the significance level 0.05, reject H0. The correl

# e.
summary(fit)
```

```
##
## Call:
## lm(formula = income ~ happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.94796 -0.57730  0.02277  0.55661  2.23185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.90533    0.10039   9.018   <2e-16 ***
## happiness    1.04973    0.02726  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8708 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

```r
## The coefficient of determination is 0.7493
## p-value for coefficient of happiness is less than 2.2e-16, which is less than the significance level

# f. ANOVA table of the regression
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: income
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## happiness   1 1124.32 1124.32  1482.6 < 2.2e-16 ***
## Residuals 496  376.13    0.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

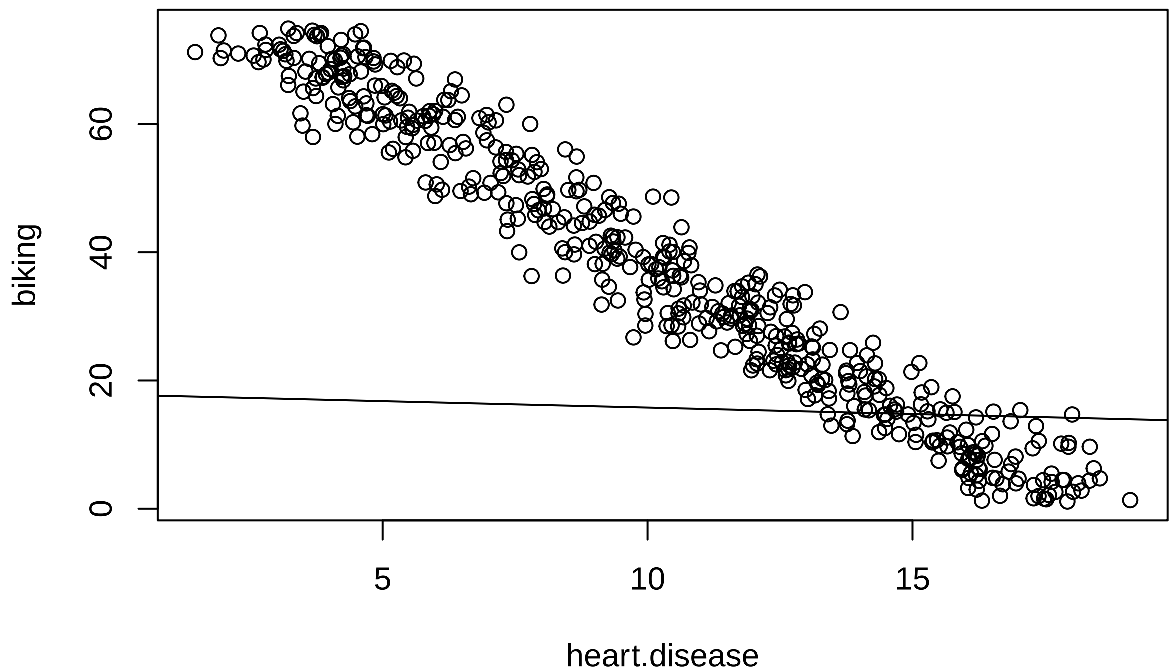
```
### The p-value for ANOVA F-test is less than 2.2e-16, which is less than the significance level 0.05, 

detach(income_data)
```

Question 02:

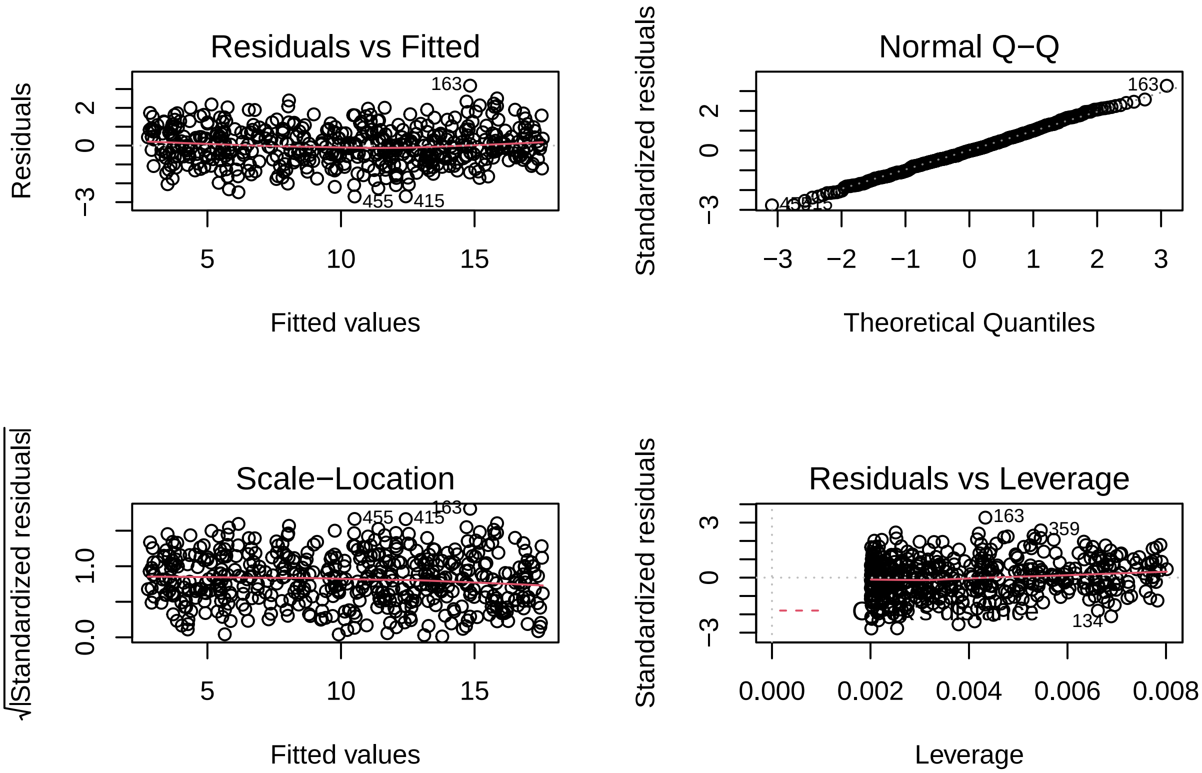
```
heart_data <- read.csv('~/OneDrive - Stony Brook University/SBU/MAT + AMS/Fall 2021/AMS 380/hw/03/heart
attach(heart_data)

fit <- lm (heart.disease ~ biking)
fit
```

```
##
## Call:
## lm(formula = heart.disease ~ biking)
##
## Coefficients:
## (Intercept)       biking
##     17.7779      -0.2003
```

```
# a.The least square regression line equation:
## heart.disease = 17.7779 - 0.2003 * biking

# b. Plot
plot(heart.disease, biking)
abline(fit)
```




```
# c. Check assumptions:
par(mfrow = c(2,2))
plot(fit)
```




```
## 1. Linearity: it is satisfied because the residuals are symmetrically distributed around the 0-line

## 2. Homoscedasticity: it is satisfied because the square root of standardized residuals is symmetrica

## 3. Independence: assume it is satisfied

## 4. Normality:
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.99801, p-value = 0.8351
```

```
### p-value is 0.8351 greater than the significance level 0.10, so residuals is normal distributed, nor

# d. Sample correlation coefficient between the 2 variables:
cor(heart.disease, biking)
```

```
## [1] -0.9753352
```

```
## Sample correlation coefficient is -0.9753352
```

```r
## The corresponding population correlation test:
cor.test(heart.disease, biking)
```

```
##
##  Pearson's product-moment correlation
##
## data:  heart.disease and biking
## t = -98.409, df = 496, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9792783 -0.9706530
## sample estimates:
##        cor
## -0.9753352
```

```r
### p-value is less than 2.2e-16, which is less than the significance level 0.10, reject H0. The correl

# e.
summary(fit)
```

```
##
## Call:
## lm(formula = heart.disease ~ biking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6975 -0.6277 -0.0205  0.6482  3.1787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.777880   0.088450  200.99   <2e-16 ***
## biking      -0.200297   0.002035  -98.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9747 on 496 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9512
## F-statistic:  9684 on 1 and 496 DF,  p-value: < 2.2e-16
```

```r
## The coefficient of determination is 0.9513
## p-value for coefficient of biking is less than 2.2e-16, which is less than the significance level 0.

# f. The percentage of people in the town who have heart disease if the percentage of people who bike t
heart.disease_rate <- 17.7779 - 0.2003 * 65
heart.disease_rate
```

```
## [1] 4.7584
```

```r
## There are 4.7584% people in the town who have heart disease if the percentage of people who bike to

# g. The 90% confidence interval:
confint(fit, level = 0.90)
```

```
##                      5 %       95 %
## (Intercept) 17.6321212 17.9236392
## biking      -0.2036511 -0.1969429
```

```
detach(heart_data)
```