STAT 628 Module 2 BodyFat Group 3

Introduction:

Our motivation is to obtain a precise, concise and reasonable regression model to predict body fat percentage of a male based on his body informations.

Dataset:

Our dataset `BodyFat` is a real data set of 252 men with measurements of their percentage of body fat and various body circumference measurements.

Background:

We assume that percentage of body fat for an individual can be estimated once body density has been determined:

$$BodyFat(\%) = \frac{495}{BodyDensity(gm/cm^3)} - 450 \quad (1)$$

And the BMI can be calculated by:

$$BMI = \frac{703 \times Weight(lbs)}{Height(inches)} \quad (2)$$

1. Exploratory data analysis

A data.frame: 6 × 16

BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP
<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
12.6	1.0708	23	154.25	67.75	23.7	36.2	93.1	85.2	94.5
6.9	1.0853	22	173.25	72.25	23.4	38.5	93.6	83.0	98.7
24.6	1.0414	22	154.00	66.25	24.7	34.0	95.8	87.9	99.2
10.9	1.0751	26	184.75	72.25	24.9	37.4	101.8	86.4	101.2
27.8	1.0340	24	184.25	71.25	25.6	34.4	97.3	100.0	101.9
20.6	1.0502	24	210.25	74.75	26.5	39.0	104.5	94.4	107.8

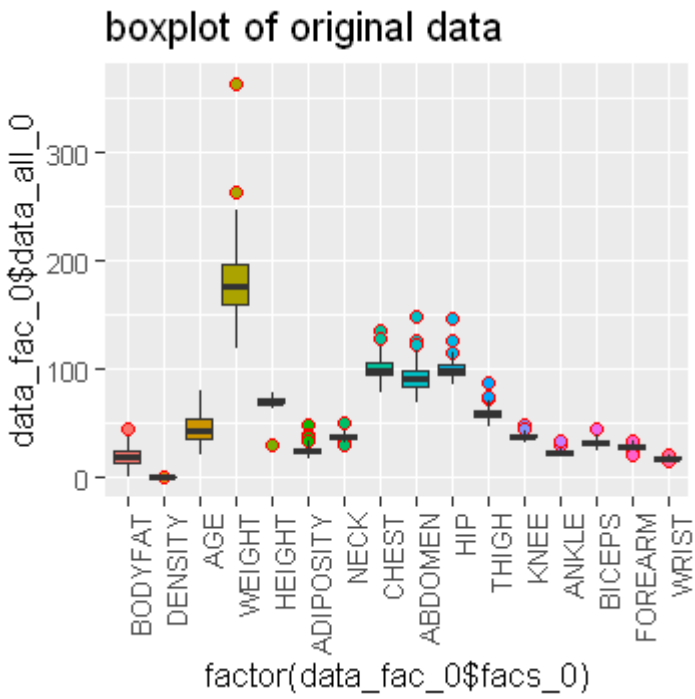
We can first intuitively analyze the data by looking at its summary tables and histograms. There are some obvious "abnormal" obsevations. Some results are as following:

```
    BODYFAT          WEIGHT          HEIGHT        ADIPOSITY    
 Min.   : 0.00   Min.   :118.5   Min.   :29.50   Min.   :18.10  
 1st Qu.:12.80   1st Qu.:159.0   1st Qu.:68.25   1st Qu.:23.10  
 Median :19.00   Median :176.5   Median :70.00   Median :25.05  
 Mean   :18.94   Mean   :178.9   Mean   :70.15   Mean   :25.44  
 3rd Qu.:24.60   3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:27.32  
 Max.   :45.10   Max.   :363.1   Max.   :77.75   Max.   :48.90  
```

From the results above, we can find some "abnormal" obsevations: For example, the minimal `BODYFAT` is 0 %, which is unlikely to be true; besides, the maximal `BODYFAT` is 36 %, which is too large for a living person. Similarly, the minimal `HEIGHT` , maximal `WEIGHT` and maxinal `ADIPOSITY` are all unlikely to be true for a nomal man. We can calculate the predicted values of these "abnormal" data using some formula or just delete these observations. Anyway, data cleaning is needed for this dataset.

2. Data cleaning

To find out outliers, we can first look at the **boxplot:**



Calculate the rages of `WEIGHT` , `HEIGHT` , `DENSITY` , `BODYFAT` between there 1% quantiles and 99% quantiles respectively, any value out of such ranges can be considered as "abnormal" observations:

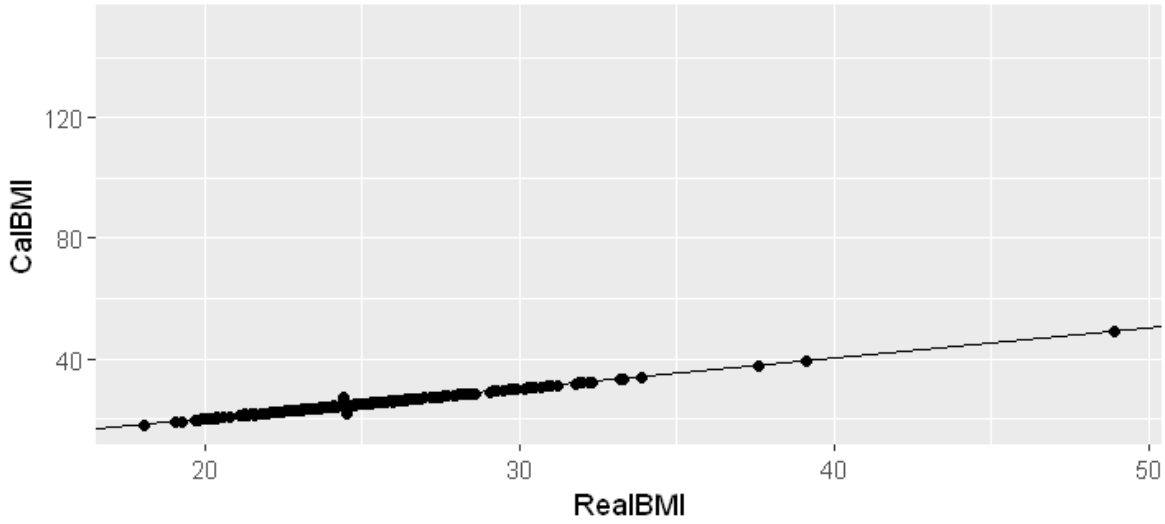
```
[1] "Quantile ranges:"
```

A matrix: 4 × 2 of type dbl

	1%	99%
weight_q	125.50500	245.720000
height_q	64.38250	76.000000
density_q	1.01604	1.095393
bodyfat_q	4.35500	35.582000

We can then compare the body fat percentages calculated from formula (1) (`CalBodyFat`) and given by the dataset (`RealBodyFat`) of each observation to see if there is any strong difference. Similarly, we can compare the BMI values calculated from formula (2) (`CalBMI`) and given by the dataset (`RealBMI`) to find some outliers:

2.1 For BMI:



The 42nd observation seems to be an outlier. Check his `WEIGHT` and `HEIGHT` to find out where is the problem:

'WEIGHT of 42nd observation: 205'

'Quantiles of WEIGHT: '

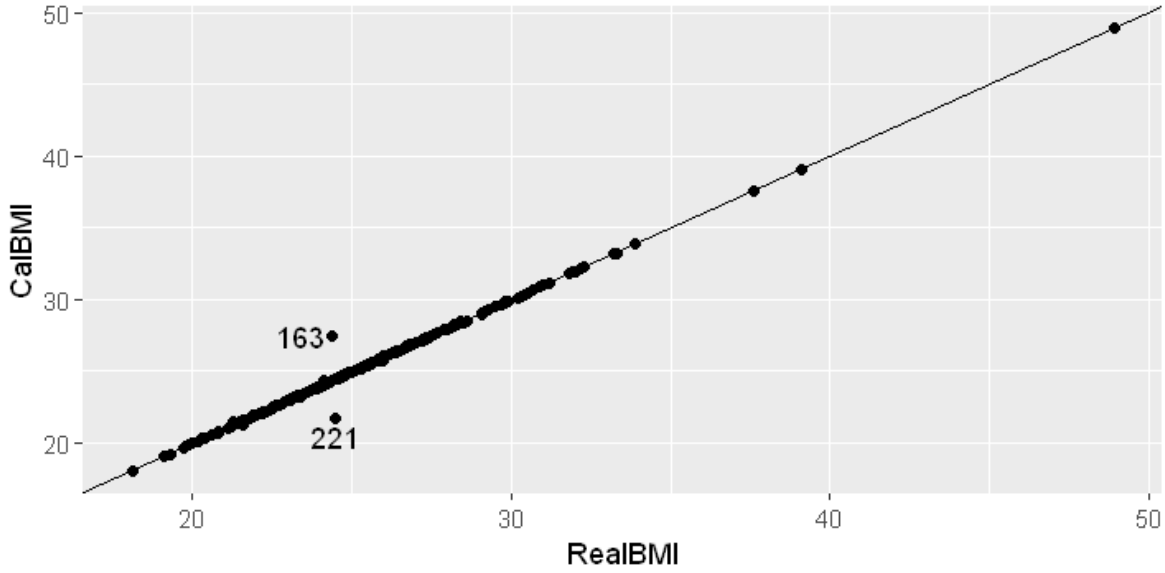
1%
125.505
99%
245.72

'HEIGHT of 42nd observation: 29.5'

'Quantiles of HEIGHT: '

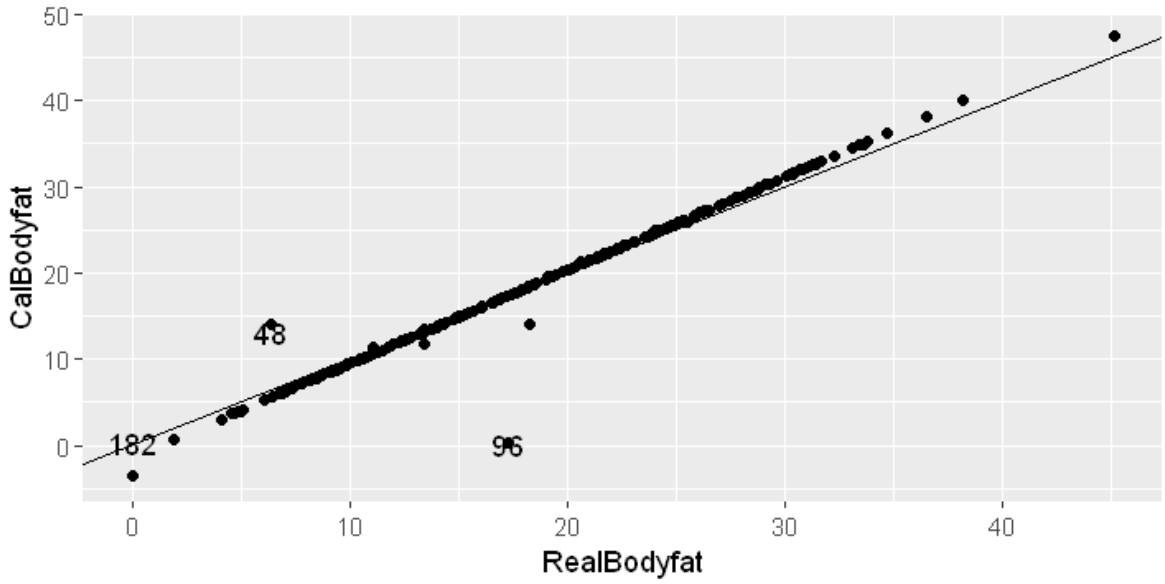
1%
64.3825
99%
76

From the results above, we can see that the *42nd* observation's height (**29.5** inches) is out of quantile range, so we use the formula (2) to calculated the height based on his weight and BMI, it turns out that his height should be **69.43** inches:



After changing the height of the *42nd* observation, we can find that the 163rd and 221st observations still don't fit the line perfectly, but after similar analysis, we find out that their heights and weights is within the quantile range above, so we won't change them.

2.2 For BodyFat:



A data.frame: 3 × 2

	RealBodyfat	CalBodyfat
	<dbl>	<dbl>
182	0.0	-3.6116873
48	6.4	14.1350211
96	17.3	0.3684833

A matrix: 2 × 2 of type dbl

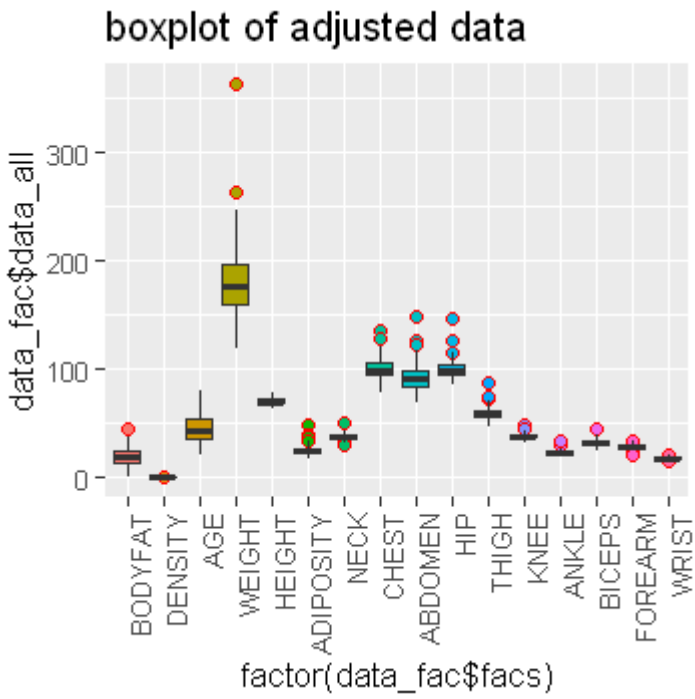
	48th	96th
density	1.0665	1.0991
bodyfat	6.4000	17.3000

From the graph above, we can see that the *182nd* observation have a **ZERO** calculated body fat percentage and a **NEGATIVE** given body fat percentage, which is unlikely to be true for a normal male, so, we will just remove this observation. The *48th* observation has normal body density and body fat percentage, so we will keep it in spite of its unfitness of the line. However, the *96th* obsrvation has a density far away from the density quantile range, so we should modify the body density of this observation or just drop this observation. Since people can not easily obtain body density in real life, we will just remove this observation (We also drop the column `DENSITY` since such data is not feasible in real world application):

2.3 Other outliers

Finally, we use the 0.01 and 0.99 quantiles to replace the values out of 0.01 ~ 0.99 ranges and save the cleaned data into a `.csv` file:

The **boxplot** of the cleaned data:



Obviously, there are less outliers after the data cleaning. Next, we will perform feature selection.

3. Feature Selection

3.1 Cross Validation

We use a **10-fold cross validation** to compute the average MSE (mean-square error) of each model:

$$MSE = E((Y - \hat{Y})^2) \tag{3}$$

We shuffle the data before every CV and repeated 1000 times.

3.2 Stepwise model selection

Perform stepwise selections: We choose **AIC** and **BIC** methods: Let k denote the number of parameters, let RSS denote the residual sum of squares, let n denote the number of observations, we have:

AIC (Akaike Information Criterion):

$$AIC = 2k + n \ln \frac{RSS}{n} \tag{4}$$

Choose the model with minimal AIC.

BIC (Bayesian information criterion):

$$BIC = k \ln n + n \ln \frac{RSS}{n} \tag{5}$$

Choose the model with minimal BIC.

```
BODYFAT ~ ABDOMEN + WEIGHT + WRIST + AGE + THIGH + NECK + FOREARM
```

```
BODYFAT ~ AGE + ABDOMEN + WRIST + HEIGHT
```

3.3 LASSO

Let β denote the parameters, we have:

LASSO (least absolute shrinkage and selection operator) regression uses an objective function as:

$$f(\beta) = RSS_\beta + \lambda ||\beta||_1 \tag{6}$$

Find the β that minimizes $f(\beta)$. Since LASSO can force some parameters to become zero, we can use LASSO to perform feature selection.

```
15 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -10.81225367
AGE           0.02143009
WEIGHT        .
HEIGHT       -0.23939696
ADIPOSITY     .
NECK          .
CHEST         .
ABDOMEN       0.60378632
HIP           .
THIGH         .
KNEE          .
ANKLE         .
BICEPS        .
FOREARM       .
WRIST        -0.55902587
```

It is not difficult to find that **LASSO** and **BIC** methods give the same feature selection result. Then, we will fit and compare the AIC and BIC models.

3.4 Model Diagnostics

AIC model:

Summary:

```
Call:
lm(formula = aic_selection, data = data_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.8430 -2.7203 -0.2356  2.6824  9.4297 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -27.18843    8.31921  -3.268  0.00124 ** 
ABDOMEN       0.79726    0.06489  12.286  < 2e-16 ***
WEIGHT       -0.08411    0.03249  -2.589  0.01022 *  
WRIST        -1.46874    0.47575  -3.087  0.00226 ** 
AGE           0.07042    0.02827   2.491  0.01342 *  
THIGH         0.22036    0.10582   2.083  0.03834 *  
NECK         -0.33554    0.21112  -1.589  0.11329    
FOREARM       0.37317    0.19006   1.963  0.05074 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.876 on 242 degrees of freedom
Multiple R-squared:  0.7418,    Adjusted R-squared:  0.7344 
F-statistic: 99.33 on 7 and 242 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

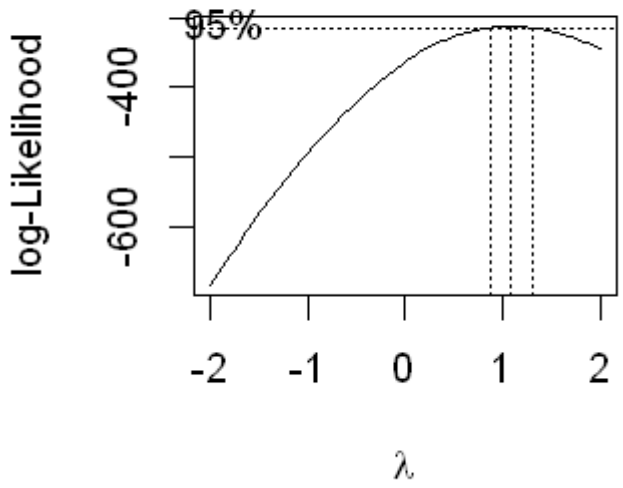
png: 2

CV MSE:

'CV MSE for AIC model: 15.4708836965418'

Box-Cox:

Test if we need a transformation or not



BIC model:

Summary:

```
Call:
lm(formula = bic_selection, data = data_new)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5911 -2.8493 -0.3755  2.9539  8.8630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94338    7.67578   0.383   0.7017
AGE          0.05205    0.02212   2.353   0.0194 *
ABDOMEN      0.69798    0.03072  22.721  < 2e-16 ***
WRIST       -1.71840    0.38130  -4.507 1.02e-05 ***
HEIGHT      -0.27799    0.11486  -2.420   0.0162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.913 on 245 degrees of freedom
Multiple R-squared:  0.7336,    Adjusted R-squared:  0.7293
F-statistic: 168.7 on 4 and 245 DF,  p-value: < 2.2e-16
```

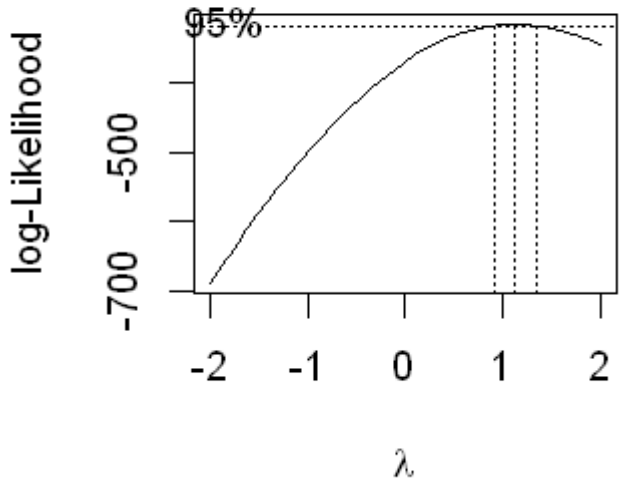
Diagnostic plots:

png: 2

CV MSE:

'CV MSE for BIC model: 15.6645935313245'

Box-Cox:
Test if we need a transformation or not



A matrix: 5 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	-12.175559711	18.06231134
AGE	0.008486528	0.09562201
ABDOMEN	0.637470941	0.75848683
WRIST	-2.469449727	-0.96735232
HEIGHT	-0.504225995	-0.05176167

3.6 Other Models

We also try some other models which we think might be useful:

$$1st: BODYFAT \sim WEIGHT + ABDOMEN$$
$$2nd: BODYFAT \sim ABDOMEN : WEIGHT$$
$$3rd: BODYFAT \sim ABDOMEN$$

Why we try these models: Since by using AIC, BIC, LASSO to select variables, the model according to BIC selection is still with 4 variables, so according to the correlation of bodyfat and other variables, abdomen and weight have relatively higher correlation, so we will try to use this two variables and build other models to see what happen.

A data.frame: 3 × 3

ID	R_square	CV_MSE
<fct>	<dbl>	<dbl>
1st	0.7180885	16.26025
2nd	0.5292770	26.92585
3rd	0.6785452	18.37374

The 1st model has the minimal CV_MSE, choose this one to perform Diagnostics:

png: 2

A matrix: 3 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	-46.3980258	-36.81650491
WEIGHT	-0.1541881	-0.07687099
ABDOMEN	0.7758684	0.98002118

3.7 Summary

1. The **R-squared** and **CV_MSE** of **AIC model** and **BIC model** are very close, so we will choose **BIC model** because it contains with less predictors.

2. Among the three new models, the **1st** model has minimal **CV_MSE**, so, we think this model is also a good choise.
3. From the residual plots we can see that linearity and equal variance assumptions are satisfied for both **BIC** model and **1st** model, because the residual points are evenly distributed on both sides of the line and they are both very close to the x-axis.
4. From the QQ plots, we can see that mornality assumptions are satisfied for both **BIC** model and **1st** model.
5. From the cook's distances in the residuals vs leverage plots, we can assume there is no outlier.

4. Strength & weakness

4.1 Strength

1. To deal with outliers, for each feature, we use 99% quantile to replace the values greater than the 99% quantile and 1% quantile to replace the values less than it, by doing so, we can prevent our model from the influences of extreme values.
2. We repeat 10-fold Cross Validation 1000 times, which can make our results more convincible.
3. Our final model is simple linear regression model, which contains only two significant predictors, and they are easy to measure during daily life in real world. That makes our model concise and easy to interpret.

4.2 Weakness

1. The methods we use in data cleaning (such as using the quantiles) may cause some internal changes of the data structure, which can affect our model results.
2. We only choose two predictors for simplicity, but there is possibility that model with more predictors may give us better results in some specific cases.

5. Conclusion

Our final models are:

Model1 (Our Choise):

$$Model1 : BODYFAT = -41.60 - 0.12 \times WEIGHT + 0.88 \times ABDOMEN$$

Possible rule of thumb: "To calculate your body fat percentage, mutiply your ABDOMEN by 0.88, minus your WEIGHT multiplied by 0.12, then minus 41.60." **Remarks**:
All the variables are significant.
With the cv MSE 16.26, R-Squared 0.7181.

Model2 (An alternative model):

$$Model2 : BODYFAT = 2.94 + 0.05 \times AGE + 0.70 * ABDOMEN - 1.72 \times WRIST - 0.28 \times HEIGHT$$

Remarks:
All the variables are significant.
With the cv MSE 15.66, R-Squared 0.7336.

6. Contribution

- Lingfeng ZHU: Implemented exploratory data analysis and copleted the Shiny App.

- Ruochen YIN: Completed presentation slides and presentation-related works.
- Jiahan LI: Implemented data cleaning and part of model diagnostic plots.
- Chong WEI: Implemented part of model selection and cross validation.