501 Project: Linear Regression on Stroke Prediction Dataset

Shruti Ramesh, Maanusri Balasubramanian, Spurthi T, Pavan Datta, Haoyuan Ren

2022-08-24

Contents

0.1 Introduction

According to the World Health Organization (WHO), stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. We have used the data set "Stroke Prediction Dataset" which is available on Kaggle. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about a patient. As part of the 501 STATS course project, our goal is focused on predicting the bmi (Body Mass Index) using average glucose level and age of a patient. And also checking the dependence between some of the categorical variables like gender, residence type and stroke.

0.2 Data

```
data <- read.csv('healthcare-dataset-stroke-data.csv')
```

Attribute Information

1) id (int, categorical): unique identifier
2) gender (str, categorical): "Male", "Female" or "Other"
3) age (int, numerical): age of the patient
4) hypertension (int, categorical): 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease (int, categorical): 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married (str, categorical): “No” or “Yes”
7) work_type (str, categorical): “children”, “Govt_jov”, “Never_worked”, “Private” or “Self-employed”
8) Residence_type (str, categorical): “Rural” or “Urban”
9) avg_glucose_level (int, numerical): average glucose level in blood
10) bmi (str, numerical): body mass index*
11) smoking_status (str, categorical): “formerly smoked”, “never smoked”, “smokes” or “Unknown”*
12) stroke (int, categorical): 1 if the patient had a stroke or 0 if not

*Note: “Unknown, NA” in smoking_status and bmi means that the information is unavailable for this patient

Preprocessing:

1. There are 11 columns/features/variables in the dataset.
2. The types of each of the variables are as follows -
   a) Gender, Ever_married, Work_type, Residence_type, smoking_status - categorical
   b) Hypertension, Heart_disease, stroke_label - boolean (0 or 1), categorical
   c) avg_glucose_level, bmi - quantitative (continuous)
   d) age - quantitative (discrete)
3. We cleaned the data set to remove NULL and Other values, and dropped ‘id’ column which are indices.
4. We converted features into their respective categorical values, and numerical values as they are read as characters from CSV

Convert bmi value from string to numeric

```
data$bmi <- as.numeric(data$bmi)
```

```
## Warning: NAs introduced by coercion
```

Omit na

```
data <- na.omit(data)
```

Filter out targeted data

```
target_data <- data %>% filter(stroke == 1)
```

```
nrow(target_data)
```

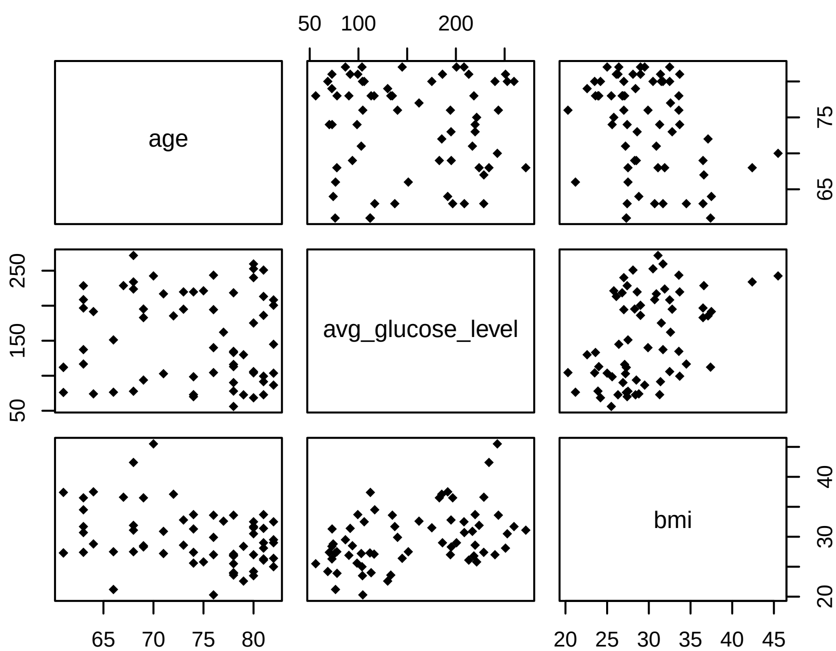
```
## [1] 209
```

```
names(target_data)
```

```
##  [1] "id"                "gender"            "age"
##  [4] "hypertension"      "heart_disease"     "ever_married"
##  [7] "work_type"         "Residence_type"    "avg_glucose_level"
## [10] "bmi"               "smoking_status"    "stroke"
```

Motivation to check the Linear Relationship strength between the following columns {age, bmi, avg_glucose_level}

```
d = target_data %>% filter(gender == 'Male' & age > 60)
d$stroke_label <- NULL
d$hypertension <- NULL
d$ever_married <- NULL
d$work_type <- NULL
d$heart_disease <- NULL
d$Residence_type <- NULL
d$smoking_status <- NULL
d$gender <- NULL
d$id <- NULL
d$stroke <- NULL
sp = pairs(d, pch=18)
```



Plot scatter graph on log bmi vs average glucose level and age [6]

0.3 Analysis

Data Analysis

The age forms two different groups on different stroke types: for those who have not had a stroke, the median of age is 43, and for those have had a stroke, the median is 70. That's the reason why we choose to only consider people who have had a stroke.

```
favstats((data %>% filter(stroke == 0))$age)
```

```
##   min Q1 median Q3 max     mean       sd    n missing
##  0.08 24     43 59  82 41.76045 22.26813 4700       0
```

```
favstats((data %>% filter(stroke == 1))$age)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##   14 58     70 78  82 67.71292 12.40285 209       0
```

Histogram shows bmi is right skewed [1]

Use log on bmi to reduce skewness [2]

```
target_data$log_bmi <- log(target_data$bmi + 1)
```

Histogram shows age is left skewed [3]

Use boxcox to reduce skewness [4]

```
target_data$bc_age <- bcPower(target_data$age, powerTransform(target_data$age)$lambda)
```

Plot the box plot of the three variables [5], there are no outliers.

0.4 Linear Regression Modeling

Fit Data

```
lr_model <- lm(log_bmi ~ avg_glucose_level + bc_age, data=target_data)
summary(lr_model)
```

```
##
## Call:
## lm(formula = log_bmi ~ avg_glucose_level + bc_age, data = target_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52430 -0.11590 -0.01195  0.11914  0.56059
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.443e+00  3.731e-02  92.298  < 2e-16 ***
## avg_glucose_level  1.148e-03  1.884e-04   6.094 5.35e-09 ***
## bc_age            -2.622e-06  4.258e-07  -6.157 3.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1689 on 206 degrees of freedom
## Multiple R-squared:  0.2497, Adjusted R-squared:  0.2424
## F-statistic: 34.28 on 2 and 206 DF,  p-value: 1.412e-13
```

1. The slope for avg_glucose_level is 1.148e-03. This means that with every unit increase in avg_glucose_level, log_bmi on average increases by 1.148e-03 units

2. The slope for bc_age is -2.622e-06. This means that with every unit increase in bc_age, log_bmi on average decreases by 2.622e-06 units
3. The value of R2 is 0.2497 (from the model summary). This means that 24.97% of the variation in log_bmi is explained by the fitted linear model using glucose_level, bc_age

Assumptions

Linearity : The red line in Residual vs Fitted graph is horizontal at zero point, therefore the assumption is satisfied. [7]

Normality : The points in Norm Q-Q graph fall approximately along the reference line of the Q-Q plot, therefore the assumption is satisfied. [8]

Homoscedasticity : The points of Scale Location graph are equally distributed, while the red line is approximately horizontal, therefore the assumption is satisfied. [9]

Independence : Because the samples are randomly collected from representative groups of people, independence is assumed. Also, the sum of residuals is $-1.19e-15 \approx 0$, which also indicates independence.

```
sum(lr_model$residuals)
```

```
## [1] -1.554312e-15
```

Remove Outliers

```
hv <- hatvalues(lr_model)
sort(hv, decreasing = TRUE)[1:2]
```

```
##        207         95
## 0.03503033 0.03016230
```

$\hat{y} = 2(p+1)/n = 4/209 = 0.01913$, remove 207 and 150 to prevent high leverage

```
cook <- cooks.distance(lr_model)
sort(cook, decreasing = TRUE)[1:3]
```

```
##        145         95         38
## 0.06927845 0.06025519 0.05477314
```

$cook = 4/(n-p-1) = 4/62 = 0.06452$, remove 95, 145, 38 to prevent high influential

```
new_target_data <- target_data %>% filter(!row_number() %in% c(207, 150, 95, 145, 38))
new_lr_model <- lm(log_bmi ~ avg_glucose_level + bc_age, data=new_target_data)
summary(new_lr_model)
```

```
##
## Call:
## lm(formula = log_bmi ~ avg_glucose_level + bc_age, data = new_target_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40351 -0.11679 -0.01126  0.11898  0.39935
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.444e+00  3.633e-02  94.786  < 2e-16 ***
## avg_glucose_level  1.200e-03  1.812e-04   6.624 3.13e-10 ***
## bc_age            -2.756e-06  4.169e-07  -6.611 3.37e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1595 on 201 degrees of freedom
## Multiple R-squared:  0.2839, Adjusted R-squared:  0.2767
## F-statistic: 39.84 on 2 and 201 DF,  p-value: 2.674e-15
```

The assumptions of the new model are satisfied from the graphs, and there are no significant outliers. [10]

The graph also shows strong relation between the independent and dependent variables. [11]

1. The slope for avg_glucose_level is 1.200e-03. This means that with every unit increase in avg_glucose_level, log_bmi on average increases by 1.200e-03 units
2. The slope for bc_age is -2.756e-06. This means that with every unit increase in bc_age, log_bmi on average decreases by 2.756e-06 units
3. The value of R2 is 0.2839 (from the model summary). This means that 28.39% of the variation in log_bmi is explained by the fitted linear model using glucose_level, bc_age

0.5 Conclusion

The final regression model has $p - value = 2.674e - 15$ for F-test, and $p - value < 4e - 10$ for all three coefficients, suggesting that there is a significant correlation between the independent variables and dependent variables, and the model is significant on predicting the bmi. The $R^2 = 0.2767$, suggests the model describes 27.7 of the correlation between the dependent and independents.

During the data analysis, we also noticed that there are two clusters on average blood glucose level [12], one is around 180, another one is around 200. This phenomena cannot be fully described by the data set. It might be caused by a general 'healthy status', because the fraction of the second cluster (high blood glucose level) increases as the person has heart disease [13], the person is old (age > 60) [14], or the person is under hypertension [15]. It also shows why the R^2 is low for this model. For us, it would be better to collect more data from analysis of blood samples which can be a better reference of health status, and make questionnaires asking the sport time per week, sleep time, sleep quality, meal type, etc.

0.6 Chi-Square Test

The chi-square test helps us evaluate whether there is an association between two categorical variables.

Assumptions for each of the following tests:

1. Counted Data Condition: Data is the counts for the categories of a categorical variable - Condition met
2. Independence Assumption: The counts in the cells must be independent of each other - Yes, they are independent, Condition met
3. Randomization Condition: We should have a random sample - Yes, this is a random sample, Condition met
4. Counts in individual cells should be at least 5. This is met for the below two contingency tables

```r
data <- read.csv("healthcare-dataset-stroke-data.csv")
dataset <- data %>% filter(gender != 'Other')
dataset <- dataset %>% filter(bmi != 'N/A')
dataset$bmi <- as.double(dataset$bmi)
dataset$id <- NULL
dataset <- dataset %>% mutate(gender = factor(gender),
                              hypertension = factor(hypertension),
                              heart_disease = factor(heart_disease),
                              ever_married = factor(ever_married),
                              work_type = factor(work_type),
                              Residence_type = factor(Residence_type),
                              smoking_status = factor(smoking_status),)
dataset$stroke_label <- as.factor(dataset$stroke)
dataset$log_bmi <- log(dataset$bmi)
dataset$stroke <- NULL

summary(dataset)
```

```
##     gender          age         hypertension heart_disease ever_married
##  Female:2897   Min.   : 0.08   0:4457       0:4665        No :1704    
##  Male  :2011   1st Qu.:25.00   1: 451       1: 243        Yes:3204    
##                Median :44.00                                          
##                Mean   :42.87                                          
##                3rd Qu.:60.00                                          
##                Max.   :82.00                                          
##          work_type    Residence_type avg_glucose_level      bmi       
##  children     : 671   Rural:2418     Min.   : 55.12    Min.   :10.30  
##  Govt_job     : 630   Urban:2490     1st Qu.: 77.07    1st Qu.:23.50  
##  Never_worked :  22                  Median : 91.68    Median :28.10  
##  Private      :2810                  Mean   :105.30    Mean   :28.89  
##  Self-employed: 775                  3rd Qu.:113.50    3rd Qu.:33.10  
##                                      Max.   :271.74    Max.   :97.60  
##          smoking_status stroke_label    log_bmi     
##  formerly smoked: 836   0:4699       Min.   :2.332  
##  never smoked   :1852   1: 209       1st Qu.:3.157  
##  smokes         : 737                Median :3.336  
##  Unknown        :1483                Mean   :3.328  
##                                      3rd Qu.:3.500  
##                                      Max.   :4.581
```

0.6.1 Residence_type vs stroke (Chi-square test)

0.6.2 Hypothesis: The place of residence of an individual(Urban/Rural) doesn't have any impact on them having a stroke at .05 significance level i.e Residence_type and stroke are independent variables.

Contingency Table & Chi-square statistic:

```r
# Residence vs Stroke
residence_stroke_data = table(dataset$Residence_type, dataset$stroke)
print(residence_stroke_data)
```

```
##
##           0    1
##   Rural 2318  100
##   Urban 2381  109
```

```
print(chisq.test(residence_stroke_data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  residence_stroke_data
## X-squared = 0.12169, df = 1, p-value = 0.7272
```

- As the p-value 0.7272 is greater than the 0.05, we don't reject the hypothesis and conclude that the place of residence of an individual doesn't have any impact on them having a stroke, and that Residence_type and stroke are independent variables. Hence we infer that there is a weak or no correlation between these two variables, Residence_type and stroke.

0.6.3 gender vs stroke (Chi-square test)

0.6.4 Hypothesis: The gender of an individual doesn't have any impact on them having a stroke at .05 significance level i.e gender and stroke are independent variables.

Contigency Table & Chi-square statistic:

```
# Gender vs Stroke
gender_stroke_data = table(dataset$gender, dataset$stroke)
print(gender_stroke_data)
```

```
##
##            0    1
##   Female 2777  120
##   Male   1922   89
```

```
print(chisq.test(gender_stroke_data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_stroke_data
## X-squared = 0.16955, df = 1, p-value = 0.6805
```

As the p-value 0.6805 is greater than the 0.05, we don't reject the claim and conclude that the gender of an individual doesn't have any impact on them having a stroke and that gender and stroke are independent variables. Hence we infer that there is a weak or no correlation between these two variables, gender and stroke.

0.7 Graphs and Plots

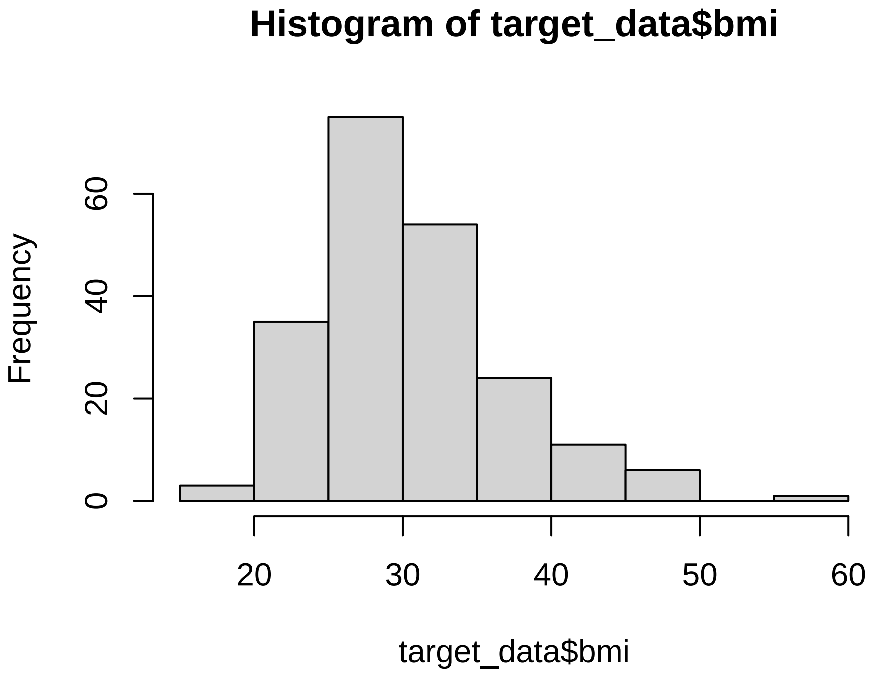


Figure 1: Histogram of BMI

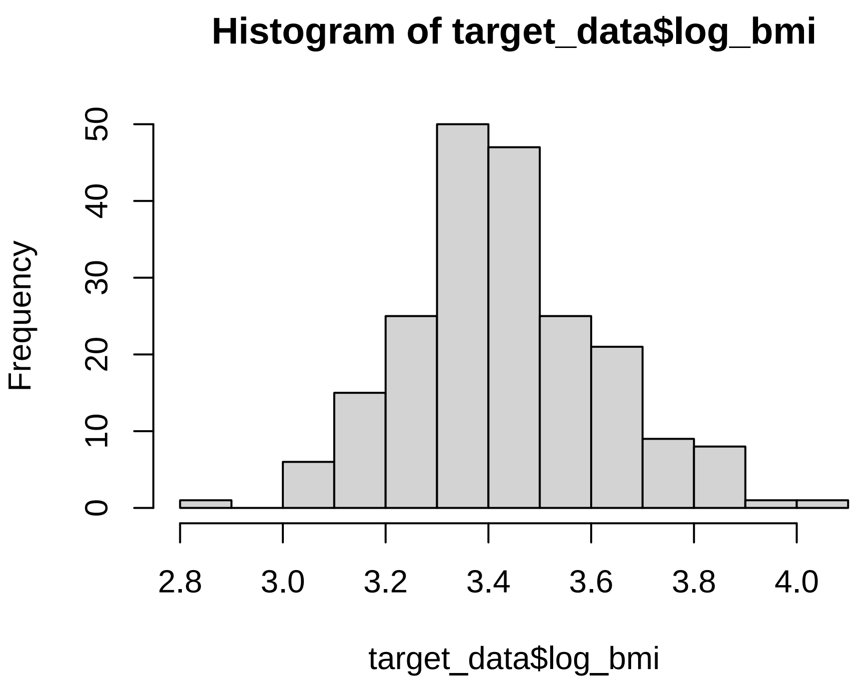


Figure 2: Histogram of Log BMI

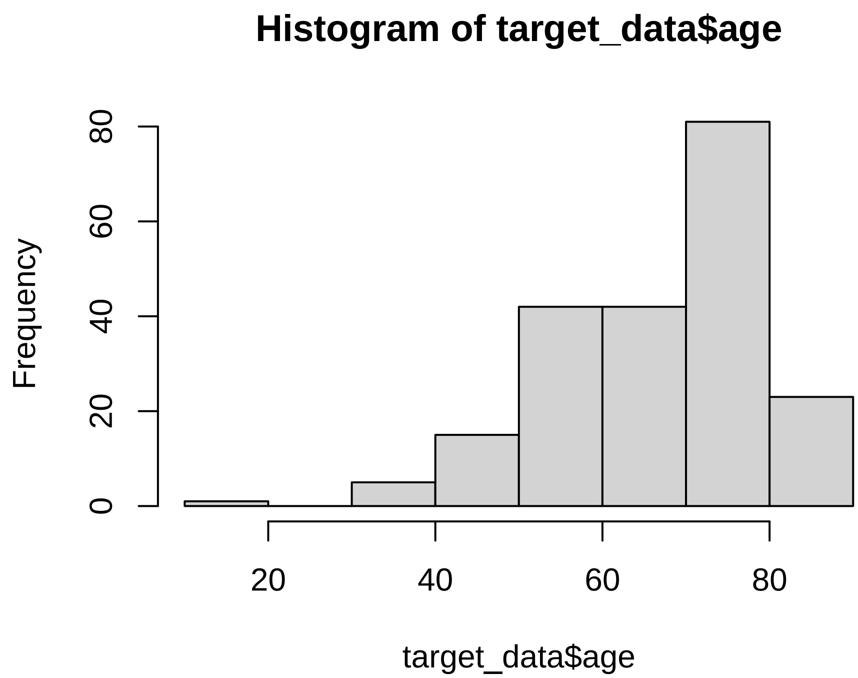


Figure 3: Histogram of Age

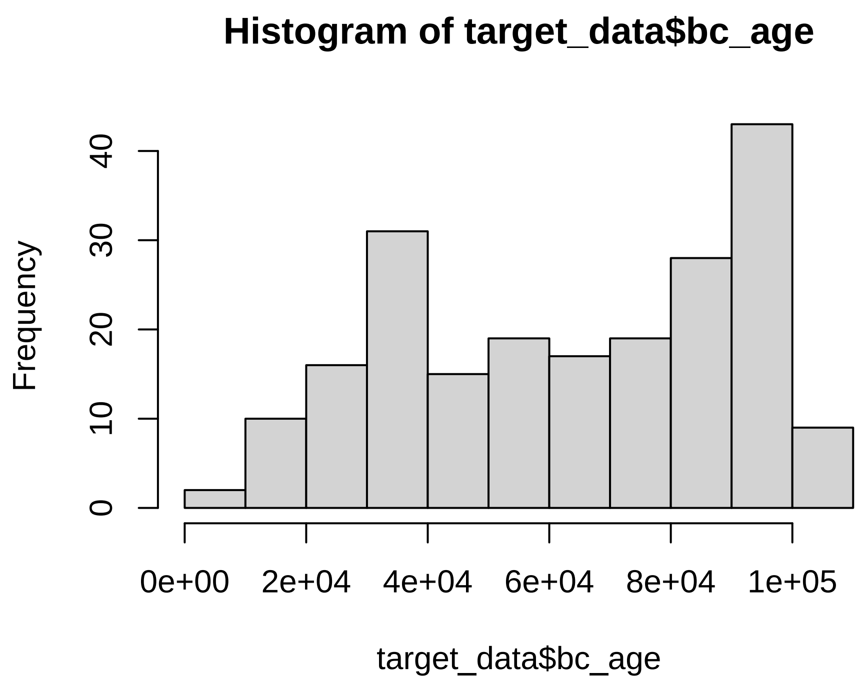


Figure 4: Histogram of BoxCox Age

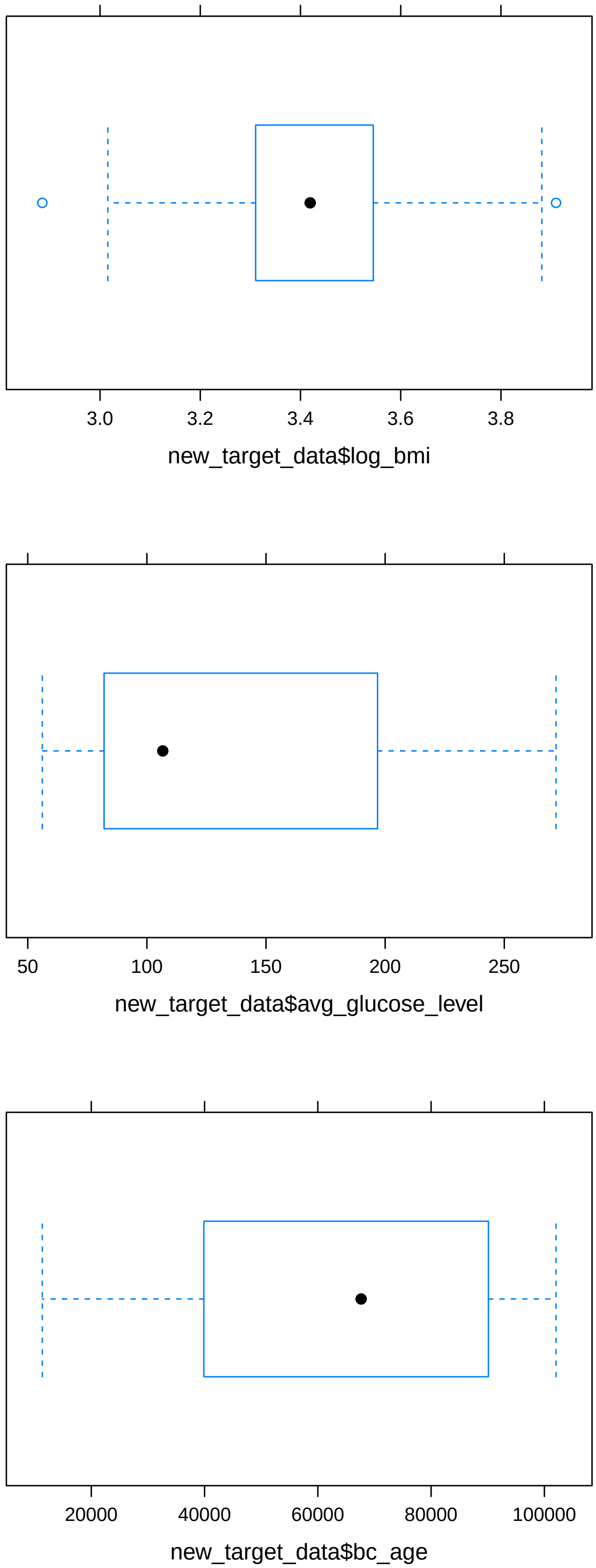


Figure 5: Box plots

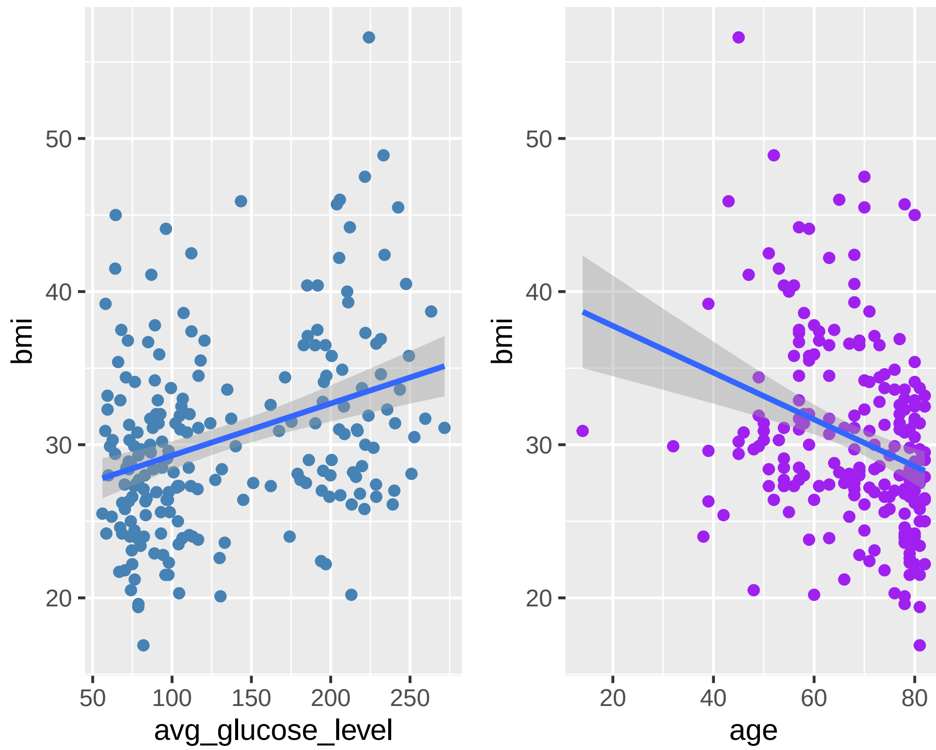


Figure 6: BMI vs Avg Glucose Level and Age

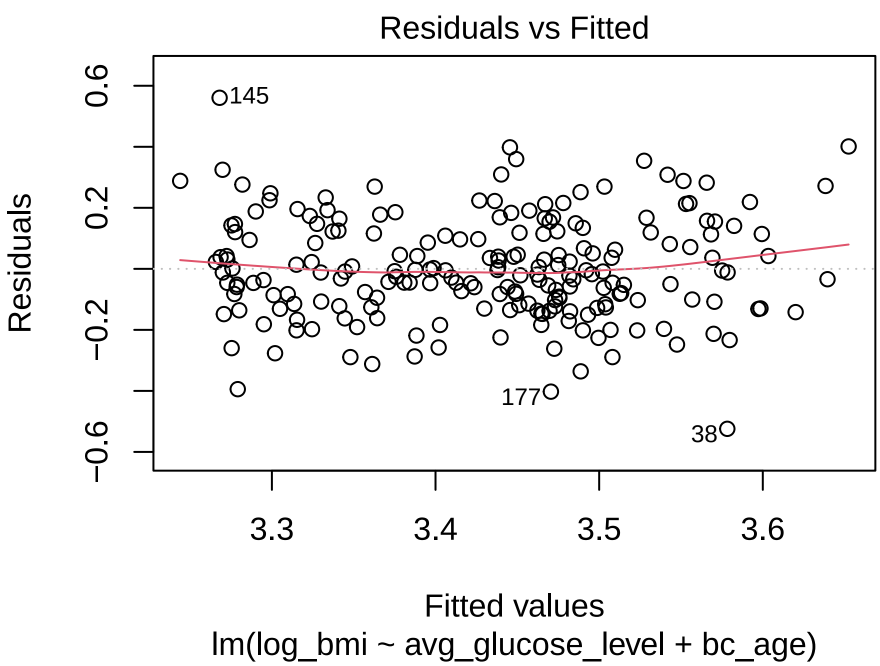


Figure 7: Residual vs Fitted

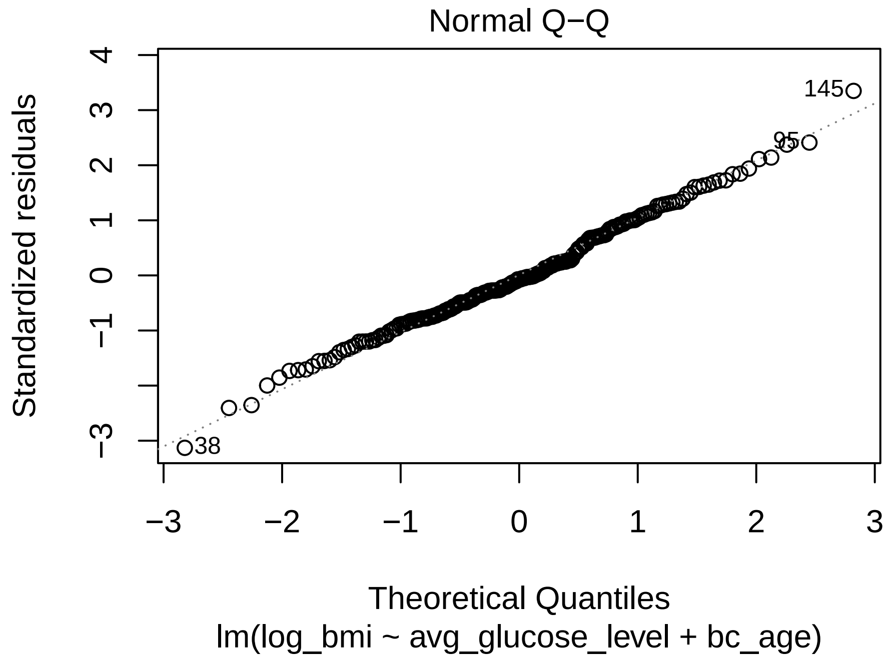


Figure 8: Norm Q-Q

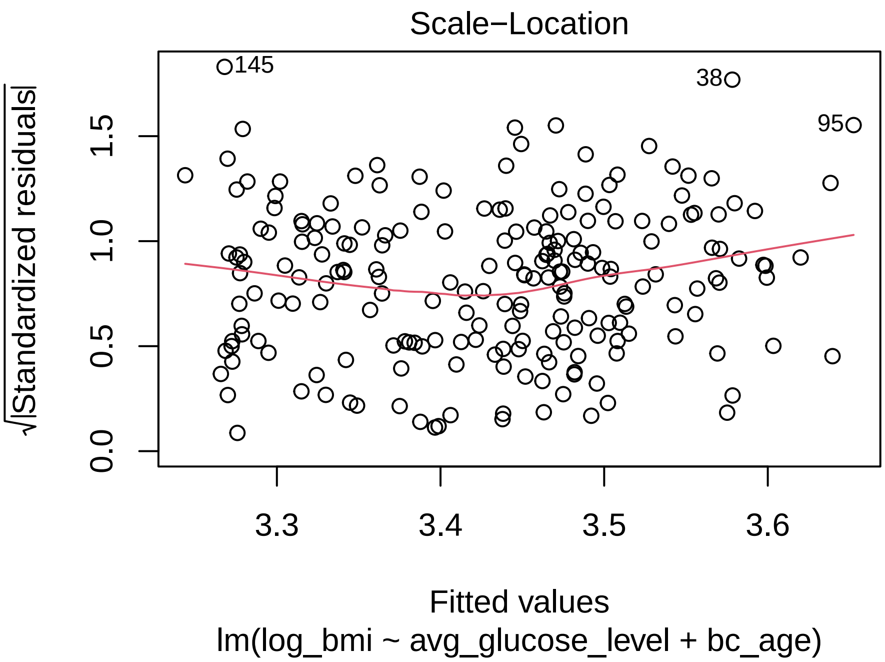


Figure 9: Scale Location

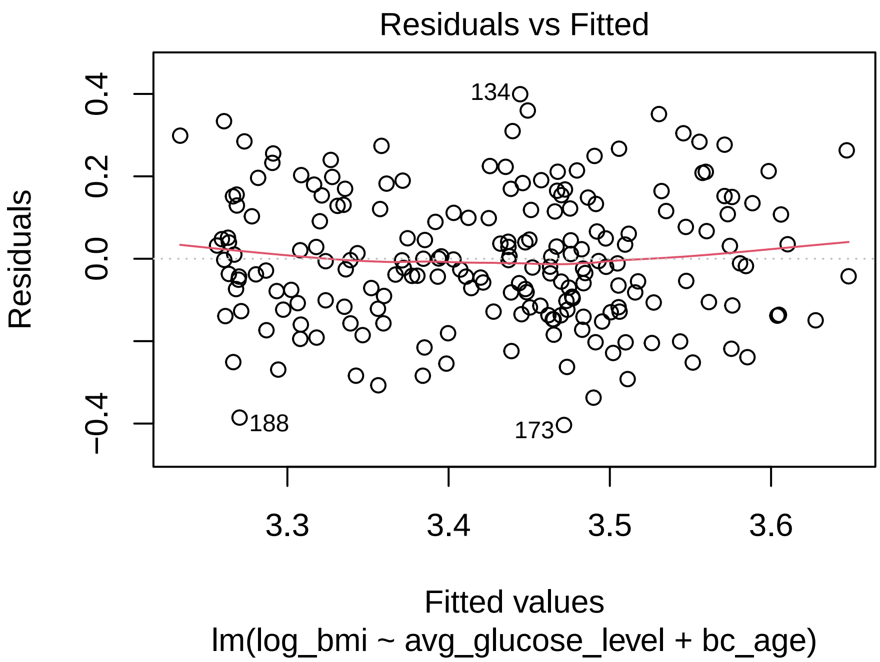
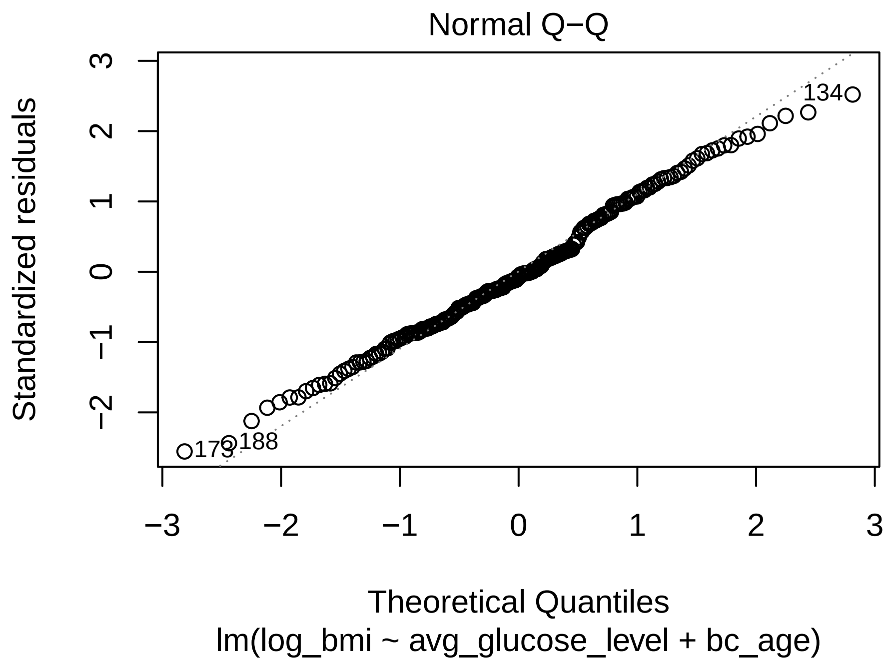
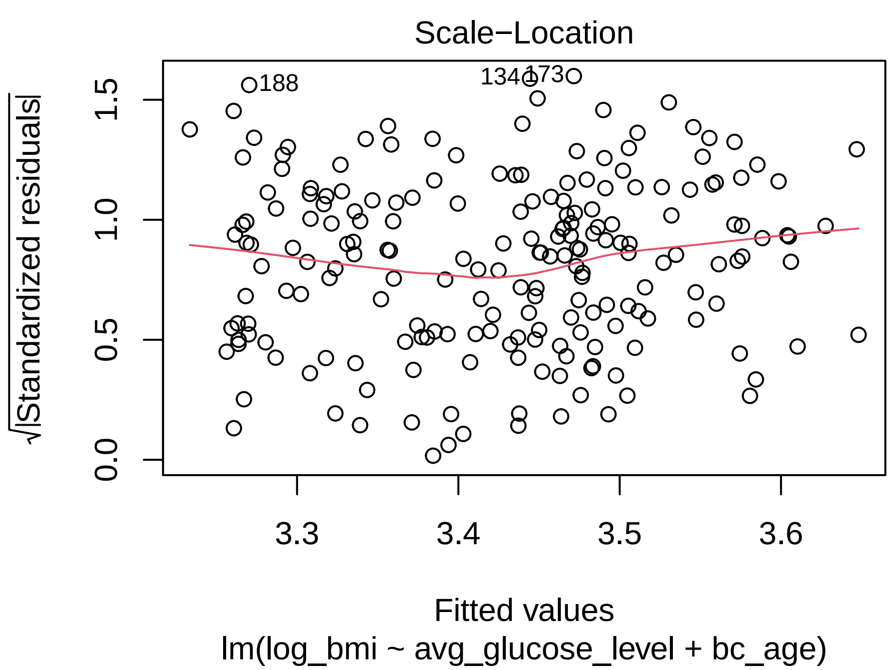


Figure 10: Lm Assumption Graphs

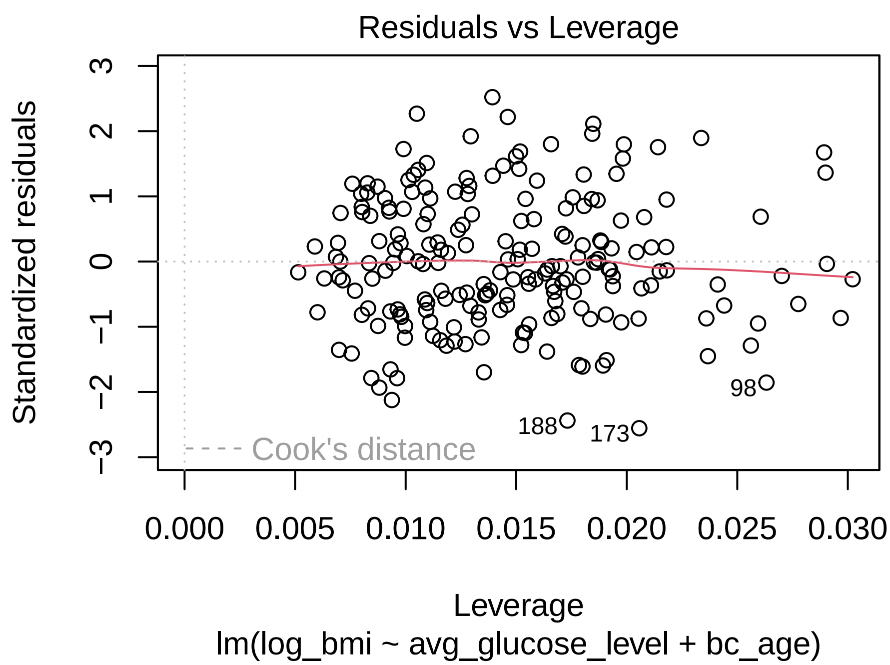
```
plot(new_lr_model, 2)
```




```
plot(new_lr_model, 3)
```



```r
plot(new_lr_model, 5)
```



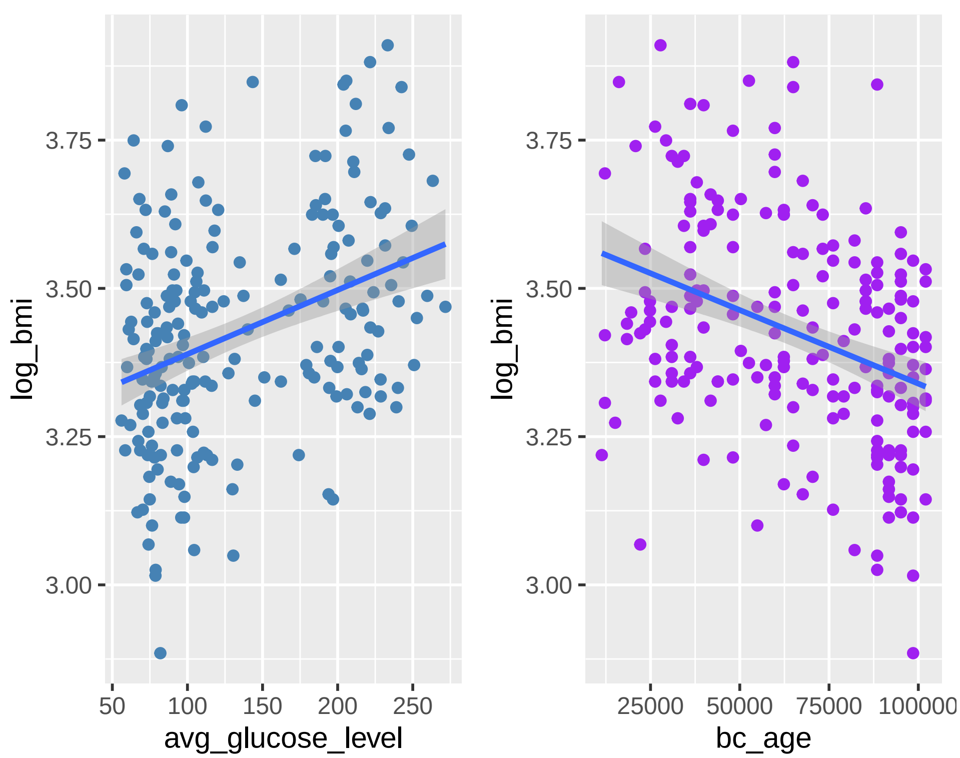


Figure 11: Lm Regression Lines

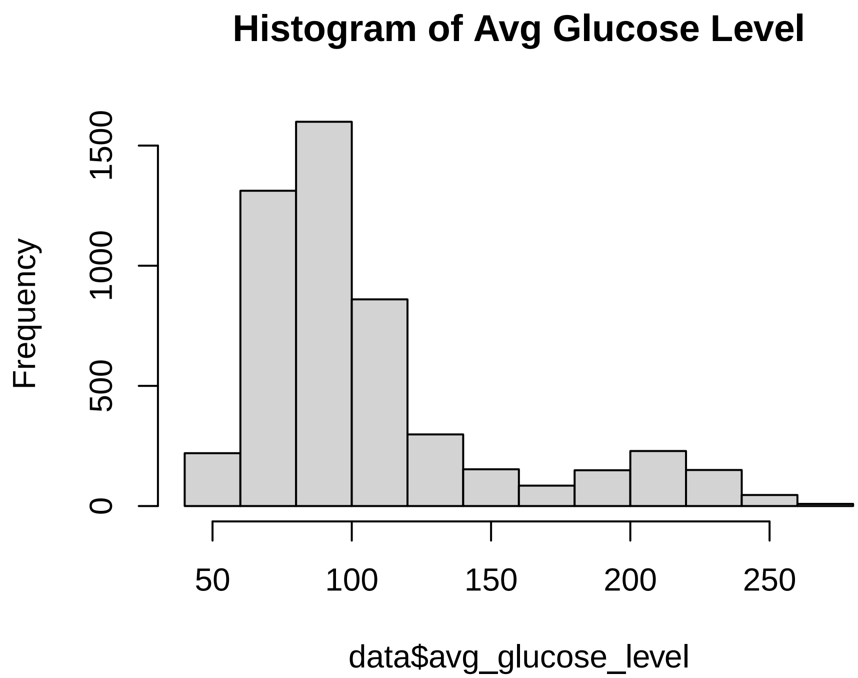


Figure 12: Histogram of Avg Glucose Level

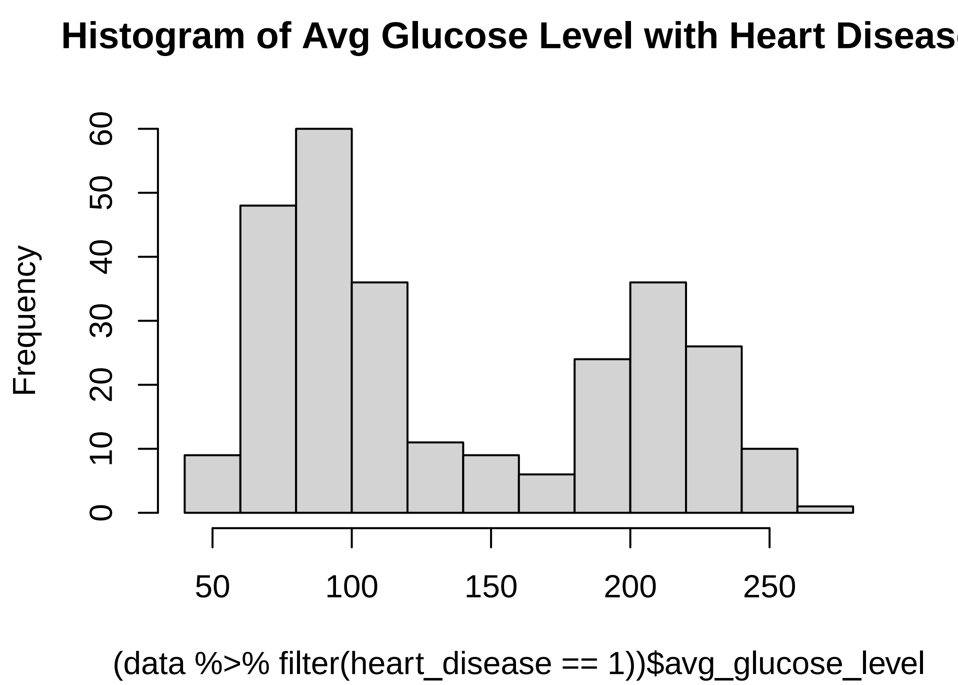


Figure 13: Histogram of Avg Glucose Level with Heart Disease

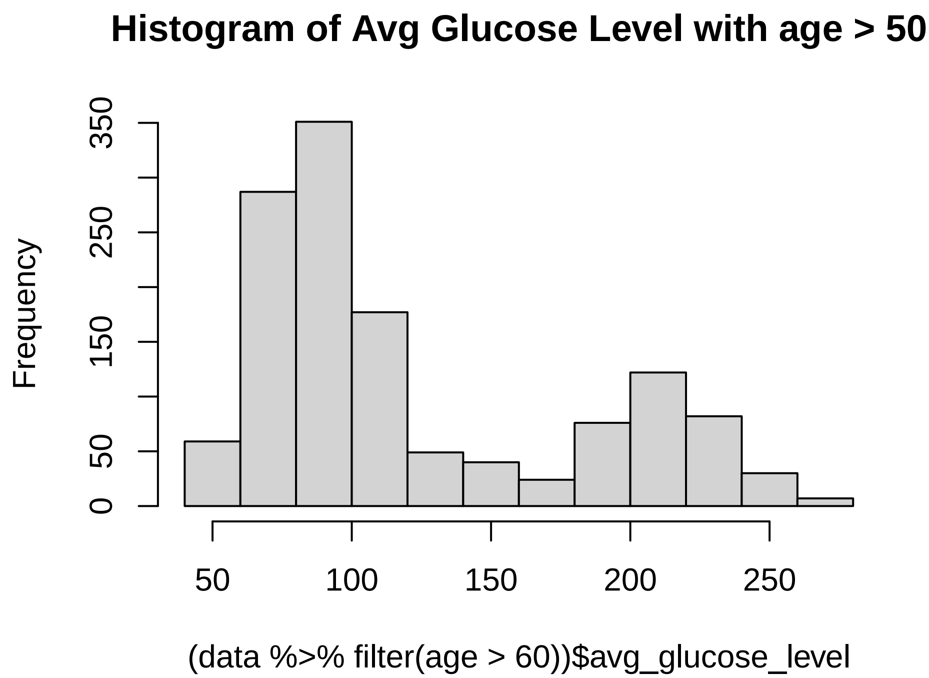


Figure 14: Histogram of Avg Glucose Level with age > 50

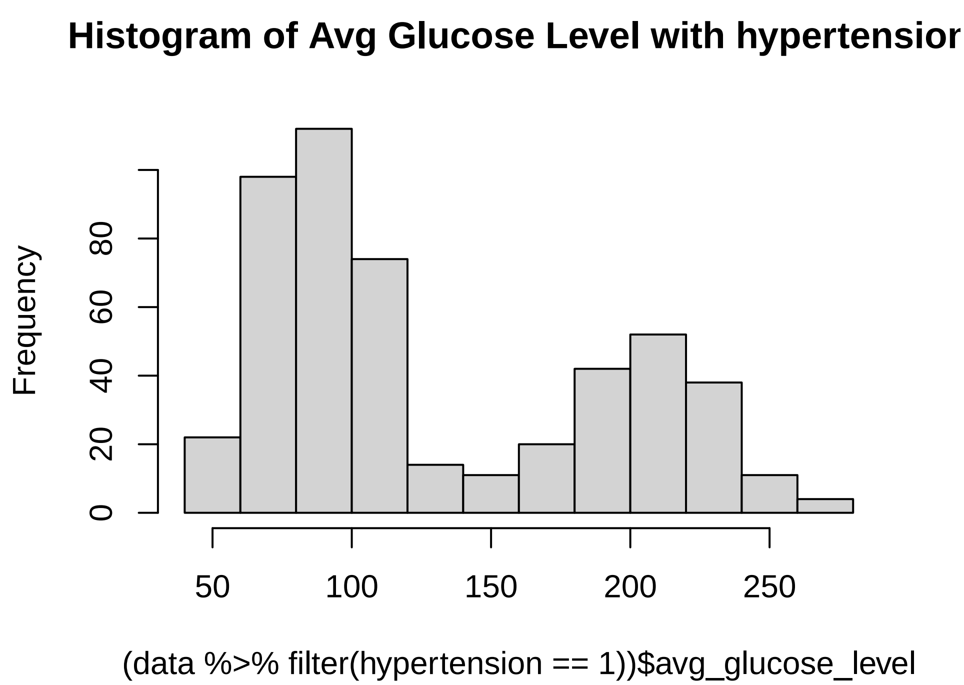


Figure 15: Histogram of Avg Glucose Level with hypertension