ISyE 6414 Regression Analysis

Computer Project 2

Instructor: Jye-Chyi, Lu

GTID:903450732 Chen-Yang, Liu Nov. 5, 2019

Contents

1 Problem 1: Motivation Studies

1.1 Summary: Forecasting in Supply Chain Management

With development of time, statistical methods are updated accordingly. For a field of supply chain management, it also needs some new forecasting techniques to mitigate risks. In this article, uncertainty might create loss, so for risk management, supply chain experts would like to discover new techniques that could address inventory problems.

The article first mentions demand amplification due to globalization. The trend of globalization leads to new obstacles in a supply chain field. The problems is how to handle increasing uncertainty with growing demands from customers. Then the author offer some quantitative methods to try to solve these problems. The first method is a classical 4-stage beer-game simulation, which shows demand information in each stage. But this method does not lower negative effect from demand amplification. Meanwhile, the author offers second traditional method, Exponential Smoothed Moving Average, which only shows heightened demand amplification.

From the above two traditional methods, which might not be a good choice to solve supply chain problems nowadays, the author provides three advanced method. One of them is autoregressive moving average models, which might tame the Bullwhip effect.

The second method is Classical Linear Regression Models(CLRM). This model is based on the assumption of homoskedasticity, which contradicts the Bullwhip Effect, which shows different variance in different stages. Different variance means that CLRM could not work for this field. Therefore, the last method is GARACH, which requires high volume data to be effectively utilized and generate results with higher accuracy levels.

In a nutshell, the author provides some methods to deal with supply chain problems, but each method has its drawbacks and mentions the possibility to use these methods to lower uncertainty.

1.2 Summary: Forecasting in Logistics

The author tries to solve the problem of forecasting daily demands of large freight transportation applications. To forecast future demands, most models, such as ARIMA or least-squares, are based on historical data, but as for freight flow, there is no history, so historical method is not ideal for forecasting freight flows.

To address this problem, the author offers exponential smoothing-based models that can update model when new information is coming. What's more, the model is refined and called damped trend multi-calendar exponential smoothing method which combines periodic and non-periodic calendar effects.

Then the author concludes that DTMC model provides better forecasts than the best fitted ARIMA models, but those two models are not suitable for irregular patterns due to calendar effects.

The formula is as follows:

Gardner's Model 7-4(Gardner (1985)) has a model with damped trend and multiplicative season for one season with period p. The equation to forecast m periods ahead from period t is

$$\hat{X}_t(m) = \left(S_t + \sum_{i=1}^{m} \phi^i T_t\right) I_{t-p+(m \mod p)} \tag{5}$$

The seasonal multipliers represent elements in a cycle such as months of the year (I_{t-11} to I_t). The subscript $t - p + (m \text{ mod } p)$ refers to the forecasted month.

and the updating equations are:

Since the calendar effects and trend multiplier add two more parameters to be determined, we choose to use the Brown model. The updating equations for the baseline, trend and seasonal multipliers, in this case, are (from Model 7-4 in Gardner (1985)):

$$S_t = S_{t-1} + \phi T_{t-1} + \alpha(2-\alpha)e_t/I_{t-p} \tag{6}$$

$$T_t = \phi T_{t-1} + \alpha(\alpha - \phi + 1)e_t/I_{t-p} \tag{7}$$

$$I_t = I_{t-p} + \delta[1 - \alpha(2-\alpha)]e_t/S_t \tag{8}$$

to include multiple calendar effects, the author provides more notation to show the models.

J = Set of calendar attributes representing weekdays, months, etc

$y_j(t) = \begin{cases} 1 & \text{if attribute } j \in J \text{ is active in period } t \\ 0 & \text{otherwise} \end{cases}$

$I_t^m(j)$ = Calendar factor for attribute $j \in J$ for period $t+m$ as of period t,
$= \exp(a_{j,t} \cdot y_j(t+m))$

I_t^m = Calendar factor across all attributes for period $t+m$ as of period t,
$= \prod_{j \in J} I_t^m(j) = \exp(\sum_{j \in J} a_{j,t} \cdot y_j(t+m))$

$a_{j,t}$ = Coefficient used to calculate calendar factor $I_t^m(j)$

λ = Smoothing parameter for the residuals

2 Problem 2: Model Fitting

2.1 Overview

In the second problem, we will use data from the Myers's experiment. Using the collected data, we will fit a second-order polynomial regression model for prediction of Mercaptobenzothiazole(MBT).

2.2 Data Exploration

Myer conducted the experiment on MBT to explore a function of reaction time and temperature. Time and temperature are observed and then we will create coded variables, which make it easier to interpret the relationship among variables, for Central Composite Design(CCD):

Coded Variable Formula as follows:

$$x_1 = \frac{time(hrs)-12}{5.6} = 5.6x_1 + 12$$
$$x_2 = \frac{temp.(Celsius)-250}{20} = 20x_2 + 250$$

By using the above formula, we create coded table for CCD:

Figure 1 ***Coded Variable Table***

	y	time	temp	x1	x2	x1x1	x2x2	x1x2
1	81.3	6.4	230	-1	-1	1	1	1
2	85.3	17.6	230	1	-1	1	1	-1
3	83.1	6.4	270	-1	1	1	1	-1
4	72.7	17.6	270	1	1	1	1	1
5	82.9	4.0816	250	-1.414	0	1.999396	0	0
6	81.7	19.9184	250	1.414	0	1.999396	0	0
7	84.7	12	221.72	0	-1.414	0	1.999396	0
8	57.9	12	278.28	0	1.414	0	1.999396	0
9	82.9	12	250	0	0	0	0	0
10	81.2	12	250	0	0	0	0	0
11	82.4	12	250	0	0	0	0	0

2.3 Model Fitting

Second-order polynomial regression model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1 x_1 + \beta_{22} x_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

After CCD design, fitted-value model is:

$$Y = 82.166 - 1.012 x_1 - 0.88 x_2 + 1.018 x_1 x_1 - 4.484 x_2 x_2 - 3.6 x_1 x_2$$

And, Using R to run regression models will yield a regression summary and, at the same time, ANOVA table is shown to explain the regression model.

Figure 2 ***Regression Table***

```
Call:
lm(formula = y ~ x1 + x2 + x1x1 + x2x2 + x1x2, data = p2d)

Residuals:
      1       2       3       4       5       6       7       8       9      10      11
-0.8998 -2.0753  5.8758  4.7003 -2.7321 -1.0695  2.8910 -6.6926  0.7341 -0.9659  0.2341

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   82.166      2.887  28.461    1e-06 ***
x1            -1.012      1.768  -0.573   0.5917
x2            -6.088      1.768  -3.443   0.0184 *
x1x1           1.018      2.105   0.484   0.6491
x2x2          -4.484      2.105  -2.130   0.0864 .
x1x2          -3.600      2.500  -1.440   0.2094
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5 on 5 degrees of freedom
Multiple R-squared:  0.8011,    Adjusted R-squared:  0.6023
F-statistic: 4.029 on 5 and 5 DF,  p-value: 0.07619
```

Figure 3 ***ANOVA Table***

```
> anova = aov(p2lm)
> summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
x1           1   8.20    8.20   0.328 0.5917
x2           1 296.45  296.45  11.856 0.0184 *
x1x1         1  33.71   33.71   1.348 0.2980
x2x2         1 113.49  113.49   4.539 0.0864 .
x1x2         1  51.84   51.84   2.073 0.2094
Residuals    5 125.02   25.00
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

From the Regression table, some key points should be explained:

- Residual Standard Error(RSE): The RSE is 5 where degrees of freedom is 5. RSE means the average amount that the response variable(y) will deviate from the true regression line.

- Multiple R-Squared: Multiple R-Squared can check how well the model fits the data. In the regression table, the value of this is 0.8011 which shows this model is not bad for prediction.

- Adjusted R-Squared: Although Multiple R-Squared checks models, this is not suitable for many explanatory variables. Therefore, Adjusted R-Squared normalizes Multiple R-Squared by taking into account the numbers of samples and variables. In this regression table, Adjusted R-squared is 0.6023 which means that this model is not bad for prediction.

- F-statistic: the reason for this test is based on the fact that if the regression models run multiple hypothesis tests (namely, on coefficients), some variables might be included but they are not actually significant. In the example, the value of F-statistic is 4.029 which could be used to determine whether all coefficients are equal to zero.

- P-value: For this regression table, the p-value is 0.07619. Thus, if the $\alpha = 0.05$, the p-value shows that null hypothesis cannot be rejected, but if the $\alpha = 0.1$, the p-value shows that null hypothesis can be rejected.

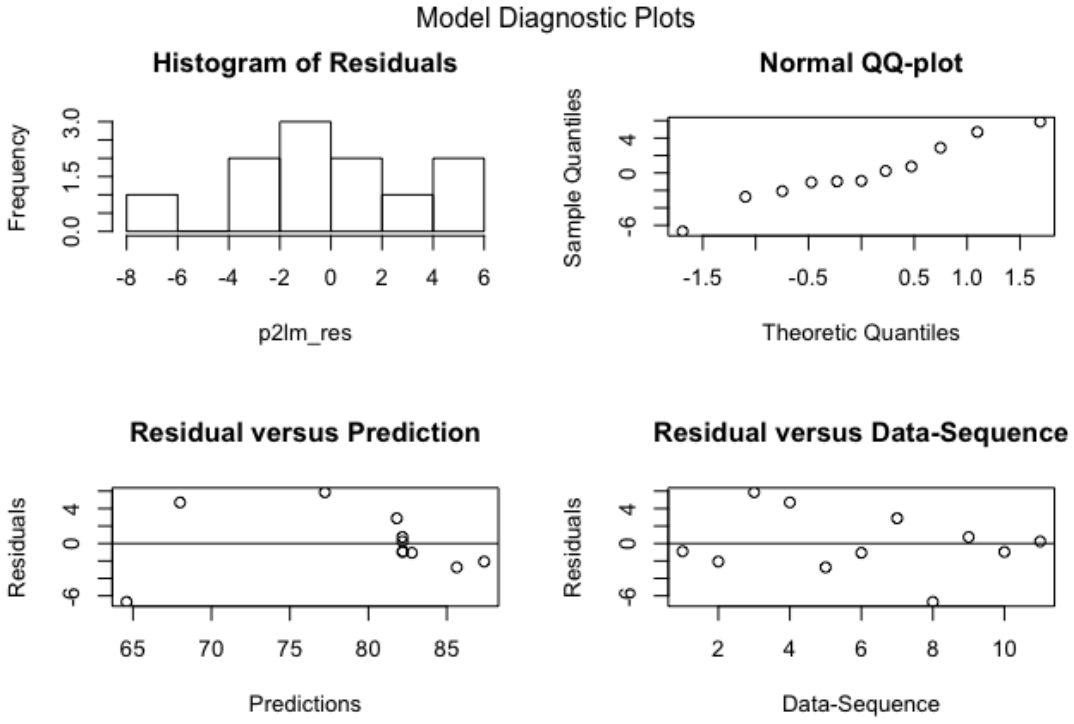
From the ANOVA table, there are two independent variables that are significant. The p-value of x_2 is 0.0184, which is smaller than α which is 0.05 and the p-value of x_2x_2 is 0.0864, which is smaller that α which is 0.1. The remaining variables are not significant enough.

2.4 Diagnostic Model

After running the regression model, some key points are needed to check whether this model is good enough for making prediction. Usually, histogram of Residuals, QQ-Plot, Residual against Prediction and Residual against Data-Sequence could explain the quality of models.

From the experiment, Model Diagnostic Plots is below:

Figure 4 ***Model Diagnostic Plots***



Now, a Normal QQ-Plot and histogram of residuals can check normal assumption.

- Normal QQ-Plot: QQ-Plot assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. From the figure, it seems that the data set is from normal distribution because these data points could be linked like a straight line.

- Histogram of Residuals: The histogram of residuals could check whether residuals are outliers and show the distribution of residuals. From the figure, because the sample sizes of residuals are generally small (¡50) which means this experiment has limited treatment combinations, the histogram is not

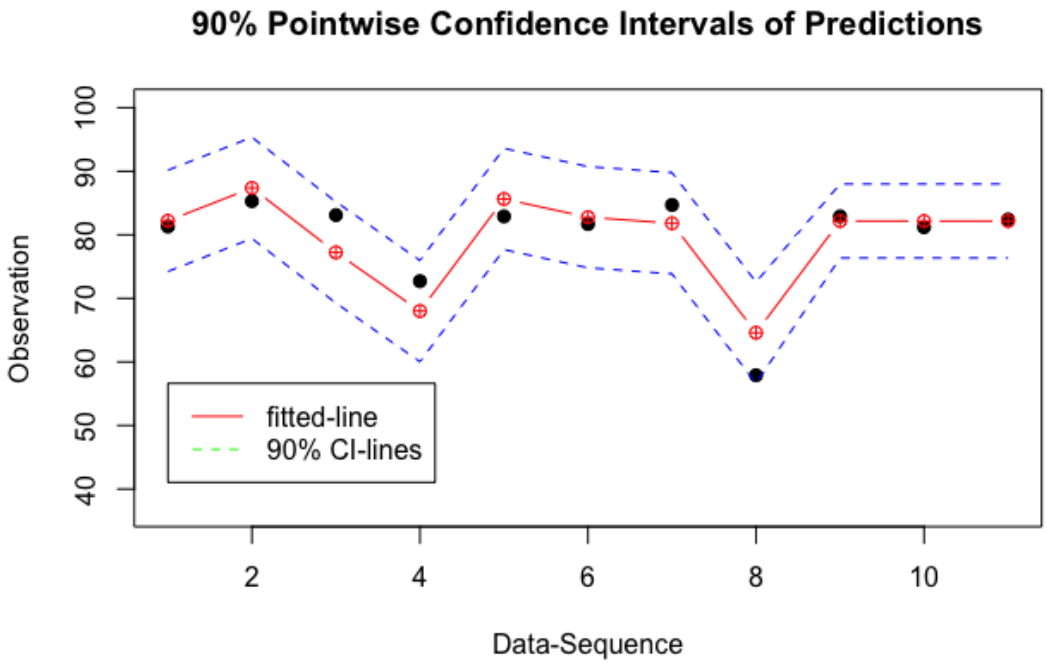
be the best choice for judging the distribution of residuals. However, if sample sizes are large enough, a histogram of residuals provides a way to test normality of data.

Then, Residual versus Prediction can check whether the predicted values are good enough. As for Residual versus Data-Sequence, it is a way of detecting a particular form of non-independence of the error terms, namely serial correlation. Therefore, from the figure, Residual versus Prediction shows there are two points far away from zero. Those two points might be outliers. And in the Residual versus Data-Sequence, the plot shows that each error term is independent and no serial correlation.

2.5 Confidence Intervals

Point-wise Confidence Interval of predictions is employed since it can check whether actual data points are within confidence intervals. That is, it could be used to check possible outliers of data sequence.

Figure 5 ***90% Pointwise Confidence Intervals***



From the Figure 5, one observed point is on the line of 90% pointwise-CI intervals.

2.6 Regression for Original Variables

So far some important regression models' points have been discussed, but the problem would not be fully analyzed without giving true description about each variables. As a result, ANOVA and Regression tables are presented to let readers fully understand the relationship among MBT, Time and temperature.

Figure 6 ***Original Variable's ANOVA Table***

```
> aov(p2olm_summary)
Call:
   aov(formula = p2olm_summary)

Terms:
                     time      temp    timesq    tempsq  timetemp Residuals
Sum of Squares    8.19601 296.44758  33.71495 113.48974  51.84000 125.01900
Deg. of Freedom         1         1         1         1         1         5

Residual standard error: 5.00038
Estimated effects may be unbalanced
```

Figure 7 ***Original Variable's Regression Table***

```
Residuals:
      1       2       3       4       5       6       7       8       9      10      11
-0.8998 -2.0753  5.8758  4.7003 -2.7321 -1.0695  2.8910 -6.6926  0.7341 -0.9659  0.2341

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.319e+02  3.373e+02  -1.874   0.1199
time         7.076e+00  5.817e+00   1.216   0.2781
temp         5.686e+00  2.646e+00   2.149   0.0843 .
timesq       3.245e-02  6.711e-02   0.484   0.6491
tempsq      -1.121e-02  5.262e-03  -2.130   0.0864 .
timetemp    -3.214e-02  2.232e-02  -1.440   0.2094
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5 on 5 degrees of freedom
Multiple R-squared:  0.8011,    Adjusted R-squared:  0.6023
F-statistic: 4.029 on 5 and 5 DF,  p-value: 0.07619
```

Original's ANOVA table is the same as the previous ANOVA table, but Original's Regression table is a little different from previous one. It is because for convenience, changing some variables into easily seen numbers will change the value of each term.

2.7 Conclusion

From the regression formula,

$Y = 82.166 - 1.012time - 0.88temp + 1.018timesq - 4.484tempsq - 3.6timetemp$

From the Regression table, the p-value of temperature lies in $\alpha = 0.1$. Therefore, temperature variable is significant enough and shows that it has influence on MBT.

3 Simulation Studies

3.1 Overview

In the Problem 3, generating errors from different variance could give readers a basic understanding about influence of variance. In this problem, given that $Y = 2.5 - 1.8x + e$ where $e \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma)$. There are two scenarios. One of them is $\sigma = 1.0$ and the other is $\sigma = 5$.

3.2 Scenario 1: $\sigma = 1.0$

In this part, under $\sigma = 1.0$, 10 random errors are generated. The x and y are listed below:

Figure 8 $X - Y_1$ ***Table***

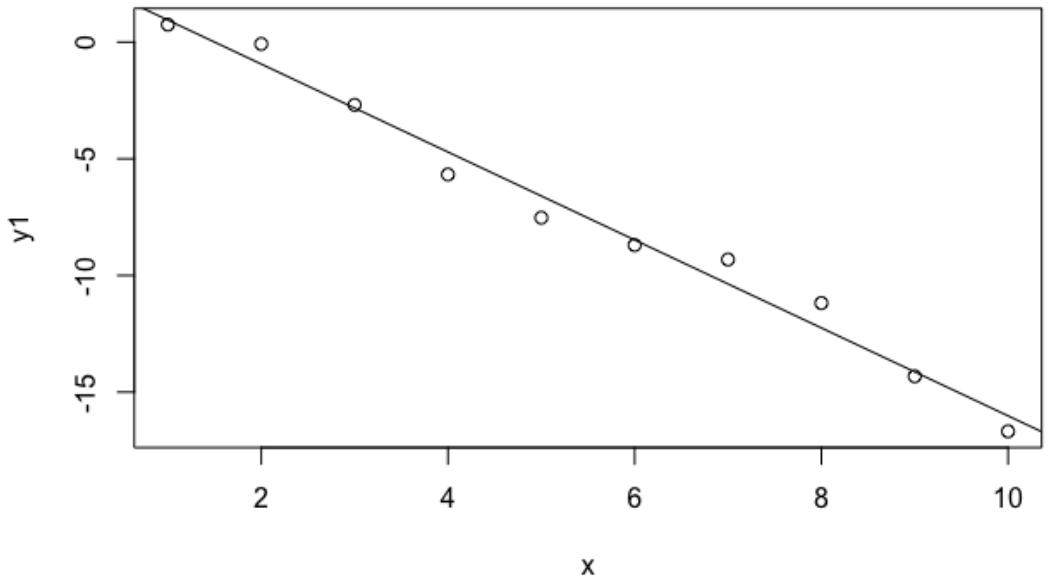
x	y1
1	0.747358694140217
2	-0.0808386561363088
3	-2.69041880160602
4	-5.67480283426056
5	-7.52719980300307
6	-8.69926750286043
7	-9.31710616835129
8	-11.1839574197496
9	-14.3227426017508
10	-16.6795363224712

And, a Regression Model on x and y_1 is ran. The Regression Table and the regression line is shown below:

Figure 9 $X - Y_1$ ***Regression Table***

```
Call:
lm(formula = y1 ~ x, data = p3d1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9601 -0.5425 -0.1876  0.6818  1.0726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.82722    0.55852   5.062 0.000975 ***
x           -1.88547    0.09001 -20.946 2.83e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8176 on 8 degrees of freedom
Multiple R-squared:  0.9821,    Adjusted R-squared:  0.9799
F-statistic: 438.7 on 1 and 8 DF,  p-value: 2.832e-08
```

Figure 10 $X - Y_1$ ***Regression Line***



From the summary of Regression table, the regression model on x and y_1 is derived:

$$Y_1 = 2.82722 - -1.88547x$$

From the regression table, the p-value of F-test is really small, which means that the model is significant. As for R^2, is 0.9821. It means 98.21% of variation can be explained by this model.

3.3 Scenario 2: $\sigma = 5.0$

In this part, under $\sigma = 5.0$, 10 random errors are generated. The x and y are listed below:

Figure 11 $X - Y_1$ ***Table***

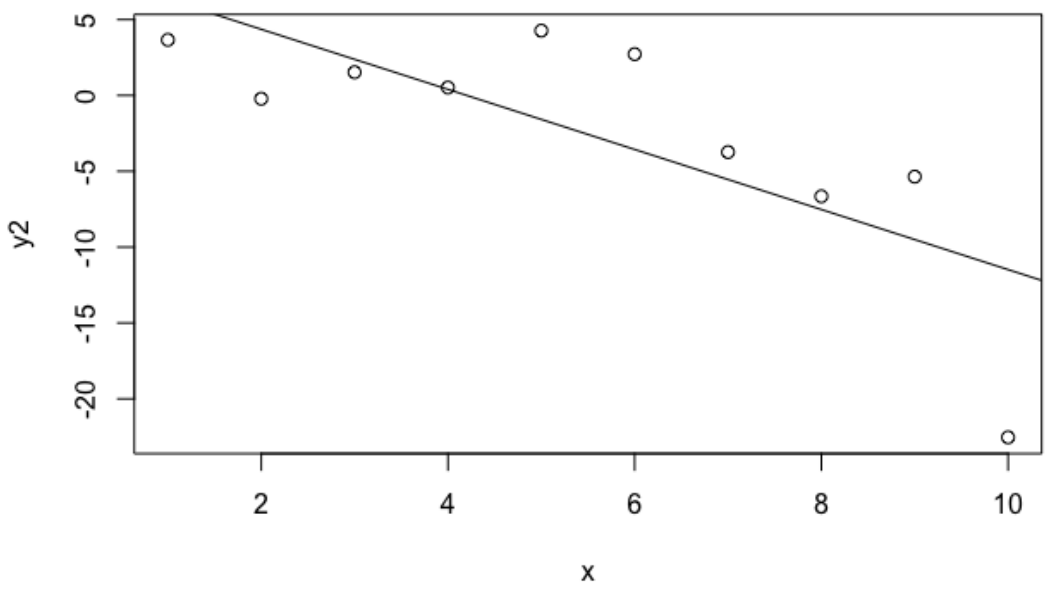
x	y2
1	3.66098682258303
2	-0.220513919845174
3	1.52907160782632
4	0.518605454345307
5	4.26825352440688
6	2.72057963405782
7	-3.73654984525632
8	-6.65216925442298
9	-5.35570498598529
10	-22.538226357855

And, a Regression Model on x and y_2 is ran. The Regression Table and the regression line is shown below:

Figure 12 $X - Y_2$ ***Regression Table***

```
Call:
lm(formula = y2 ~ x, data = p3d2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0406  -2.2177   0.5046   3.5744   6.2919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.3181     3.7882   2.196   0.0594 .
x            -1.9816     0.6105  -3.246   0.0118 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.545 on 8 degrees of freedom
Multiple R-squared:  0.5684,    Adjusted R-squared:  0.5144
F-statistic: 10.53 on 1 and 8 DF,  p-value: 0.01178
```

Figure 13 $X - Y_2$ ***Regression Line***



From the summary of Regression table, the regression model on x and y_1 is derived:

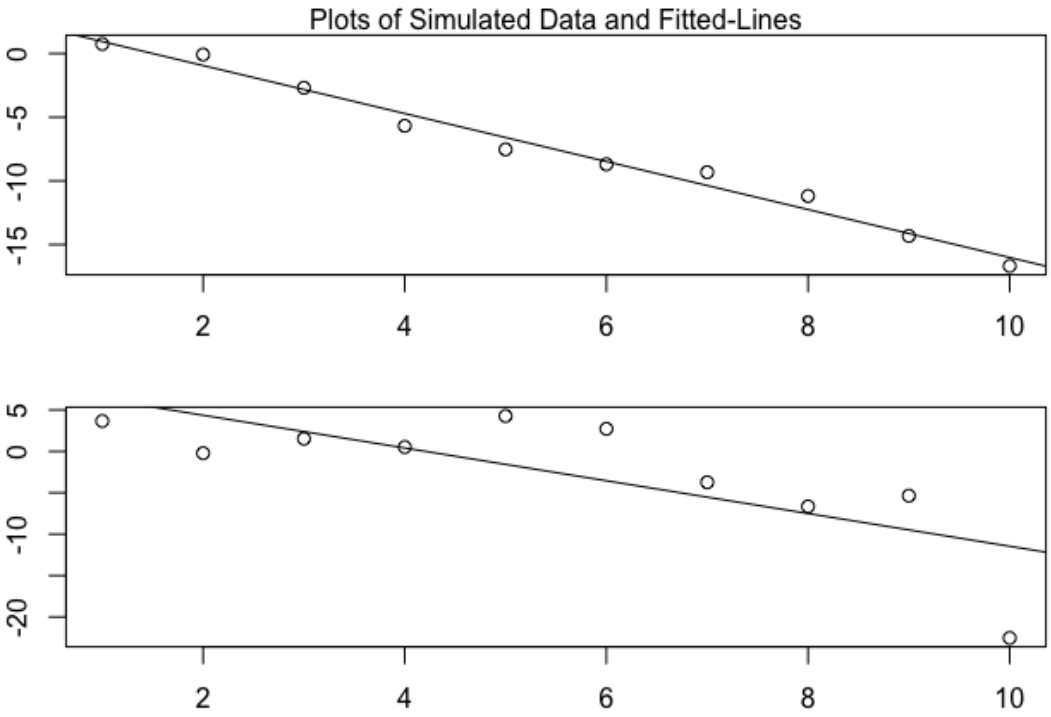
$$Y_2 = 8.3181 - 2.2177x$$

From the regression table, the p-value of F-test is 0.1178(around 0.1), which means that the model is not significant under $alpha = 0.05$ but is nearly significant under $\alpha = 0.1$. As for R^2, is 0.5684. It means 56.84% of variation can be explained by this model. In other words, this model is not good enough for prediction.

3.4 Conclusion

A combined figure is below:

Figure 14 ***Plots of Simulated Data and Fitted-Lines***



Clearly, $e \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma = 5)$ is less straight than $e \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma = 1)$. Moreover, from the above figure, data points with $\sigma = 5$ spread wider than ones with $\sigma = 1$. Therefore, σ has influence on quality of linear models.

4 Real-life Data Analysis for Variable Selections

4.1 Overview

The dataset fram200.xls includes sex, systolic blood pressure (sbp), diastolic blood pressure (dbp), age, serum cholesterol (scl), coronary heart disease outcome (chdfate), number of days before a follow-up (followup), age, body mass index (bmi), month of the year in which baseline exam occurred (month), and patient id (id) for 199 patients. In this report, the variables sex, dbp, scl, chdfate, followup, age, bmi, and month will be used to make a prediction on a patient's systolic blood pressure (sbp). In order to construct a model for this prediction, Box-Cox transformation, Variance Inflation Factor(VIF), Variable Selection will be employed for building a quality model.

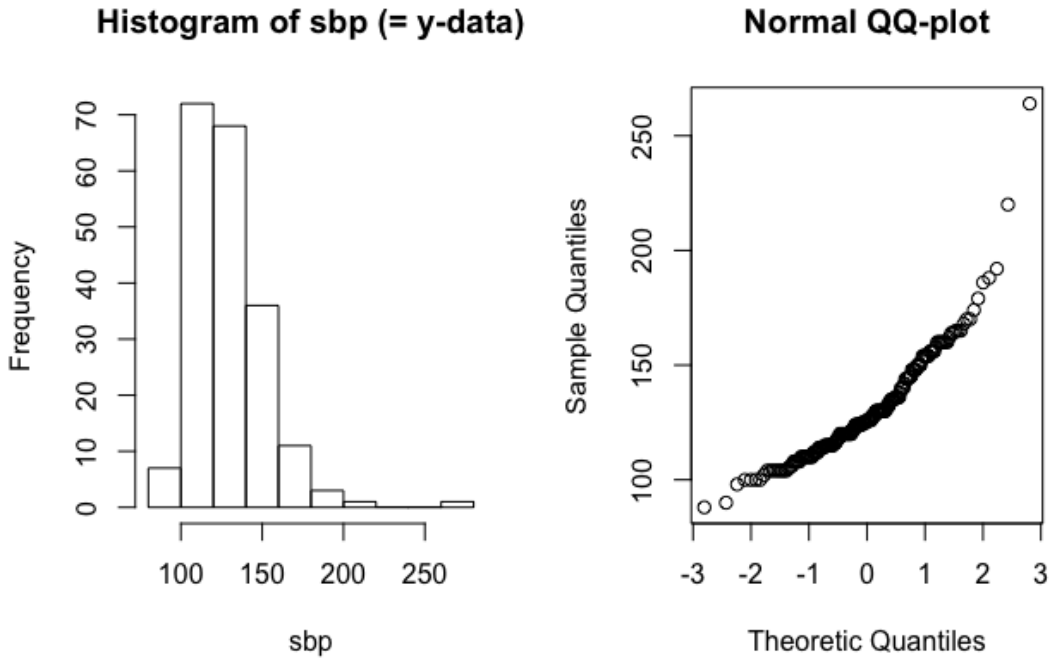
The following is the description of the columns in *fram200.xls* data set:

Sex	sex("1" is "men")
sbp	Systolic Blood Pressure
dbp	Diastolic Blood Pressure
scl	Serum Cholesterol
chdfate	Coronary Heart Disease ("1" means with CHD)
followup	Follow-up in Days
age	Age in Years
bmi	Body Mass Index
month	Study Month of Baseline Exam
id	identification number

4.2 Data Exploration

The purpose of the problem is to discover the relationship between sbp and other posiible variables. Thus, the first step is to know the distribution of sbp to get insight of data. Histogram of sbb and Normal QQ-plot are created to know whether sbp follows normal distribution.

Figure 15 ***Histogram of sbp and Normal QQ-Plot***

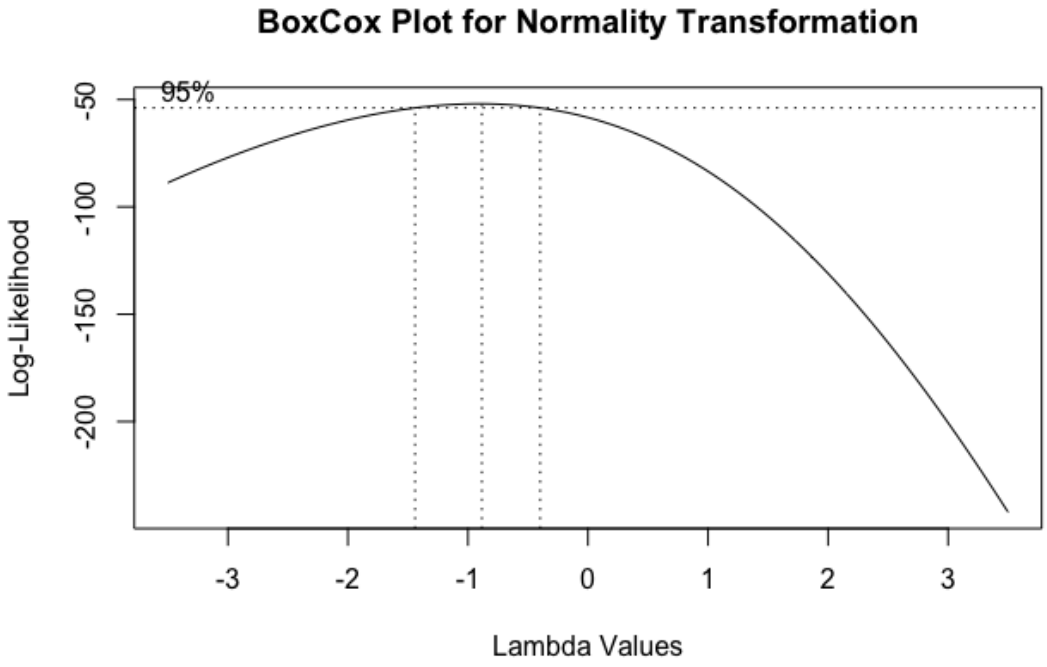


From the histogram and normal QQ-Plot, the distribution of sbp does not follows normal distribution. Obviously, the distribution of sbp is right-tailed skew. The frequency of sbp is high on the interval of 100 to 145. And, from the QQ-Plot, the data sequence of sbp does not fit a nearly diagonal line. What's more, there is one data point which is an outlier. Hence, Box-Cox transformation is needed in this problem since this transformation will change data into normally distributed data.

4.3 Data Transformation

The data transformation method used here is Box-Cox transformation, which changes the distribution to a more linear or approximated normal distribution. For this method, lambda values are chosen based on analyzers' experience. Chosen lambda value is a sequence from -3.5 to 3.5 with each interval of 1/10.

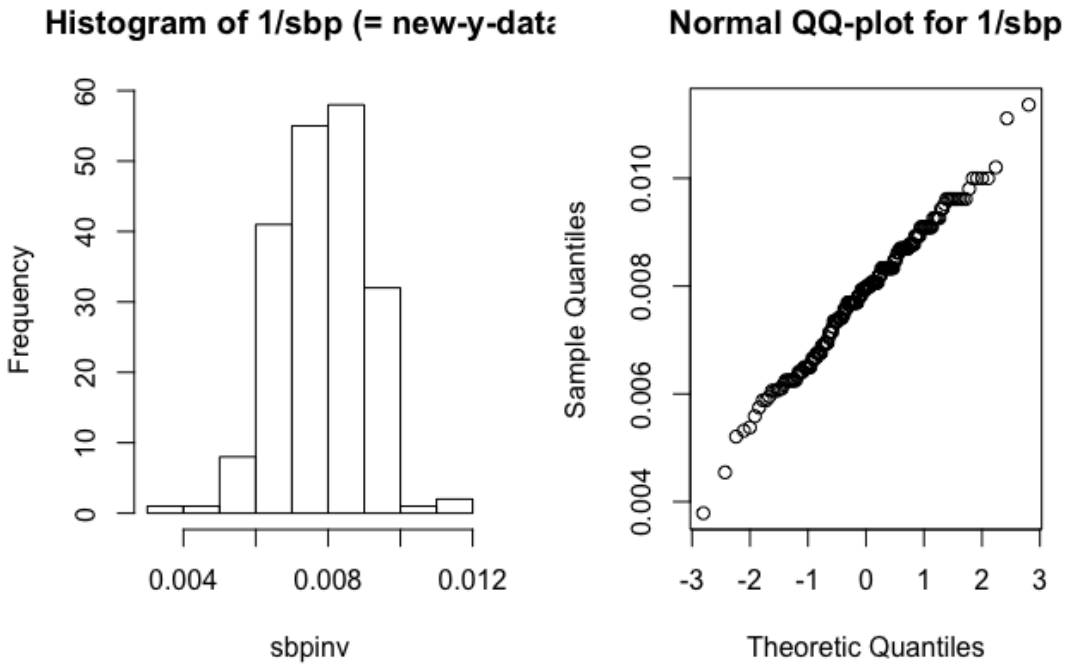
Figure 16 ***Box-Cox Plot***



From the above figure, as lambda value approaches -1, the values of Log-Likelihood reaches maximum. Therefore, $\lambda = -1$ is selected in this problem for further analysis.

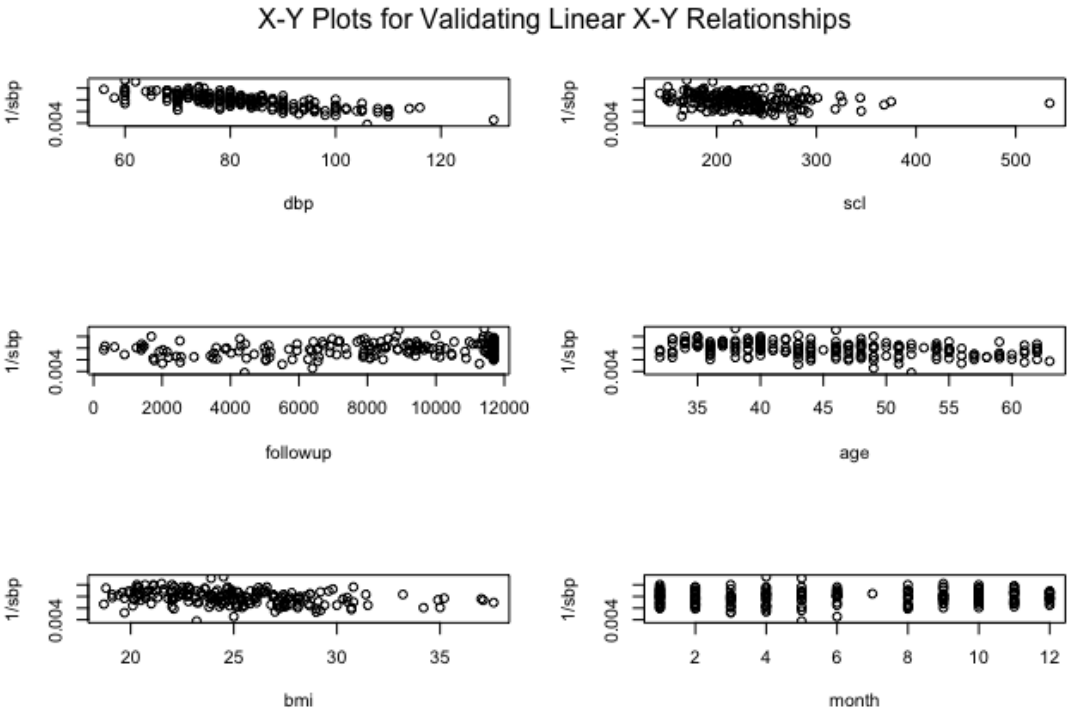
Now, after transformation, data of sbp should be checked again to ensure that the distribution of the data nearly follows normal distribution. From the Figure 17, the approximately normal distribution is seen. The frequency of the histogram is nearly bell-shaped. And, the Normal QQ-Plot is nearly a straight line and the data sequence of the center has more data points which follows a normal distribution. All in all, after transformation, some data points are changed into a nearly normal distribution which further analysis could be conducted.

Figure 17 ***Histogram of sbp and Normal QQ-Plot (After)***



After transformation, the relationship between sbp and other variables is discovered. From the Figure 18, the relationship between dbp and sbp is possibly correlated.

Figure 18 ***X-Y Plots for Validating X-Y Relationships***



4.4 VIF Analysis and Model Building

Although sbp data points are changed into more like a normal distribution, for Regression analysis, some of variables are unecessary for sbp and multicollinearity will occur if there are more exploratory variables. Hence, VIF analysis will be performed to drop unnecessary variables. There are two main parts of this problem. The first one is that VIF analysis is conducted on main effects and the other is that VIF analysis is performed on two-variable combinational-effects.

4.4.1 VIF on Two-Variable Combinational-Effects

There are 15 two-way combinations, but only continuous variables are considered in this analysis since sbp belongs to the continuous variable. The regression model will be derived and then the Regression Table will be used for determining the quality of model.

The regression model on all the main-effects and two variable interactions:

$sbpinv = \beta_0(Intercept) + \beta_1(dbp) + \beta_2(scl) + \beta_3(followup) + \beta_4(age) + \beta_5(bmi) + \beta_6(month) + \beta_7(sex) + \beta_8(chdfate) + \beta_{1,1}(dscl) + \beta_{1,2}(dfol) + \beta_{1,3}(dage) + \beta_{1,4}(dbmi) + \beta_{1,5}(dmo)$

Before regression models are created, one point should be noted that there might be multicollinearity problem in regression models, so VIF analysis should be conducted first to detect multicollinearity.

Figure 19 VIF Table

	vifmod2
dbp	124.590020466808
scl	49.0124896245254
followup	66.0206295868804
age	59.4880441531409
bmi	62.6170966439502
month	52.9322139444442
sex	1.07825080679535
chdfate	1.30944503157326
dscl	91.9858212417988
dfol	63.2191849626828
dage	122.149918683483
dbmi	181.949408334324
dmo	53.7766053948665

(a) VIF on Main-Effects/Two-Variables

	vifmodn1
dbp	1.23842543701183
scl	1.19474673863323
followup	1.33320050278732
age	1.22779144797003
bmi	1.23156470930215
month	1.04469393133536
sex	1.02418752370365
chdfate	1.27932192072123

(b) VIF on Main-Effects

From the Figure 19, the values of two-variables interactions are too large to make a regression model, so in this report, two-variables situation will not be considered, although from the regression table, Adjusted R-Square is 0.6436.
To fully understand two-variable interaction, the regression table of this is presented. From the Figure 20, Adjusted R-Squared is 0.6436 and the p-value is less than 2.2e-16 which is really small.

Figure 20 ***Main-Effects and Two-Variables Interaction Regression Table***

```
Call:
lm(formula = sbpinv ~ dbp + scl + followup + age + bmi + month +
    sex + chdfate + dscl + dfol + dage + dbmi + dmo, data = p4dataext)

Residuals:
       Min         1Q     Median         3Q        Max
-2.182e-03 -4.337e-04 -3.820e-06  4.653e-04  1.885e-03

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.589e-02  3.773e-03   6.863 9.88e-11 ***
dbp         -1.963e-04  4.509e-05  -4.354 2.21e-05 ***
scl         -8.200e-06  7.521e-06  -1.090   0.2770
followup    -7.695e-08  1.228e-07  -0.627   0.5316
age         -1.225e-04  4.762e-05  -2.572   0.0109 *
bmi         -1.565e-04  1.105e-04  -1.416   0.1583
month       -5.993e-05  1.020e-04  -0.588   0.5574
sex         -1.251e-04  1.068e-04  -1.171   0.2432
chdfate     -2.171e-04  1.250e-04  -1.738   0.0839 .
dscl         9.067e-08  8.815e-08   1.029   0.3051
dfol         8.655e-10  1.457e-09   0.594   0.5533
dage         1.091e-06  5.697e-07   1.914   0.0571 .
dbmi         1.834e-06  1.328e-06   1.382   0.1688
dmo          1.015e-06  1.253e-06   0.810   0.4189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007224 on 185 degrees of freedom
Multiple R-squared:  0.667,     Adjusted R-squared:  0.6436
F-statistic:  28.5 on 13 and 185 DF,  p-value: < 2.2e-16
```

4.4.2 VIF on Main-Effects

The regression model of Main-Effects is:

$$sbpinv = \beta_0(Intercept) + \beta_1(dbp) + \beta_2(scl) + \beta_3(followup) + \beta_4(age)$$
$$+\beta_5(bmi) + \beta_6(month) + \beta_7(sex) + \beta_8(chdfate)$$

From the Figure 19, the values of Main-Effects are suitable because all of them are not greater than 4.5. Therefore, in this report, the regression model on Main-effects will be used for prediction.

Figure 21 ***Main-Effects Regression Table***

```
Call:
lm(formula = sbpinv ~ dbp + scl + followup + age + bmi + month +
    sex + chdfate, data = p4dataext)

Residuals:
       Min         1Q     Median         3Q        Max
-2.279e-03 -4.594e-04  6.170e-06  4.556e-04  1.952e-03

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.507e-02  6.065e-04  24.849  < 2e-16 ***
dbp         -6.611e-05  4.554e-06 -14.516  < 2e-16 ***
scl         -4.396e-07  1.190e-06  -0.370   0.7121
followup    -8.328e-09  1.768e-08  -0.471   0.6381
age         -3.496e-05  6.931e-06  -5.044 1.06e-06 ***
bmi          1.366e-06  1.570e-05   0.087   0.9307
month        2.237e-05  1.451e-05   1.542   0.1248
sex         -9.420e-05  1.055e-04  -0.893   0.3730
chdfate     -2.089e-04  1.251e-04  -1.670   0.0967 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007319 on 190 degrees of freedom
Multiple R-squared:  0.649,     Adjusted R-squared:  0.6342
F-statistic: 43.91 on 8 and 190 DF,  p-value: < 2.2e-16
```

An interesting point is that the adjusted R-squared in the regression table of main effects is less than one with two-variable interaction. Does it mean that the regression table on main effects and two-variable interactions should be used? Not necessarily. Based on VIF analysis, if two-variables were taken into analysis, there would be collinearity problem. Therefore, the regression model with only main effects is used in the analysis.

4.5 Variable Selections

After VIF analysis, some of variables are not helpful for analysis. Therefore, in this section, Variable Selection will be performed to select few of variables to create the most representative model.

4.5.1 Forward Selection

Forward selection means the initial model contains no regressors but they enter the model one at a time. In this situation, a regressor enters the model if its p-value is less than a critical value, say 0.05. That is, originally assumed that the first model is $y = \beta_0 + \beta_1 x1 + \epsilon$. After the model is built, regression table and anova table will be derived to see whether p-value is less than a critical value. Then, extra variable is added into the model, which means the previous one becomes $y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \epsilon$. Now the critical value is set as 0.15. So if p-value of models goes into the critical region, the model will be accepted.
The steps are as follows:

Step	# of Variables	Selected Variable	Partial-F Value	P-Value	P-Value > 0.85 ?
1	2	dbp	N/A	< 2.2e-16	N
2	3	scl	2.4054	0.1225	N
3	4	followup	2.419	0.1215	N
4	5	age	24.346	1.727e-06	N
5	6	bmi	3e-04	0.9862	Y

From the above table, the model with 6 variables is not accepted because the p-value of 6-variable model (0.9862) enters into critical region. Hence, the variable(bmi) will not be included in this model.
Now, the final model is derived:

$$sbpinv = \beta_0 + \beta_1 dbp + \beta_2 scl + \beta_3 followup + \beta_4 age$$

Figure 22 ***Regression Table***

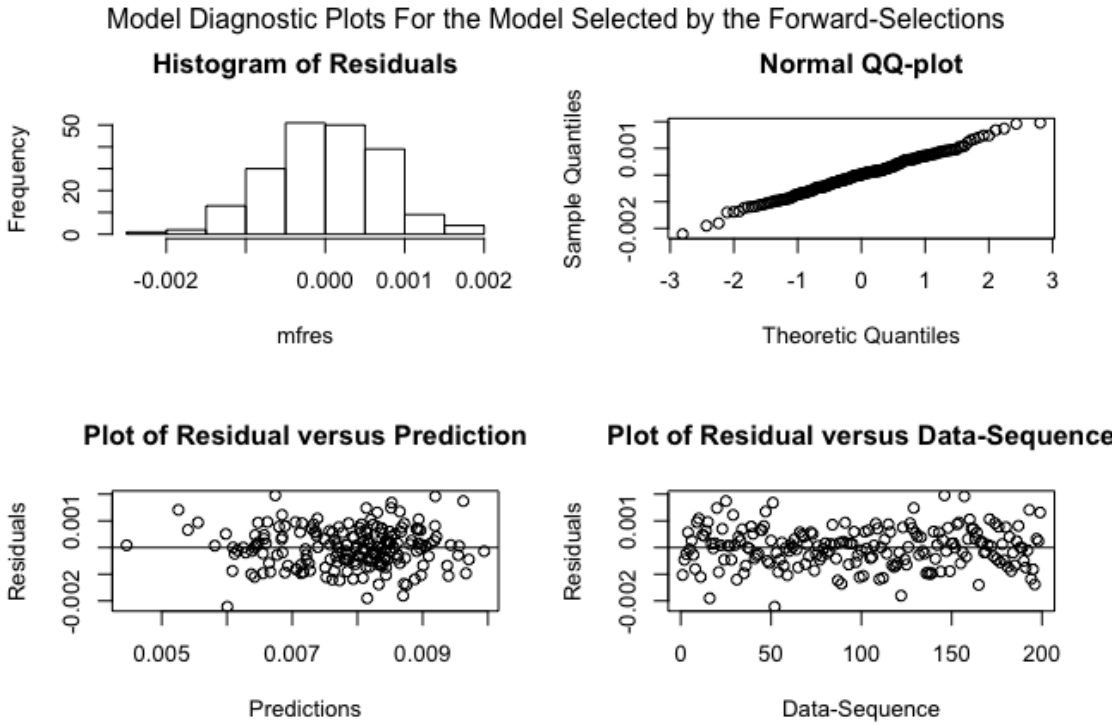
```
Call:
lm(formula = sbpinv ~ dbp + scl + followup + age, data = p4dataext)

Residuals:
       Min         1Q     Median         3Q        Max
-2.217e-03 -4.601e-04  1.518e-05  5.289e-04  1.955e-03

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.502e-02  5.352e-04  28.063  < 2e-16 ***
dbp         -6.712e-05  4.339e-06 -15.471  < 2e-16 ***
scl         -7.293e-07  1.154e-06  -0.632    0.528
followup     3.030e-09  1.687e-08   0.180    0.858
age         -3.366e-05  6.822e-06  -4.934 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007346 on 194 degrees of freedom
Multiple R-squared:  0.6389,    Adjusted R-squared:  0.6315
F-statistic: 85.82 on 4 and 194 DF,  p-value: < 2.2e-16
```

From the Figure 22, coefficients of dbp and age are not equal to zero and Adjusted R-squared is 0.6315.

Model Diagnosis of Forward Selection

After the linear model is derived, model diagnosis then is performed to ensure the quality of the model. From the Figure 23, histogram of residuals and normal QQ-plot shows the model follows normal distribution. As for plot of residuals, most points lie in -0.002 and 0,001, so there are no obvious outliers in this figure. Hence, the model is well-defined and reasonable to predict sbp.

Figure 23 ***Regression Table***



4.5.2 All-Subsets Selection

The second variable selection method is All-Subsets Selection, which means the general idea behind this is that we select the subset of predictors that do the best at meeting some well-defined objective criterion, such as having the largest R2 value or the smallest MSE.

- Adjusted R-Squared: the one with the largest adjusted R_2-value is chosen since the formula of Adjusted R-Squared makes us pay a penalty for adding more predictors to the model
- Mallows' C_p-statistic: the statistic estimates the size of the bias that is introduced into the predicted responses by having an under-specified model.
- BIC: This gives a way of considering the trade-off between a simple parsimonious model (practical significance) and a more complex and closer to reality model (statistical significance).

From Figure 25 and Figure 26, In this all-subset regression, the number of best model is set as 2 for each number of predictors and the limit on number of

variables is 10. As a result, there are 20 combinations and based on Adujusted R-squared, No. 8(1) is highest (0.6490464) among others which means 64% of the variation could explain this model, so the No.8 combination is chosen. The No.8 combination means there are eight variables which are selected for building models.

Figure 24 ***Combinations Table***

Selection Algorithm: exhaustive

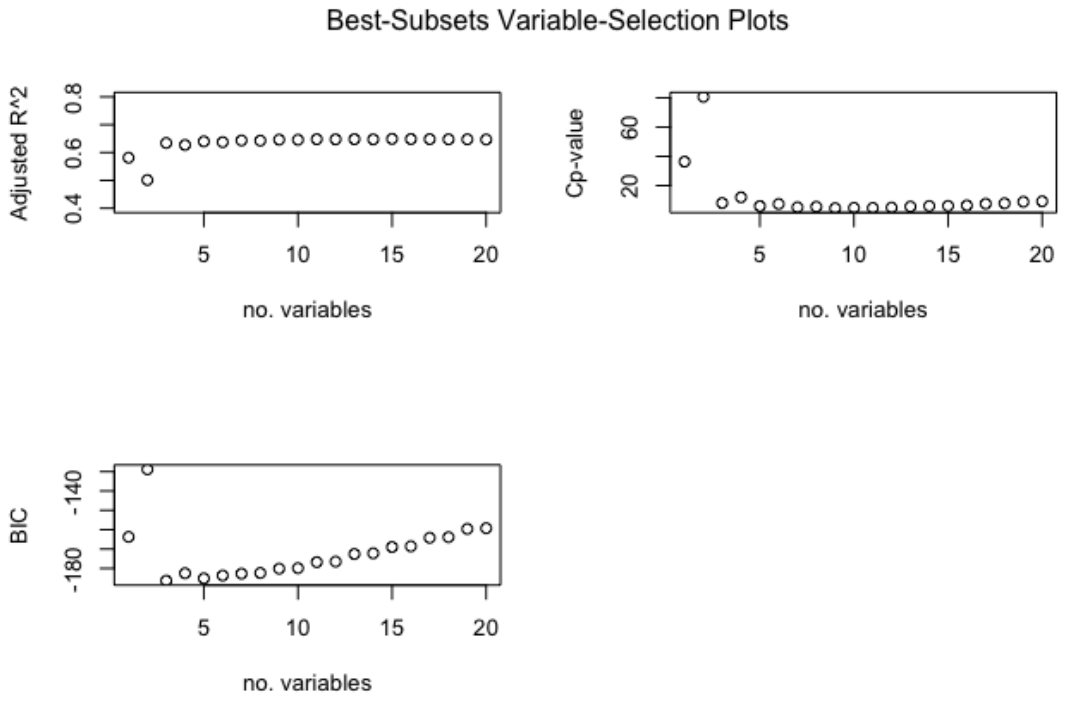
	dbp	scl	followup	age	bmi	month	sex	chdfate	dscl	dfol	dage	dbmi	dmo
1 (1)	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
1 (2)	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "
2 (1)	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "
2 (2)	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "
3 (1)	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	" "	"*"	" "	" "	" "
3 (2)	"*"	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	" "	" "	" "	" "
4 (1)	"*"	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	"*"	" "	" "	" "
4 (2)	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	" "	"*"	" "	" "	"*"
5 (1)	"*"	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	"*"	" "	" "	"*"
5 (2)	"*"	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	"*"	" "	" "	" "
6 (1)	"*"	" "	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"	" "	" "	"*"
6 (2)	"*"	" "	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"	" "	" "	" "
7 (1)	"*"	" "	" "	" "	"*"	"*"	" "	" "	" "	" "	"*"	" "	"*"	"*"
7 (2)	"*"	" "	" "	" "	"*"	"*"	"*"	" "	" "	" "	"*"	" "	"*"	" "
8 (1)	"*"	" "	" "	" "	"*"	"*"	" "	"*"	"*"	" "	"*"	" "	"*"	"*"
8 (2)	"*"	" "	" "	" "	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"	" "
9 (1)	"*"	" "	" "	" "	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"	"*"
9 (2)	"*"	" "	" "	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"	" "	"*"	"*"
10 (1)	"*"	"*"	" "	" "	"*"	"*"	" "	"*"	"*"	"*"	" "	"*"	"*"	"*"
10 (2)	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "

Figure 25 ***Adusjested R-squared Table***

	Nvar	AdjR2	Cp	BIC
1	1	0.5813975	36.372433	-163.7169
2	1	0.5012476	80.673360	-128.8544
3	2	0.6343364	8.085786	-186.3427
4	2	0.6271859	12.017973	-182.4888
5	3	0.6399515	5.987717	-185.1468
6	3	0.6372826	7.447907	-183.6772
7	4	0.6432638	5.174609	-182.7158
8	4	0.6426208	5.524641	-182.3574
9	5	0.6462563	4.553260	-180.1273
10	5	0.6457950	4.803065	-179.8680
11	6	0.6478140	4.721661	-176.7460
12	6	0.6473451	4.974218	-176.4813
13	7	0.6479270	5.672971	-172.5557
14	7	0.6472698	6.025140	-172.1846
15	8	0.6490464	6.088402	-168.9408
16	8	0.6483866	6.440126	-168.5670
17	9	0.6481359	7.586571	-164.1820
18	9	0.6475725	7.885303	-163.8636
19	10	0.6475139	8.927404	-159.5929
20	10	0.6468740	9.264953	-159.2320

Although Figure 24 and Figure 25 are presented above, there is another way to present these values through a graph.

Figure 26 ***Best-Subsets Variable Selection***



Obviously, Adjusted R-Squared becomes stable when the number of variable is 5. For C_p-value, when the number of variables is 5, the size of bias becomes stable. Lastly, for BIC, when the number is between 5 and 8, the values of this interval is smaller than other intervals. Therefore, based on three values, the number of variable that could build a better model is around 7.

Modeling from All-Subsets Selection

After the number of variables are chosen, a regression model can be built. According to Regression Table, the coefficients of each term could build a regression model:

$$sbpinv = \beta_0 + \beta_1 dbp + \beta_2 age + \beta_3 bmi + \beta_4 sex + \beta_5 chdfate + \beta_4 dage + \beta_4 dbmi + \beta_4 dmo$$

From Figure 27, the variables of dbp, age and dage are less than $\alpha = 0.05$ which infers that dbp, age, dage and chdfate are not zero. That is, these four variables possibly have influence(or linearly related) to sbp.

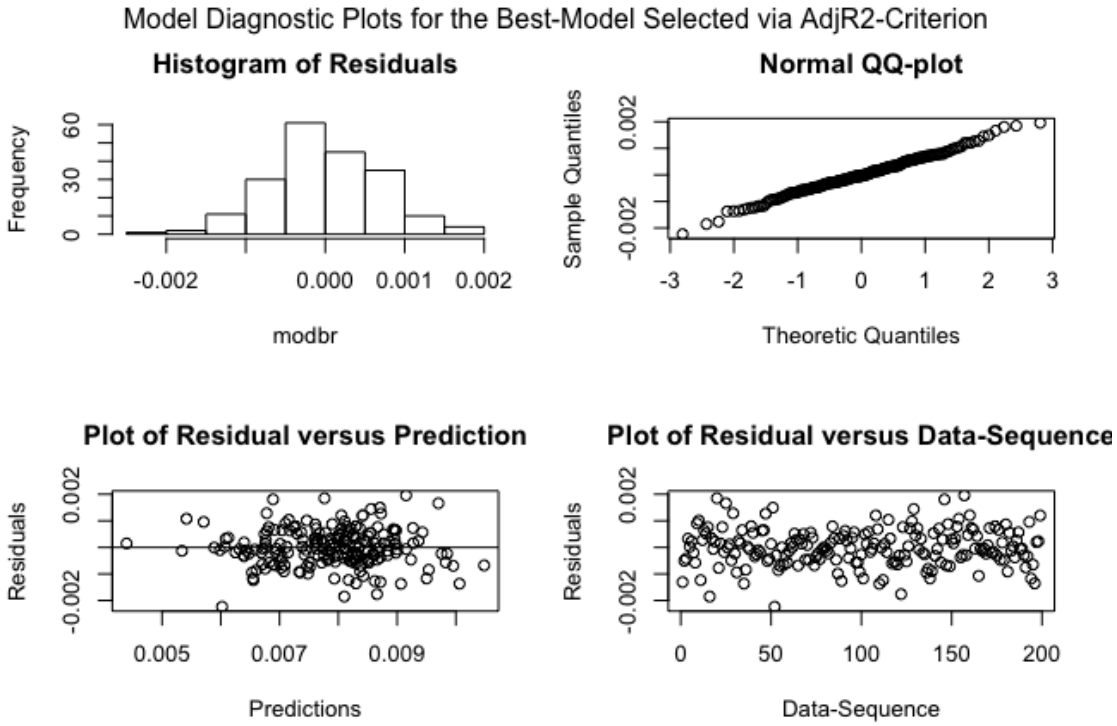
Figure 27 ***Regression Table for All-Subset Selection***

```
> modb = lm(formula = sbpinv ~ dbp + age + bmi + sex + chdfate + dage + dbmi + dmo, p4dataext)
> modb_s = summary(modb)
> modb_s

Call:
lm(formula = sbpinv ~ dbp + age + bmi + sex + chdfate + dage +
    dbmi + dmo, data = p4dataext)

Residuals:
      Min        1Q    Median        3Q       Max
-0.0022334 -0.0004508 -0.0000410  0.0004796  0.0019582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.307e-02  2.930e-03   7.875 2.55e-13 ***
dbp         -1.647e-04  3.512e-05  -4.689 5.24e-06 ***
age         -1.190e-04  4.306e-05  -2.763  0.00628 **
bmi         -1.626e-04  1.004e-04  -1.619  0.10700
sex         -1.336e-04  1.053e-04  -1.269  0.20615
chdfate     -2.315e-04  1.152e-04  -2.009  0.04593 *
dage         1.035e-06  5.188e-07   1.995  0.04752 *
dbmi         1.946e-06  1.190e-06   1.635  0.10367
dmo          2.728e-07  1.733e-07   1.574  0.11717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007169 on 190 degrees of freedom
Multiple R-squared:  0.6632,    Adjusted R-squared:  0.649
F-statistic: 46.77 on 8 and 190 DF,  p-value: < 2.2e-16
```

Model Diagnosis

After the regression model is built, Histogram of residuals and Normal QQ-Plot can be utilized to check normal distribution assumption. From Figure 28, histogram of residuals and Normal QQ-Plot follows normal distribution. As for Plot of Residuals versus Prediction and Data-Sequence, most data points lie between -0.002 and 0.002 and there is only one data point lie below -0.002, which may not greatly affect the quality of analysis.

Figure 28 ***Regression Table for All-Subset Selection***



4.6 Conclusion

In problem 4, the report went through most steps for data analysis. First, data exploration and data background are provided to let readers to know more about the data set. Second, real data points are not usually normal distributed, so data transformation is needed to change data points. But, some data points from some variables are correlated and there is a collinearity problem. Therefore, VIF analysis could address this issue. After this analysis is performed, Variable Selections are needed. In this report, Forward Selection and All-Subsets selection are used. Forward selection is time-saving compared with All-subsets selection, but All-Subsets selection is more comprehensive than the other one. All in all, in this report, we learn how to perform data analysis in a real business field.