Final Project

Katrina Liu

2022-12-10

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

BMI 715 Final Project

Motivation

Hypertension has been known to increase the risk for conditions including heart disease, heart attack, stroke, and etc ("High blood pressure (hypertension)," 2022). Diagnostic of hypertension is focused on systolic blood pressure level and the diastolic blood pressure level. The treatment of hypertension greatly depends on how the blood pressure level change. Motivated by this, in this project, we want to study which factors are associated with the systolic blood pressure level and build a predictive model for the systolic blood pressure level using the variables for the cohort that is already diagnosed with hypertension using the NHANES data (Disease Control & National Center for Health Statistics (NCHS)., 2013-2014). The result of the study would be useful in better adjusting treatment plan.

```
# Loading the NHANES Dataset
nhanes = read.csv("nhanes_13_14_subset_updated.csv", row.names = "X")
```

Systolic Blood Pressure Related Variables

There can be several different criteria of determine the hypertension cohort. However, given that we are taking into account the medication intake, the individual should be aware that they are hypertension. Therefore, we choose the variable BPQ020, a questionnaire variable indicating whether the individual has been told they have hypertension to filter out the hypertension cohort we want to investigate.

Variable Name	Variable Description
BPQ020	{Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had hypertension, also called high blood pressure?

The dependent variable would be the systolic blood pressure level and is well characterized into the variable BPXSY1 in the NHANES data set.

Dependent Variable Name	Variable Description
BPXSY1	Systolic: Blood pressure (first reading) mm Hg

To select the predictors, we first included common demographic variables including age, sex, BMI. Commonly seen lab test variables such as total protein level, serum glucose level, and total cholesterol level are included as well. Blood pressure related examination variables such as 60s pulse, diastolic blood pressure, and maximum inflation levels.

One study, in particular, investigated whether medication would affect the systolic blood pressure level (Naci *et al.*, 2019). Indeed, taking hypertension treatment medications could potentially lead to changes in the systolic blood pressure. Therefore, it is being included as one indicator predictor variable.

Predictor Variable Name	Variable Description
RIAGENDR	Gender of the participant
RIDAGEYR	Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
BMXBMI	Body Mass Index
LBXSTP	Total protein (g/dL)
LBXSGL	Glucose, refrigerated serum (mg/dL)
LBXTC	Total Cholesterol(mg/dL)
BPXPLS	60 sec. pulse (30 sec. pulse * 2)
BPXDI1	Diastolic: Blood pressure (first reading) mm Hg
BPXML1	MIL: maximum inflation levels (mm Hg)
BPQ100D	(Are you/Is SP) now following this advice to take prescribed medicine?

Data Processing and Exploratory Analysis

Remove Invalid Data Point

We first filter out all of the datapoint that include at least one NA value for the variables we selected. While pre-processing the data, we realized that there are noticeable invalid data that should be left out to avoid their misleading effect on the result. The data being left out of the study includes: age = 80 (all age > 80 are being recorded as 80), systolic blood pressure = 0 (should not be true), diastolic blood pressure = 0 (should not be true).

```
# Filter the cohort with hypertension
nhanes_ht = nhanes %>% filter(BPQ020 == 2, BPXDI1>0, BPXSY1>0, RIDAGEYR<80)
nhanes_ht_related = nhanes_ht %>%
  dplyr::select(c("BMXBMI", "RIDAGEYR", "LBXTC","LBXSGL","LBXSTP",
                  "BPXSY1","BPXML1","BPXPLS", "BPXDI1", "BPQ100D",
                                                  "RIAGENDR")) %>%
  drop_na %>% mutate(BPQ100D = BPQ100D-1, RIAGENDR = RIAGENDR-1)
head(nhanes_ht_related)
```

```
##   BMXBMI RIDAGEYR LBXTC LBXSGL LBXSTP BPXSY1 BPXML1 BPXPLS BPXDI1 BPQ100D
## 1   24.2       69   203     67    7.0    116    140     60     68       0
## 2   24.0       62   171    143    6.7    126    160     64     50       0
## 3   26.3       32   172     62    7.2    120    140     88     74       1
## 4   37.8       57   271     92    7.2    132    160     72     74       1
## 5   27.7       61   127    106    7.2    112    130     74     46       0
## 6   28.0       54   146     87    7.1    118    140     62     70       0
##   RIAGENDR
## 1        1
## 2        1
## 3        0
## 4        1
## 5        1
## 6        0
```

Scaling the Predictors

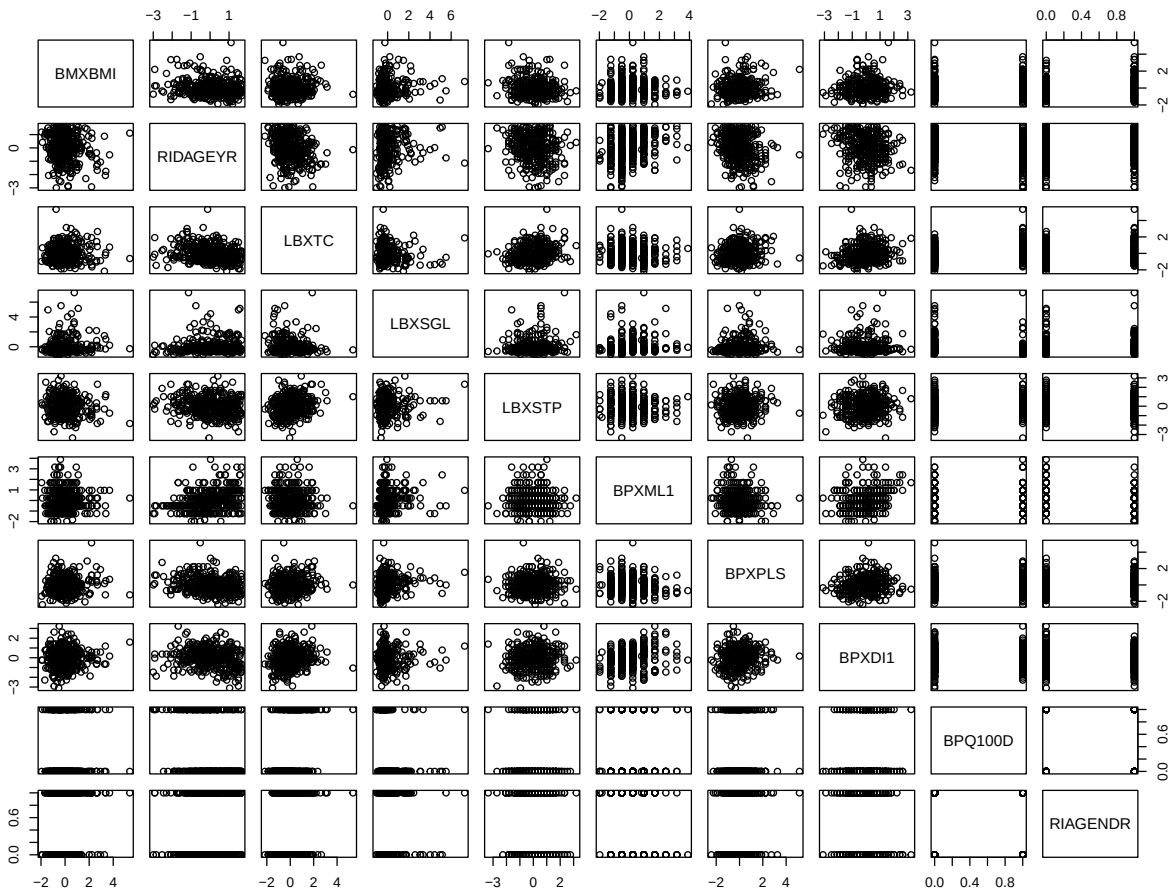
We scaled the continuous variables so that they are on the same scale of mean 0 and variance 1 and leave the binary indicator variable untouched.

```
# Scale continuous variables
nhanes_ht_ready = nhanes_ht_related[1:9] %>% scale %>% as.data.frame
nhanes_ht_ready = cbind(nhanes_ht_ready, nhanes_ht_related[10],
                        nhanes_ht_related[11])
```

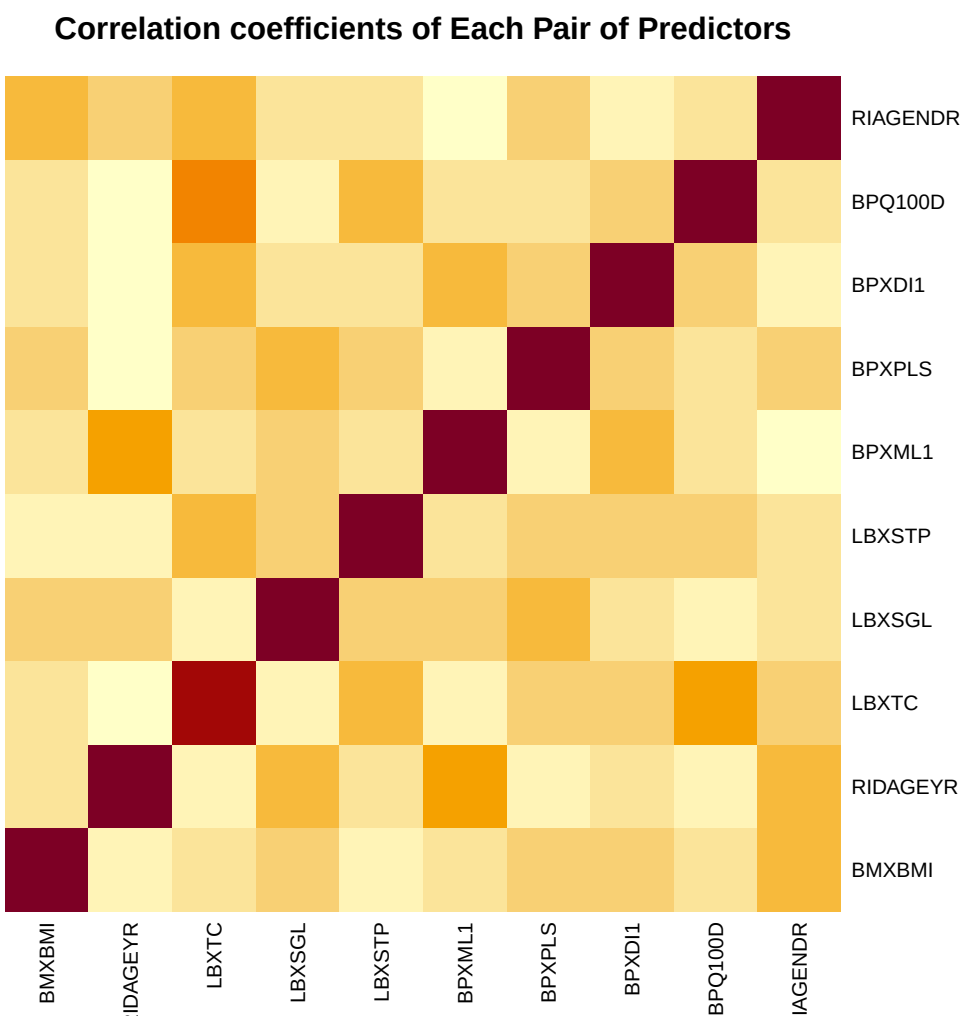
Relationships Between Variables

Lastly, we want to verify that there is no correlation or dependence between the predictor variables. Here, we plot out the relationships and correlation coefficients between each pair of predictor variables. Based on the figures, it seems that there is no relationship between the predictor variables and we can proceed to building the predicting model.

```
# Plot collinearity and correlation between predictors
plot(nhanes_ht_ready[,-6],
     main="Relationships between Each Pair of Predictors")
```




```
heatmap(cor(nhanes_ht_ready[,-6]),
        main = "Correlation coefficients of Each Pair of Predictors",Colv=NA,
        Rowv=NA)
```

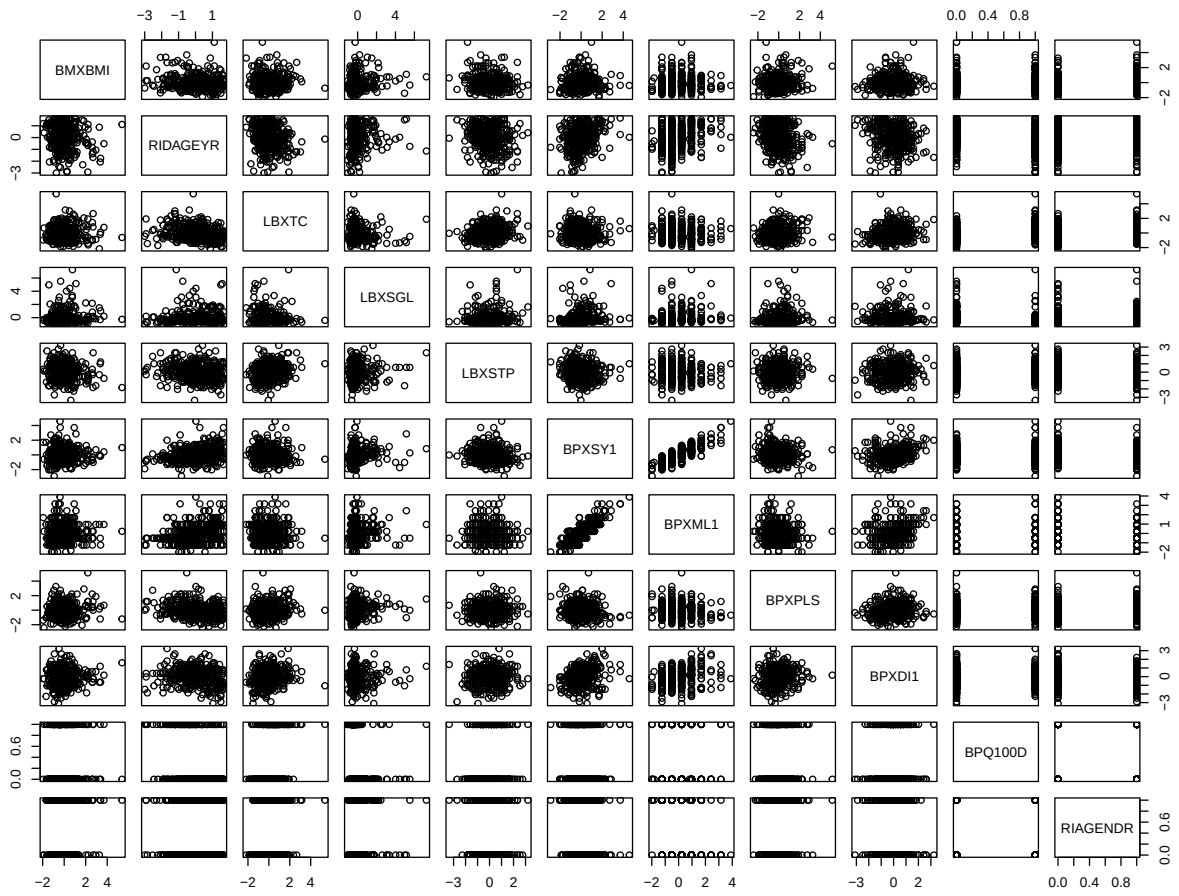


Linear Regression Model

Assessing the Relationship between the Predictors and the Outcome

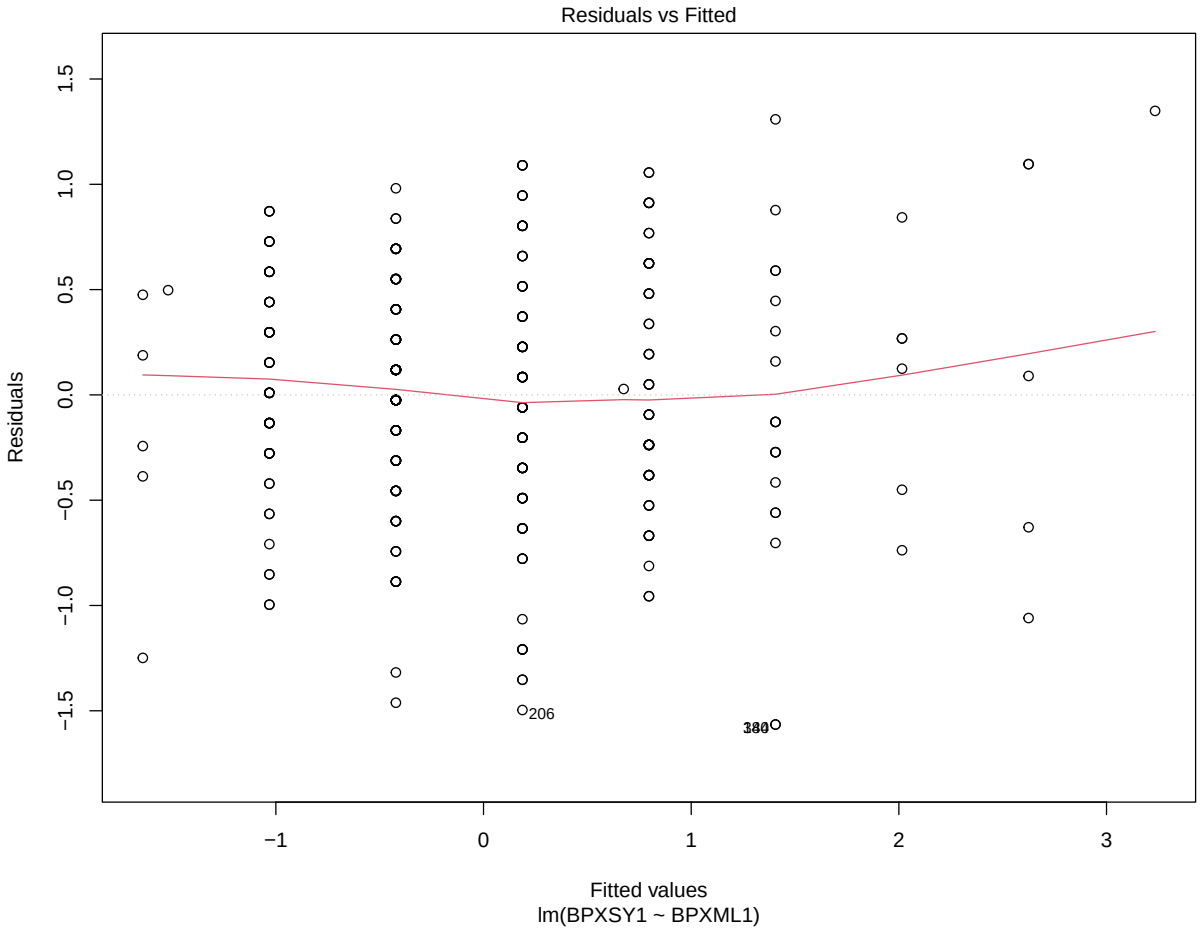
If we look specifically for the row and column with BPXSY1, we see that seemlingly the relationship of it and BPXML1 looks linear. Other variables seem to not have a linear relationship with the systolic blood pressure and therefore violates the assumption of linear regression. However, we decide to still include them in the model so that their impact could be assessed.

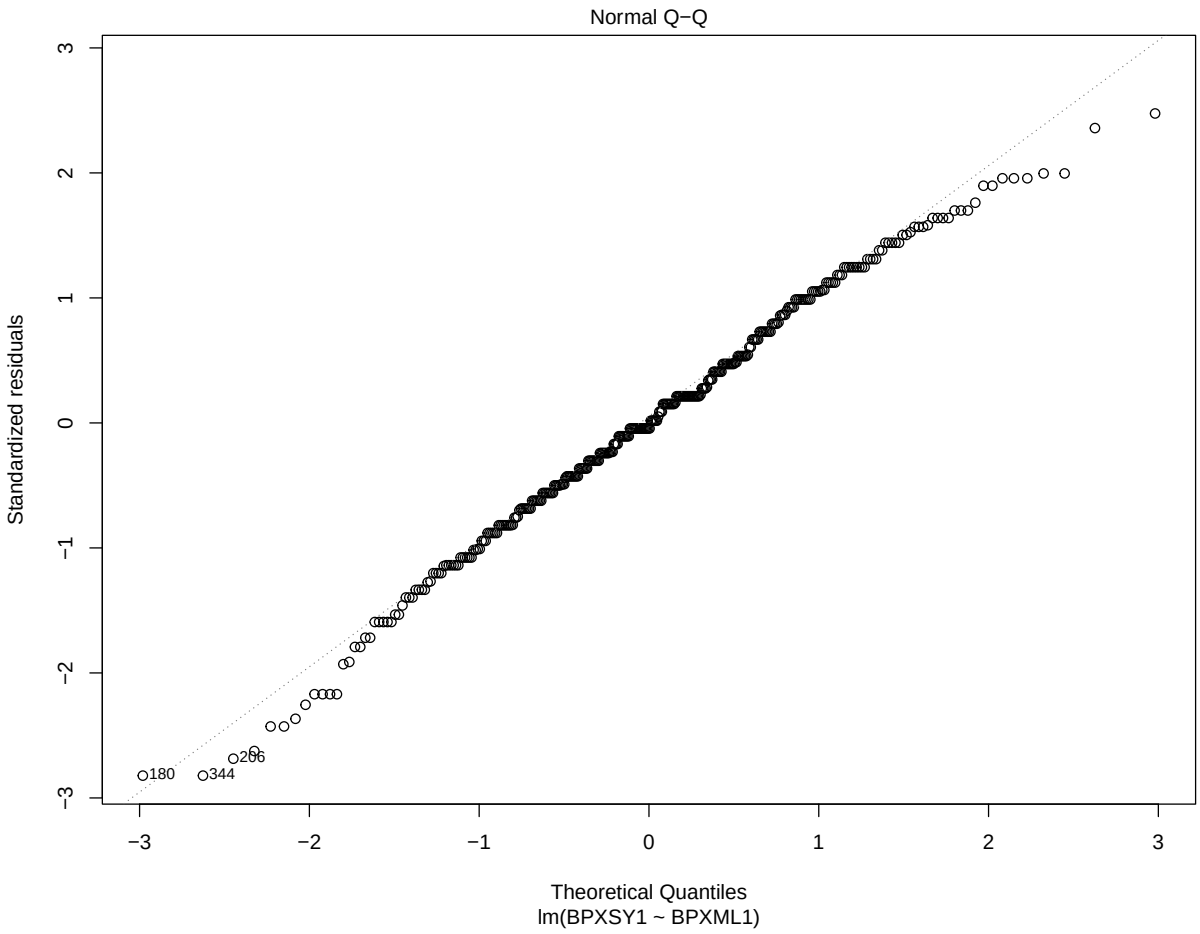
```
# Plot collinearity and coefficient of all variables
plot(nhanes_ht_ready, main="Relationships between Each Pair of Variables")
```

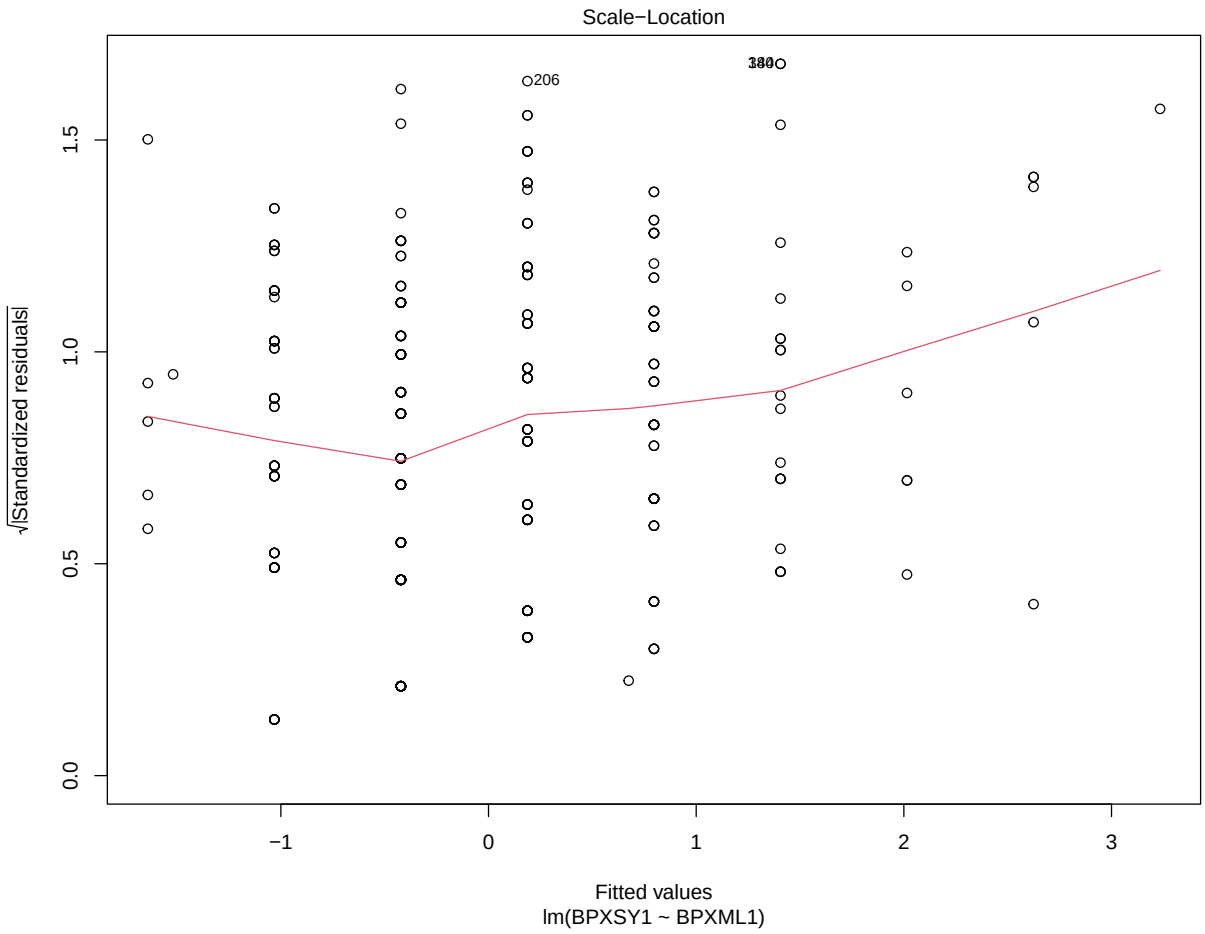


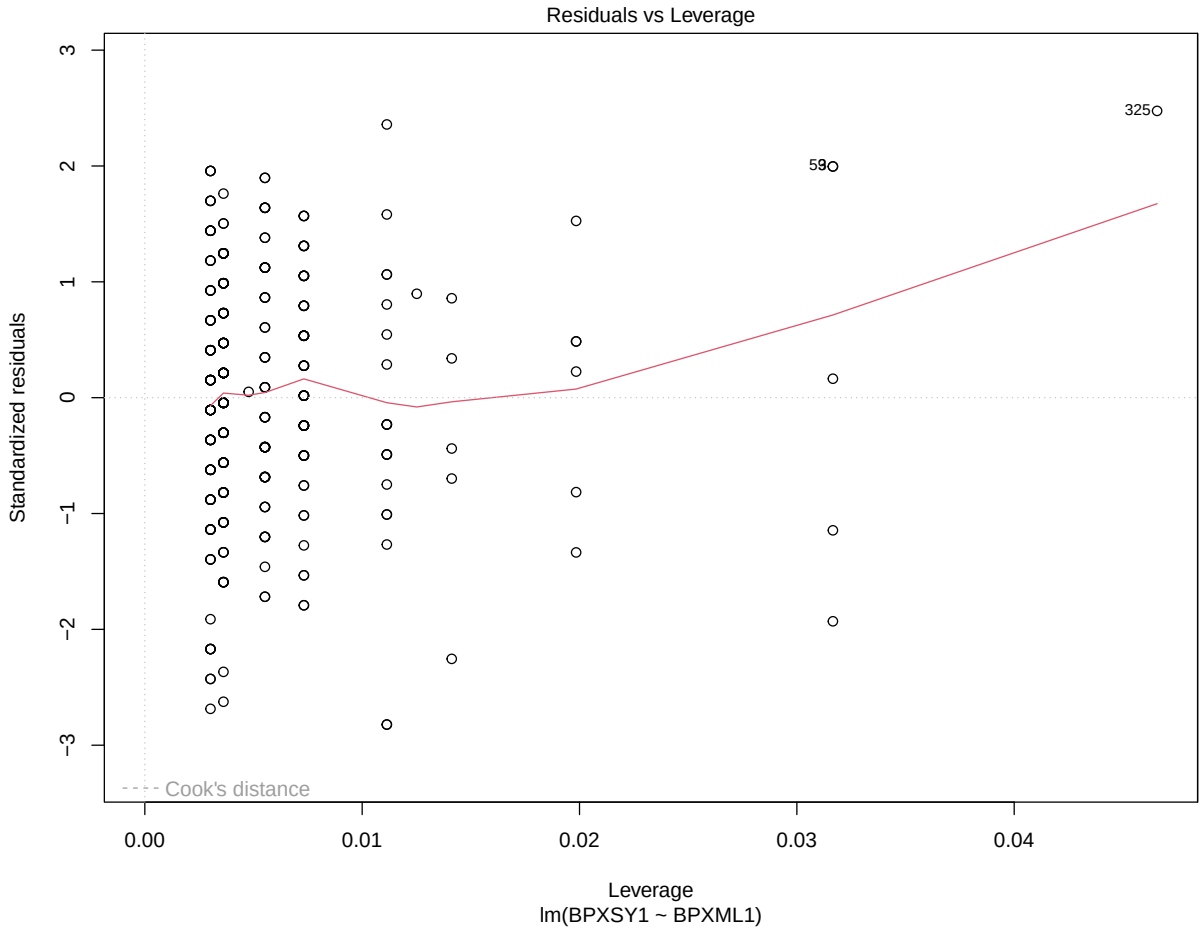
We would further assess if BPXML1 and BPXSY1 follows all of the constraints of a linear relationship. We see that the residuals have a normal distribution centered at 0 and have similar variances of 1 for each fitted values. Each data point is independent. Therefore, BPXML1 and BPXSY1 have a linear relationship.

```
# Assees the linear relationship between BPXSY1 and BPXML1
plot(lm(BPXSY1~BPXML1, data=nhanes_ht_ready))
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
206
−1
0
1
2
3
Fitted values
lm(BPXSY1 ~ BPXML1)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(BPXSY1 ~ BPXML1)
180
344
266
−3
−2
−1
0
1
2
3


Scale−Location
206
√|Standardized residuals|
0.0
0.5
1.0
1.5
−1
0
1
2
3
Fitted values
lm(BPXSY1 ~ BPXML1)



Since the dependent variable, systolic blood pressure, is continuous, I use the linear regression model to characterize its relationship with the predictors. I start by including all of the variables in a linear regression model to see how the model performs.

```
# Full model
lm.full = lm(BPXSY1~.,data=nhanes_ht_ready)
summary(lm.full)
```

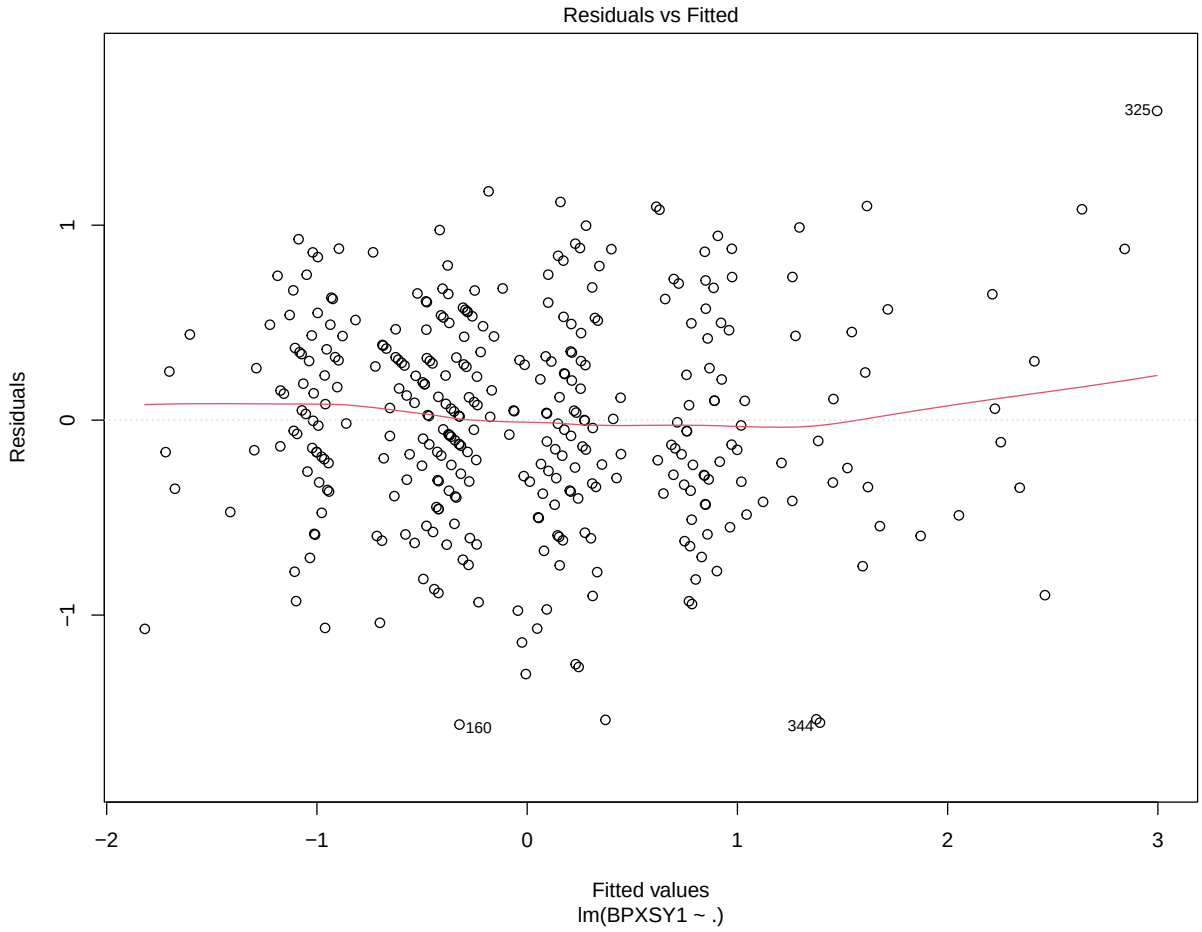
```
##
## Call:
## lm(formula = BPXSY1 ~ ., data = nhanes_ht_ready)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56158 -0.34290 -0.00111  0.36411  1.58622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.075048   0.048798  -1.538 0.125001
## BMXBMI       0.031927   0.030783   1.037 0.300397
## RIDAGEYR     0.053422   0.035248   1.516 0.130553
## LBXTC        0.002592   0.034422   0.075 0.940031
## LBXSGL       0.026578   0.031007   0.857 0.391966
```

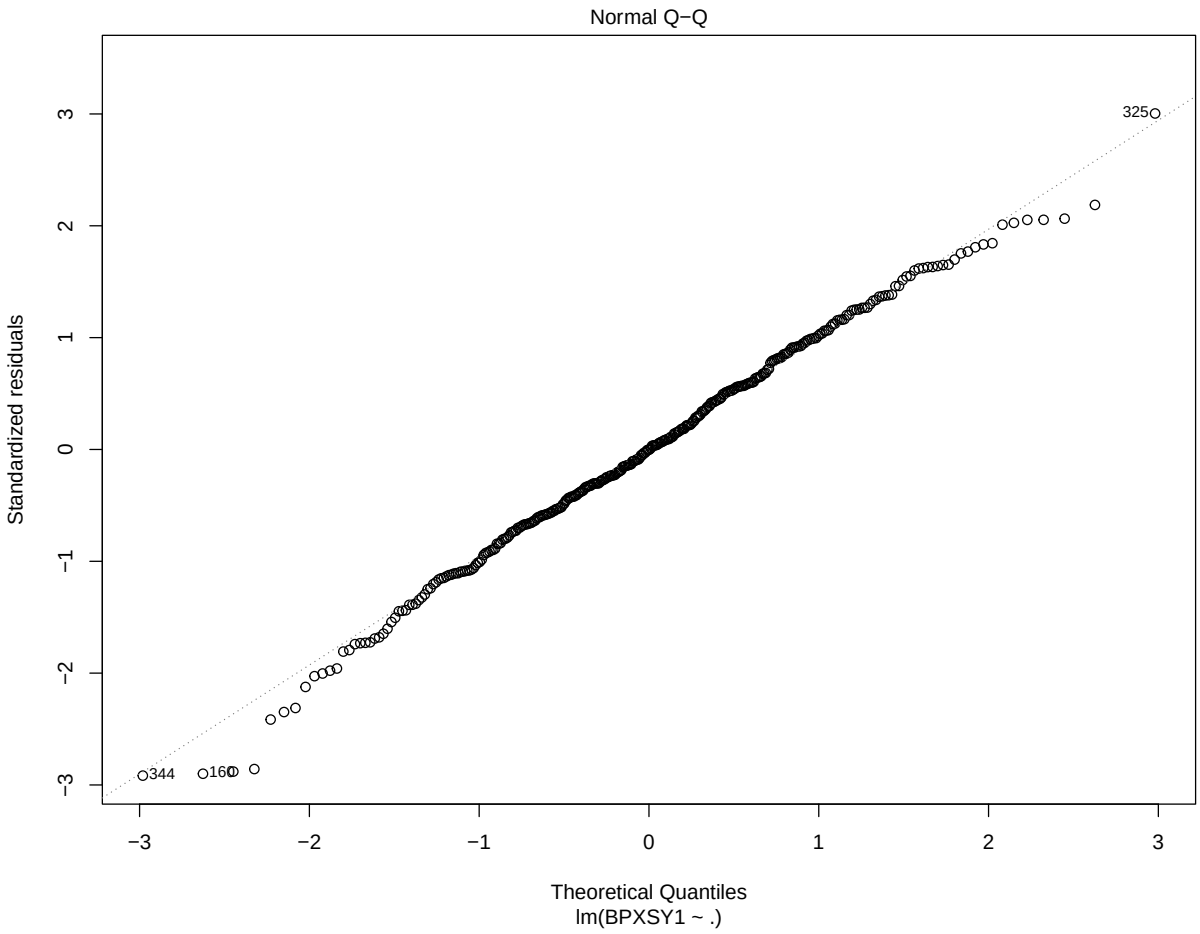
```
## LBXSTP      -0.035528   0.030858  -1.151 0.250412
## BPXML1       0.792749   0.033526  23.646  < 2e-16 ***
## BPXPLS      -0.042004   0.031028  -1.354 0.176716
## BPXDI1       0.112297   0.031645   3.549 0.000442 ***
## BPQ100D      0.029008   0.070281   0.413 0.680057
## RIAGENDR     0.136010   0.062772   2.167 0.030957 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5462 on 337 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.7016
## F-statistic: 82.59 on 10 and 337 DF,  p-value: < 2.2e-16
```

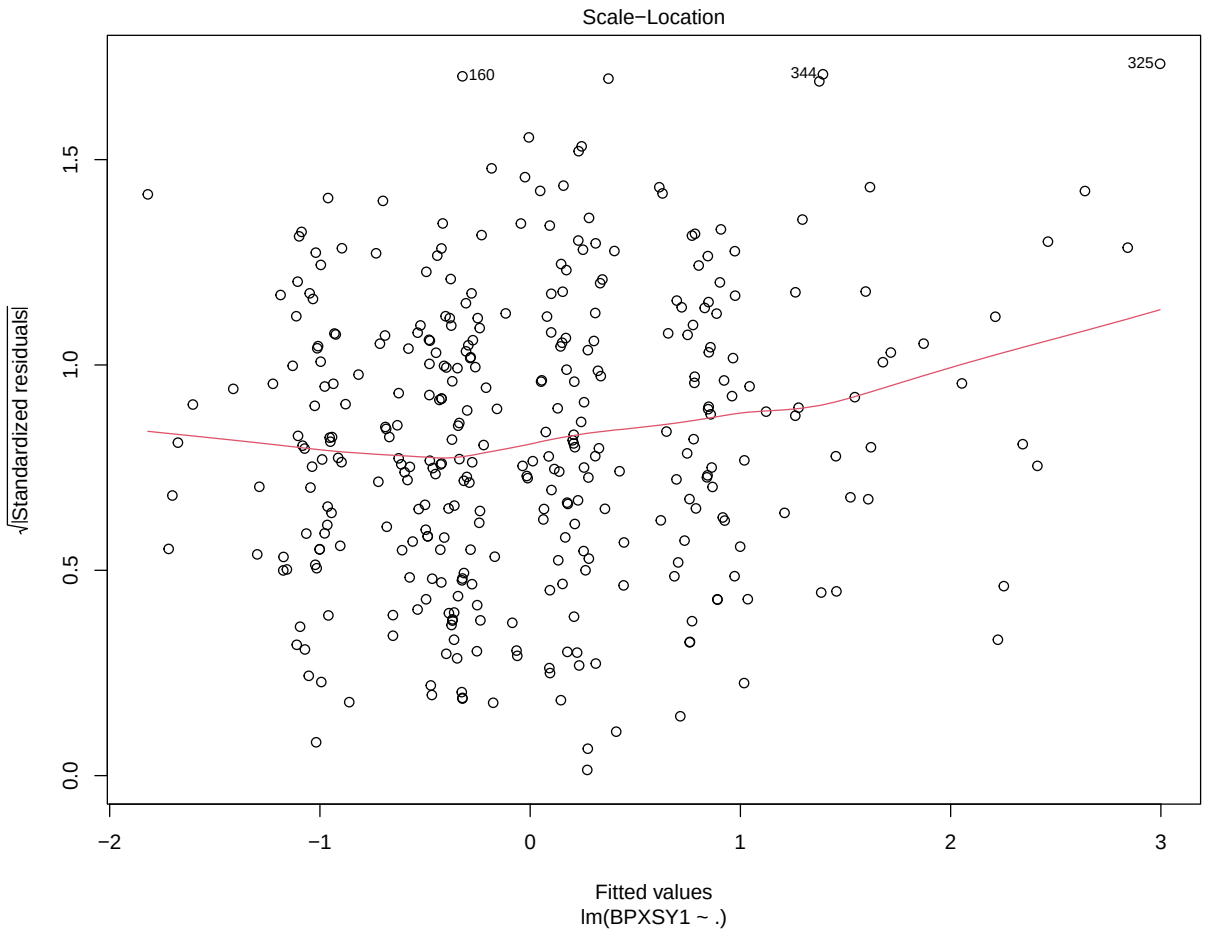
From the summary of the full model, we see that there are three variables with coefficients of significant P-values: the estimated coefficient of variable BPXML1 (maximum inflation level) is 0.793 with a very significant P-value less than 2e-16 indicating that the systolic blood pressure level increases with the maximum inflation level (0.793 unit of systolic blood pressure level per 1 unit of maximum inflation level); the estimated coefficient of variable BPXDI1 (diastolic blood pressure) is 0.112 with a significant P-value 0.000442 indicating that the systolic blood pressure level increases with the diastolic blood pressure level (0.112 unit of systolic blood pressure level per 1 unit of diastolic blood pressure level); and the estimated coefficient of variable RIAGENDR (gender) is 0.136 with a significant P-value less than 0.031 indicating that the systolic blood pressure level of the samples of gender group 0 is less than that of gender 1 (after processing, original gender group 1 = gender group 0, original gender group 2 = gender group 1).

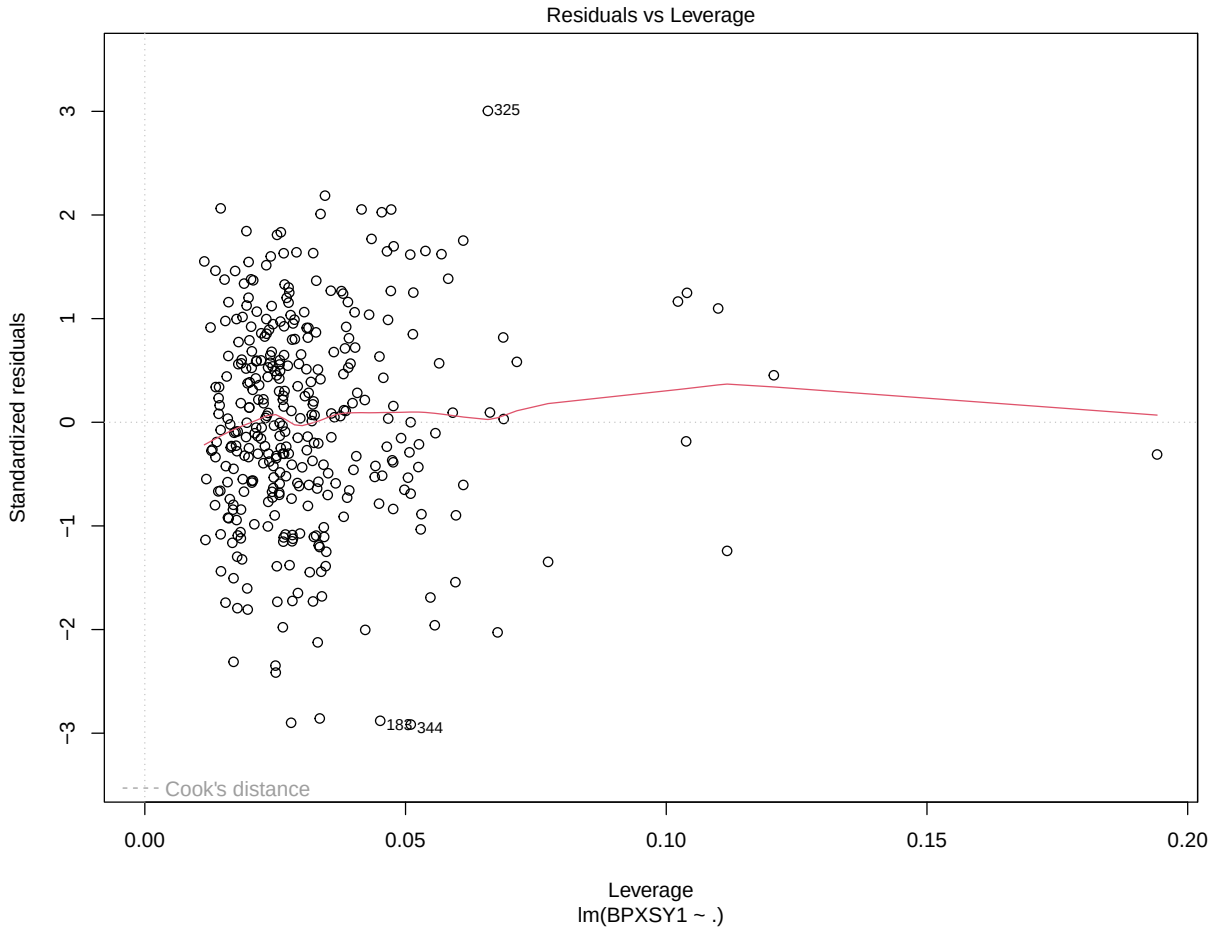
Together, the model explains 71% of the variance of the changes in systolic blood level of the individuals who has been informed of having hypertension.

```
# Assess the full model
plot(lm.full)
```


Residuals vs Fitted
Residuals
Fitted values
lm(BPXSY1 ~ .)
325
160
344
−2
−1
0
1
2
3
−1
0
1


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(BPXSY1 ~ .)
325
344
160
−3
−2
−1
0
1
2
3


Scale−Location
160
344
325
√|Standardized residuals|
0.0
0.5
1.0
1.5
−2
−1
0
1
2
3
Fitted values
lm(BPXSY1 ~ .)



Plotting out the full model, we see that residuals is distributed centered at 0 with variance approximately 1. The variances look approximately the same for each fitted value. All of data points are independent. The Q-Q plot of the residuals and the normal distribution indicates that the standardized residuals fitted almost perfectly to the normal distribution. Therefore, the full model seems to be capturing the changes of the systolic blood pressure level well through a linear relationship with the predictor variables.

Variable Selection

I performed the step-wise model selection of both directions elimination by AIC to find the model with the most relevant variables contributing to systolic blood pressure.

```
# Base model
lm.base = lm(BPXSY1~1,data=nhanes_ht_ready)
summary(lm.base)
```

```
##
## Call:
## lm(formula = BPXSY1 ~ 1, data = nhanes_ht_ready)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8895 -0.5905 -0.1594  0.5591  4.5825
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.280e-16  5.361e-02       0        1
##
## Residual standard error: 1 on 347 degrees of freedom
```

```
# Step-wise model selection
lm.step =stepAIC(lm.full, scope=list(lm.base, lm.full), direction = "both")
```

```
## Start:  AIC=-410.04
## BPXSY1 ~ BMXBMI + RIDAGEYR + LBXTC + LBXSGL + LBXSTP + BPXML1 +
##     BPXPLS + BPXDI1 + BPQ100D + RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## - LBXTC     1     0.002 100.56 -412.03
## - BPQ100D   1     0.051 100.61 -411.86
## - LBXSGL    1     0.219 100.78 -411.28
## - BMXBMI    1     0.321 100.88 -410.93
## - LBXSTP    1     0.396 100.95 -410.67
## - BPXPLS    1     0.547 101.10 -410.15
## <none>                  100.56 -410.04
## - RIDAGEYR  1     0.685 101.24 -409.68
## - RIAGENDR  1     1.401 101.96 -407.23
## - BPXDI1    1     3.757 104.31 -399.27
## - BPXML1    1   166.835 267.39  -71.70
##
## Step:  AIC=-412.03
## BPXSY1 ~ BMXBMI + RIDAGEYR + LBXSGL + LBXSTP + BPXML1 + BPXPLS +
##     BPXDI1 + BPQ100D + RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## - BPQ100D   1     0.065 100.62 -413.81
## - LBXSGL    1     0.218 100.78 -413.28
## - BMXBMI    1     0.320 100.88 -412.93
## - LBXSTP    1     0.401 100.96 -412.65
## - BPXPLS    1     0.546 101.10 -412.15
## <none>                  100.56 -412.03
## - RIDAGEYR  1     0.687 101.24 -411.67
## - RIAGENDR  1     1.472 102.03 -408.98
## - BPXDI1    1     3.830 104.39 -401.02
## - BPXML1    1   167.338 267.89  -73.04
##
## Step:  AIC=-413.81
## BPXSY1 ~ BMXBMI + RIDAGEYR + LBXSGL + LBXSTP + BPXML1 + BPXPLS +
##     BPXDI1 + RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## - LBXSGL    1     0.186 100.81 -415.16
## - BMXBMI    1     0.315 100.94 -414.72
## - LBXSTP    1     0.368 100.99 -414.54
## - BPXPLS    1     0.570 101.19 -413.84
## <none>                  100.62 -413.81
## - RIDAGEYR  1     0.622 101.24 -413.66
## - RIAGENDR  1     1.512 102.14 -410.62
## - BPXDI1    1     3.818 104.44 -402.85
```

```
## - BPXML1    1   169.138 269.76  -72.63
##
## Step:  AIC=-415.16
## BPXSY1 ~ BMXBMI + RIDAGEYR + LBXSTP + BPXML1 + BPXPLS + BPXDI1 +
##     RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## - LBXSTP    1     0.319 101.13 -416.06
## - BMXBMI    1     0.375 101.18 -415.87
## - BPXPLS    1     0.474 101.28 -415.53
## <none>                  100.81 -415.16
## - RIDAGEYR  1     0.694 101.50 -414.78
## - RIAGENDR  1     1.495 102.30 -412.04
## - BPXDI1    1     3.733 104.54 -404.51
## - BPXML1    1   173.360 274.17  -68.98
##
## Step:  AIC=-416.06
## BPXSY1 ~ BMXBMI + RIDAGEYR + BPXML1 + BPXPLS + BPXDI1 + RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## - BMXBMI    1     0.463 101.59 -416.48
## - BPXPLS    1     0.503 101.63 -416.34
## <none>                  101.13 -416.06
## - RIDAGEYR  1     0.847 101.97 -415.16
## - RIAGENDR  1     1.509 102.64 -412.91
## - BPXDI1    1     3.700 104.83 -405.56
## - BPXML1    1   173.465 274.59  -70.45
##
## Step:  AIC=-416.48
## BPXSY1 ~ RIDAGEYR + BPXML1 + BPXPLS + BPXDI1 + RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## - BPXPLS    1     0.429 102.02 -417.01
## <none>                  101.59 -416.48
## - RIDAGEYR  1     0.674 102.27 -416.17
## - RIAGENDR  1     1.906 103.50 -412.01
## - BPXDI1    1     3.696 105.29 -406.04
## - BPXML1    1   176.649 278.24  -67.85
##
## Step:  AIC=-417.01
## BPXSY1 ~ RIDAGEYR + BPXML1 + BPXDI1 + RIAGENDR
##
##            Df Sum of Sq    RSS     AIC
## <none>                  102.02 -417.01
## - RIDAGEYR  1     0.912 102.93 -415.91
## - RIAGENDR  1     1.762 103.78 -413.05
## - BPXDI1    1     3.567 105.59 -407.05
## - BPXML1    1   177.430 279.45  -68.34
```

```
summary(lm.step)
```

```
##
## Call:
## lm(formula = BPXSY1 ~ RIDAGEYR + BPXML1 + BPXDI1 + RIAGENDR,
##     data = nhanes_ht_ready)
```

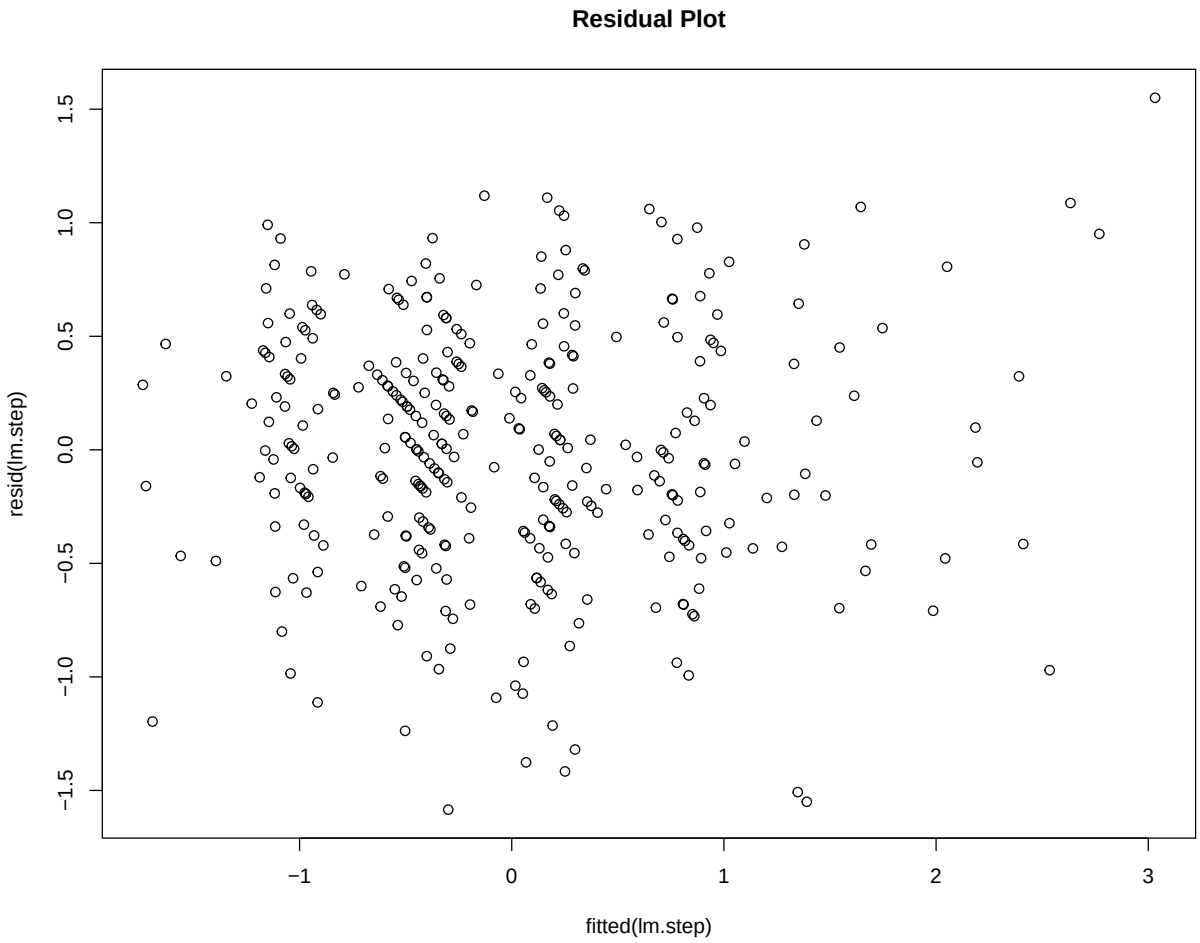
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58452 -0.36750  0.00438  0.38056  1.55018
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06977    0.04094  -1.704 0.089268 .
## RIDAGEYR     0.05623    0.03210   1.751 0.080777 .
## BPXML1       0.80293    0.03287  24.424  < 2e-16 ***
## BPXDI1       0.10826    0.03126   3.463 0.000602 ***
## RIAGENDR     0.14627    0.06009   2.434 0.015440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5454 on 343 degrees of freedom
## Multiple R-squared:  0.706,  Adjusted R-squared:  0.7026
## F-statistic: 205.9 on 4 and 343 DF,  p-value: < 2.2e-16
```

The final model selected by the step-wise AIC metric contains variables RIDAGEYR (age), RIAGENDR (gender), BPXDI1 (diastolic blood pressure level), and BPXML1 (maximum inflation level). We see that the coefficients of BPXML1 and RIAGENDR slightly increase and that of BPXDI1 slight decreases. From the previous analysis of the full model, the variable RIDAGEYR was not included in one of the significant variables based on the P-value, and still it does not have a significant P-value when the threshold is at 0.05 but it is included in the final model selected by step-wise iterative selection. The multiple R-squared of the selected model is of 0.706 which is slightly lower than that of the full model of 0.7102 and the adjusted R-squared is 0.7026 of the selected model which is higher the adjusted R-squared of the full model of 0.7016, indicating adding the rest of the variables does not improve the model as expected.

Evaluate Model Fit

Plotting the Residuals

```r
# Residual plot
plot(fitted(lm.step), resid(lm.step), main="Residual Plot")
```



The residuals of the selected model is normally distributed and centered at 0 with variance about 1 for all fitted values, indicating the model captures well the changes of systolic blood pressure. ### Evalution with RMSE and MAE RMSE (root mean squared error) and MAE (mean absolute error) are metrics to assess the the performances of linear regression model. Both of them assess the errors of the models, implying the smaller the value of the metrics are, the better the model performs.

To evaluate the model with the metric RMSE and MAE, We first partition our data set into the training data set and the testing data set with a proportion of 4:1. We first create a training model on the selected variables on the training data and predict on the testing data set to obtain RMSE and MAE on the testing data.

```
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 4.2.2
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:Metrics':
```

```
##
##     precision, recall
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Partition training/testing data
random_sample = createDataPartition(nhanes_ht_ready$BPXSY1, p=0.8, list=FALSE)
training_data = nhanes_ht_ready[seq(random_sample),]
testing_data = nhanes_ht_ready[-seq(random_sample),]

# Train selected model and predict
lm.train = lm(BPXSY1~RIDAGEYR+RIAGENDR+BPXDI1+BPXML1, data=training_data)
predictions = predict(lm.train, testing_data)

# Output results
data.frame( RMSE = rmse(predictions, testing_data$BPXSY1),
            MAE = mae(predictions, testing_data$BPXSY1))
```

```
##        RMSE      MAE
## 1 0.5755291 0.465692
```

We want to train a full model to obtain the RMSE and MAE of the full model to compare them with the selected model.

```
# Train full model and predict
lm.train.full = lm(BPXSY1~., data=training_data)
full_predictions = predict(lm.train.full, testing_data)

data.frame( RMSE = rmse(full_predictions, testing_data$BPXSY1),
            MAE = mae(full_predictions, testing_data$BPXSY1))
```

```
##        RMSE       MAE
## 1 0.5823277 0.4710284
```

We see that the RMSE and MAE of the selected training model are slightly lower than those of the full training model, indicating the selected model perform better on strange data set after training and the previous higher multiple R-squared of the full model is partially affected by over fitting.

Alternative Models: Regularization

Regularization is another popular regression model to fit the data and prevent over-fitting. We want to see how each regularization penalty performs with the model.

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.2
```

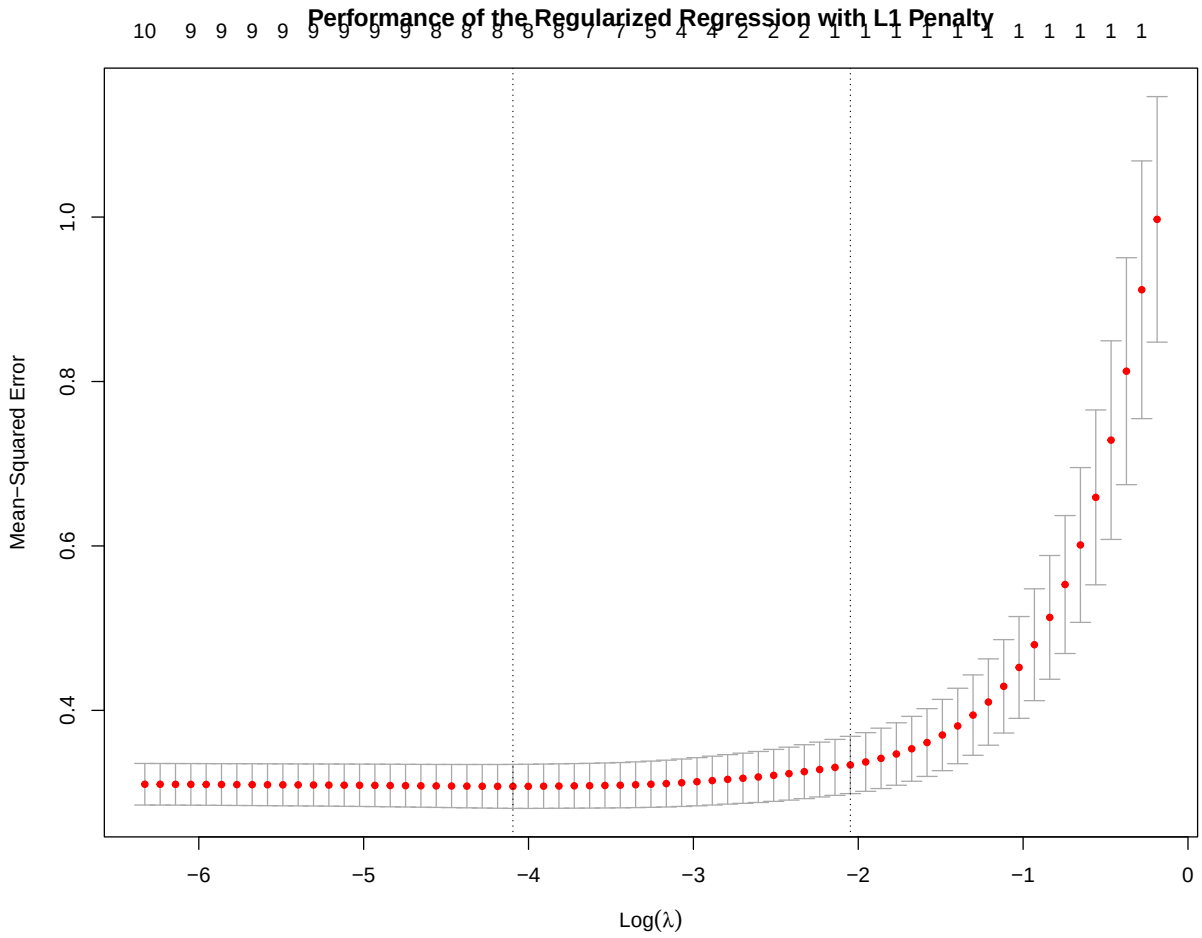
```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

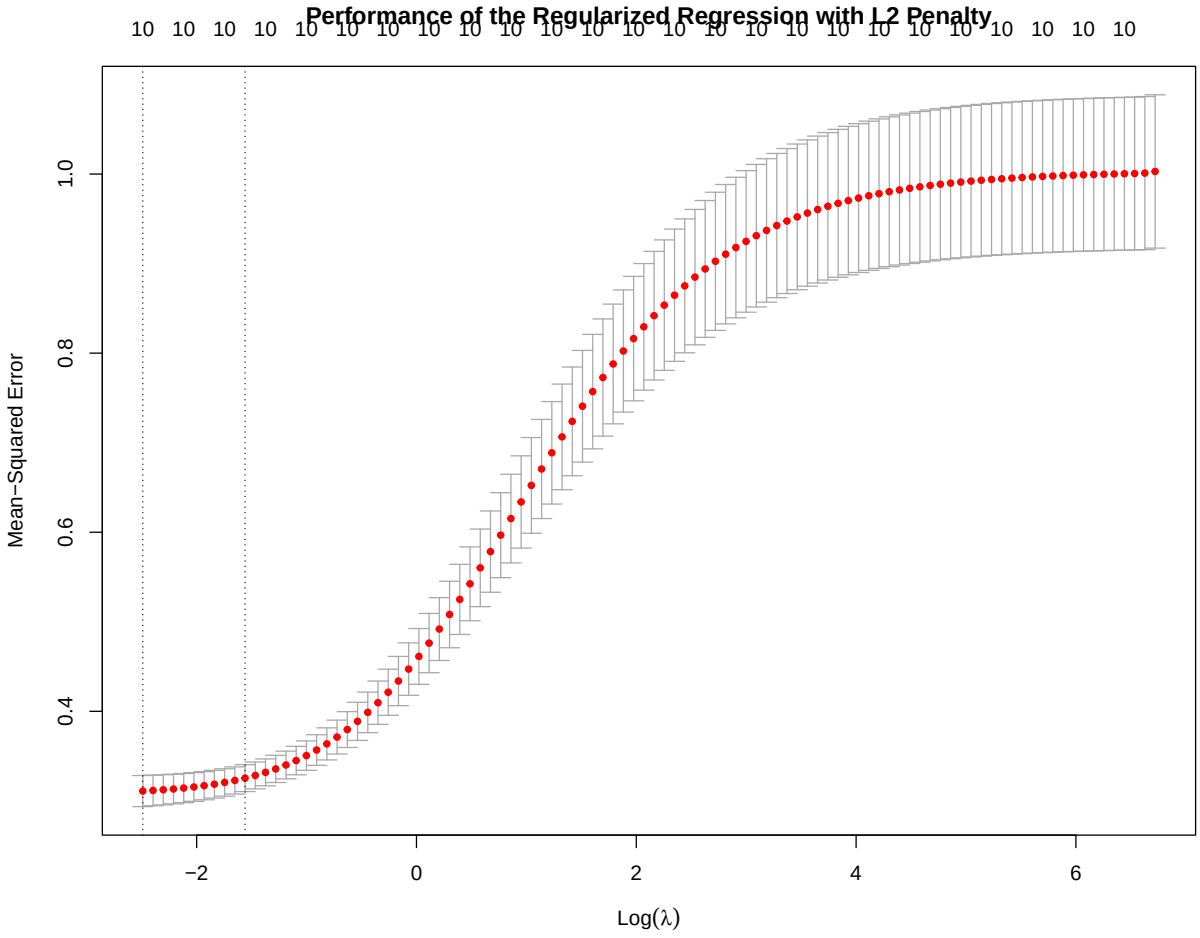
```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-4
```

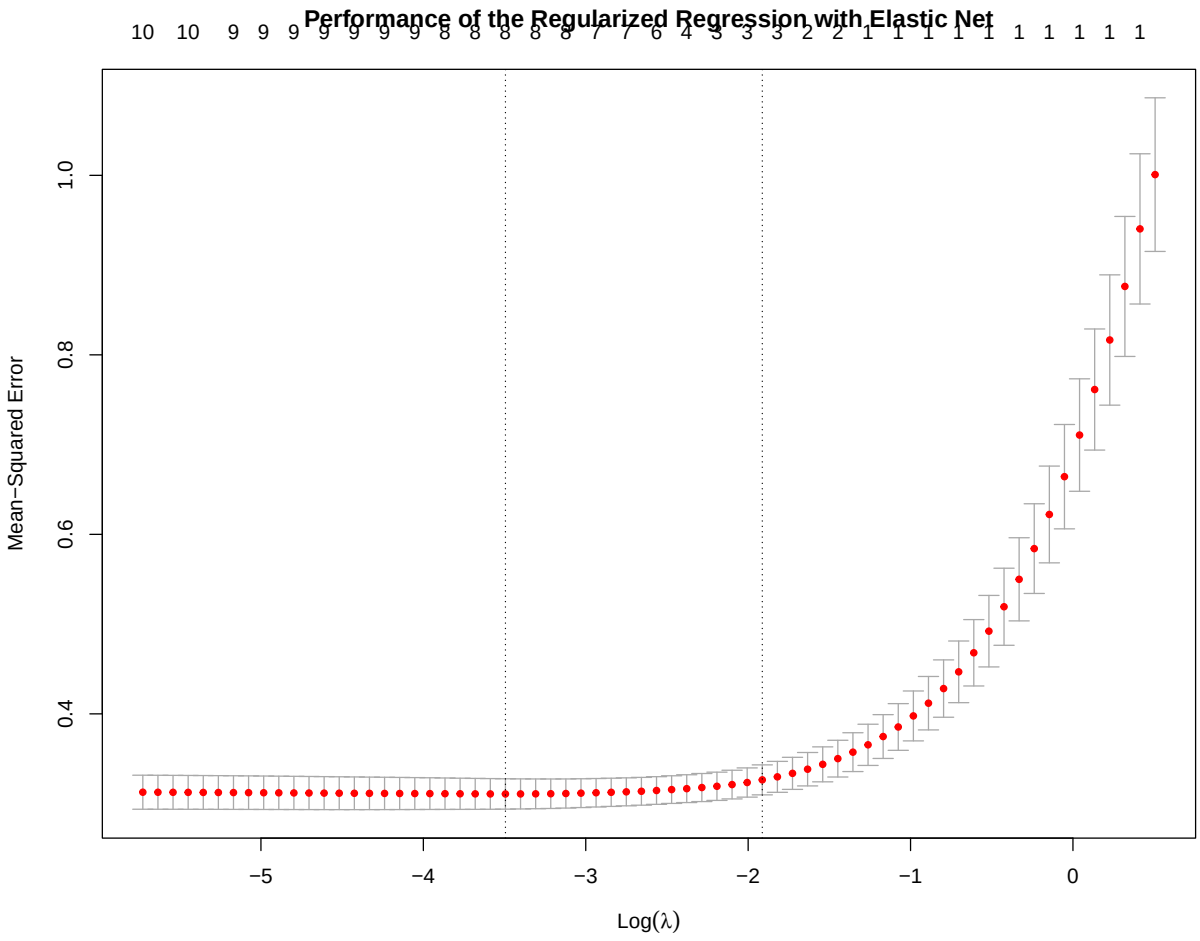
```r
# Process the ht as a matrix
ht_mat = nhanes_ht_ready %>% as.matrix

# Perform different penalties
cv_model_lasso = cv.glmnet(ht_mat[,-6], ht_mat[,6], nfolds=10,alpha=1)
cv_model_ridge = cv.glmnet(ht_mat[,-6], ht_mat[,6], nfolds=10,alpha=0)
cv_model_elastic = cv.glmnet(ht_mat[,-6], ht_mat[,6], nfolds=10,alpha=0.5)

# Plot the results
plot(cv_model_lasso,
     main="Performance of the Regularized Regression with L1 Penalty")
```




```r
plot(cv_model_ridge,
     main="Performance of the Regularized Regression with L2 Penalty")
```




```
plot(cv_model_elastic,
     main="Performance of the Regularized Regression with Elastic Net")
```



We train the models on the same training data and predict on the test data set to see their performance.

```
train_mat = training_data %>% as.matrix
test_mat = testing_data %>% as.matrix

# Train regularized regression model with optimal lambda
cv_final_lasso = glmnet(train_mat[,-6], train_mat[,6], nfolds=10,
                        alpha=1, lambda=cv_model_lasso$lambda.min)
cv_final_ridge = glmnet(train_mat[,-6], train_mat[,6], nfolds=10,
                        alpha=0, lambda=cv_model_ridge$lambda.min)
cv_final_elastic = glmnet(train_mat[,-6], train_mat[,6], nfolds=10,
                          alpha=0.5, lambda=cv_model_elastic$lambda.min)

# Predict
predictions_lasso = cv_final_lasso %>% predict(test_mat[,-6]) %>% as.vector()
predictions_ridge = cv_final_ridge %>% predict(test_mat[,-6]) %>% as.vector()
predictions_elastic = cv_final_elastic %>% predict(test_mat[,-6])%>%
  as.vector()

# Output results
data.frame(RMSE = rmse(predictions_lasso, test_mat[,6]),
           MAE = mae(predictions_lasso, test_mat[,6]))
```

```
##        RMSE       MAE
## 1 0.5841248 0.4700955
```

```
data.frame(RMSE = rmse(predictions_ridge, test_mat[,6]),
           MAE = mae(predictions_ridge, test_mat[,6]))
```

```
##       RMSE       MAE
## 1 0.602861 0.4803285
```

```
data.frame(RMSE = rmse(predictions_elastic, test_mat[,6]),
           MAE = mae(predictions_elastic, test_mat[,6]))
```

```
##        RMSE       MAE
## 1 0.5879759 0.4722171
```

To summarize, the regression model with lasso regularization performs the best with the lowest RMSE and MAE among the three regularization method. However, they are still worse than the performance of the selected model.

Follow-up Study:

Discarding Age Variable

We wonder for the selected model, the included indicator variable age does not have a significant P-value for its coefficient. Whether we should include it in the model needs to be further investigated.

```
# Build model without age
lm.no_age = lm(BPXSY1~RIAGENDR+BPXDI1+BPXML1, data=nhanes_ht_ready)
summary(lm.no_age)
```

```
##
## Call:
## lm(formula = BPXSY1 ~ RIAGENDR + BPXDI1 + BPXML1, data = nhanes_ht_ready)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64786 -0.38408 -0.00089  0.38751  1.47093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07565    0.04093  -1.848  0.06541 .
## RIAGENDR     0.15859    0.05986   2.649  0.00844 **
## BPXDI1       0.09346    0.03019   3.096  0.00212 **
## BPXML1       0.82422    0.03064  26.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.547 on 344 degrees of freedom
## Multiple R-squared:  0.7034, Adjusted R-squared:  0.7008
## F-statistic: 271.9 on 3 and 344 DF,  p-value: < 2.2e-16
```

```
# Train and predict
lm.train.no_age = lm(BPXSY1~RIAGENDR+BPXDI1+BPXML1, data=training_data)
predictions_no_age = predict(lm.train, testing_data)
data.frame( RMSE = rmse(predictions_no_age, testing_data$BPXSY1),
            MAE = mae(predictions_no_age, testing_data$BPXSY1))
```

```
##        RMSE      MAE
```

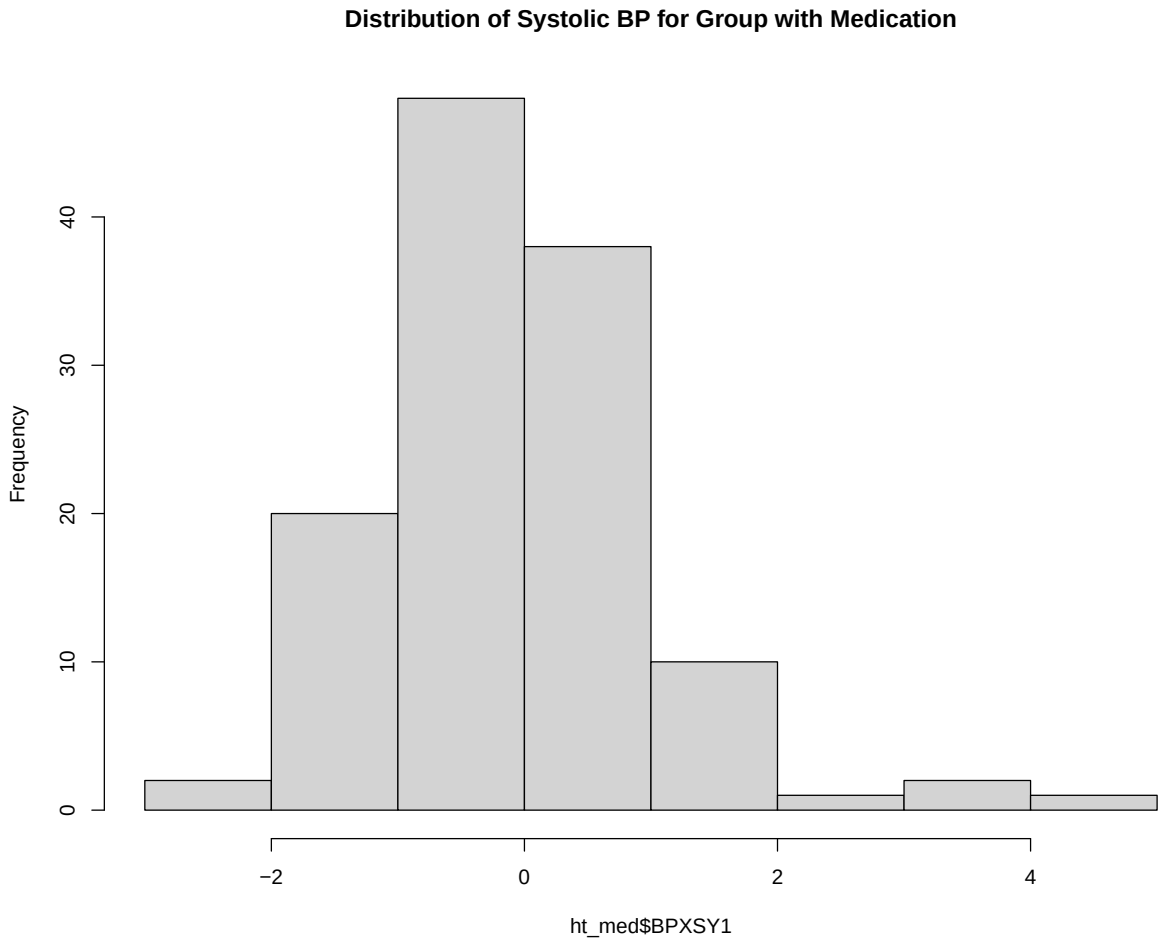
```
## 1 0.5755291 0.465692
```

The adjusted R squared of the model without age is slightly lower than it of the model with age, indicating that adding the age variable improves the model as expected. We see that for the trained model with the variable age, the RMSE and MAR are the same with them of the trained model without the variable age, indicating adding the variable age does not damage the trained model's performance on strange data set. However, from this partition of training data and testing data, the model with age does not improve the performance as well to reflect the more variance it explained comparing with the model without age.

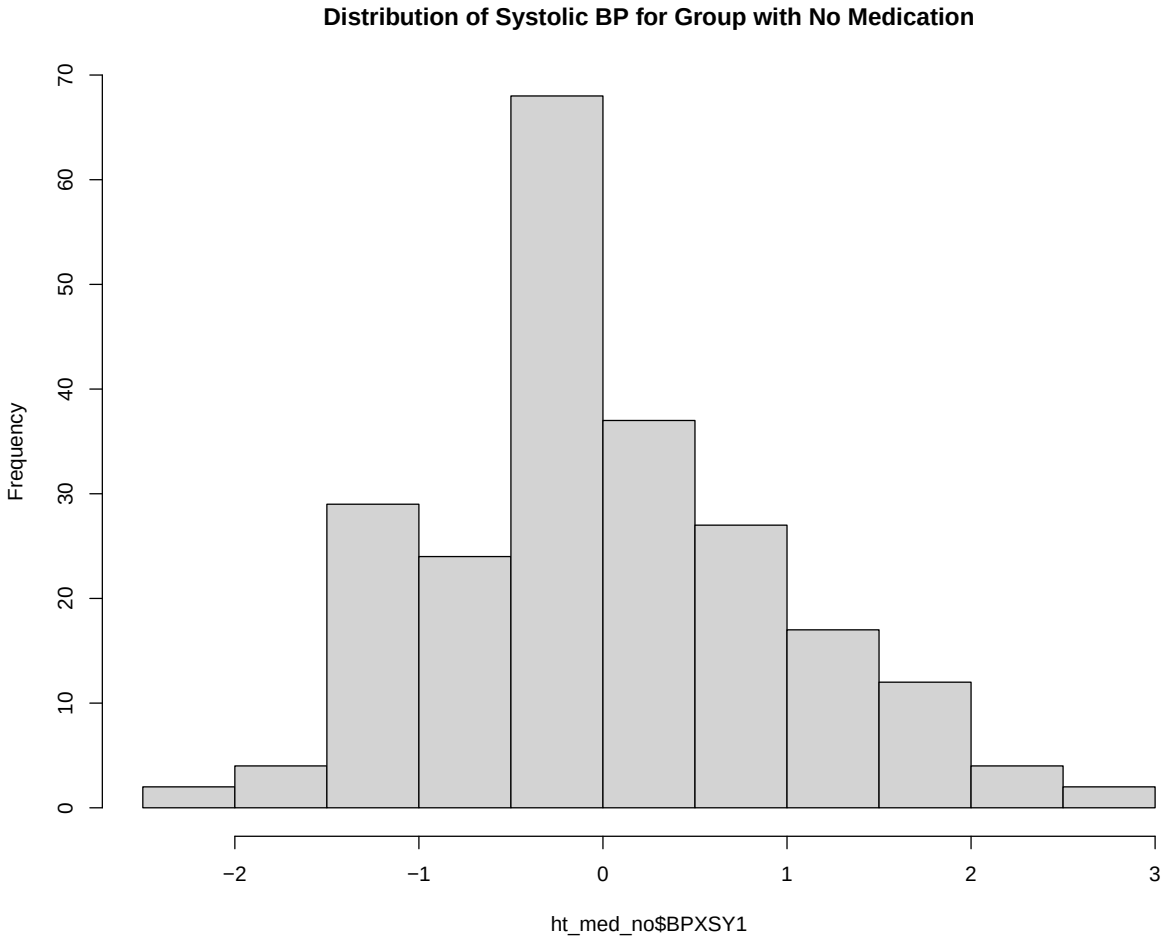
Effect of BP Medications

We are specific interested in the relationship between the impact of use of medications on the systolic blood pressure in this group with diagnosed hypertension. Looking directly at the histograms, the distributions look identical.

```
# Separate the cohort by medication use
ht_med = nhanes_ht_ready %>% filter(BPQ100D==1)
ht_med_no = nhanes_ht_ready %>% filter(BPQ100D==0)
# Plot distribution
hist(ht_med$BPXSY1,
     main="Distribution of Systolic BP for Group with Medication")
```

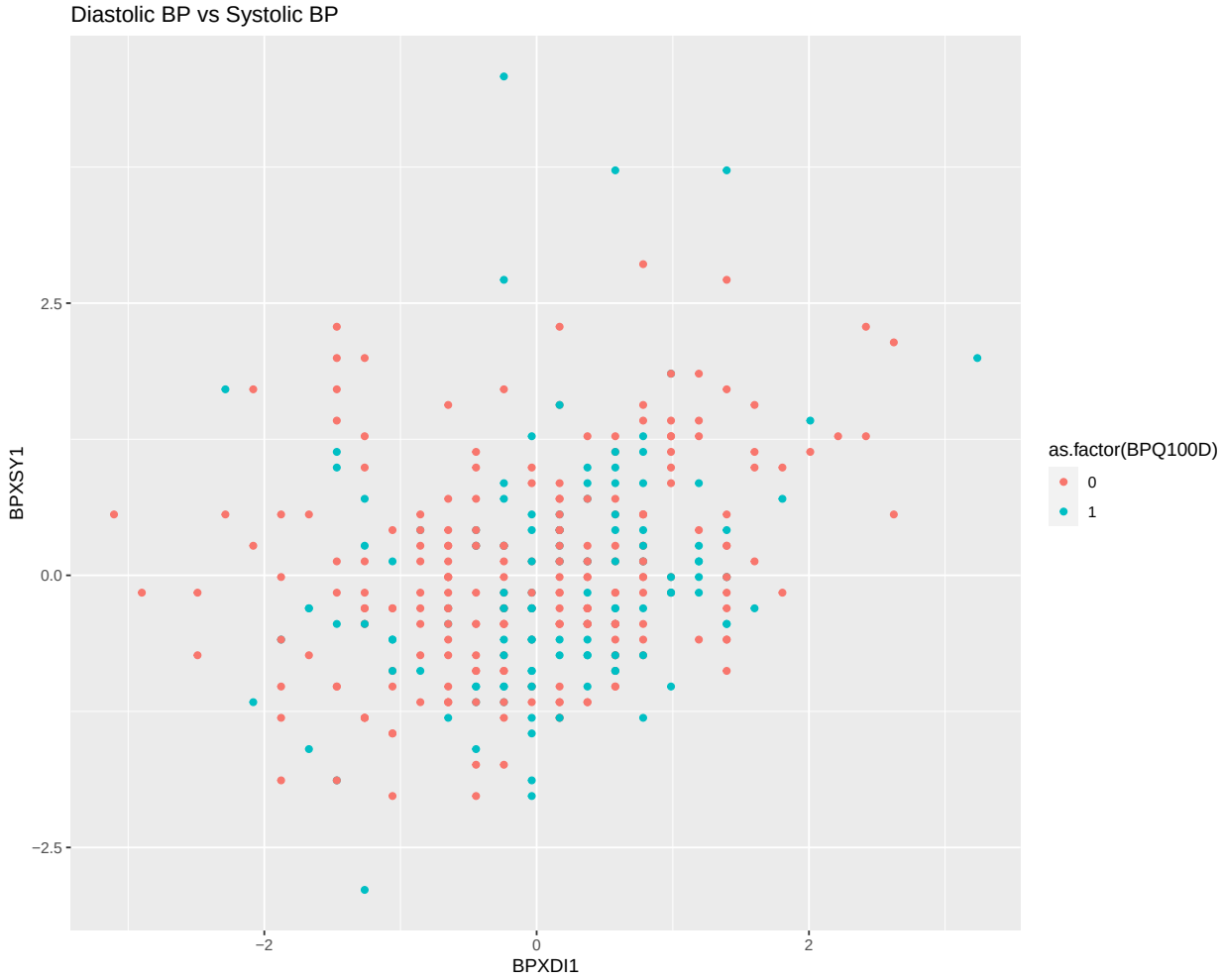



```
hist(ht_med_no$BPXSY1,
     main="Distribution of Systolic BP for Group with No Medication")
```



If we look at the plot of the diastolic BP and systolic BP colored with if taking medication, we do not observe any specific pattern for each group.

```
library(ggplot2)
# Plot out diastolic BP and systolic BP
ggplot(data=nhanes_ht_ready)+geom_point(aes(BPXDI1, BPXSY1,
                                            color=as.factor(BPQ100D)))+
  labs(title="Diastolic BP vs Systolic BP")
```



If we conduct a Pearson correlation test between the systolic BP and the medication status, the result is not significant with a P-value of 0.5584.

```
# Conduct a pearson correlation test
cor.test(nhanes_ht_ready$BPQ100D, nhanes_ht_ready$BPXSY1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  nhanes_ht_ready$BPQ100D and nhanes_ht_ready$BPXSY1
## t = -0.5858, df = 346, p-value = 0.5584
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13615741  0.07389857
## sample estimates:
##         cor
## -0.03147698
```

```
# Assess the linear relationship between systolic BP and
# medication use alone
lm.bp = lm(BPXSY1~BPQ100D,data=nhanes_ht_ready)
summary(lm.bp)
```

```
##
```

```
## Call:
## lm(formula = BPXSY1 ~ BPQ100D, data = nhanes_ht_ready)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8468 -0.6136 -0.1825  0.5360  4.6253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02309    0.06658   0.347    0.729
## BPQ100D     -0.06587    0.11245  -0.586    0.558
##
## Residual standard error: 1.001 on 346 degrees of freedom
## Multiple R-squared:  0.0009908,  Adjusted R-squared:  -0.001897
## F-statistic: 0.3432 on 1 and 346 DF,  p-value: 0.5584
```

To conclude the BP Medication analysis, we do not observe any significant difference between the group who takes medication and the group who does not take the medication, indicating the effect of the medication is not seemingly.

Conclusion

In this study, using step-wise iteration selection, we have identified four predictor variables to be included in a linear regression model from ten variables we selected which best captures the variances in systolic blood pressure. The four variables in the selected model included are age, gender, maximum inflation level, and diastolic blood pressure. We have evaluated the performance of the model by RMSE and MAE. We also included different regularized regression full model with lasso penalty, ridge penalty, and elastic net. However, the selected model still outperforms them based on the RMSE and MAE. In the selected model, age has a coefficient with non-significant P-value. We further validate its inclusion in the model using a model without age. We also conducted an analysis specifically on the effect of medication use on the systolic blood pressure level. However, no specific correlation was found.

Reference

Disease Control C. for & National Center for Health Statistics (NCHS). P. (CDC). (2013-2014). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

High blood pressure (hypertension) (2022). In: Mayo Clinic. Mayo Foundation for Medical Education; Research. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410.

Naci H., Salcher-Konrad M., Dias S., Blum M.R., Sahoo S.A., Nunan D. & Ioannidis J.P.A. (2019). How does exercise treatment compare with antihypertensive medications? A network meta-analysis of 391 randomised controlled trials assessing exercise and medication effects on systolic blood pressure. Br J Sports Med 53 (14): 859–869.