

I. Reading in the data

```
fhs.df <- read.csv('framingham.csv')
#fhs.df[ 1:10, 1:16 ]
#fhs.df
```

dimension of the matrix

```
dim( fhs.df )
```

```
## [1] 4240 16
```

The data frame is a 16 by 4240 matrix.

There are 15 variables (columns), excluding the index.

There are 4240 subjects.

variable names

```
# what are the variable names
names( fhs.df )
```

```
## [1] "male"          "age"           "education"     "currentSmoker"
## [5] "cigsPerDay"    "BPMeds"        "prevalentStroke" "prevalentHyp"
## [9] "diabetes"      "totChol"       "sysBP"         "diaBP"
## [13] "BMI"          "heartRate"     "glucose"       "TenYearCHD"
```

attach dataset

```
# link to dataset
attach( fhs.df )
```

II. Model building

We will start with a forward selection strategy first. If this does not lead to the model with the best goodness of fit, we will consider alternative model along the way.

The variables we have are :

sex, age

Education

CurrentSmoker, cigsPerday, heartRate

PrevalentStroke, prevalentHyp, BPMeds

diabetes

glucose, totChol, BMI

We will build a simple linear model for the our defined response variable, using only the explanatory variables (or functions of them):

CurrentSmoker, cigsPerday

Once we have a model in term of these two confounding confounding variables, we will add:

BPMeds

- response variable

Blood pressure is an health measure of cardiovascular function. In general, high blood pressure indicates poor vascular health, and lower blood pressure indicates otherwise. The ideal blood pressure reading is less than 120/80 mmHg.

General guidelines:

Normal: less than 120/80 mmHg

Elevated: 120/129 / <80 mmHg

Hypertension status 1: 130-139 / 80-90 mmHg

Hypertension status 2: 140 or higher / 90 or higher mmHg

we decide to make our response variable as the difference between of the systolic blood pressure and the diastolic blood pressure.

When blood pressure is taken, systolic and diastolic blood pressures are measured.

1. Systolic pressure (the top number in the blood pressure reading) measures the pressure of the arteries.
2. Dystolic pressure (the bottom number in the blood pressure reading) measures the pressure of the arteries.

```
# define the difference bet. systolic and dystolic pressure
bpd = sysBP - diaBP
#bp.df

# create the log.bpd variable
log.bpd = log(bpd)
```

- age model

Does age has to do with cardiovascular health?

In theory, yes, because age is a risk factor for age-related cardiovascular declines.

```
age.model <- lm ( log.bpd ~ age )
summary( age.model )
```

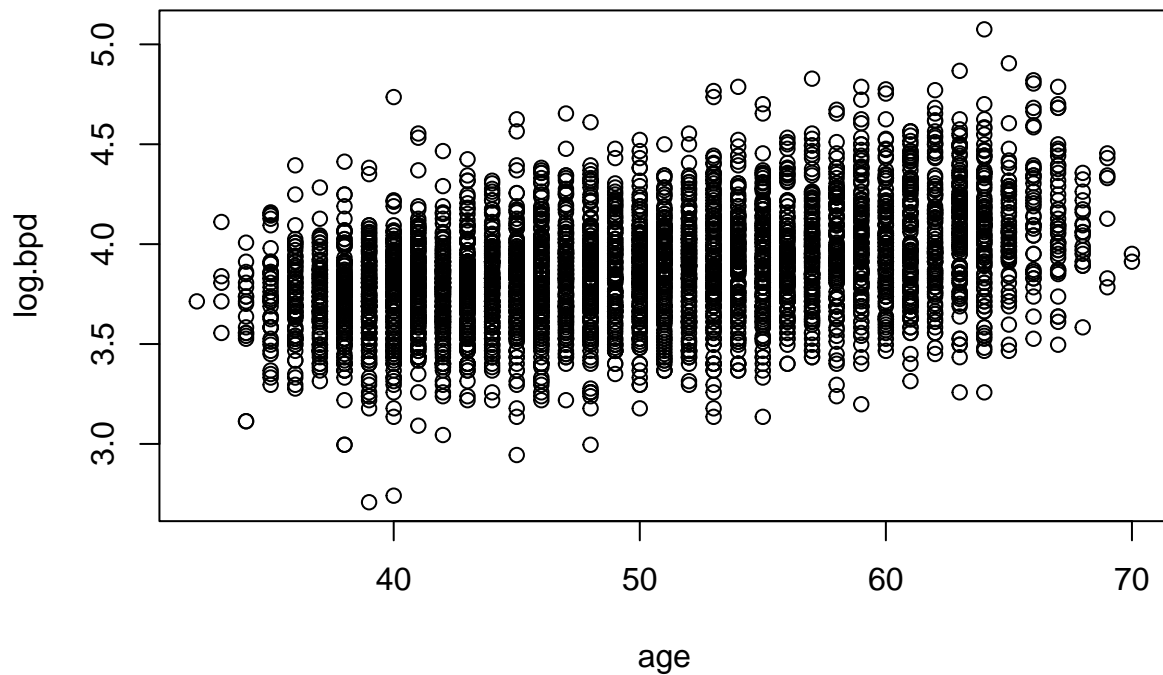
```
##
## Call:
## lm(formula = log.bpd ~ age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01065 -0.16238 -0.00306  0.15327  1.01834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.1912082  0.0225468  141.54  <2e-16 ***
## age          0.0135254  0.0004481   30.18  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2501 on 4238 degrees of freedom
## Multiple R-squared:  0.1769, Adjusted R-squared:  0.1767
## F-statistic:  911 on 1 and 4238 DF,  p-value: < 2.2e-16
```

```
anova( age.model )
```

```
## Analysis of Variance Table
##
## Response: log.bpd
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age         1  56.993   56.993  911.04 < 2.2e-16 ***
## Residuals 4238  265.122    0.063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

plot

```
plot( age, log.bpd )
```



- gender model

Does gender contribute to heart health?

```
# is.female equals to 0 if the gender is female, 1 otherwise.
is.female <- ifelse( male == 1, 0, 1 )
#is.female
```

```
gender.model <- lm( log.bpd ~ is.female )
summary( gender.model )
```

```
##
## Call:
## lm(formula = log.bpd ~ is.female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17545 -0.19462 -0.01523  0.15955  1.19167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.832940   0.006436  595.580 < 2e-16 ***
## is.female    0.050564   0.008519   5.936 3.16e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2746 on 4238 degrees of freedom
## Multiple R-squared:  0.008245,    Adjusted R-squared:  0.008011
## F-statistic: 35.23 on 1 and 4238 DF,  p-value: 3.16e-09
```

```
anova( gender.model )
```

```
## Analysis of Variance Table
##
## Response: log.bpd
##           Df Sum Sq Mean Sq F value    Pr(>F)
## is.female   1   2.66  2.65584   35.233 3.16e-09 ***
## Residuals 4238 319.46  0.07538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#plot <- (is.female, log.bpd )
```

- bmi model

```
BMI.model <- lm( log.bpd ~ fhs.df$BMI )
summary(BMI.model)
```

```
##
## Call:
## lm(formula = log.bpd ~ fhs.df$BMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16472 -0.18097 -0.01419  0.15927  1.05504
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.546993   0.026661  133.04  <2e-16 ***
## fhs.df$BMI   0.012188   0.001021   11.94  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2705 on 4219 degrees of freedom
## (19 observations deleted due to missingness)
## Multiple R-squared:  0.03269,    Adjusted R-squared:  0.03246
## F-statistic: 142.6 on 1 and 4219 DF,  p-value: < 2.2e-16
```

```
anova(BMI.model)
```

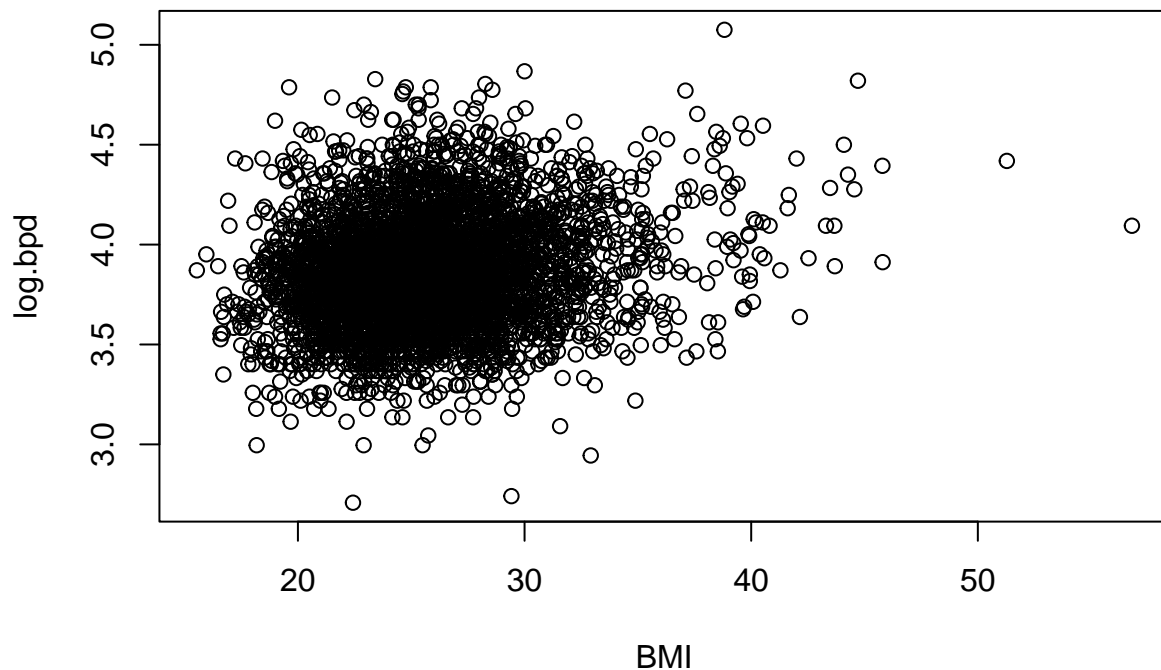
```
## Analysis of Variance Table
##
```

```
## Response: log.bpd
##           Df Sum Sq Mean Sq F value    Pr(>F)
## fhs.df$BMI     1  10.434  10.4343   142.6 < 2.2e-16 ***
## Residuals 4219 308.714   0.0732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 2e-16.

```
#fhs.df$BMI
```

```
plot(BMI, log.bpd)
```



- diabetets model

```
# is.diabetes equals to 1 if the diabetic status is yes, 0 otherwise.
is.diabetes <- ifelse( diabetes == 1, 1, 0 )
is.diabetes
```

```
##      [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##      [38] 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
##      [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
##     [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

##	[149]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[186]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[223]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
##	[260]	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
##	[297]	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[334]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
##	[371]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[408]	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
##	[445]	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
##	[482]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[519]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[556]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
##	[593]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[630]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[667]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[704]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[741]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
##	[778]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[815]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
##	[852]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[889]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
##	[926]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

[illegible]


```
fhs.df$diabetes
```

9

[illegible]

```
## [3775] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0
## [3812] 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
## [3849] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [3886] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [3923] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [3960] 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [3997] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [4034] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
## [4071] 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
## [4108] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [4145] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [4182] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [4219] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
diabetic.model <- lm( log.bpd ~ fhs.df$diabetes )
summary(diabetic.model)
```

```
##
## Call:
## lm(formula = log.bpd ~ fhs.df$diabetes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14865 -0.18040 -0.01725  0.15968  1.21847
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.856703   0.004262  905.005 < 2e-16 ***
## fhs.df$diabetes 0.198257   0.026579   7.459 1.05e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2739 on 4238 degrees of freedom
## Multiple R-squared:  0.01296,    Adjusted R-squared:  0.01273
## F-statistic: 55.64 on 1 and 4238 DF,  p-value: 1.05e-13
```

```
anova(diabetic.model)
```

```
## Analysis of Variance Table
##
## Response: log.bpd
##              Df Sum Sq Mean Sq F value    Pr(>F)
## fhs.df$diabetes    1    4.17   4.1742    55.64 1.05e-13 ***
## Residuals       4238  317.94   0.0750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Assuming fhs.df$diabetes is a binary variable indicating diabetic status (0 for non-diabetic, 1 for d
```

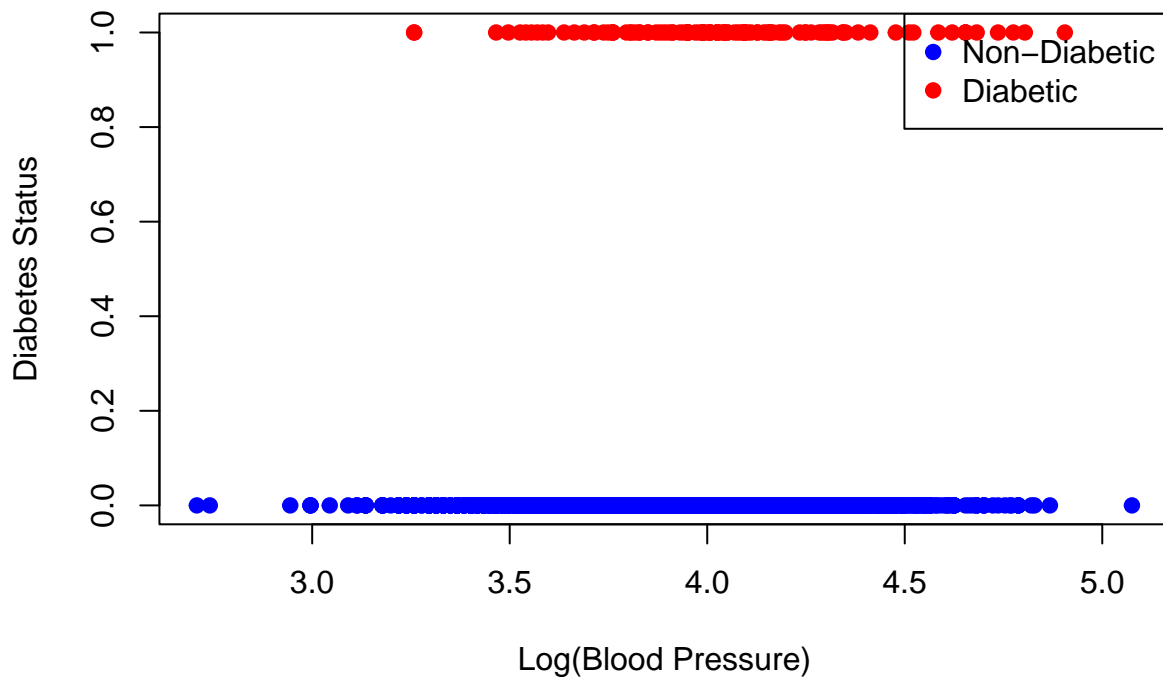
```
# Assign colors based on diabetic status
```

```
colors <- c("non-diabetic" = "blue", "diabetic" = "red")
```

```
point_colors <- colors[fhs.df$diabetes + 1] # Adding 1 to convert 0/1 to 1/2 for indexing colors
```

```
# Plot log.bpd against diabetes status
plot(log.bpd, fhs.df$diabetes, col = point_colors, pch = 19, xlab = "Log(Blood Pressure)", ylab = "Diabetes Status")

# Add legend
legend("topright", legend = c("Non-Diabetic", "Diabetic"), col = colors, pch = 19)
```



```
diabetic.count <- sum(fhs.df$diabetes == 1 )
diabetic.count
```

```
## [1] 109
```

```
nondiabetic.count <- sum(fhs.df$diabetes == 0 )
nondiabetic.count
```

```
## [1] 4131
```

1/40 of people are diabetic in this study

```
# Example data
```

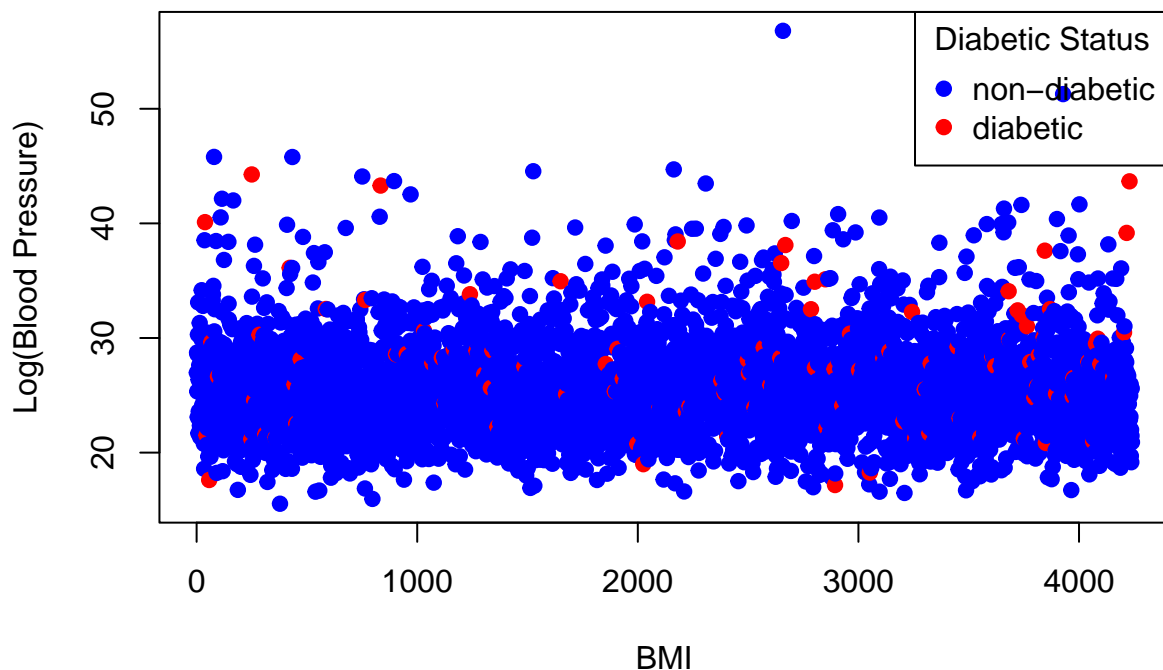
```
# Convert diabetes to a factor for color assignment
fhs.df$diabetes_factor <- factor(fhs.df$diabetes, levels = c(0, 1), labels = c("non-diabetic", "diabetic"))
```

```

# Create a color vector based on diabetes_factor
colors <- c("non-diabetic" = "blue", "diabetic" = "red")
point_colors <- colors[fhs.df$diabetes_factor]

# Plot BMI against log.bpd, color by diabetic status
plot(fhs.df$BMI, fhs.df$log.bpd, col = point_colors, pch = 19, xlab = "BMI", ylab = "Log(Blood Pressure)",
legend("topright", legend = levels(fhs.df$diabetes_factor), col = colors, pch = 19, title = "Diabetic Status"))

```



glu. model

```

glucose.model <- lm( heartRate ~ glucose )
summary(glucose.model)

```

```

##
## Call:
## lm(formula = heartRate ~ glucose)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.832  -8.170  -0.782   6.594  67.594
##

```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.875266   0.682222 105.355 < 2e-16 ***
## glucose      0.047075   0.007989   5.892 4.14e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.88 on 3849 degrees of freedom
## (389 observations deleted due to missingness)
## Multiple R-squared:  0.008939, Adjusted R-squared:  0.008682
## F-statistic: 34.72 on 1 and 3849 DF, p-value: 4.14e-09
```

```
anova( glucose.model)
```

```
## Analysis of Variance Table
##
## Response: heartRate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## glucose      1   4897   4896.7   34.717 4.14e-09 ***
## Residuals 3849 542890    141.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cigar.model <- lm( heartRate ~ cigsPerDay )
summary(cigar.model)
```

```
##
## Call:
## lm(formula = heartRate ~ cigsPerDay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.322  -8.180  -0.865   6.820  67.820
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 75.17976   0.23148 324.777 < 2e-16 ***
## cigsPerDay   0.07617   0.01549   4.916 9.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.99 on 4208 degrees of freedom
## (30 observations deleted due to missingness)
## Multiple R-squared:  0.00571, Adjusted R-squared:  0.005474
## F-statistic: 24.17 on 1 and 4208 DF, p-value: 9.176e-07
```

```
anova(cigar.model)
```

```
## Analysis of Variance Table
##
## Response: heartRate
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## cigsPerDay    1    3471  3471.4  24.166 9.176e-07 ***
## Residuals    4208 604477   143.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

sysBP

smoking model

```
# is.smoker equals to 1 if the the individual currently smokes, 0 otherwise.
is.smoker <- ifelse( currentSmoker == 1, 1, 0 )
#is.smoker
```

```
smoking.model <- lm ( log.bpd ~ is.smoker )
summary( smoking.model )
```

```
##
## Call:
## lm(formula = log.bpd ~ is.smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14821 -0.18775 -0.01618  0.15509  1.18612
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.889052   0.005923  656.624 < 2e-16 ***
## is.smoker    -0.055155   0.008426  -6.546 6.62e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2743 on 4238 degrees of freedom
## Multiple R-squared:  0.01001,    Adjusted R-squared:  0.009776
## F-statistic: 42.85 on 1 and 4238 DF,  p-value: 6.62e-11
```

```
anova( smoking.model )
```

```
## Analysis of Variance Table
##
## Response: log.bpd
##              Df Sum Sq Mean Sq F value    Pr(>F)
## is.smoker      1   3.22   3.2241   42.848 6.62e-11 ***
## Residuals 4238 318.89   0.0752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- cholestrol model

```
chol.model <- lm( log.bpd ~ totChol )
summary(chol.model)
```

```
##
## Call:
## lm(formula = log.bpd ~ totChol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17427 -0.17793 -0.01424  0.16260  1.19618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.602e+00  2.267e-02  158.91  <2e-16 ***
## totChol      1.094e-03  9.411e-05   11.63  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2716 on 4188 degrees of freedom
## (50 observations deleted due to missingness)
## Multiple R-squared:  0.03128,    Adjusted R-squared:  0.03105
## F-statistic: 135.2 on 1 and 4188 DF,  p-value: < 2.2e-16
```

```
anova(chol.model)
```

```
## Analysis of Variance Table
##
## Response: log.bpd
##           Df Sum Sq Mean Sq F value    Pr(>F)
## totChol      1   9.977   9.9768  135.25 < 2.2e-16 ***
## Residuals 4188 308.943   0.0738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```