REGRESSION ANALYSIS

REAL-LIFE DATA ANALYSIS FOR VARIABLE SELECTIONS

Wenchen Guo

Contents

I REAL-LIFE DATA ANALYSIS FOR VARIABLE SELECTIONS

problem 1. *Analyze the dataset, which is a real-life data set. Apply the box-cox transformation to the output variable, sbp. Model the transformed sbp against significant input variables selected by the two variable selection methods: (i) stepwise selection, and (ii) all-subsets selection. Note that some of the input variables might require transformation and there are possible significant combination effects from input variables (second-order terms such as dbp*dbp is not needed). Provide regression analyses and model checking plots.*

1.1 Data Description

Description of the columns in the data set:

- sbp (Response Variable): Systolic Blood Pressure
- sex: sex ("1" is "men")
- 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 (not relevant to your analysis)

"id" is not relevant to our analysis thus we remove it out of the model. Then we have 8 explanatory variables "sex", "dbp", "scl", "chdfate", "followup", "age", "bmi", "month".
The response variable is "sbp".

1.2 Transform the Response Variable

1.2.1 Analyze the Data

First, let's check whether normality assumption is valid for the y-data = "sbp". Figure 1 is the

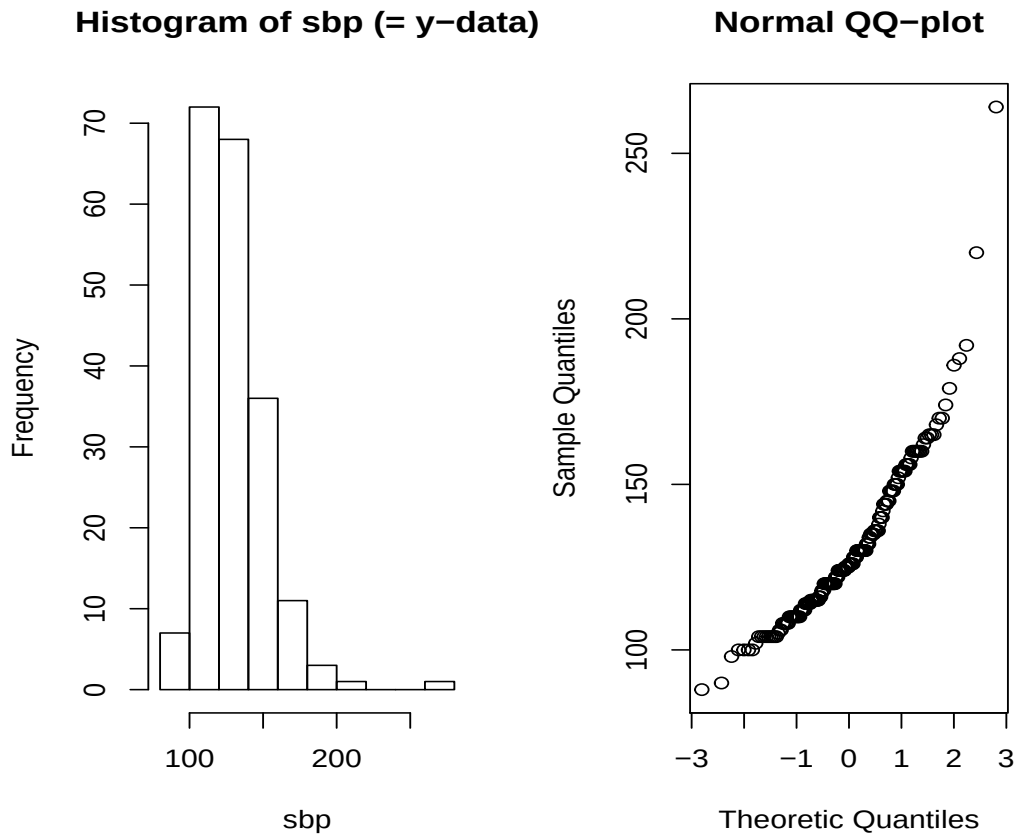

Figure 1: Model-diagnostics Plots for "sbp"

model-diagnostics plots includes the histogram and Normal QQ-plot of sbp. It is clear that the residuals are not normally distributed since the data-distribution is skewed-to-the-right with a thin-tail in the right-side. Therefore, a Box-Cox transformation has to be applied to "sbp".

1.2.2 Box-Cox Transformation for Response Variable

The Box-Cox transformation is a method for finding the best fit of transformation for the response. The formula for Box-Cox transformation is:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & if\ \lambda \neq 0; \\ log(y), & if\ \lambda = 0 \end{cases}$$

Let's use the BoxCox transformation to transform "sbp" to a near-normal distribution.

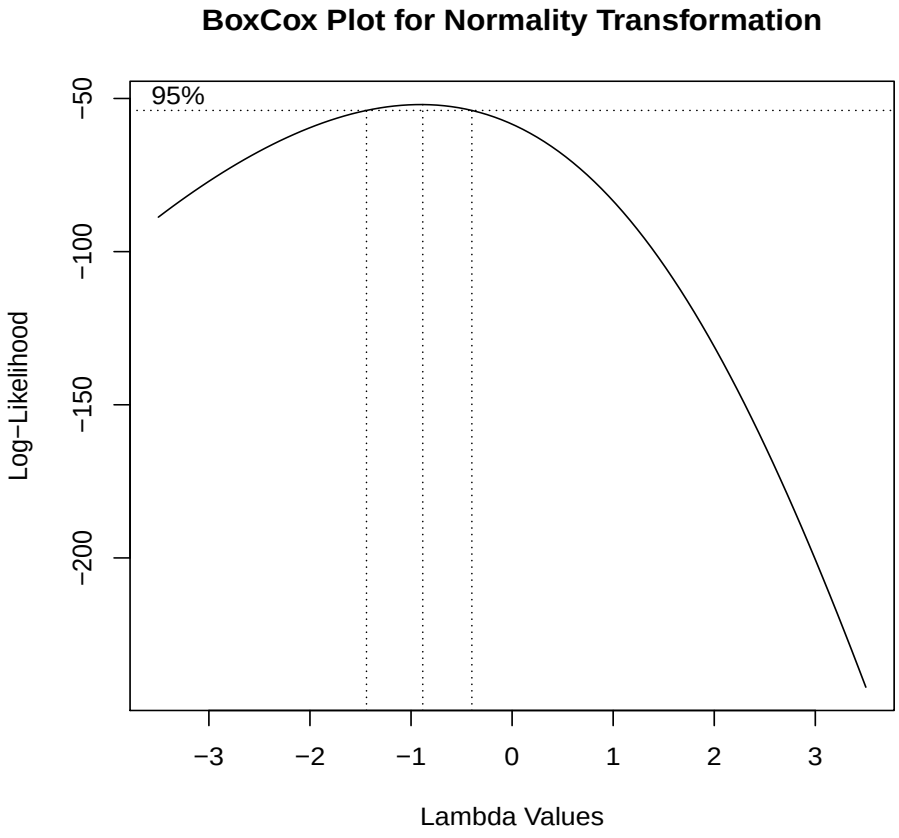

Figure 2: BoxCox Plot

Figure 2 is the BoxCox Plot. We could choose $\lambda = -1$ for transformation. It indicates that "sbp" should be transformed into "1/sbp" for a better normality-y-data modeing.
Let us check its histogram and Normal QQ-plot to see whether the inverse-transformation "1/sbp" works well.

In Figure 3, the normality-validation-plots look well. We will use "1/sbp" as our response variable now.

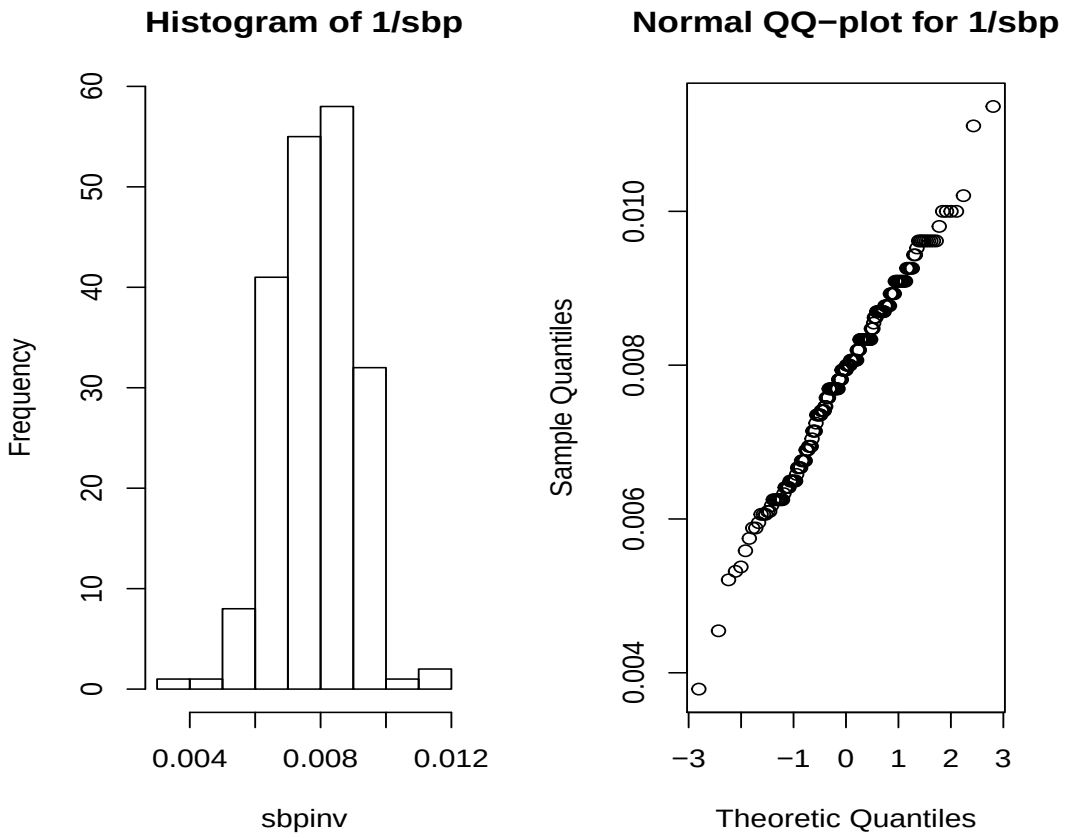

Figure 3: Model-diagnostics Plots for "1/sbp"

1.3 Transform the Input Variable

1.3.1 Analyze the Data

Now, let's plot data to examine the relationship within response and variables. Figure 4 is the combine-plots for variable "dbp", "scl", "followup", "age", "bmi", "month" against "1/sbp". There is no obvious nonlinear patterns between x-variables and inversely-transformed y. Thus, we will stay with the original x-variables for modeling.

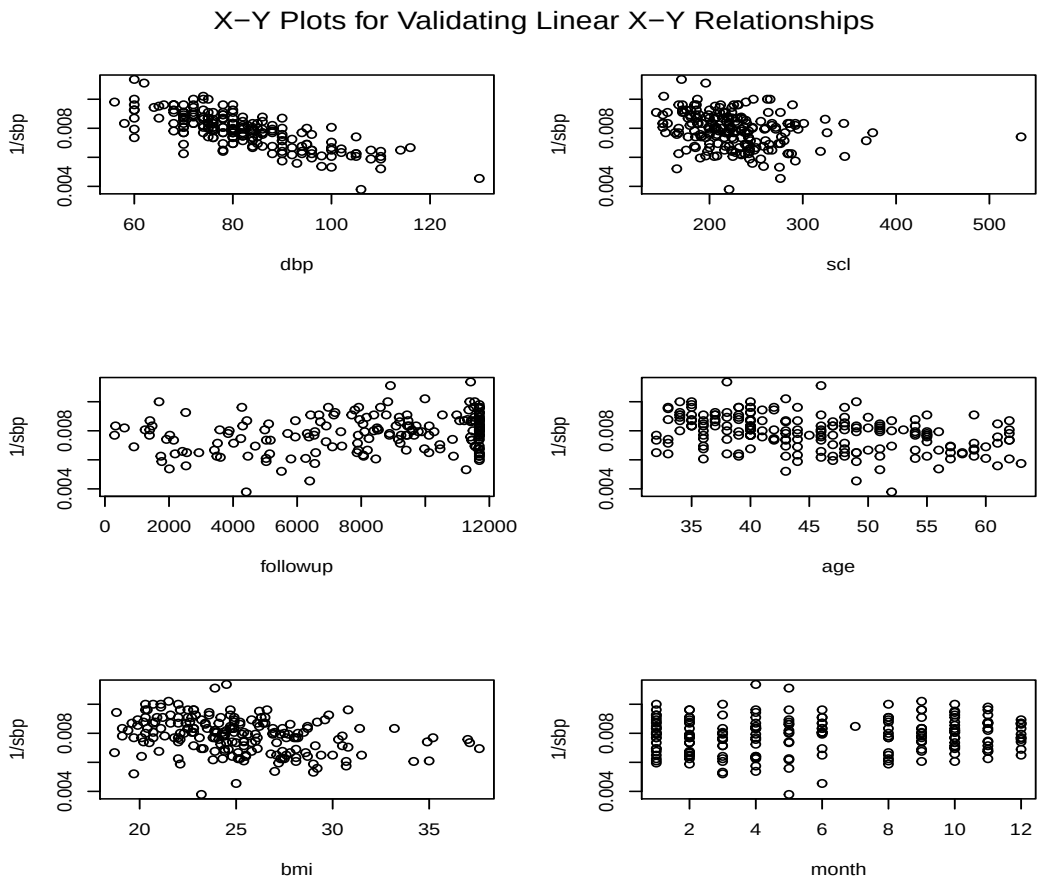

Figure 4: X-Y Plots for Validating Linear X-Y Relationships

1.4 Exploring Combinational-Effects

From Figure 4, since only "dbp" has a visualized clear x-y linear relationship, we will only consider "dbp"'s combinational effects with other continuous variables.
Let us define the following two-variable combinational-effects:

- dscl = dbp*scl
- dfol = dbp*followup
- dage = dbp*age
- dbmi = dbp*bmi
- dmo = dbp*month

Model 1

sbpinv ~ dbp + scl + followup + age + bmi + month + sex + chdfate

Let us run the regression model just for all main-effects without any 2-factor-interactions.
Figure 5 is the regression analysis result of this model.

```
> modn1_summary

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
```

Figure 5: Regression Analysis Result for Model 1

Model 2

1/sbp ~ dbp + scl + followup + age + bmi + month + sex + chdfate + dscl + dfol + dage + dbmi + dmo

Then, we can run a regression model with all main-effects and 2-factor-interactions.

Figure 6 is the regression analysis result of this model.

```
> mod2_summary

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
```

Figure 6: Regression Analysis Result for Model 2

1.5 Multicollinearity in Input Variables

Then we need to check the collinearity among the explanatory variables.
Variance Inflation Factor (VIF) is used to examine the multicollinearity for input variables in regression model. If VIF ≥ 5, we should exclude the variable from model due to high collinearity.
Now, we perform VIF analysis for Model 1 and Model 2 to remove unnecessary variables.

```
> vifmodn1
     dbp      scl followup      age      bmi    month      sex  chdfate
1.238425 1.194747 1.333201 1.227791 1.231565 1.044694 1.024188 1.279322
> vifmod2
       dbp        scl   followup        age        bmi      month        sex
124.590020  49.012490  66.020630  59.488044  62.617097  52.932214   1.078251
   chdfate       dscl       dfol       dage       dbmi        dmo
  1.309445  91.985821  63.219185 122.149919 181.949408  53.776605
```

Figure 7: VIF of Variables in Model 1 and Model 2

Comparing Model 1 Figure 5 and Model 2 Figure 6, the value for R-squared and Adjusted R-squared are increase slightly. Thus, adding the 2-factor-interactions does improve the model-quality slightly.
However, the VIF analysis in Figure 7 indicates that the two-factor-interactions might have strong linear correlation against dbp. Thus, these 2-factor-variables should not be included in the model-ftting exercise.
Next step we can launch variable selection in two ways (stepwise selection and all-subsets selection) to conduct variable selection.
Here, we will still include these 2-factor-variables in our model in order to have more x-variables in the candidate-set for selections and model-comparisons in next step.

1.6 Stepwise selection

Let us work on a partial-F-test based Forward-Selection.
Stepwise selection stops at the point that the p value of added variables is larger than 0.15.

One-variable Model: 1/sbp ~ dbp
Let us start with the x-variable, which is most correlated to y = sbpinv

```
> cor_yax
       sex      sbpinv         dbp         scl     chdfate    followup         age         bmi
0.04391205  1.00000000 -0.76387935 -0.24270558 -0.28175816  0.26916668 -0.39383804 -0.31733896
     month          id       1/sbp         dbp         dbp         dbp         dbp         dbp
0.09215455  0.07548513  1.00000000 -0.56448868 -0.01059398 -0.70976513 -0.65462151 -0.09020080
```

Figure 8: y-x Correlations

Figure 8 is correlation matrix for all x-y. We can see that x-variable "dbp" is most correlated to "1/sbp". Thus, Model sbpinv ~ dbp will be our one-variable model.

Two-variable Model: 1/sbp ~ dbp + scl
let us add next variable "scl" to our model to create a two-variable model and do the regression.

```
Analysis of Variance Table

Model 1: sbpinv ~ dbp
Model 2: sbpinv ~ dbp + scl
  Res.Df         RSS Df Sum of Sq      F Pr(>F)
1    197 0.00012075
2    196 0.00011929  1 1.464e-06 2.4054 0.1225
```

Above is the result of ANOVA comparing one-variable model and two-variable model.
Note that the ANOVA shows that the partial-F-test has a p-value = 0.1225, which is less than the "default" alpha-to-enter = 0.15. Thus, “scl” is added into the model to formulate a Two-Variable Model.

Three-variable Model: 1/sbp ~ dbp + scl + followup
add variable "followup" to our model.

```
Analysis of Variance Table

Model 1: sbpinv ~ dbp + scl
```

```
Model 2: sbpinv ~ dbp + scl + followup
  Res.Df        RSS Df  Sum of Sq     F Pr(>F)
1    196 0.00011929
2    195 0.00011783  1 1.4617e-06 2.419 0.1215
```

Above is the result of ANOVA comparing two-variable model and three-variable model. Note that the p-value = 0.1215, which is less than the "default" alpha-to-enter = 0.15. Thus, “followup” is added into the model to formulate a Two-Variable Model.

Four-variable Model: 1/sbp ~ dbp + scl + followup + age

add variable "age" to our model.

The p-value is still less than 0.15. We add "age" and continue.

```
Analysis of Variance Table

Model 1: sbpinv ~ dbp + scl + followup
Model 2: sbpinv ~ dbp + scl + followup + age
  Res.Df        RSS Df  Sum of Sq      F
1    195 0.00011783
2    194 0.00010469  1 1.3138e-05 24.346
     Pr(>F)
1
2 1.727e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Five-variable Model: 1/sbp ~ dbp + scl + followup + age + bmi

Add variable "bmi" to our model.

```
Analysis of Variance Table

Model 1: sbpinv ~ dbp + scl + followup + age
Model 2: sbpinv ~ dbp + scl + followup + age + bmi
  Res.Df        RSS Df  Sum of Sq     F Pr(>F)
1    194 0.00010469
2    193 0.00010469  1 1.6186e-10 3e-04 0.9862
```

Now, the partial-F-test has a p-value = 0.9862, which is large than alpha-to-enter = 0.15. Thus,

the forward selection is stop at the Four-Variable-Model. We will not include the variable "bmi" in our model.

Summary for the Final Model

Figure 9 is the regression analysis result for the final model from the forward selections.

```
> mod_forward_s

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
```

Figure 9: Regression Analysis Result

The regression model is

$$1/sbp = 0.015 - 0.000067dbp - 0.0000007scl + 0.000000003followup - 0.000034age$$

Then, let's examine the model-diagnostic plots to make sure that all model assumptions are valid.

Figure 10 are the results of the model-diagnostic plots. The residuals are random and

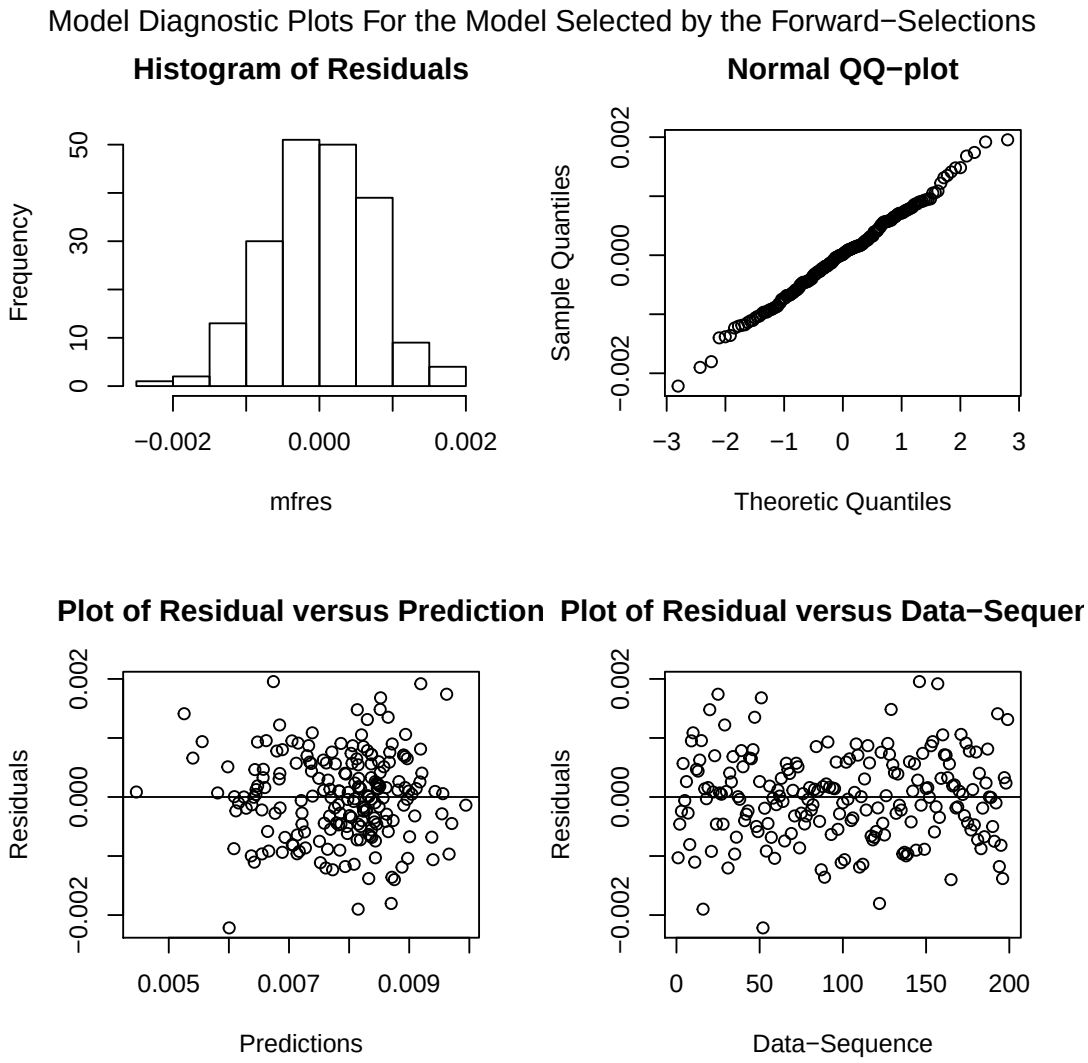

Figure 10: Model-diagnostic Plots

normally distributed.

1.7 All-subsets Selection

All-subset regression searches through all subsets and finds the best subset considering a variety of criterion.

```
> modas
Subset selection object
Call: regsubsets.formula(sbpinv ~ dbp + scl + followup + age + bmi +
    month + sex + chdfate + dscl + dfol + dage + dbmi + dmo,
    data = p4dataext, nbest = 2, nvmax = 10)
13 Variables  (and intercept)
         Forced in Forced out
dbp          FALSE      FALSE
scl          FALSE      FALSE
followup     FALSE      FALSE
age          FALSE      FALSE
bmi          FALSE      FALSE
month        FALSE      FALSE
sex          FALSE      FALSE
chdfate      FALSE      FALSE
dscl         FALSE      FALSE
dfol         FALSE      FALSE
dage         FALSE      FALSE
dbmi         FALSE      FALSE
dmo          FALSE      FALSE
2 subsets of each size up to 10
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 11: All-subsets Selection

The result is in Figure 11.

Let us get the Adjusted-R^2, Cp and BIC for each model calculated:

	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

Figure 12 is the best-subsets variable-selection plots.
In best subset regression, the chosen criterions are:

- The adjusted R-square should be as higher as better.
- For Mallow's Cp criterion, the minimum Cp should ≤ p. The closer it is to p, the better.
- BIC is Bayesian Information Criterion . The model with the lowest BIC is preferred.

In consideration of all the informations above, the 15th model with variables (dbp + age + bmi + sex + chdfate + dage + dbmi + dmo) is preferred

For this model, the adjusted R-square is 0.6490464 (best), Cp criterion is 6.088402 (satisfied), and BIC is -168.9408 (satisfied).

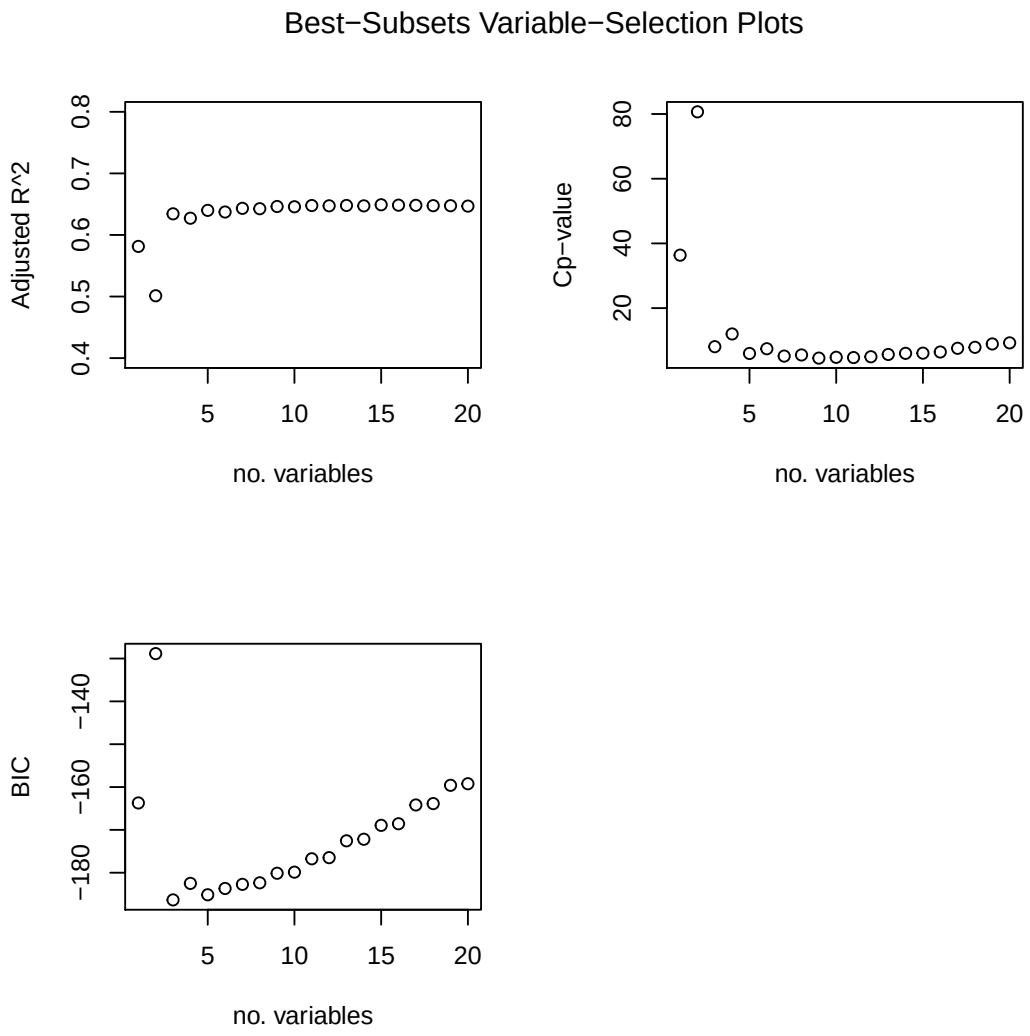

Figure 12: Best-Subsets Variable-Selection Plots

Then, run a regular regression for the selected model

$$1/sbp \sim dbp + age + bmi + sex + chdfate + dage + dbmi + dmo$$

```
> 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
```

Figure 13: Regression Analysis Result

Figure 13 is the regression analysis result for the final model from the all-subsets selection. The regression model is

$$1/sbp = 0.023 - 0.0002dbp - 0.0001age - 0.0002bmi - 0.0001sex$$

$$-0.0002chdfate + 0.000001dage + 0.000002dbmi + 0.000003dmo$$

Then, let's examine the model-diagnostic plots to make sure that all model assumptions are valid.

Figure 14 are the results of the model-diagnostic plots. The residuals are random and normally distributed.

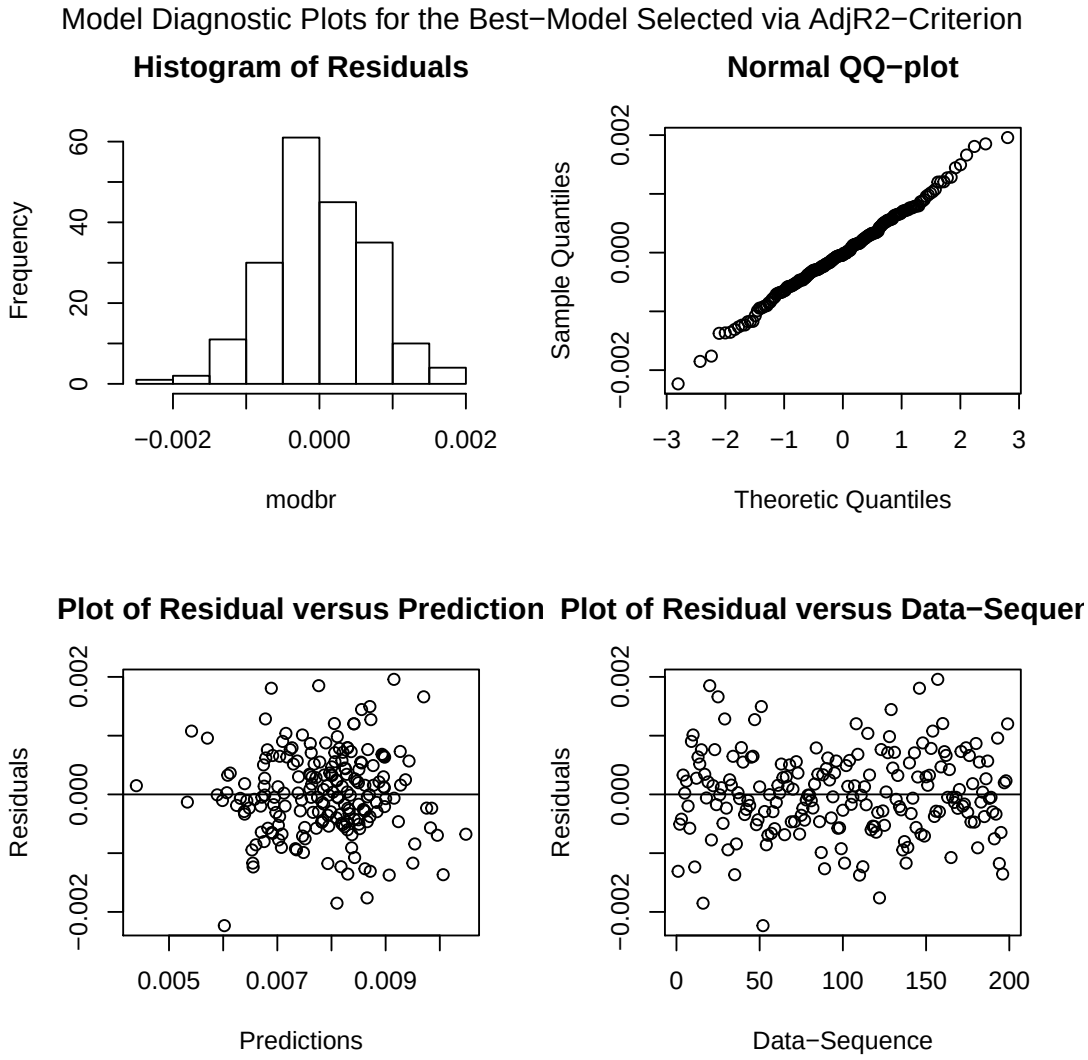

Figure 14: Model-diagnostic Plots

1.8 Final Model and Regression Analysis

The final model from the forward selection:

$$1/sbp \sim dbp + scl + followup + age$$

The final model from the all-subsets selection:

$$1/sbp \sim dbp + age + bmi + sex + chdfate + dage + dbmi + dmo$$

After comparison the two models above, we can get the final model and check the quality of the regression model.
Figure 9 is the regression analysis result for the final model from the forward selections.
We can see, the p-value for variable "scl" is 0.528 and the p-value for variable "followup" is 0.858. Thus, both "scl" and "followup" is not significant. We can consider to remove this two variables out of our model. Figure 13 is the regression analysis result for the final model from the all-subsets selection.
Similarly, we can consider to remove variables "bmi","sex", "dbmi", "dmo".
The remaining variables are "dbp", "age", "chdfate", "dage".

The regression model is

$$1/sbp \sim dbp + age + chdfate + dage$$

and VIF values are:

```
      dbp        age    chdfate       dage
33.948514 45.841293  1.084548 95.425143
```

```
> modb_r

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

Residuals:
       Min         1Q     Median         3Q        Max
-0.0024349 -0.0004602 -0.0000103  0.0004918  0.0019189

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.876e-02  1.932e-03   9.711  < 2e-16 ***
dbp         -1.142e-04  2.355e-05  -4.851 2.51e-06 ***
age         -1.187e-04  4.182e-05  -2.838  0.00503 **
chdfate     -1.907e-04  1.138e-04  -1.676  0.09526 .
dage         1.041e-06  5.038e-07   2.066  0.04013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007228 on 194 degrees of freedom
Multiple R-squared:  0.6505,    Adjusted R-squared:  0.6433
F-statistic: 90.26 on 4 and 194 DF,  p-value: < 2.2e-16
```

Figure 15: Regression Analysis Result

Figure 15 is the regression analysis result for this model. All variables are significant.
But, the VIF value for varaible "dage" is 95.425143, the VIF value for varaible "dbp" is 33.948514, and the VIF value for varaible "age" is 45.841293. They are all greater than 5.
Since dage = dbp*age, this two-factor-interaction "dage" might have strong linear correlation against "dbp" and "age". Thus, we romove variable "dage" from our model and do the regression agian to check.

The new regression model is

$$1/sbp \sim dbp + age + chdfate$$

Figure 16 is the regression analysis result for the new model.

```
> modb_f

Call:
lm(formula = sbpinv ~ dbp + age + chdfate, data = p4dataext)

Residuals:
       Min         1Q     Median         3Q        Max
-2.307e-03 -4.794e-04  3.369e-05  5.120e-04  1.897e-03

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.486e-02  4.070e-04  36.512  < 2e-16 ***
dbp         -6.637e-05  4.266e-06 -15.559  < 2e-16 ***
age         -3.328e-05  6.451e-06  -5.159 6.09e-07 ***
chdfate     -1.846e-04  1.147e-04  -1.610    0.109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007288 on 195 degrees of freedom
Multiple R-squared:  0.6428,    Adjusted R-squared:  0.6373
F-statistic:   117 on 3 and 195 DF,  p-value: < 2.2e-16

> vif(model_f)
     dbp      age  chdfate
1.095794 1.072796 1.083818
```

Figure 16: Regression Analysis Result

The VIF corresponding to each variable is not larger than 5. Thus we can conclude the multicollinearity of the model does not exist. The R-squared is 0.6428 and the adjusted R-squared is 0.6373, which implies that it is a good fitting of the model to the data.
Thus, our final model is:

$$1/sbp \sim 0.01486 - 0.000066dbp - 0.000033age - 0.000185chdfate$$

Then, let's examine the model-diagnostic plots to make sure that all model assumptions are valid.

Figure 17 are the results of the model-diagnostic plots. The residuals are random and

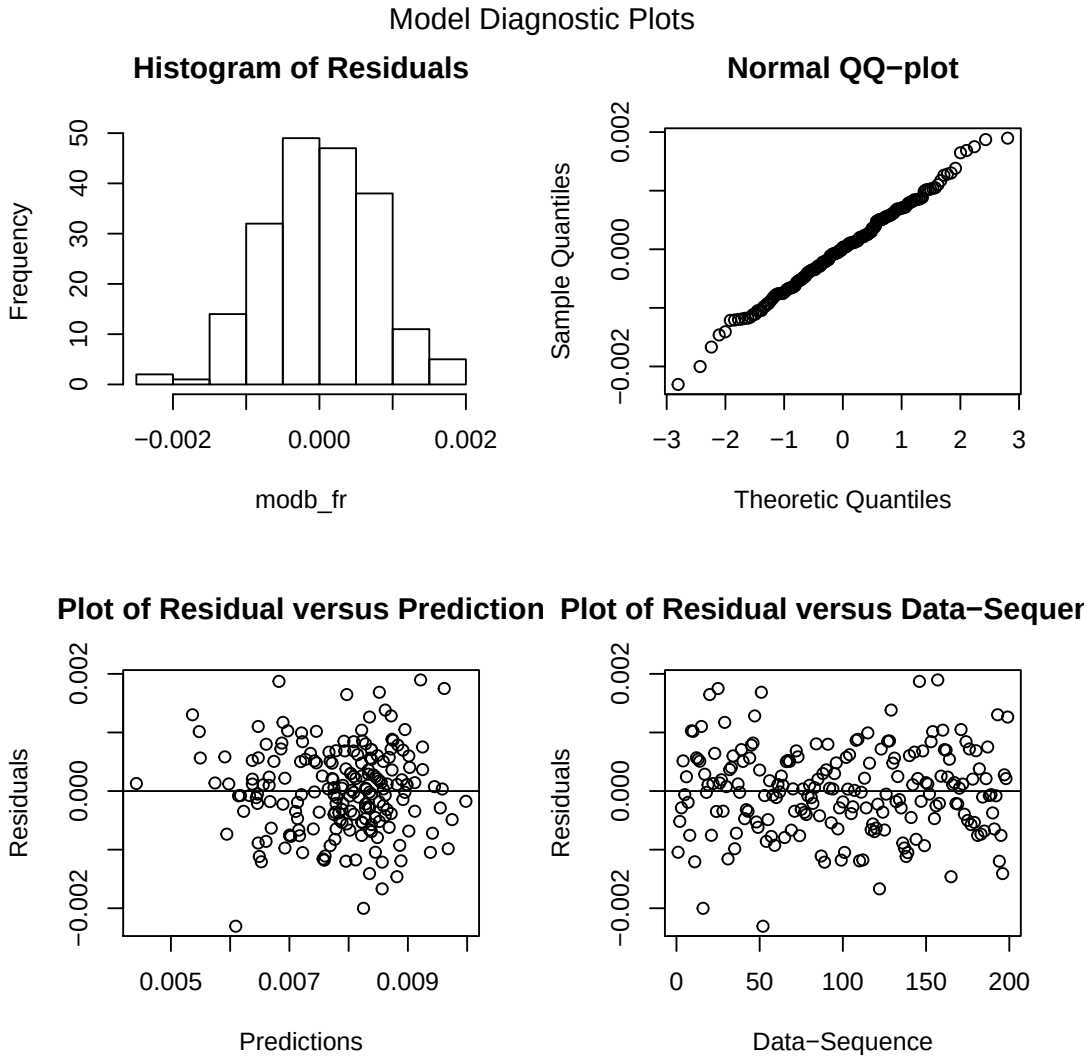

Figure 17: Model-diagnostic Plots

normally distributed. It satisfies the linear assumption and the homoscedastic assumption.