STAT8111 Assigment1

Phan Vinh Phu 45747989

2023-08-18

Question 1

a.Examine first graphically and numerically correlation between variables, then comment :

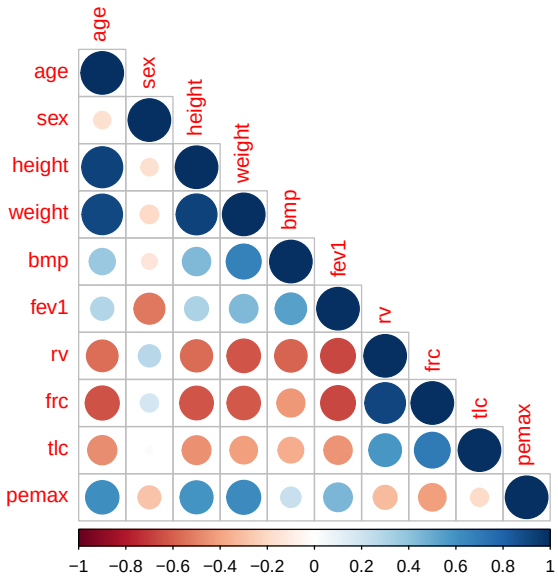
Numerical correlation

```
num_cor <- round(cor(data), 4)
num_cor
```

```
##            age     sex  height  weight     bmp    fev1      rv     frc     tlc
## age     1.0000 -0.1671  0.9261  0.9059  0.3778  0.2945 -0.5519 -0.6394 -0.4694
## sex    -0.1671  1.0000 -0.1675 -0.1904 -0.1376 -0.5283  0.2714  0.1836  0.0242
## height  0.9261 -0.1675  1.0000  0.9207  0.4408  0.3167 -0.5695 -0.6243 -0.4571
## weight  0.9059 -0.1904  0.9207  1.0000  0.6725  0.4488 -0.6215 -0.6173 -0.4185
## bmp     0.3778 -0.1376  0.4408  0.6725  1.0000  0.5455 -0.5824 -0.4344 -0.3649
## fev1    0.2945 -0.5283  0.3167  0.4488  0.5455  1.0000 -0.6659 -0.6651 -0.4430
## rv     -0.5519  0.2714 -0.5695 -0.6215 -0.5824 -0.6659  1.0000  0.9106  0.5891
## frc    -0.6394  0.1836 -0.6243 -0.6173 -0.4344 -0.6651  0.9106  1.0000  0.7044
## tlc    -0.4694  0.0242 -0.4571 -0.4185 -0.3649 -0.4430  0.5891  0.7044  1.0000
## pemax   0.6135 -0.2886  0.5992  0.6352  0.2295  0.4534 -0.3156 -0.4172 -0.1816
##          pemax
## age     0.6135
## sex    -0.2886
## height  0.5992
## weight  0.6352
## bmp     0.2295
## fev1    0.4534
## rv     -0.3156
## frc    -0.4172
## tlc    -0.1816
## pemax   1.0000
```

Graphical correlation

```
corrplot(num_cor, type="lower")
```



Comment :

From the correlation plot, it can be seen that :

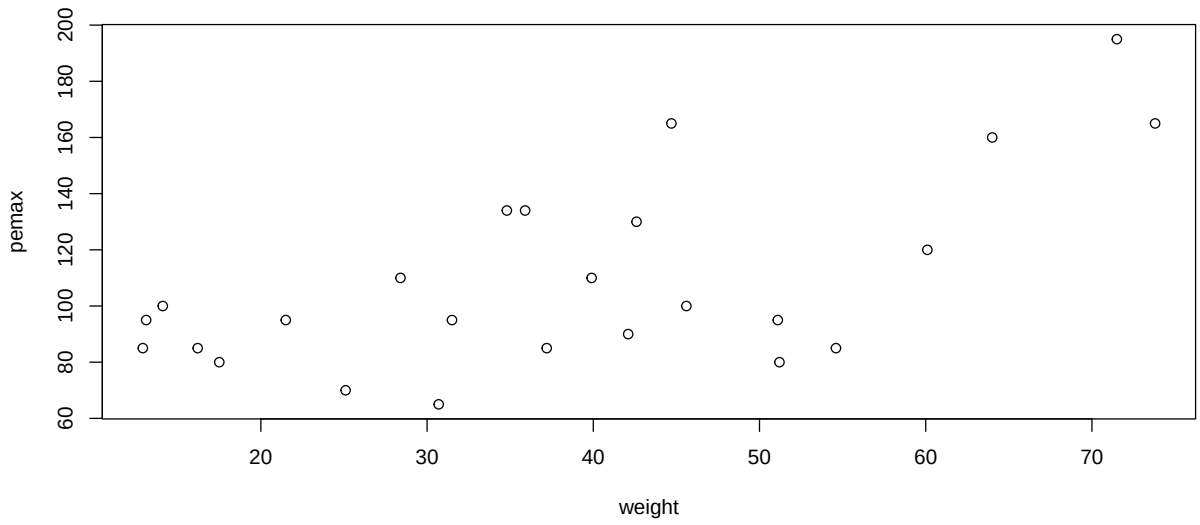
- There might be a strong positive linear relationship between `age` and `height`, `weight` and `pemax`. Beside that, `age` is negatively correlated with `frc`.
- `height` and `weight` are strongly correlated. In addition, `weight` has moderate negative correlation with `rv`, `frc`, `tlc` and positive correlation with `bmp`, `pemax`.
- Similarly, `fev1` has moderate negative correlation with `rv`, `frc`, `tlc`.
- Lastly, `rv` highly positively correlated with `frc`.

b. the relationship between `weight` and `pemax`

In this part, the relationship between `weight` and `pemax` will be examined.Specifically, `pemax` is the dependent variable (Y) and `weight` is the independent variable (X).

Scatter Plot

```
plot(x = data$weight, y = data$pemax, xlab="weight", ylab = "pemax")
```



The graph illustrates that `weight` is positively related with `pemax`. The trend is that when `weight` increase, `pemax` will increase.

Linear Model

```
model <- lm(pemax ~ weight, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = pemax ~ weight, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.30 -22.69   2.23  15.91  48.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.5456    12.7016   5.003 4.63e-05 ***
## weight        1.1867     0.3009   3.944 0.000646 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.38 on 23 degrees of freedom
## Multiple R-squared:  0.4035, Adjusted R-squared:  0.3776
## F-statistic: 15.56 on 1 and 23 DF,  p-value: 0.0006457
```

The model equation $\hat{pemax} = 63.5456 + 1.1867 weight + \epsilon \ \ (\epsilon \sim N(0, \sigma^2))$

Model Fit

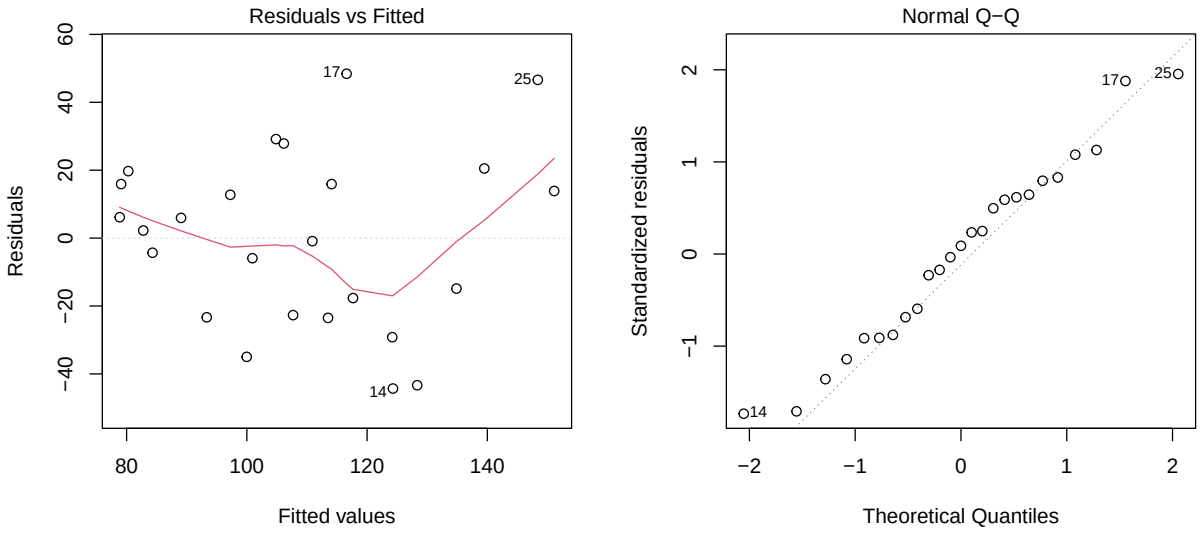
- The $R^2 = 0.4035$ which means 40.35% of variation in `pemax` can be explained by `weight`. This show that the model is not fit well.

Model interpretation

- According to the equation, for each unit increase in weight, the pemax will increase about 1.1867.
- `weight` is significant predictor since the p-value = 0.000646 (< 0.001).

```
par(mfrow = c(1,2))
plot(model, which = c(1,2))
```

Diagnostic Checking



- The standardized Residuals versus Fitted values plot appears to be a random scatter about zero, so the model is adequate.This graph also show some residuals which are 14, 17, 25 are low and high. This can be evidence of small amount of heteroscedasticity.
- the Normal Q-Q plot is approximately linear, so it can be said that the normality assumption holds.

c. Include in the previous model the sex variable

Model 2

```r
model_2 <- lm(pemax ~ weight + sex, data = data)
summary(model_2)
```

```
##
## Call:
## lm(formula = pemax ~ weight + sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.388 -16.850   0.073  13.168  43.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.9719    14.4644   4.907 6.61e-05 ***
## weight        1.1248     0.3056   3.681  0.00131 **
## sex         -11.4776    10.7963  -1.063  0.29926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.31 on 22 degrees of freedom
## Multiple R-squared:  0.4327, Adjusted R-squared:  0.3811
## F-statistic: 8.388 on 2 and 22 DF,  p-value: 0.00196
```

The model 2 equation

$\hat{pemax} = 70.9719 + 1.1248 weight + (-11.4776) sex + \epsilon \quad (\epsilon \sim N(0, \sigma^2))$

Model 3

```r
model_3 <- lm(pemax ~ sex + weight, data = data)
summary(model_3)
```

```
##
## Call:
## lm(formula = pemax ~ sex + weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.388 -16.850   0.073  13.168  43.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.9719    14.4644   4.907 6.61e-05 ***
## sex         -11.4776    10.7963  -1.063  0.29926
## weight        1.1248     0.3056   3.681  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.31 on 22 degrees of freedom
## Multiple R-squared:  0.4327, Adjusted R-squared:  0.3811
## F-statistic: 8.388 on 2 and 22 DF,  p-value: 0.00196
```

The model 3 equation

$\hat{pemax} = 70.9719 + (-11.4776)sex + 1.1248weight + \epsilon \quad (\epsilon \sim N(0, \sigma^2))$

Analyse the two proposed models

- Model_2 and model_3 are seem to be similar. The only difference is the order of `weight` and `sex`. The R^2 of both model are 0.4327 which mean 43.27% of variant in `pemax` can be explained by `weight` and sex. On the one hand, in both two model, `sex` is insignificant predictor with p-value = 0.29926. On the other hand, `weight` is still significant predictor.
- For one unit increase in `weight` , `pemax` will increase 1.1248. The coefficient of `sex` represents the difference in `pemax` between females and males, while `weight` is same. In this case, models indicate that, on average, females have a `pemax` that is lower by 11.4776 units compared to males.

In conclusion, it appears that the order of variables (`weight` and `sex`) doesn't affect the results in this case. Model_2 and model_3 are better than the first model on question (b) but these two still are not good model.

Choose one model

In comparison of three models, model 2 and model 3 are appear to explain `pemax` better with higher R^2 and adjusted R^2. Beside that, there are no collinearity issue between `weight` and `sex`. Therefore, I choose model 2.

d. Construct a statistical model for the response variables pemax based on the normal response distribution and the weight, bmp, fev1, rv,frc.

We are interested in predicting `pemax` (maximal expiratory pressure) by using `weight`, `bmp`, `fev1`, `rv`,`frc`. In this case, the stepwise backward selection will be used to find the best model. In stepwise backward selection, people first regress with all predictor variables in the model. Then, the predictors with the largest p-value in the t-test will be dropped. Next step is fitting the reduced model. This progess will be run iteratively until all variables in the model are significant.Let's start with the full model.

Full Model

```
full_model <- lm(pemax ~ weight + bmp + fev1 + rv + frc , data = data)
summary(full_model)
```

```
##
## Call:
## lm(formula = pemax ~ weight + bmp + fev1 + rv + frc, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.72 -12.17   4.83  15.29  34.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.18640   54.73550   1.173 0.255423
## weight       1.73556    0.42529   4.081 0.000637 ***
## bmp         -1.35105    0.66763  -2.024 0.057303 .
## fev1         1.53087    0.62948   2.432 0.025078 *
```

```
## rv           0.13612    0.15668   0.869 0.395787
## frc         -0.02477    0.31278  -0.079 0.937703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 19 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.5127
## F-statistic:  6.05 on 5 and 19 DF,  p-value: 0.001637
```

It can be seen that `frc` has the largest P-value which is 0.937703. Therefore, `frc` explains the least variation when added to the model. Now , `frc` will be dropped.

The reduced model without `frc`

```
reduced_model <- lm(pemax ~ weight + bmp + fev1 + rv , data = data)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = pemax ~ weight + bmp + fev1 + rv, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.77 -11.74   4.33  15.66  35.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 63.94669   53.27673   1.200 0.244057
## weight       1.74891    0.38063   4.595 0.000175 ***
## bmp         -1.37724    0.56534  -2.436 0.024322 *
## fev1         1.54770    0.57761   2.679 0.014410 *
## rv           0.12572    0.08315   1.512 0.146178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.75 on 20 degrees of freedom
## Multiple R-squared:  0.6141, Adjusted R-squared:  0.5369
## F-statistic: 7.957 on 4 and 20 DF,  p-value: 0.000523
```

Similarly, `rv` is the insignificant predictor and has the largest P-value which is 0.146178. Hence, I drop `rv` and fit the reduced model without `frc` and `rv`.

The reduced model without `frc` and `rv`

```
reduced_model_2 <- lm(pemax ~ weight + bmp + fev1 , data = data)
summary(reduced_model_2)
```

```
##
## Call:
## lm(formula = pemax ~ weight + bmp + fev1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -42.388 -13.496   3.991  14.856  40.373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126.3336    34.7199   3.639 0.001536 **
## weight        1.5365     0.3644   4.216 0.000387 ***
## bmp          -1.4654     0.5793  -2.530 0.019486 *
## fev1          1.1086     0.5144   2.155 0.042893 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.44 on 21 degrees of freedom
## Multiple R-squared:   0.57,  Adjusted R-squared:  0.5086
## F-statistic: 9.279 on 3 and 21 DF,  p-value: 0.000418
```

At this stage, all predictors are significant, therefore, selection process stops here. Nevertheless, there is moderate multi-collinearity since `bmp`, `weight`, `fev1` are positively correlated.

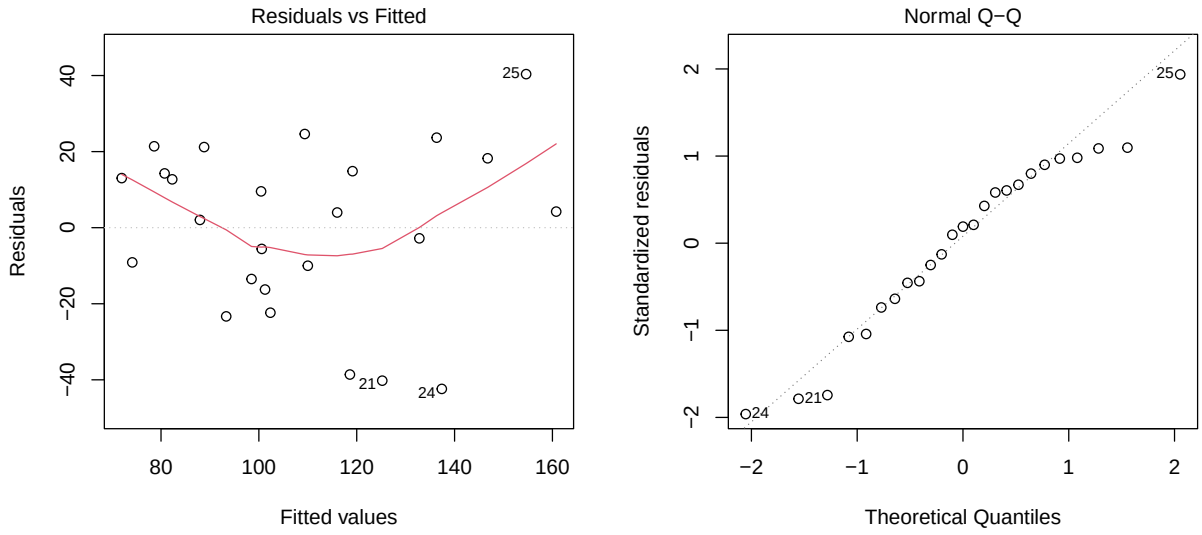
```
# Calculate the correlation matrix
correlation_matrix <- cor(data[, c("weight", "bmp", "fev1")])

# Print the correlation matrix
print(correlation_matrix)
```

```
##           weight       bmp      fev1
## weight 1.0000000 0.6725463 0.4488393
## bmp    0.6725463 1.0000000 0.5455204
## fev1   0.4488393 0.5455204 1.0000000
```

Diagnostic Checking

```
par(mfrow = c(1,2))
plot(reduced_model_2, which = 1:2)
```



- The Residuals vs Fitted plot look like random scatter around 0. Therefore, there is no obvious pattern in any of the residual plots so it appears the linearity and constant variance assumptions of the multiple linear model are justified.
- The quantile plot of residuals look approximately linear, suggesting the normality assumption for residuals is appropriate.

The final model equation

$\hat{pemax} = 126.3336 + 1.5365weight + (-1.4654)bmp + 1.1086fev1 + \epsilon \quad (\epsilon \sim N(0, \sigma^2))$