University of Toronto Scarborough

STAC67 Case Study: Data Set 2

Factors Affecting Systolic Blood Pressure (SBP)

By Group 12

Donglin Que(1006741233): Format report, Background and significance, Result, Discussion, Conclusion and References

Yichen Bao(1005773254): Data cleaning, Model building, diagnostics and selection

Yifan Cui(1006227995): Model diagnostics, Result, Model validation and explanation

Yuxin Zhang(1006747004): Background and significance, Discussion

Last compiled on April 04, 2023

Contents

1 Background and Significance

Diastolic blood pressure was once considered the most important component of blood pressure and the main target of antihypertensive therapy. However, important epidemiological studies over the decades have pointed to the importance of systolic blood pressure (SBP)(CDC 2021). Unlike diastolic blood pressure, systolic blood pressure increases gradually with age and is the most common form of hypertension in aging societies. If the characteristic changes in systolic and diastolic blood pressure with age lead to elevated pulse pressure, the greater your risk of other health problems, such as heart disease, heart attack, and stroke (Strandberg and Pitkala 2003).

In a 1990 study, the Cardiovascular Health Study recruited and examined 5,888 people aged 65 and over and followed them for seven years. After adjusting for potential confounders, the study found that systolic blood pressure is a better predictor of cardiovascular events than diastolic blood pressure, which is important data for studying cardiovascular events(Psaty et al. 2001).

This project aims to find out the factors that affect SBP by studying the blood pressure data of populations and to find the possible causes of high blood pressure statistically. It has a good warning effect on the prevention of high blood pressure. This study is based on data samples provided by UTSC. The data sample contains 18 variables for 500 observations. First, we use correlation to judge whether every two variables are related, and select uncorrelated factors from 18 variables for research. After cleaning the data, we built a linear regression model to show a linear relationship between the response variable SBP and the manipulated variable. Finally, we use stepwise selection based on AIC values to select the best model involving the following variables: '***Exercise***', '***Age***', '***Alcohol***', '***Treatment***', '***BMI***' and '***Smoking***'.

2 Exploratory data analysis

2.1 Data cleaning

This data set include 500 observations, the response variable is SBP and there are 18 related variables. After loading data, we mutate data and build the correlation matrix as Figure 1.

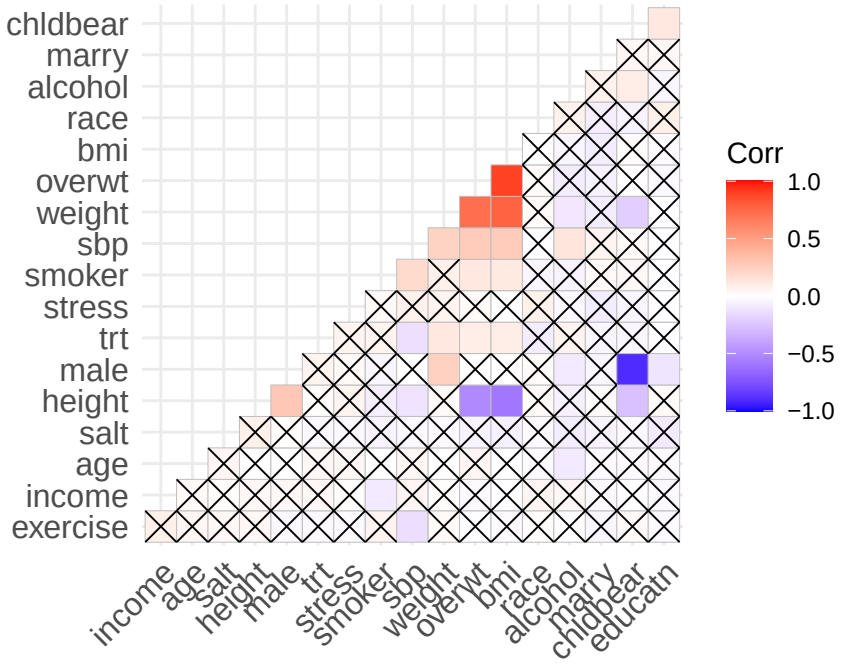

Figure 1: Correlation matrix of all variables.

2.2 Collinearity check

In the first stage we test the correlationship between each variable and we delete those variables who have a highly correlationship. For example, the correlation between '***weight***' and '***overwt***' is 0.717, which is really high. If both variables are involved in the model, the model may suffer from multicollinearity. The significance for both variables will drop. We choose to remove '***'height***', '***weight***' and '***overwt***' in our model, because they can be represented by '***bmi***'. This is proved by the correlation matrix as well. Therefore, there is no need to involve all these four variables in the model.

Base on the full model, some of the p-values are bigger than 0.05,which is not significant,

such that they will not effect to our response variable. Therefore, we keep these variables are significant and should be involved in the model and remove the rest from the model to improve the overall significance.

3 Model

3.1 Model building

We first build our model as

```
# build full model
full_model <- lm(sbp ~ factor(exercise) + age + factor(race) + factor(alcohol) + factor(trt) + bmi + factor(stress) + factor(salt) +
                   factor(chldbear) + factor(income) + factor(educatn) + factor(male) + factor(marry) + factor(smoker), data)
```

The summary of our model is

```
## MODEL FIT:
## F(22,477) = 5.90, p = 0.00
## R² = 0.21
## Adj. R² = 0.18
##
## Standard errors: OLS
## --------------------------------------------------------
##                             Est.   S.E.   t val.      p
## ----------------------- -------- ------ -------- ------
## (Intercept)               105.95   6.96    15.21   0.00
## factor(exercise)2         -10.98   2.90    -3.78   0.00
## factor(exercise)3         -10.52   2.71    -3.88   0.00
## age                         0.13   0.09     1.53   0.13
## factor(race)2               0.70   2.94     0.24   0.81
## factor(race)3               1.72   5.36     0.32   0.75
## factor(race)4              -6.90   5.86    -1.18   0.24
## factor(alcohol)2            1.42   2.87     0.50   0.62
## factor(alcohol)3           12.19   2.89     4.22   0.00
## factor(trt)1              -14.14   2.91    -4.86   0.00
## bmi                         0.94   0.14     6.88   0.00
## factor(stress)2             3.00   2.87     1.05   0.30
## factor(stress)3             5.67   2.87     1.98   0.05
## factor(salt)2               2.45   2.88     0.85   0.40
## factor(salt)3               1.67   2.82     0.59   0.55
## factor(chldbear)2          -3.48   2.75    -1.26   0.21
## factor(chldbear)3           0.97   2.90     0.33   0.74
## factor(income)2             1.31   2.79     0.47   0.64
## factor(income)3             4.30   2.86     1.51   0.13
## factor(educatn)2            0.20   2.83     0.07   0.94
## factor(educatn)3            0.07   2.81     0.03   0.98
## factor(male)1
## factor(marry)1              2.82   2.33     1.21   0.23
## factor(smoker)1            10.72   2.35     4.57   0.00
## --------------------------------------------------------
```

But there is still too much variables and we need to do the model selection.

3.2 Model selection

We use the stepwise selection based on AIC value

variable	step1	step2	step3	step4	step5	step6	step7	step8
male	-	-	-	-	-	-	-	-
race		-	-	-	-	-	-	-
educatn			-	-	-	-	-	-
salt				-	-	-	-	-
chldbear					-	-	-	-
income						-	-	-
marry							-	-
stress								-
exercise								
alcohol								
smoker								
trt								
bmi								
age								
AIC	3256.6	3252.34	3248.35	3244.98	3243.4	3242.6	3242.16	3241.55

After the stepwise selection, our final model involves these variables '***exercise***', '***age***', '***alcohol***', '***trt***', '***bmi***', and '***smoker***'. Which is shown below.

```
## MODEL INFO:
## Observations: 500
## Dependent Variable: sbp
## Type: OLS linear regression
##
## MODEL FIT:
## F(8,491) = 14.72, p = 0.00
## R² = 0.19
## Adj. R² = 0.18
##
## Standard errors: OLS
## -------------------------------------------------------
##                              Est.   S.E.   t val.      p
## ----------------------- -------- ------ -------- ------
## (Intercept)               113.11   5.73    19.74   0.00
## factor(exercise)2         -10.30   2.85    -3.61   0.00
## factor(exercise)3         -10.27   2.67    -3.84   0.00
## age                         0.14   0.09     1.58   0.11
## factor(alcohol)2            0.65   2.81     0.23   0.82
## factor(alcohol)3           11.82   2.81     4.21   0.00
```

```
## factor(trt)1              -13.43   2.87    -4.68   0.00
## bmi                         0.91   0.13     6.77   0.00
## factor(smoker)1            10.91   2.30     4.75   0.00
## ------------------------------------------------------
```

3.3 Model diagnostics

Since the p value for '***age***' is still insignificant. Therefore, we try to remove this variable from the model and regress again.

Comparing the two model with and without age in Page 13, we found that even though the p value shows that '***age***' is insignificant, but when it is removed, the adjusted R square drops, which means that we should not remove it. The same as '***alcohol***'. We can see that alcohol2 is not significant in our model but for the part of highly usage of alcohol (alcohol3) is significant, so that we cannot drop the whole variable. In conclusion, our final model is:

```
##
## Call:
## lm(formula = sbp ~ factor(exercise) + age + factor(alcohol) +
##     factor(trt) + bmi + factor(smoker), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.344 -17.733  -1.147  16.480  69.701
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       113.11227    5.73018  19.740  < 2e-16 ***
## factor(exercise)2 -10.30448    2.85115  -3.614 0.000332 ***
## factor(exercise)3 -10.27216    2.67461  -3.841 0.000139 ***
## age                 0.13572    0.08593   1.579 0.114882
## factor(alcohol)2    0.65455    2.81069   0.233 0.815952
## factor(alcohol)3   11.81848    2.81015   4.206 3.09e-05 ***
## factor(trt)1      -13.42877    2.86645  -4.685 3.63e-06 ***
## bmi                 0.91330    0.13486   6.772 3.64e-11 ***
## factor(smoker)1    10.90908    2.29620   4.751 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.35 on 491 degrees of freedom
## Multiple R-squared:  0.1934, Adjusted R-squared:  0.1803
## F-statistic: 14.72 on 8 and 491 DF,  p-value: < 2.2e-16
```

4 Result

For the distribution of response variable SBP, there is a approximately normal distribution, but with a slightly left-skewed(Fig. 2).

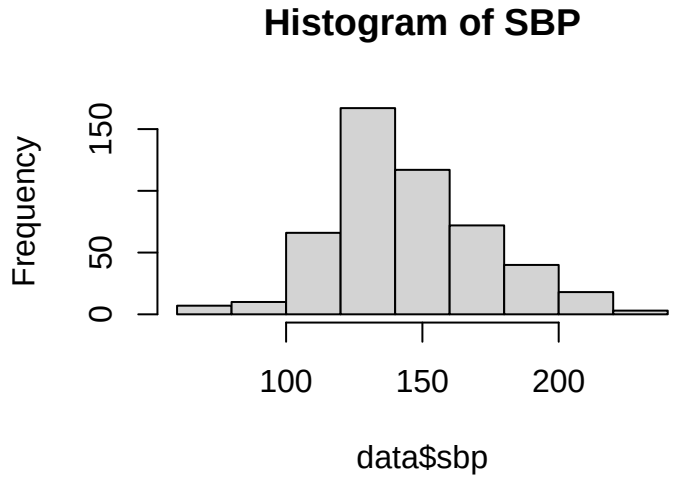

Figure 2: Histogram of data SBP

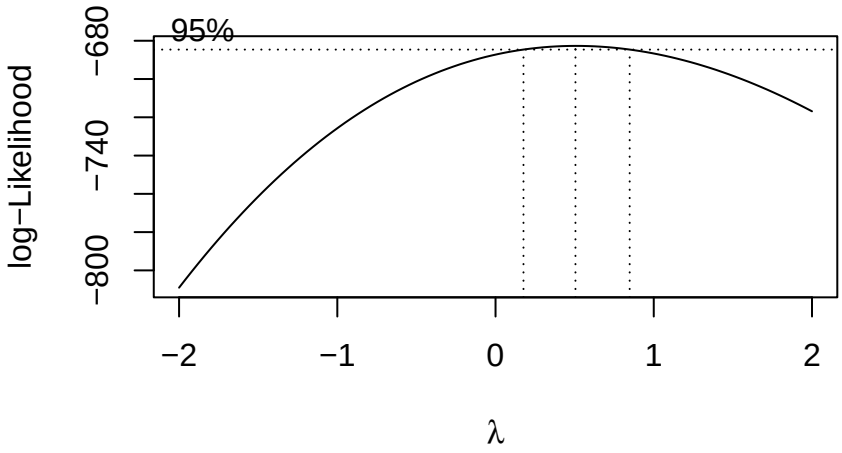

Figure 3: boxcox, lamba = 0.51

By the normal QQ plot of residuals, we can see that in Fig. 4, most of the points are NOT follow the QQ lines. There is a concern of non-normal assumption. By the figure of residuals vs fitted, there is a concern of non-constant variance (slightly tramper shape), also the red line is not parallel to x-axis. So we transfer to boxcox (Fig. 3).

After transformation, the residual plot looks better and satisfy the constant variance assumption, but by the new QQ plot (Fig. 5), it may exists some outliers and few plots are not follow the qqline, such that it failed normality assumption.

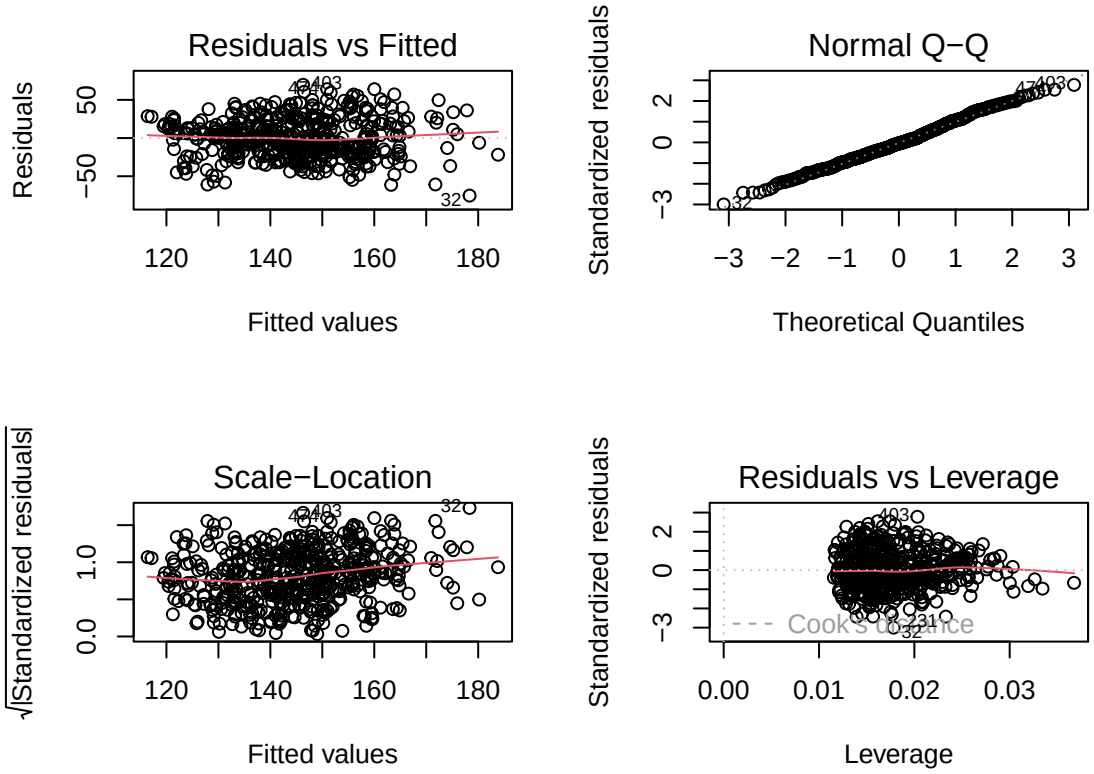

Figure 4: normal QQ plot

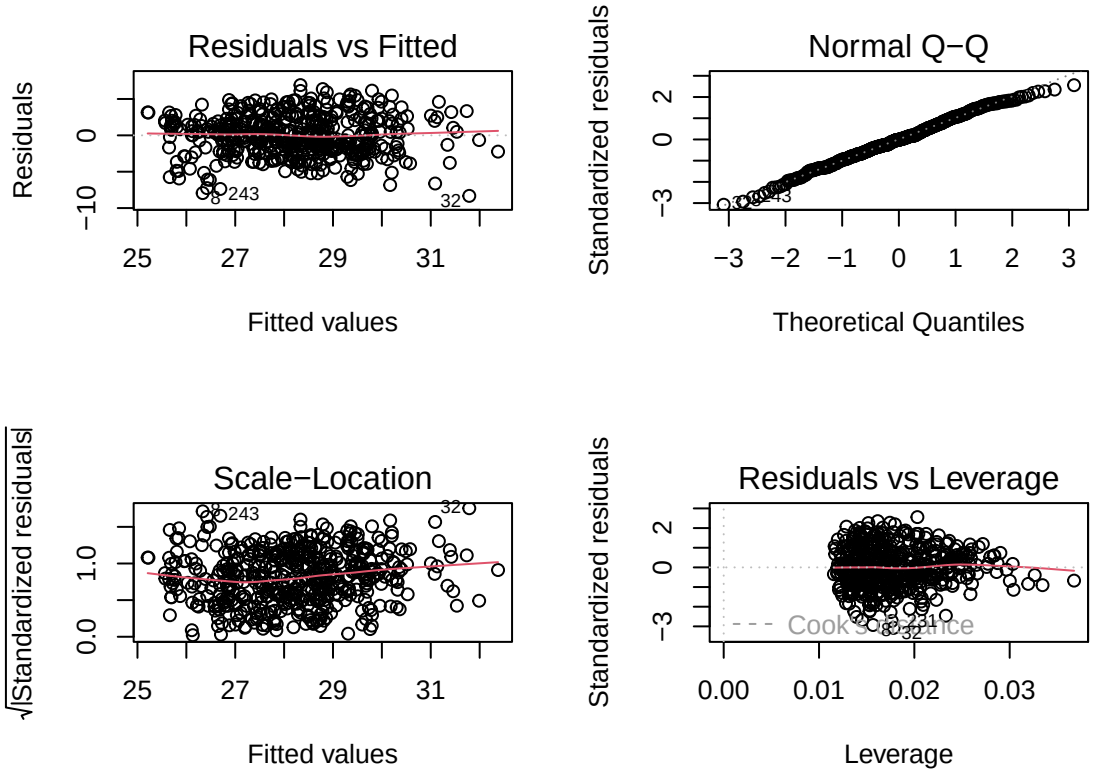

Figure 5: normal QQ plot

```r
p = length(Final_model$coefficients)
n = nrow(data)
d = qf(0.5,p,n-p)
DFFITTS = dffits(Final_model)
which(DFFITTS > 1)
```

```
## named integer(0)
```

```r
D = cooks.distance(Final_model)
which(D > d)
```

```
## named integer(0)
```

```r
DFBETA = dfbetas(Final_model)
which(DFBETA > 1)
```

```
## integer(0)
```

Check:

- $|DFFITS| > 2 \times \sqrt{\frac{p}{n}}$
- $|DFBETA| > \frac{2}{\sqrt{n}}$ 3. $D_i < 10_{th}$to 20_{th} percentile

There is no influential points

```r
crit = qt(1-0.05/(2*n), n-p-1)
which(abs(rstudent(Final_model)) > crit)
```

```
## named integer(0)
```

There is no outliers.

```r
which(hatvalues(Final_model) > 2*p/n)
```

```
## 4
## 4
```

Check: $hatvaule > \frac{2p}{n}$ There is a leverage point.

5 Discussion

After getting the final model through stepwise AIC, the p-value of '***age***' in the summary table is still very low, which probably shows that '***age***' is insignificant. But when we remove '***age***', the R^2 of the model decreases from 0.193 to 0.189 (Page 13). In this regard, we believe that the variable '***age***' should not be removed from the model, not only because the fitting degree decreases in statistics, but also in reality, SBP will increase with age(Chrysant 2018).For example, our model in Figure 6 shows that SBP increases with age and unhealthy living habits.

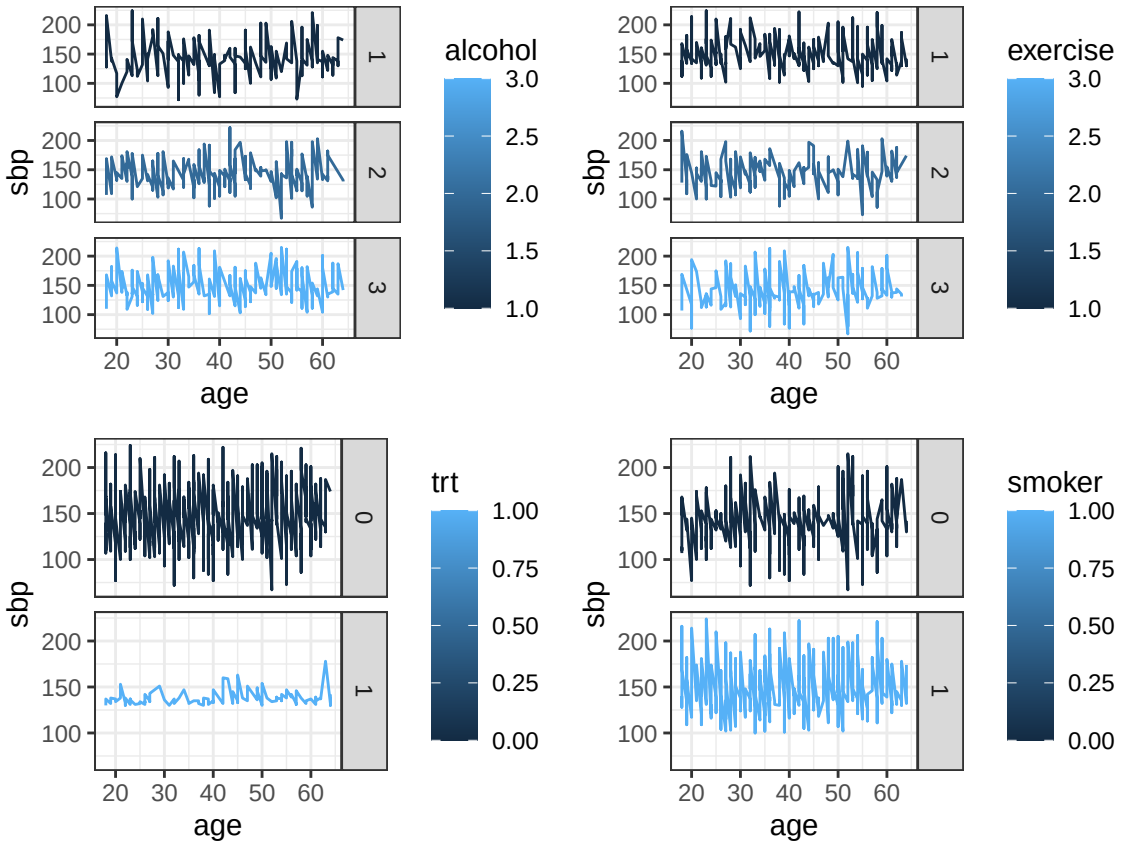

Figure 6: Comparison of SBP to related factors. How does alcohol, bmi, smoke and execrise effect on SBP

The survey is not comprehensive enough in terms of data collection. The database only provided blood pressure data for people aged 18-64, however high blood pressure is not unique to this age group. In addition, the Sample data size is too small to find the provenance, with an average of only 10 samples per age. We can draw more precise conclusions with more

samples.

In reality, target SBP below 120 mmHg is associated with lower cardiovascular events and mortality(Okin, Kjeldsen, and Devereux 2018). CDC also states that unhealthy lifestyle choices such as not getting enough regular physical activity, and alcohol and nicotine intake can all lead to elevated SBP. Another non-randomized epidemiological study suggested that lower SBP may be associated with higher mortality in the very old(Ravindrarajah et al. 2017). Therefore, in addition to our uncontrollable age, living a healthy life and exercising more, staying away from tobacco and alcohol will greatly reduce our risk of high blood pressure and have a longer life(Bundy et al. 2017).

6 Conclusion

In this case study, we focus on the factors that affect systolic blood pressure. We first use '***BMI***' to represent '***'Height***', '***Weight***' and '***Overwt***' through the strength of correlation among 18 variables, which simplifies the difficulty of the model. Then we used stepwise AIC to remove variables with low correlation to the model. Finally, our test of the model also proves the validity of our model, which is consistent with the general situation of reality. Our final model contains '***Exercise***', '***Age***', '***Alcohol***', '***Treatment***', '***BMI***', and '***Smoke***', which means that exercise, age, alcohol consumption, targeted therapy, BMI, and smoking all have a large impact on SBP.

7 Reference

Bundy, Joshua D., Changwei Li, Patrick Stuchlik, Xiaoqing Bu, Tanika N. Kelly, Katherine T. Mills, Hua He, Jing Chen, Paul K. Whelton, and Jiang He. 2017. "Systolic Blood Pressure Reduction and Risk of Cardiovascular Disease and Mortality: A Systematic Review and Network Meta-analysis." *JAMA Cardiology* 2 (7): 775–81. https://doi.org/10.1001/jamacardio.2017.1421.

CDC. 2021. "High Blood Pressure Symptoms, Causes, and Problems | cdc.gov." Centers for Disease Control and Prevention. May 18, 2021. https://www.cdc.gov/bloodpressure/about.htm.

Chrysant, Steven G. 2018. "Aggressive Systolic Blood Pressure Control in Older Subjects: Benefits and Risks." *Postgraduate Medicine* 130 (2): 159–65. https://doi.org/10.1080/00325481.2018.1433434.

Okin, Peter M., Sverre E. Kjeldsen, and Richard B. Devereux. 2018. "The Relationship of All-Cause Mortality to Average on-Treatment Systolic Blood Pressure Is Significantly Related to Baseline Systolic Blood Pressure: Implications for Interpretation of the Systolic Blood Pressure Intervention Trial Study." *Journal of Hypertension* 36 (4): 916–23. https://doi.org/10.1097/HJH.0000000000001620.

Psaty, Bruce M., Curt D. Furberg, Lewis H. Kuller, Mary Cushman, Peter J. Savage, David Levine, Daniel H. O'Leary, R. Nick Bryan, Melissa Anderson, and Thomas Lumley. 2001. "Association Between Blood Pressure Level and the Risk of Myocardial Infarction, Stroke, and Total Mortality: The Cardiovascular Health Study." *Archives of Internal Medicine* 161 (9): 1183–92. https://doi.org/10.1001/archinte.161.9.1183.

Ravindrarajah, Rathi, Nisha C. Hazra, Shota DrPH Hamada, Judith Charlton, Stephen H. D. Jackson, Alex Dregan, and Martin C. Gulliford. 2017. "Systolic Blood Pressure Trajectory, Frailty, and All-Cause Mortality >80 Years of Age: Cohort Study Using Electronic Health Records." *Circulation* 135 (24): 2357–68. https://doi.org/10.1161/CIRCULATIONAHA.116.026687.

Strandberg, Timo E., and Kaisu Pitkala. 2003. "What Is the Most Important Component of Blood Pressure: Systolic, Diastolic or Pulse Pressure?" *Current Opinion in Nephrology and Hypertension* 12 (3): 293–97. http://journals.lww.com/co-nephrolhypertens/Fulltext/2003/05000/What_is_the_most_important_component_of_blood.11.aspx.

	(1)	(2)
(Intercept)	113.112 ***	118.714 ***
	(5.730)	(4.508)
factor(exercise)2	-10.304 ***	-10.336 ***
	(2.851)	(2.855)
factor(exercise)3	-10.272 ***	-10.061 ***
	(2.675)	(2.675)
age	0.136	
	(0.086)	
factor(alcohol)2	0.655	0.434
	(2.811)	(2.811)
factor(alcohol)3	11.818 ***	11.399 ***
	(2.810)	(2.802)
factor(trt)1	-13.429 ***	-13.243 ***
	(2.866)	(2.868)
bmi	0.913 ***	0.913 ***
	(0.135)	(0.135)
factor(smoker)1	10.909 ***	10.858 ***
	(2.296)	(2.299)
N	500	500
R2	0.193	0.189
logLik	-2321.245	-2322.512
AIC	4662.490	4663.023

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.