HW3 Solution Q1

Wookyeong Song (most of them from Yan-Yu Chen)

2022/10/20

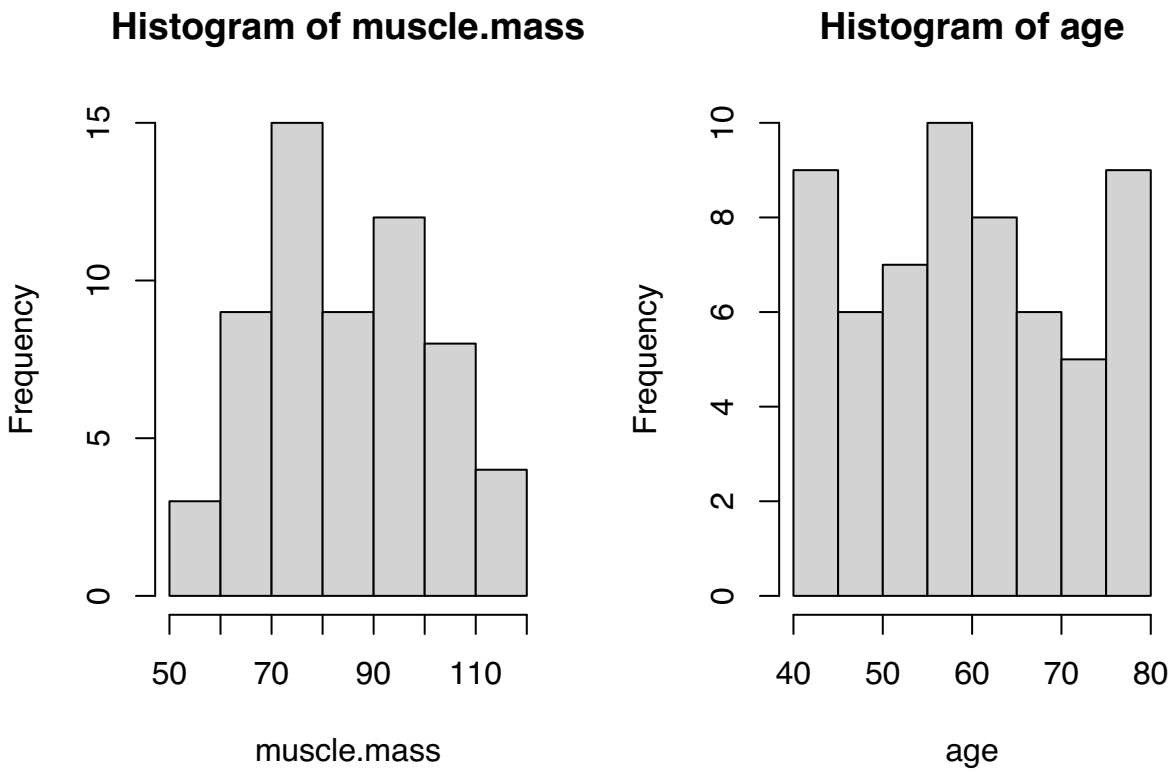
1 A simple linear regression case study by R.

You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

A person's muscle is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each of the four 10-year age groups, beginning with age 40 and ending with age 79. Two variables being measured are: age (X) and the amount of muscle mass (Y).

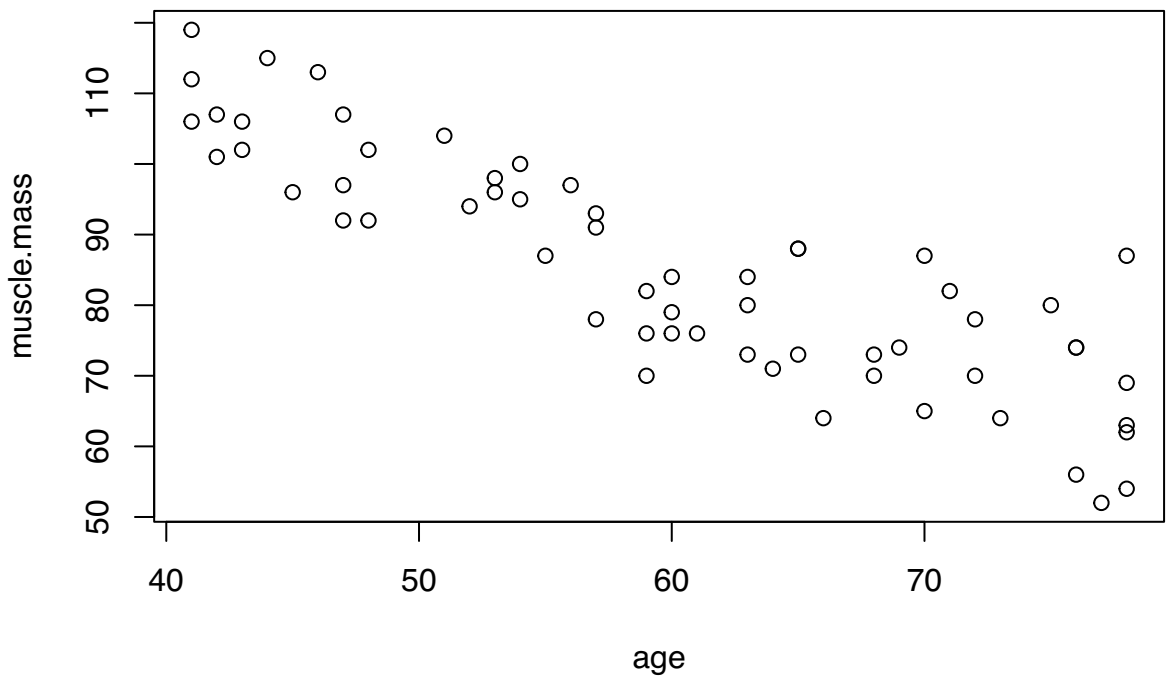
(a) Read data into R. Draw histogram for muscle mass and age, respectively. Comment on their distributions. Draw the scatter plot of muscle mass versus age. Do you think their relation is linear? Does the data support the anticipation that the amount of muscle mass decreases with age?

```
df<-read.table("muscle.txt",col.names = c("muscle mass","age"))
par(mfrow=c(1,2))
with(df,{
  hist(muscle.mass)
  hist(age)
})
```



The histogram of muscle mass is approximately bell-shaped; The histogram of age is pretty flat, which means the individuals are relatively evenly distributed throughout all ages.

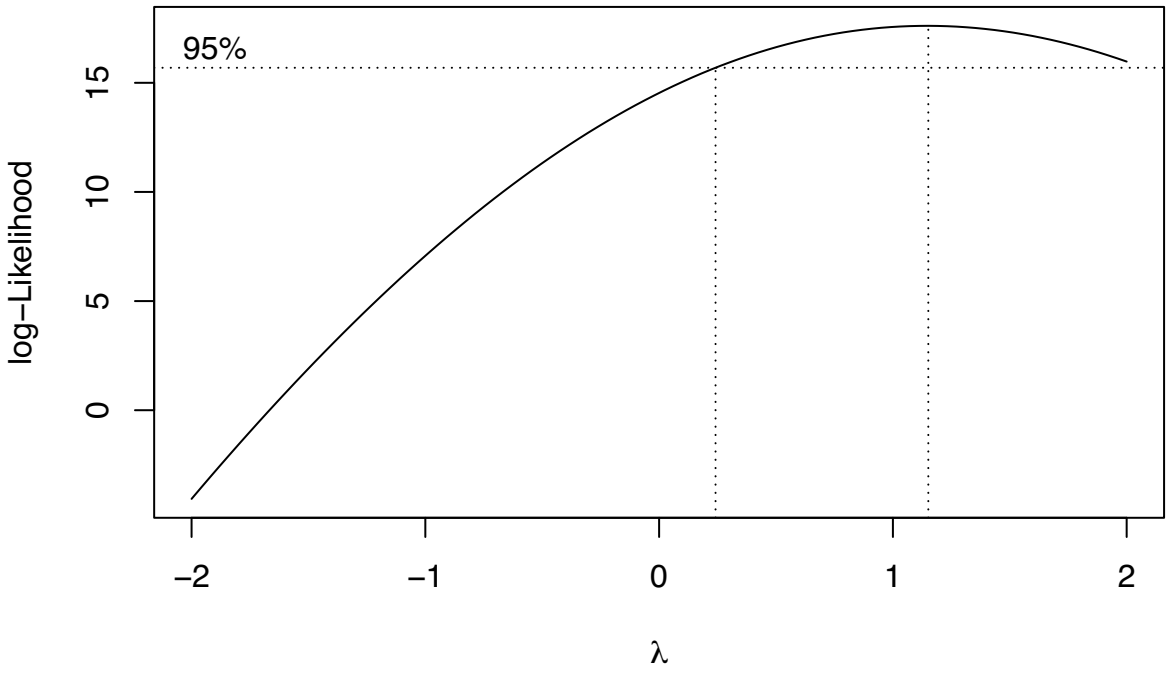
```
with(df, plot(age,muscle.mass))
```



Yes, the relation looks linear and it supports the anticipation that the amount of muscle mass decreases with age.

(b) Use the Box-Cox procedure to decide whether a transformation of the response variable is needed.

```
MASS::boxcox(muscle.mass~age, data=df)
```



The suggested transformation by box-cox is a λ slightly bigger than 1 which means basically no transformation is needed.

(c) Perform linear regression of the amount of muscle mass on age and obtain a summary. From the summary, obtain the estimated regression coefficients and their standard errors, the mean squared error (MSE) and its degrees of freedom.

```
summary(df.lm<-lm(muscle.mass~age, data=df))
```

```
##
## Call:
## lm(formula = muscle.mass ~ age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
```

```
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

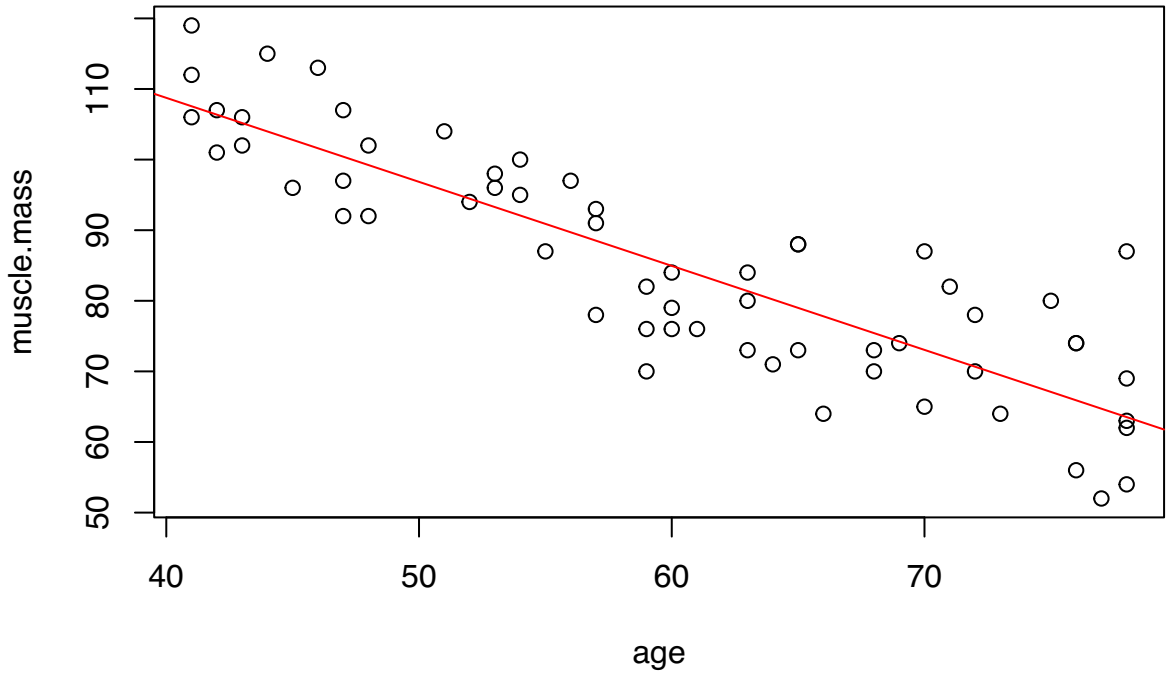
$\hat{\beta}_0 = 156.3466$, $s\{\hat{\beta}_0\} = 5.5123$, $\hat{\beta}_1 = -1.19$, $s\{\hat{\beta}_1\} = 0.0902$, $MSE = 8.173^2 = 66.7979$, with 58 degrees of freedom.

(d) Write down the fitted regression line. Add the fitted regression line to the scatter plot. Does it appear to fit the data well?

Fitted regression line:

$$y = 156.3466 - 1.19x.$$

```
with(df, plot(age,muscle.mass))
abline(a=df.lm$coef[1],b=df.lm$coef[2],col="red")
```



The regression line seems to fit the data well.

(e) Obtain the fitted values and residuals for the 6th and 16th cases in the data set.

```
df.lm$fitted.values[c(6,16)]
```

```
##         6        16
## 107.55675  90.89681
```

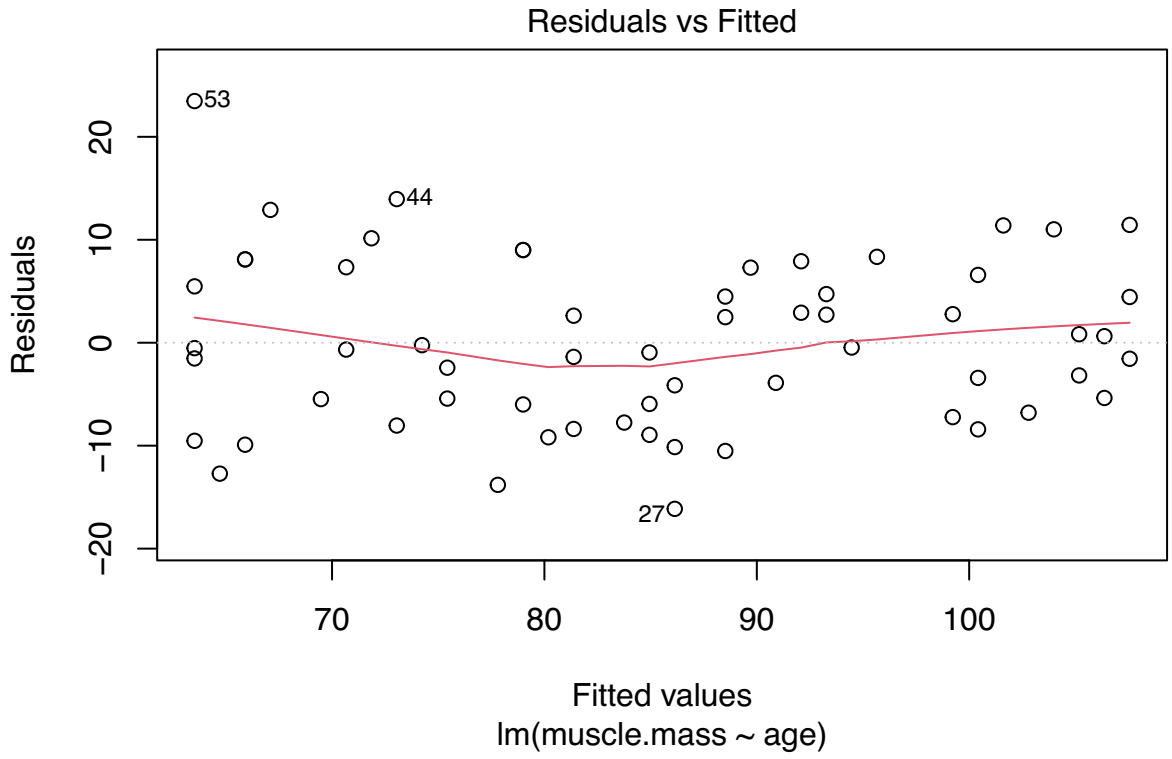
```
df.lm$residuals[c(6,16)]
```

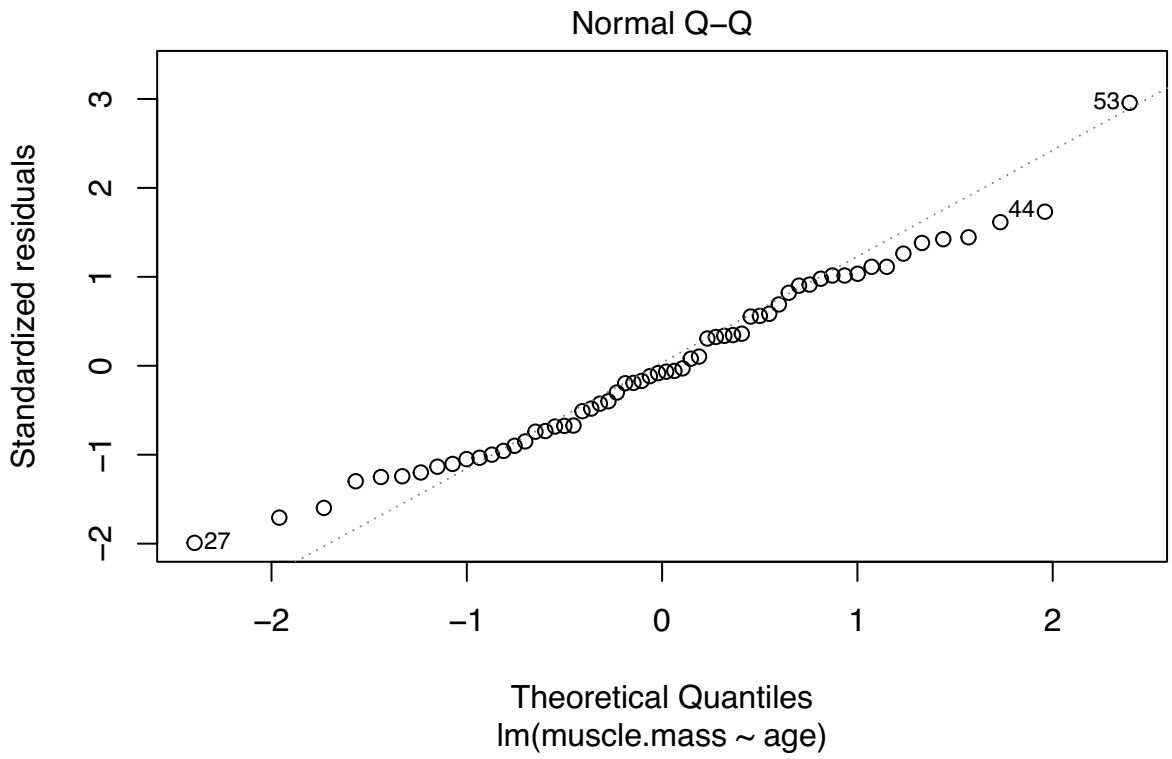
```
##         6        16
## 11.443252 -3.896811
```

$\widehat{Y}_6 = 107.5568$, $\widehat{Y}_{16} = 90.8968$, $e_6 = 11.4433$, $e_{16} = -3.8968$

(f) Draw the residuals vs. fitted values plot and the residuals Normal Q-Q plot. Write down the simple linear regression model with Normal errors and its assumptions. Comment on these assumptions based on the residual plots.

```
# help("plot.lm")
plot(df.lm, which=c(1,2))
```





$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Model assumptions: The random error terms ε_i are independently and identically distributed (i.i.d.) as $N(0, \sigma^2)$. From the residuals vs fitted values plot we can see that the expected value of the residuals is approximately zero and the variance is approximately constant. From the residuals Normal Q-Q plot we can see that the residuals are slightly light-tailed compared to Normal.

(g) Construct a 99% confidence interval for the regression intercept. Interpret your confidence interval.

```
confint(df.lm,parm= 1, level=0.99) # intercept is the 1st param
```

```
##                0.5 %   99.5 %
## (Intercept) 141.6658 171.0273
```

(h) Conduct a test at level 0.01 to decide whether or not there is a negative linearassociation between the amount of muscle mass and age. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. (Hint: Which form of alternatives should you use?)

```
qt(0.01,58)
```

```
## [1] -2.392377
```

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 < 0$$

$$T^* = \frac{\hat{\beta}_1}{s\{\hat{\beta}_1\}} = -13.19$$

Under null hypothesis, $T^* \sim t_{58}$. Since $T^* = -13.19 < -2.39 = t(0.01, 58)$, we reject the null hypothesis and conclude that there is significant negative linear association between the amount of muscle mass and age.

(i) Construct a 95% prediction interval for the muscle mass of a woman aged at 60. Interpret your prediction interval.

```
predict(df.lm, data.frame(age=60), interval="predict")
```

```
##        fit      lwr     upr
## 1 84.94683 68.45067 101.443
```

We are 95% confident that the muscle mass of a woman aged at 60 is in between 68.45 and 101.44.

(j) Obtain the ANOVA table for this data. Test whether or not there is a linear association between the amount of muscle mass and age by an F test at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

```
anova(df.lm)
```

```
## Analysis of Variance Table
##
## Response: muscle.mass
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 11627.5$	d.f.$(SSR) = 1$	$MSR = 11627.5$	$F^* = MSR/MSE$
Error	$SSE = 3874.4$	d.f.$(SSE) = 58$	$MSE = 66.8$	$=174.06$
Total	$SSTO = 15501.9$	d.f.$(SSTO) = 59$		

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

Test statistic

$$F^* = 174.06$$

Under null hypothesis, $F^* \sim F_{1,58}$. Since the p-value is less than 0.01, we reject H_0 and conclude that there is a significant linear association between the amount of muscle mass and age.

(k) What proportion of the total variation in muscle mass is "explained" by age? What is the correlation coefficient between muscle mass and age?

Since $R^2 = 0.7501$, about 75% of the total variation in muscle mass is "explained" by age. Since $\hat{\beta}_1 < 0$, the correlation coefficient r between the amount of muscle mass and age is $-\sqrt{R^2} = -0.8661$.

Statistics 206

Homework 3 (Solution)

Due : Oct. 20, 2022, 11:59PM

Instructions:

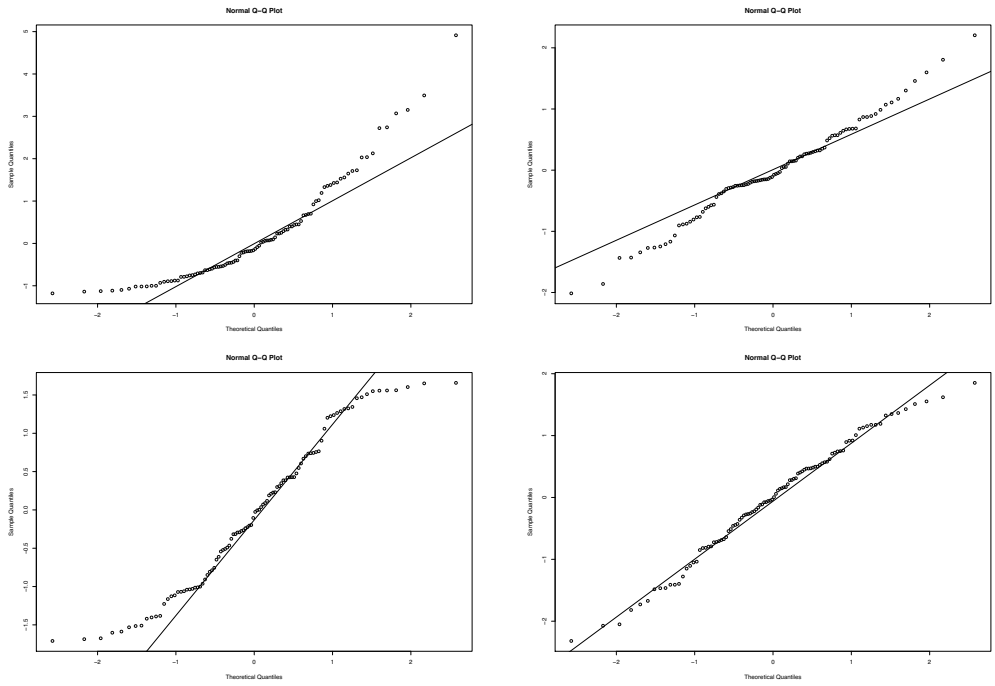
- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.rmd".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. See soluton to Question 1.
2. **Q-Q plots.** For each of the Q-Q plot in Figure 1, describe the distribution of the data (whether it is Normal or heavy tailed, etc.).

   Looking at it in anticlockwise fashion,

   * Top left: right skewed
   * Bottom left: light tailed
   * Bottom right: approximately normal
   * Top right : heavy tailed

Figure 1: Q-Q plots



3. **Coefficient of determination.** Show that

$$R^2 = r^2, \quad r = \text{sign}\{\hat{\beta}_1\}\sqrt{R^2},$$

where R^2 is the coefficient of determination when regressing Y onto X and r is the sample correlation coefficient between X and Y.

Proof.

$$R^2 = SSR/SSTO = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 / \sum (y_i - \bar{y})^2$$

$$= (\sum (x_i - \bar{x})(y_i - \bar{y}))^2 / \sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2 = r^2$$

□

4. Confirm the formula for inverting a 2×2 matrix.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Check if the following equality holds.

$$\begin{aligned} & \tfrac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ = \; & \tfrac{1}{ad-bc}\begin{bmatrix} da-bc & db-bd \\ -ca+ac & -cb+ad \end{bmatrix} \\ = \; & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

5. **Projection matrices**.

   (a) $\mathbf{I}_n - \mathbf{H}$

$$(\mathbf{I}_n - \mathbf{H})' = \mathbf{I}_n' - \mathbf{H}' = \mathbf{I}_n - \mathbf{H}$$
$$(\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - \mathbf{I}_n\mathbf{H} - \mathbf{H}\mathbf{I}_n + \mathbf{H}^2 = \mathbf{I}_n - \mathbf{H}$$

   Its rank is $n - p = n - 2$.

   (b) $\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$

$$(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)' = \mathbf{I}_n' - \frac{1}{n}\mathbf{J}_n' = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$$
$$(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)^2 = \mathbf{I}_n^2 - \mathbf{I}_n\frac{1}{n}\mathbf{J}_n - \frac{1}{n}\mathbf{J}_n\mathbf{I}_n + \frac{1}{n^2}\mathbf{J}_n^2 = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$$

   Its rank is $n - 1$.

   (c) $\mathbf{H} - \frac{1}{n}\mathbf{J}_n$

$$(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)' = \mathbf{H}' - \frac{1}{n}\mathbf{J}_n' = \mathbf{H} - \frac{1}{n}\mathbf{J}_n$$

   For the rest, notice that $\mathbf{H}\mathbf{J}_n = \mathbf{J}_n$ because $\mathbf{H}$ is the projection matrix onto the column space of X and every column of $\mathbf{J}_n$, namely $\mathbf{1}_n$, is in the column space of X. So,

$$(\mathbf{H}-\frac{1}{n}\mathbf{J}_n)^2 = \mathbf{H}-\frac{1}{n}\mathbf{J}_n\mathbf{H}-\mathbf{H}\frac{1}{n}\mathbf{J}_n+\frac{1}{n^2}\mathbf{J}_n^2 = \mathbf{H}-\frac{1}{n}\mathbf{J}_n\mathbf{H}-\mathbf{H}\frac{1}{n}\mathbf{J}_n+\frac{1}{n}\mathbf{J}_n = \mathbf{H}-\frac{1}{n}\mathbf{J}_n,$$

   where $\mathbf{J}_n = \mathbf{J}_n\mathbf{H}$ follows from

$$\mathbf{J}_n\mathbf{H} = \mathbf{J}_n^t\mathbf{H}^t = (\mathbf{H}\mathbf{J}_n)^t = \mathbf{J}_n^t = \mathbf{J}_n.$$

   Its rank is $p - 1 = 1$.

6. (a) *Proof.*

$$e = (I - \mathbf{H})Y, \quad \hat{\beta} = (X'X)^{-1}X'Y$$
$$\mathrm{Cov}(e, \hat{\beta}) = (I - \mathbf{H})\mathrm{Cov}(Y)((X'X)^{-1}X')' = \sigma^2(I - \mathbf{H})X(X'X)^{-1} = 0,$$

   since $(I - \mathbf{H})X = X - X = 0$. Therefore $\hat{\beta}$ and the residuals e are uncorrelated.

   - $\hat{Y} = X\hat{\beta}$. Hence, $\mathrm{Cov}(\hat{Y}, e) = \mathrm{Cov}(X\hat{\beta}, e) = X\mathrm{Cov}(\hat{\beta}, e) = 0$. Therefore $\hat{Y}$ and the residuals e are uncorrelated.

- (Alternative)

$$\hat{Y} = \mathbf{H}Y, \quad e = (I - \mathbf{H})Y$$

$$\text{Cov}(\hat{Y}, e) = \text{Cov}(\mathbf{H}Y, (I - \mathbf{H})Y) = H\text{Cov}(Y)(I - \mathbf{H})^t = \sigma^2\mathbf{H}(I - \mathbf{H}) = 0$$

since $\mathbf{H}(I - \mathbf{H}) = \mathbf{H} - \mathbf{H} = 0$. Therefore $\hat{Y}$ and the residuals e are uncorrelated.

□

(b) *Proof.* By Hint, $e = (I_n - \mathbf{H})Y$ and $d = (\mathbf{H} - \frac{1}{n}\mathbf{J}_n)Y$ are jointly normally distributed.

$$Cov(e, d) = (I_n - \mathbf{H})Cov(Y)(\mathbf{H} - \frac{1}{n}\mathbf{J}_n) = \sigma^2(I_n - \mathbf{H})(\mathbf{H} - \frac{1}{n}\mathbf{J}_n) = 0.$$

Thus, we know they are uncorrelated, and they are mutually independent by their normal distributions. Because $SSE = e^T e$ and $SSR = d^T d$ are functions of e and d, they are also independent. From part (a), e and $\hat{\beta}$ are uncorrelated. By Hint, they are jointly normal. Hence, e and $\hat{\beta}$ are independent, and so is $SSE = e^T e$ and $\hat{\beta}$, SSE being a function of e. □