Heart Attack Model Analysis Report

1.Introduction

Heart attacks are a significant concern for many people, with blood pressure being a critical factor in predicting heart disease. This research focuses on the relationship between resting blood pressure and 11 other variables, including age, ST depression peak, and fasting blood sugar levels, as analyzed using data from the "*Heart Attack Analysis & Prediction dataset*" on Kaggle.

Supporting this research, the *Journal of Hypertension* highlights that blood pressure tends to rise with age. Additionally, a study titled *"Fasting Blood Glucose is Independently Associated with Resting and Exercise Blood Pressures and Development of Elevated Blood Pressure"* found a strong correlation between fasting blood glucose levels and both resting and exercise blood pressure in healthy, non-diabetic, and non-hypertensive men. This relationship persisted over a 7-year follow-up period, underscoring the importance of closely monitoring fasting glucose metabolism in understanding the development of hypertension.

The research may encounter several challenges, including issues with multicollinearity, data quality and accuracy, and the determination of causality. Interrelationships among variables like age, weight, and blood sugar can complicate the analysis of their independent effects on blood pressure. High-quality data is crucial, but clinical data often includes measurement errors, incomplete datasets, or subjective reports, which may affect accuracy and reliability. Additionally, the complex interactions between blood pressure and factors such as lifestyle, genetics, and environment make it difficult to establish causality, with observational studies typically revealing correlations rather than clear cause-and-effect relationships.

To study this problem, we first built a model with 11 predictors. Then, we used tests such as forward and backward selection, AIC, BIC, adjusted R-squared, T-test, and transformations to obtain the final model. There are three significant predictors required to predict blood pressure: age, ST depression peak, and fasting blood sugar levels. The model shows that all three predictors have a positive linear relationship with blood pressure.

This project applies what we learned in STA302. During the process of doing the project, we became familiar with the steps involved in studying a statistical problem and integrated all the knowledge from lecture 1 to11. This involves collecting data, building models, testing those models, and ultimately summarizing the final model to make predictions for the response variable.

2.Method

We tried to find the best multiple linear regression model that predicts trtbps by using age, sex, ca, chol, thalach,oldpeak, cp, rest_ecg, exang, fbs and target.

Response variable(Y):
trtbps : resting blood pressure (in mm Hg)

Predictors(X):
Numerical valuables:
1.Age : Age of the patient
2.Sex : Sex of the patient
3.ca: number of major vessels (0-3)
4.chol : cholestoral in mg/dl fetched via BMI sensor
5.thalach : maximum heart rate achieved
6.oldpeak: the ST depression observed on an electrocardiogram (ECG) during exercise testing(in mm)
7.cp : Chest Pain type chest pain type
    Value 1: typical angina
    Value 2: atypical angina
    Value 3: non-anginal pain
    Value 4: asymptomatic
8.rest_ecg : resting electrocardiographic results
    Value 0: normal
    Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

Dummy variables:
9.exang: exercise induced angina (1 = yes; 0 = no)
10.fbs : (fasting blood sugar if > 120 mg/dl) (1 = true; 0 = false)
11.target : 0= less chance of heart attack 1= more chance of heart attack

First, we split the data into training dataset (75%) and testing dataset (25%) from the collected data. We split data to make sure that our model is accurate and not overfitting. Then we build our first multiple linear regression model with resting blood pressure(trtbps) as y and all other variables as x which is the full model.

After carrying out our full model, we apply both backward and forward AIC as well as BIC to our model. AIC and BIC can help us determine which model fits data better. The lower AIC/BIC is, the better the model is. We used AIC, BIC and adjusted R squared to do the preliminary model selection and reduced some predictors. Since the result model from the AIC test has one more predictor than the BIC test, we used the T test for that additional predictor to check whether it is significant. If the t-test shows p-value less than the significant level of 0.05, we consider the predictor to be significant. After all the tests, we form the model 2.

We used correlation plot and VIF testing to make sure that our best model doesn't have multicollinearity between variables. If VIF is less than 5 and correlation is less than 0.2,

we consider two variables to have no obvious correlation. Then we check the 4 assumptions of model 2, since we find out the QQ plot shows the errors not following normal distribution. We applied log transformation to our model 2 to get our final model, so that the Normal QQ plot gets closer to the diagonal line, which also makes it easier for us to conclude outliers and leverage points. At last we conclude our final model.

Finally, we use test data from the collected data to fit the model again and compare it with our final model to do the model validation.

3.Results

Full model:
Response(Y) variable: trtbps
Predictors(X): age, sex, ca, chol, thalach,oldpeak, cp, rest_ecg, exang, fbs and target.

```
Call:
lm(formula = trtbps ~ age + sex + cp + chol + fbs + restecg +
    thalachh + oldpeak + exng, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-34.292 -10.485  -2.124  10.297  61.551

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.601703  15.536666   5.831 2.11e-08 ***
age          0.511189   0.152645   3.349 0.000966 ***
sex         -2.610392   2.651085  -0.985 0.325957
cp           0.467549   1.257889   0.372 0.710505
chol         0.008373   0.022729   0.368 0.712963
fbs          6.231011   3.339515   1.866 0.063492 .
restecg     -1.455935   2.229069  -0.653 0.514386
thalachh     0.061622   0.062567   0.985 0.325836
oldpeak      2.751817   1.033724   2.662 0.008383 **
exng         1.558807   3.000988   0.519 0.604020
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.04 on 205 degrees of freedom
Multiple R-squared:  0.1362, Adjusted R-squared:  0.09829
F-statistic: 3.592 on 9 and 205 DF,  p-value: 0.0003575
```

From the backward and forward AIC testing, the model with lowest AIC 1832.41 is trtbps ~ ***age + oldpeak + fbs***.(see table 1). Although BIC testing suggests a different model,(with one less variable) the model with lowest BIC 1847.702 is ***trtbps ~ age + oldpeak.*** (see table 2).

We second compared the adjusted R-squared values for these two models. The model: ***trtbps*** ~ ***age + oldpeak + fbs*** yielded an adjusted R-squared of 0.1102, which is higher than the adjusted R-squared of 0.09862 for the model: trtbps~age+oldpeak (see table 3).

Finally we use the T test for predictor fbs, since the p-value is 0.0239 which is less than 0.05, we can reject that the predictor fbs is not significant. So we retained the predictor fbs.(see table 4)

(Table 1)

```
AIC(lm(`trtbps` ~ `age`+`fbs`+`oldpeak`, data=train_data))
```

```
## [1] 1832.41
```

(Table 2)

```
BIC(lm(`trtbps` ~ `age`+`oldpeak`, data=train_data))
```

```
## [1] 1847.702
```

(Table3)

```
Call:
lm(formula = trtbps ~ +oldpeak + age + fbs, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-32.804 -10.671  -1.629   9.729  63.820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.8533     7.2700  14.010  < 2e-16 ***
oldpeak       2.4762     0.9715   2.549 0.011520 *
age           0.4875     0.1341   3.636 0.000348 ***
fbs           6.3364     3.2629   1.942 0.053472 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.92 on 211 degrees of freedom
Multiple R-squared:  0.1227, Adjusted R-squared:  0.1102
F-statistic: 9.839 on 3 and 211 DF,  p-value: 4.216e-06
```

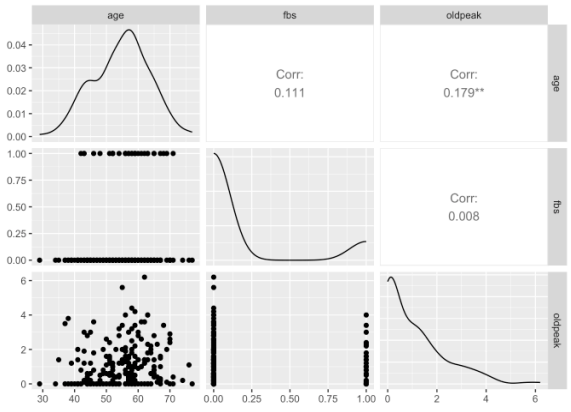
```
Call:
lm(formula = trtbps ~ age + oldpeak, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-33.648 -10.909  -0.797   9.596  62.872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.2338     7.3103  13.848  < 2e-16 ***
age           0.5166     0.1341   3.852 0.000155 ***
oldpeak       2.4532     0.9778   2.509 0.012858 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.03 on 212 degrees of freedom
Multiple R-squared:  0.107,  Adjusted R-squared:  0.09862
F-statistic: 12.71 on 2 and 212 DF,  p-value: 6.138e-06
```

(Table4)

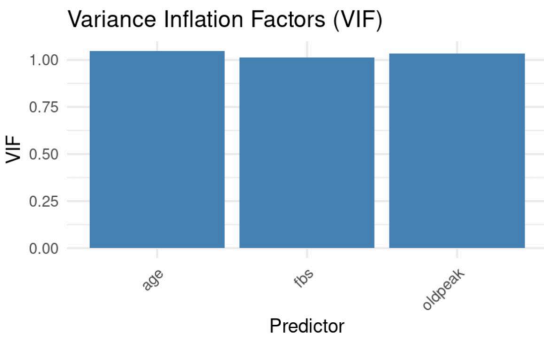
```
Call:
lm(formula = trtbps ~ +fbs, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-37.656 -10.913  -0.913   9.087  61.344

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  130.913      1.313  99.674   <2e-16 ***
fbs            7.744      3.404   2.275   0.0239 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.77 on 213 degrees of freedom
Multiple R-squared:  0.02371,    Adjusted R-squared:  0.01913
F-statistic: 5.174 on 1 and 213 DF,  p-value: 0.02393
```

In the correlation plot we found there is some correlation between the age and oldpeak which is 0.179.(see table 5). The results from the VIF statistic show that there is very low multicollinearity between the covariates, which is below the level of concern.(see table 6).

(Table 5)



(Table6)

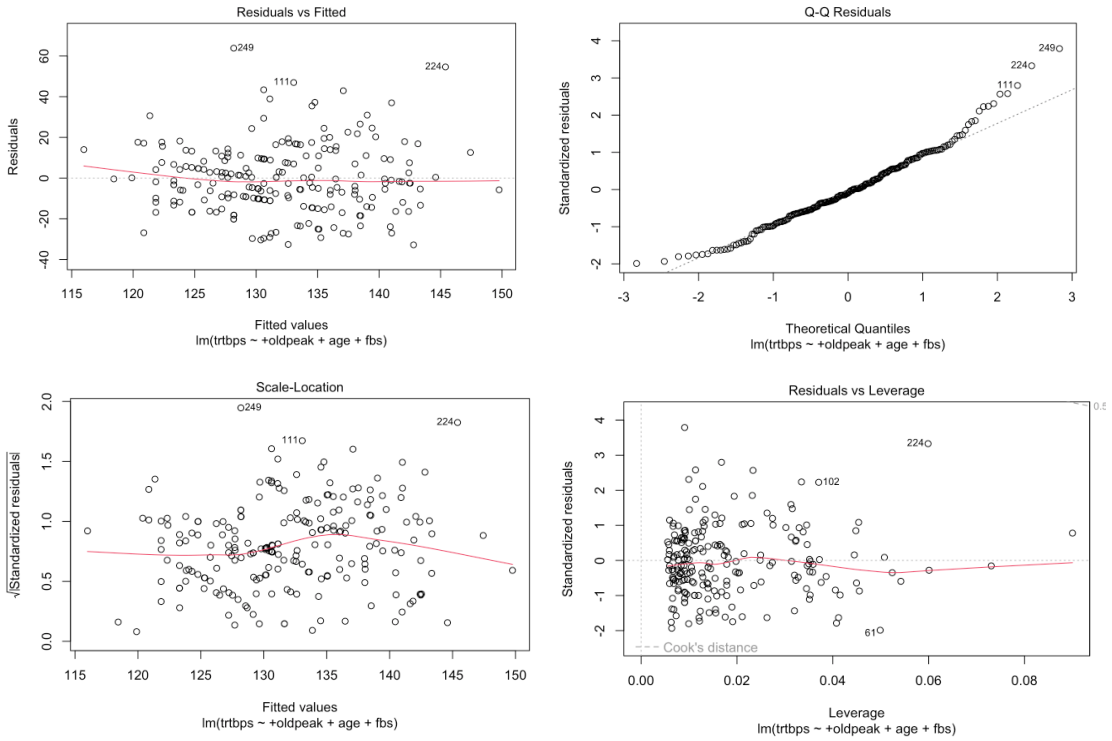


To check the 4 assumption of model: ***trtbps ~ age + oldpeak + fbs***

From the result, the other three plots are randomly distributed but the normal QQ plot shows the data are a little bit not normal.(see figure 7). So we add the log transformation to the response variable. After the transformation most data are on the straight diagonal line except few on both tails, proving the assumptions of normality.(see figure 8)

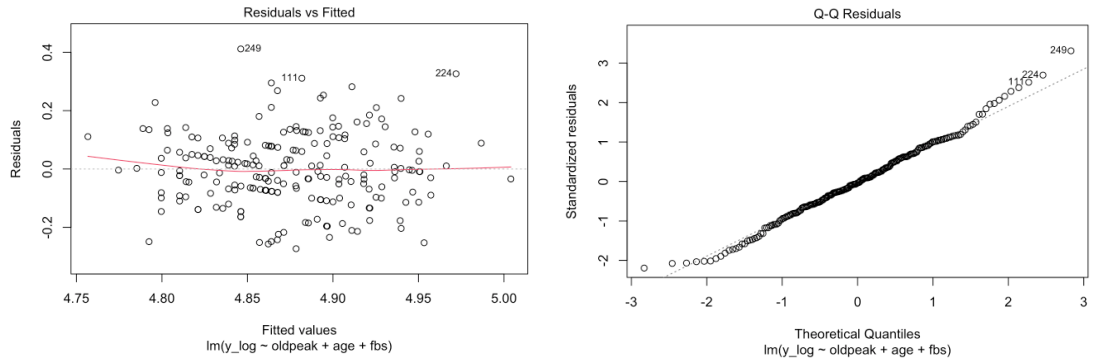
(Figure 7)

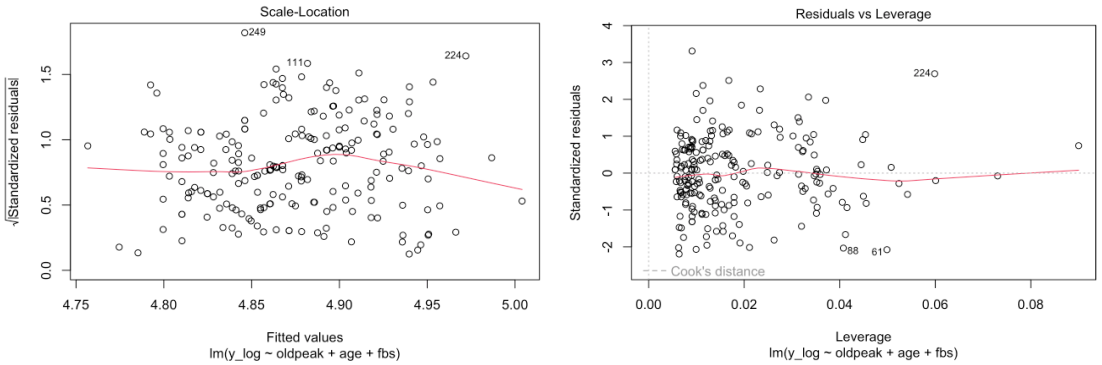
For trtbps~age+oldpeak+fbs



(Figure 8)

For log(trtbps) ~ age + oldpeak + fbs





We checked the outliers (does not have standardized residual between [-2,2] and leverage points (checking fitted value that is larger than the testing formula). As result, outliers are : {9th, 47th, 51th, 61th, 67th, 68th, 83th 88th, 90th, 111th, 125th, 156th, 160th, 169th, 174th, 185th, 192th, 224th, 229th, 242th, 249th, 267th, 274th}, (see table 9)

leverage points : {1th, 2th, 47th, 54th, 58th, 61th, 67th, 69th, 73th, 77th, 79th, 88th, 91th, 97th, 104th, 107th, 137th, 138th, 141th, 149th, 154th, 156th, 167th, 176th, 180th, 205th, 206th, 213th, 216th, 222th, 224th, 239th, 252th, 260th, 292th, 301th}. (see table 10)

(Table 9)

```
#outliers
studentized_residuals<-rstudent(log_mod)
outliers<-which(abs(studentized_residuals)>2)
print(outliers)
```

```
  9  61  67  88  90 111 125 224 229 242 249 267 274
  9  47  51  67  68  83  90 156 160 169 174 185 192
```

(Table 10)

```
#leverage points
hat_values<- hatvalues(log_mod)
leverage_points<- which(hat_values>2*(length(coef(log_mod))/nrow(data)))
print(leverage_points)
```

```
##   1   2  61  73  77  88  91 104 107 138 198 205 216 222 224 239 252 260 292 301
##   1   2  47  54  58  67  69  77  79  97 137 141 149 154 156 167 176 180 206 213
```

We fit the same model using the testing dataset and obtain estimated coefficients of the final model as : β_0=4.700968, β_1=0.009253, β_2=0.002678, and β_3=0.068339. (see table 11) The difference in beta coefficients between the training model and testing model is minimal, and overall, they are acceptable.

(Table 11)

```
Call:
lm(formula = y_log2 ~ oldpeak + age + fbs, data = test_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.294228 -0.080456 -0.008828  0.088054  0.229618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.700968   0.073394  64.051   <2e-16 ***
oldpeak     0.009253   0.013060   0.709   0.4806
age         0.002678   0.001410   1.899   0.0610 .
fbs         0.068339   0.036647   1.865   0.0657 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1206 on 84 degrees of freedom
Multiple R-squared:  0.1074,    Adjusted R-squared:  0.07552
F-statistic: 3.369 on 3 and 84 DF,  p-value: 0.02229
```

So the final model we built is a multiple linear regression model with response variable is log(trtbps) and predictors x_1: age, x_2: oldpeak, and x_3: fbs. (see Data Summary Table)

Data Summary Table

	intercept	oldpeak	age	fbs
Estimate	4.6529997	0.0180709	0.0035752	0.0464656
Std.Error	0.0536494	0.0071694	0.0009895	0.0240788
p-Value	< 2e-16	0.012457	0.000378	0.054981
VIF		1.033389	1.046242	1.012654

Predictors	Adjusted R^2	AIC	Outliers	Leverage points
3	0.1086	-278.4745	13	20

4.Conclusion

The final model ***log(trtbps)*** ~ ***age + oldpeak + fbs*** indicates that if we fix age and fasting blood sugar unchanged, the ST depression observed on an electrocardiogram (ECG) during exercise testing increased by 1mm, the blood pressure will increase by 1.8% on average.If we fix oldpeak and fasting blood sugar unchanged, the age increases by 10 years old, the blood pressure will increase by 3.6% on average.If we fix age and oldpeak unchanged, if the person has fasting blood sugar greater than 120 mg/dl, the blood pressure will increase 4.64% more than the person who has fasting blood sugar less than 120 mg/dl. Since the age and oldpeak are not likely to be 0, the intercept 4.65 has no meaning.

This research also has some limitations. First, we have some multicollinearity, where the independent variables, age and oldpeak, are weakly correlated with each other. This could make it difficult to isolate the unique effect of each variable on blood pressure. Second, the study relies on clinical data, which may contain measurement errors, incomplete datasets, or subjectively reported variables. These issues could affect the accuracy and reliability of the model's predictions. Finally, the study uses an observational design, which limits the ability to draw causal inferences. The relationships identified in the model may reflect correlations rather than direct cause-and-effect relationships.

For improvement, adding a penalty to the regression might address some multicollinearity. To improve data collection methods and minimize errors and missing values, we can implement rigorous data cleaning and validation processes to ensure the accuracy and reliability of the dataset. We can also consider using more advanced statistical techniques or experimental designs, such as instrumental variables or randomized controlled trials, to better establish causal relationships between the variables.

5.Acknowledgement

Name	Contribution
Ying*** (ID Removed)	Method and conclusion parts, R code for full model and final model, assumption check.
Si*** (ID Removed)	Reference and choose topics, part of introduction,leverage points and outlier, log transformation
Qi*** (ID Removed)	Finding data related to the topic, constructing the full model, codings and excel work for results visualization.
Shao*** (ID Removed)	R code for testing methods and explanations of them in method and results.
Zhaopeng Zhang (Pierre Zhang, ID Removed)	Basic codes for models. Constructed ‘Introduction’ part. Produced visualized tables and finished the reference part.

6.Reference

Data Base:
https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Rahman, R. (Ed.). (2021, March 22). *Heart attack analysis & prediction dataset*. Kaggle. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data

Bjørnholt, J. V., Erikssen, G., Kjeldsen, S. E., Boård, J., Thaulow, E., & Erikssen, J. (2003). Fasting blood glucose is independently associated with resting and exercise blood pressures and development of elevated blood pressure. *Journal of hypertension*, *21*(7), 1383–1389. https://doi.org/10.1097/00004872-200307000-00029

Association of age and blood pressure among 3.3 million adults: insights from China PEACE million persons project. Journal of Hypertension 39(6):p 1143-1154, June 2021. | https://doi.org/10.1097/HJH.0000000000002793