STAT8111 Assigment1

Phan Vinh Phu 45747989

2023-08-18

Question 1

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

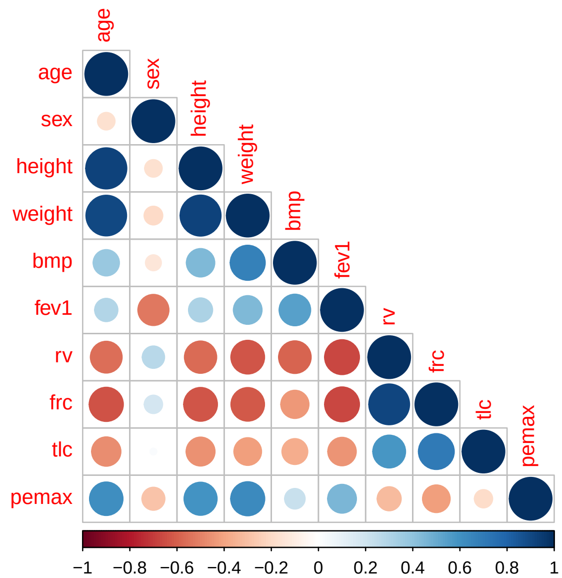
Numerical correlation

```r
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 :

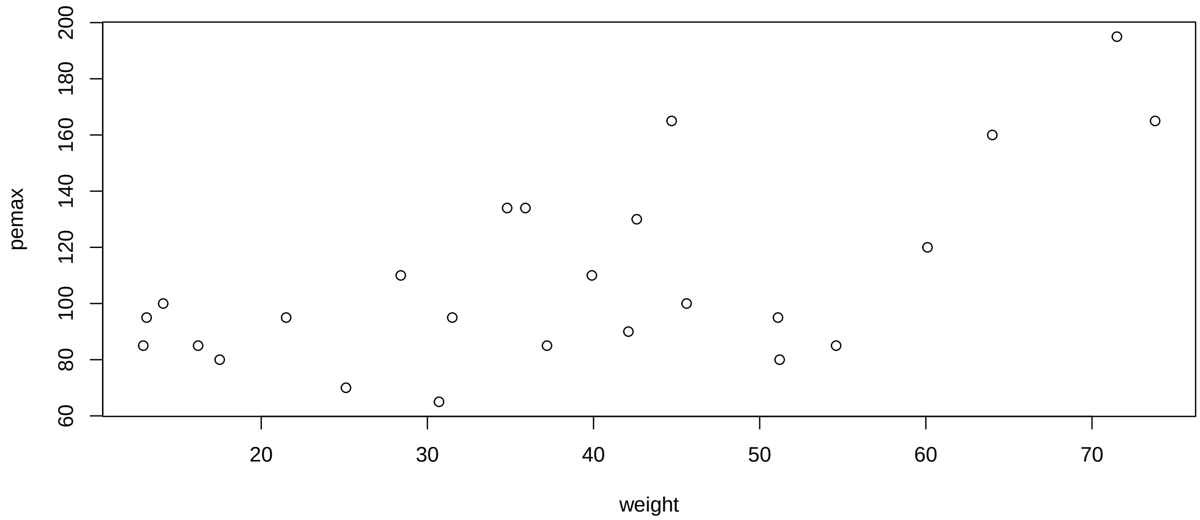
- There might be a strong positive linear relationship between `age` and `height`, `weight` and `pemax`. Beside that, `age` is negatively correlated with `frc`.
- `height` and `weight` are strongly correlated. In addition, `weight` has moderate negative correlation with `rv`, `frc`, `tlc` and positive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



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 $\hat{pmax} = 63.5456 + 1.1867 weight + \epsilon$

Model Fit

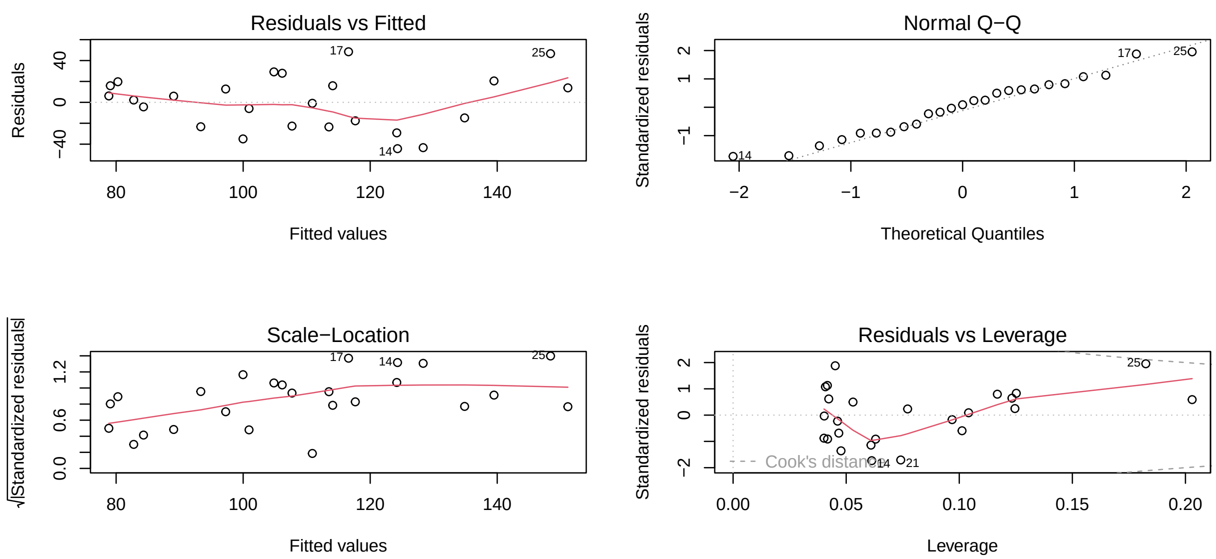
- The $R^2 = 0.4035$ which means 40.35% of variation in `pemax` can be explained by `weight`. This show that the model is not fit well.

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(2,2))
plot(model)
```

Diagnostic



- 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

```
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

$p\hat{max} = 70.9719 + 1.1248 weight + (-11.4776) sex + \epsilon$

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

$p\hat{max} = 70.9719 + (-11.4776) sex + 1.1248 weight + \epsilon$

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 `pmax` 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` , `pmax`will increase 1.1248. The coefficient of `sex` represents the difference in `pmax` 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. 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 `pmax` better with higher R^2 and adjusted R^2. Beside that, there are no multicollinearity 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.