Blood Pressure Analysis

STAT306 Group project

Hongyu Xu	21000187
Omer Tahir	15885593
Xiao Hu	69067437
Yifan Wang	45753621

University of British Columbia

2020-12-03

Contents

1 Introduction

Hypertension (or elevated blood pressure) is a very common medical condition. Surveys show that about one-third of the population suffers from hypertension. It can create several complications as it significantly increases the risk of heart, brain, kidney and other diseases. This motivates one to think what the main causes are behind elevated blood pressure and try to establish links between clinical factors and blood pressure. This proves to be very meaningful to target-oriented treatment by exploring how the risk factors perform in predicting blood pressure.
In this project, we aim to build a model that will help to predict how the resting blood pressure of a patient varies based on clinical signs and age of the patient. Therefore, we will be taking the Resting Blood pressure of the patient as the response variable, and the clinical signs as well as the age as explanatory variables.

2 Data Description

2.1 Data Source

The data set was donated to UCI in 1988, and it was collected by Andras Janosi,William Steinbrunn,Matthias Pfisterer and Robert Detrano with the goal that refers to the presence of heart disease in patients. This database contains 76 attributes, but all published experiments only used a subset of 14 of them. Details of 303 patients information were included in this data set. The data set can be found here.

2.2 Variable Description

1. **Resting Blood Pressure (mm Hg)**: This is chosen as our response variable. It is measured on the admission of the patient to the hospital. The unit is millimeters of mercury.

2. **Age (in years)**: This will be used as one of the numerical explanatory variables.

3. **Fasting Blood Sugar (> 120 mg/dl)**: This is a categorical explanatory variable which portrays the fasting blood sugar of the patient and is denoted by z_i. Since it has 2 categories, when $z_i = 1$, it represents the fasting blood sugar with a value greater than 120 milligrams per deciliter and $z_i = 0$ otherwise.

4. **Cholesterol Level (mg/dl)**: The serum cholesterol level of the patients measured in the hospital. It is one of the numerical explanatory variables. It is measured in milligrams per deciliter.

5. **Thalach (bpm)**: This is the maximum heart rate achieved by the patient. It is measured in beats per minute and is another numerical explanatory variable.

6. **Old peak (mm)**: ST depression induced by exercise relative to rest. ST segment depression may be determined by measuring the vertical distance between the patient's trace and the isoelectric line at a location 2-3 millimeters from the QRS complex. In a cardiac stress test, an ST depression of at least 1 mm after adenosine administration indicates a reversible ischaemia, while an exercise stress test requires an ST depression of at least 2 mm to significantly indicate reversible ischaemia. This is also one of the continuous explanatory variables.

3 Data analysis

3.1 Interaction

First of all, we began with our full model which includes all the explanatory variables and the dummy variable "fbs" which has interaction with all the other terms.

```
Call:
lm(formula = trestbps ~ fbs * (age + chol + thalach + oldpeak),
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-35.297 -11.428  -0.874  10.055  58.596

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.36762   12.23909   6.730 8.93e-11 ***
fbs1          0.18237   35.12832   0.005 0.995861
age           0.48308    0.12497   3.866 0.000136 ***
chol          0.02346    0.02032   1.154 0.249249
thalach       0.09391    0.05147   1.825 0.069081 .
oldpeak       2.01212    0.93639   2.149 0.032469 *
fbs1:age      0.28036    0.38790   0.723 0.470402
fbs1:chol    -0.03245    0.05280  -0.615 0.539288
fbs1:thalach -0.04349    0.12924  -0.336 0.736738
fbs1:oldpeak  5.29076    2.55113   2.074 0.038963 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 16.4 on 293 degrees of freedom
Multiple R-squared:  0.1519,    Adjusted R-squared:  0.1259
F-statistic: 5.832 on 9 and 293 DF,  p-value: 1.769e-07
```

From the output we can see the only significant interaction term is the interaction between fbs and oldpeak, so we only include this interaction term in our model:

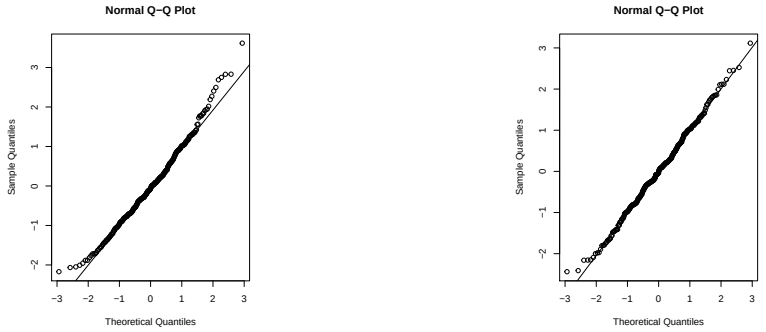
$$\text{trestbps} \sim \text{age} + \text{chol} + \text{thalach} + \text{oldpeak} + \text{fbs*age}$$

3.2 Residuals

The points in Q-Q plot forms a curve instead of a straight line, which indicates that there is a right skewness in residual distribution, so we take log of the response variable and refit the model as:

$$\log(\text{trestbps}) \sim \text{age} + \text{chol} + \text{thalach} + \text{oldpeak} + \text{fbs*age}$$

After transformation, We can see the points lying perfectly on the straight line in the Q-Q plot



(a) before transformation (b) after transformation

Figure 1: Q-Q plots of residuals

Also, there is no obvious pattern in residual versus fitted value plot.

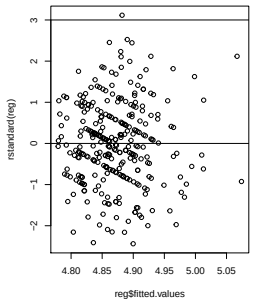
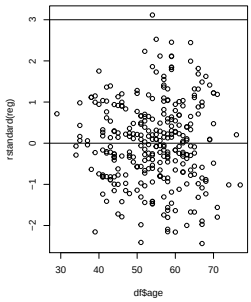
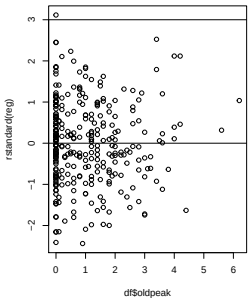


Figure 2: residual plot of fitted value

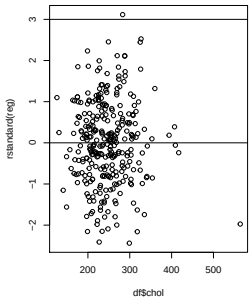
When we look at the residual plots versus explanatory variables, there is no obvious pattern for both age and thalach. For chol, although there is an oultlier with much larger chol value than the others, the other points are distributed randomly, so the residual plot is still valid. However, there is a left skewed pattern in residual plot of oldpeak, we may need to take transformation of oldpeak. Since the transformation can not increase adjusted R squared value in the model, and by consideration of the cost of interpretation, we still keep the original value of oldpeak in our model.



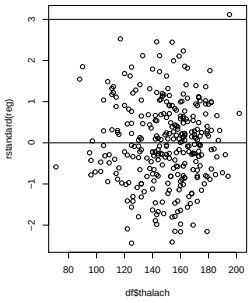
(a) residual plot of age



(b) residual plot of oldpeak



(c) residual plot of chol



(d) residual plot of thalach

Figure 3: Distribution of different schemes

3.3 Variable correlation

Figure below shows the correlation between each numerical explanatory variable, we can see there is no large correlation between each other, so there is no collinearity problem in our model.

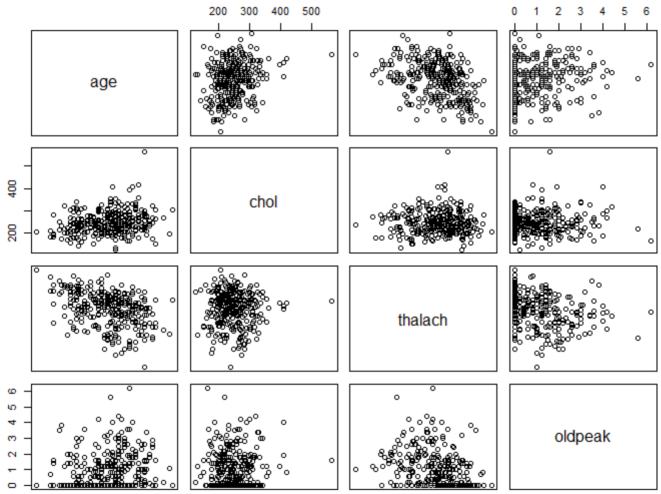


Figure 4: variable scatter plot

At this point, we have six variables in total, but since some of the variables are not significant, we need to reduce the number of variables in next step.

```
Call:
lm(formula = logtrestbps ~ . + fbs * oldpeak, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29469 -0.08464 -0.00410  0.08019  0.37546

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5108920  0.0846149  53.311  < 2e-16 ***
age          0.0037648  0.0008748   4.304 2.29e-05 ***
chol         0.0001336  0.0001391   0.961   0.3375
thalach      0.0006668  0.0003506   1.902   0.0582 .
oldpeak      0.0149132  0.0068893   2.165   0.0312 *
fbs1         0.0129746  0.0277770   0.467   0.6408
oldpeak:fbs1 0.0363722  0.0183526   1.982   0.0484 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.1218 on 296 degrees of freedom
Multiple R-squared:  0.142,     Adjusted R-squared:  0.1246
F-statistic: 8.165 on 6 and 296 DF,  p-value: 3.501e-08
```

3.4 Variable Selection

We used R to find best model for all sizes up to maximum number of parameter.

```
(Intercept)  age  chol thalach oldpeak  fbs1 oldpeak:fbs1
1        TRUE TRUE FALSE   FALSE   FALSE FALSE        FALSE
2        TRUE TRUE FALSE   FALSE   FALSE FALSE         TRUE
3        TRUE TRUE FALSE   FALSE    TRUE FALSE         TRUE
4        TRUE TRUE FALSE    TRUE    TRUE FALSE         TRUE
5        TRUE TRUE  TRUE    TRUE    TRUE FALSE         TRUE
6        TRUE TRUE  TRUE    TRUE    TRUE  TRUE         TRUE
```

Then, for each best model of different sizes, we calculate the Cp values, comparing with the number of parameter (p). From the graph, it is obvious that model 4 and model 5 have Cp values closest to p. Here

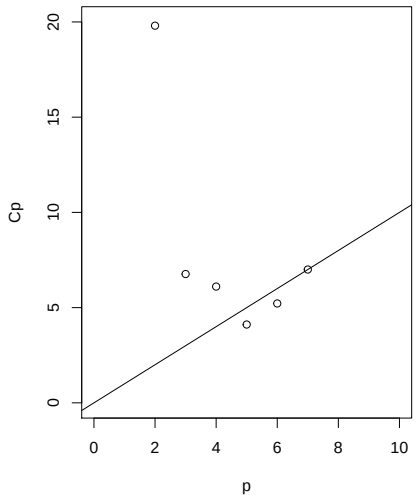


Figure 5: Cp vs p

we didn't consider the full model, since it always has Cp value equals to number of parameters. What's more, model 4 not only has the smallest Cp value, but also has the largest adjusted R squared value. Since model 4 and model 5 both have better performance in fitness, we used Cross-validation to estimate how these two model are expected to perform in predicting data that is not used during training model. By 5-fold Cross-Validation, RMSE of model 4 (0.1219) is less than that of model 5 (0.1221). Therefore, we achieved the result that model 4 is the best model for our data set:

$$\log(\text{trestbps}) \sim \text{age} + \text{thalach} + \text{oldpeak} + \text{oldpeak:fbs}$$

```
Call:
lm(formula = logtrestbps ~ age + oldpeak + thalach + oldpeak:fbs , data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29614 -0.08296  0.00252  0.08035  0.37746

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5280897  0.0823952  54.956  < 2e-16 ***
age          0.0039980  0.0008451   4.731 3.46e-06 ***
oldpeak      0.0143123  0.0066593   2.149   0.0324 *
thalach      0.0006977  0.0003487   2.001   0.0463 *
oldpeak:fbs1 0.0425028  0.0130928   3.246   0.0013 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.1216 on 298 degrees of freedom
Multiple R-squared:  0.1388,    Adjusted R-squared:  0.1272
F-statistic:    12 on 4 and 298 DF,  p-value: 4.66e-09
```

After transformation and variable selection, all the remaining coefficients in the model are all significant, and the adjusted R squared value is larger than the initial full model (0.1259).

4 Conclusion and Discussion

4.1 Results

After conducting statistical analyses of all possible models under the full model in the R environment, it was concluded that the best fitting model included the age, old peak, thalac, and the interaction between the old peak and fbs variables. This suggests that all the aforementioned variables are statistically significant and so may be potentially useful features in predicting the patient's resting blood pressure. We end up with the following best model :

$$\begin{aligned} log(Y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 z x_2 + \beta_3 x_3 \\ &= 4.5281 + 0.0040 x_1 + 0.0143 x_2 - 0.0425 z x_2 - 0.000698 x_3 + \epsilon \end{aligned}$$

where Y is blood pressure of a patient, x_1 is age, x_2 is the ST depression induced by exercise relative to rest, x_3 is the maximum heart rate achived by the patient, and z_i is the dummy variable which equals to 1 when the ith patient has fasting blood sugar (>120mg/dl), and equals to 0 when the ith patient has blood sugar less than 120mg/dl. The error term ϵ is a random variable in normal distribution with mean 0 and same variance as y, $\epsilon \sim (0, \sigma^2)$.
Therefore, we can predict the blood pressure of patient i by

$$\begin{aligned} log(\hat{y}) &= 4.5281 + 0.0040 x_1 + 0.0143 x_2 - 0.0425 z x_2 - 0.000698 x_3 \\ \hat{y} &= e^{4.5281 + 0.0040 x_1 + 0.0143 x_2 - 0.0425 z x_2 - 0.000698 x_3} \end{aligned}$$

4.2 Limitations

Firstly, since the correlated relationship between blood pressure and explanatory variables are not very significant, the adjusted R squared value is relatively low, even in our best model. There is only over 12 percent of variation in blood pressure can be captured by our model. Hence, we may need to find more variable that has stronger correlation with blood pressure and collect more data, in order to predict the value of blood pressure using more accurate model. What's more, there are very small amount of outliers and skewnesses in our residual plots, but we didn't transform explanatory variables in light of keeping easier interpretation, so these residuals might caused some bias in prediction. Additionally, since the data was collected in 1988, the information is a bit outdated, if we can collect data in recent years, the results would be more reliable.

5 Appendix

5.1 R Script

```
df = heart[c("age","chol","thalach","oldpeak","fbs","trestbps")]
cor(df)
df$fbs <- as.factor(df$fbs)
#variable correlation
pairs(df)
#full model
reg <- lm(trestbps ~ fbs*(age + chol + thalach + oldpeak),data = df)
summary(reg)
reg <- lm(trestbps~.+fbs*oldpeak,data=df)
summary(reg)
#qqplot
qqnorm(rstandard(reg))
qqline(rstandard(reg))
#light heavy tail, so we need to take log of trestbps

df$logtrestbps <- log(df$trestbps)
reg <- lm(logtrestbps~age + chol + thalach + oldpeak+fbs*oldpeak,data=df)
qqnorm(rstandard(reg))
qqline(rstandard(reg))
#residual plot against fitted value
plot(rstandard(reg)~reg$fitted.values)
abline(0,0)
#check outlier
abline(3,0) #one outlier

#residual plots against every variable
plot(rstandard(reg)~df$age)
abline(0,0)
abline(3,0)
plot(rstandard(reg)~df$chol)
abline(0,0)
abline(3,0)
plot(rstandard(reg)~df$thalach)
abline(0,0)
abline(3,0)
plot(rstandard(reg)~df$oldpeak)
abline(0,0)
abline(3,0)

# model: log(trestbps)~age+chol+thalach+oldpeak+fbs
summary(reg) #adjr2 = 0.1246

#varibale selection
library(leaps)
sreg <- regsubsets(logtrestbps~age+chol+thalach+oldpeak+fbs*oldpeak,data = df)
summary(sreg)$which
cp <- summary(sreg)$cp
plot(2:7,cp,xlab="p",ylab="Cp",ylim=c(0,20),xlim=c(0,10))
abline(0,1)
#Cp value of model4 and model5 are closest to p, and smallest
which.max(summary(sreg)$adjr2)
#model4 has largest adjr2 value

#C-V using model4 and model5
library(caret)
# Define train control for k fold cross validation
set.seed(123)
train_control <- trainControl(method="cv", number=5)
# Fit Model
model <- train(logtrestbps~age+oldpeak+thalach+oldpeak*fbs-fbs, data=df, trControl=train_control, method="lm")
# Summarise Results
print(model)

set.seed(123)
train_control <- trainControl(method="cv", number=5)
model <- train(trestbps~age+oldpeak+thalach+oldpeak*fbs+chol-fbs, data=df, trControl=train_control, method="lm")
print(model)
#model4 has smaller RMSE, model4 is the best model

model <- lm(trestbps~age+oldpeak+thalach+oldpeak*fbs-fbs,data = df)
summary(model)
#every coefficient is significant, and adjusted R-squared is larger than the full
```

```
#model (0.1272 > 0.1246)
```