Tachycardia data analysis project

Tetsuya Chau

2022-03-09

Problem Background

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects.

The dataset Tachycardia.txt contains a subset of data from the Framingham Heart Study for individuals who had 3 follow-up visits monitoring their heart health. Specifically, the variables included in the dataset are as follows:

Variable Name Description:

*RANDID - Patients Random ID

*SEX - Patients sex (1=M, 2=F)

*TOTCHOL - Patients total blood cholesterol

*AGE - Patients Age

*SYSBP - Patients Systolic Blood Pressure

*DIABP - Patients Diastoloic Blood Pressure

*CURSMOKE - Patients smoking status (1=Smoker)

*BMI - Patients body mass index

*DIABETES - Patients diabetes status (1=Yes)

*BPMEDS - Patients medication stats (1=Taking meds)

*HEARTRTE - Patients heart rate

*GLUCOSE - Patients glucose level

*PERIOD - The visit number (3 visits per patient)

In this analyses, we will be specifically looking at risk factors for Tachycardia (hence the name of the dataset) which refers to fast resting heart rate, usually over 100 beats per minute. Tachycardia can be dangerous, depending on its underlying cause and on how hard the heart has to work. Tachycardia significantly increases the risk of stroke, sudden cardiac arrest, and death. To analyze this dataset, do the following:

Analysis Questions:

1. Create exploratory plots of looking at the relationship between log(HEARTRTE) (the response variable) and some of the explanatory variables. Comment on any general relationships you see from the data.
2. Fit an independent MLR model with a linear effect of all variables except RANDID and PERIOD. Explore the residuals to see if there is evidence of correlation within a patients from period to period (visit to visit).

3. To determine an appropriate correlation structure to use, fit a longitudinal MLR model with an AR1, MA1 and general symmetric correlation matrix within each patient but independent across patients. Compare the model fits using AIC (which can be extracted from a gls() object using AIC()).
4. Write out your model for analyzing the Tachycardia data in terms of parameters. Explain and interpret any parameters associated with the model.
5. Fit your longitudinal model and validate any assumptions you made to fit the model.
6. Is DIABETES a risk factor for Tachycardia? Justify your answer and explain any effect of DIABETES on heart rate (include uncertainty in your conclusions).
7. What is the expected difference in heart rate for a female patient with at age 35 who is a smoker vs. an older female of 45 but not a smoker (assume the other characteristics are the same)? What does this say about the effect of smoking?

```
library(car)
```

```
## Loading required package: carData
```

```
#install.packages("ggplot2")
library(ggplot2)
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

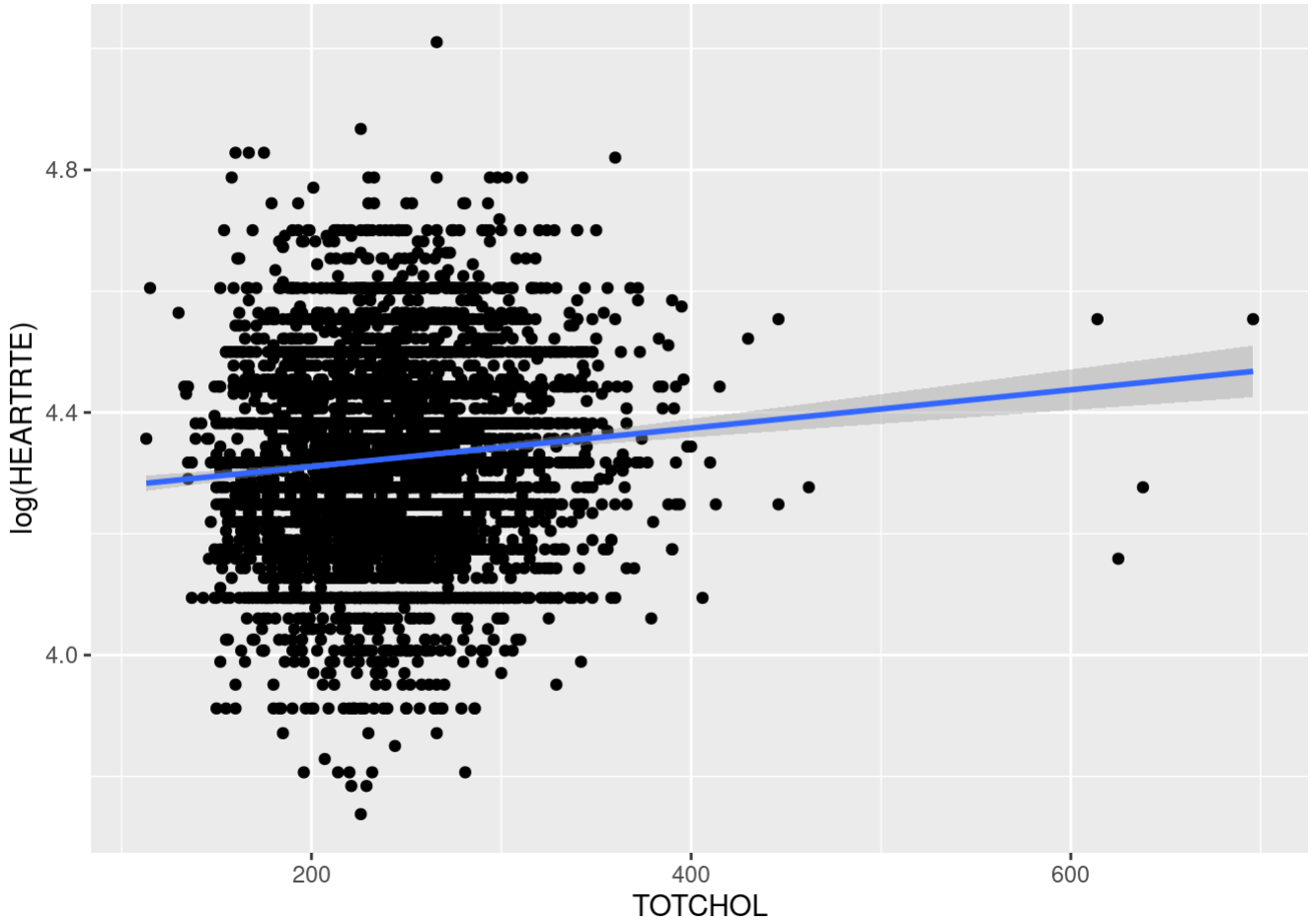
```
library(MASS)
library(mvtnorm)
library(nlme)
```

```
t <- read.csv("/cloud/project/Tachycardia.txt", sep="")
```

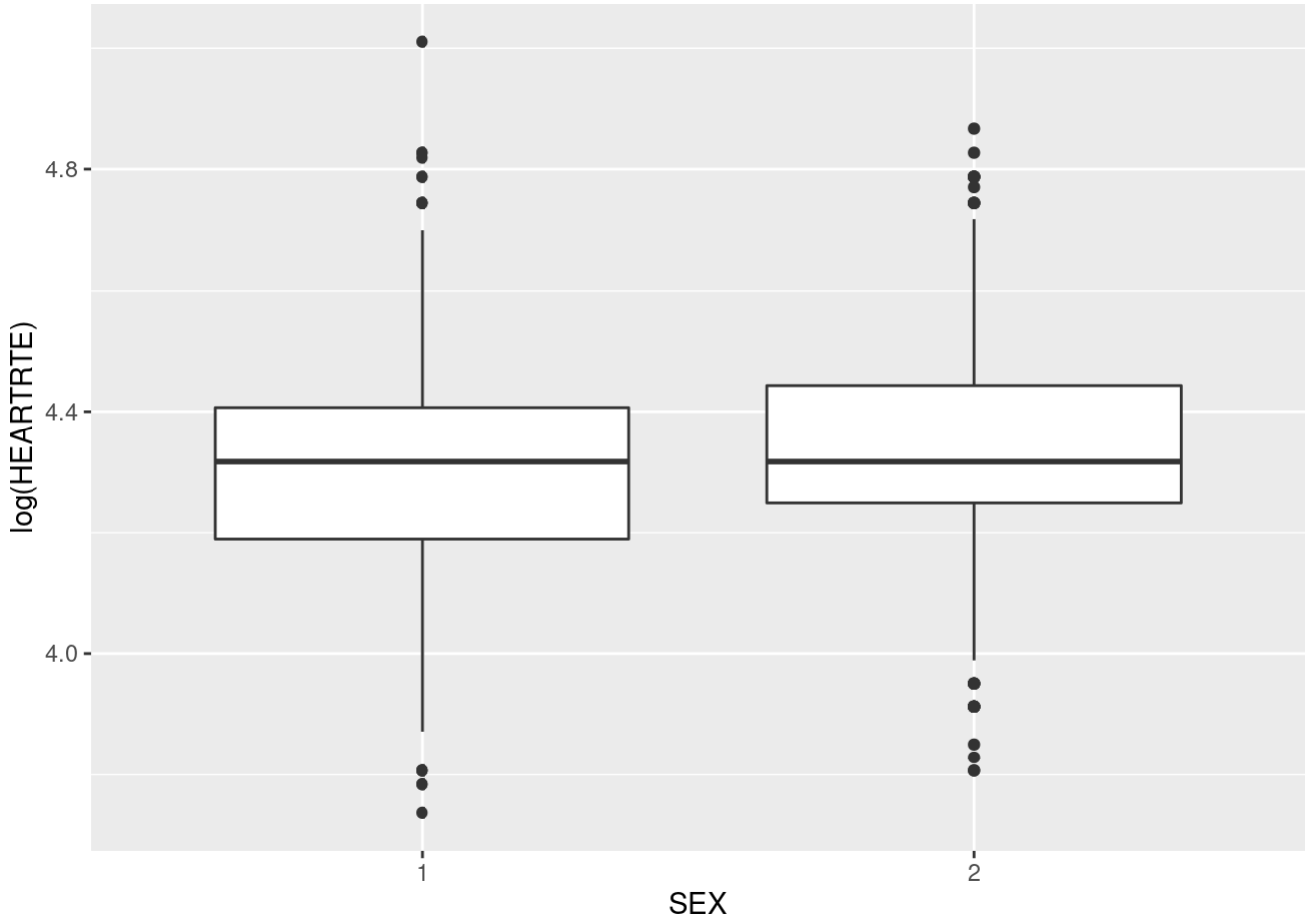
```
t$RANDID <- as.factor(t$RANDID)
t$SEX <- as.factor(t$SEX)
t$CURSMOKE <- as.factor(t$CURSMOKE)
t$DIABETES <- as.factor(t$DIABETES)
t$BPMEDS <- as.factor(t$BPMEDS)
#t$PERIOD <- as.factor(t$PERIOD)
```

1. Create exploratory plots of looking at the relationship between log(HEARTRTE) (the response variable) and some of the explanatory variables. Comment on any general relationships you see from the data.

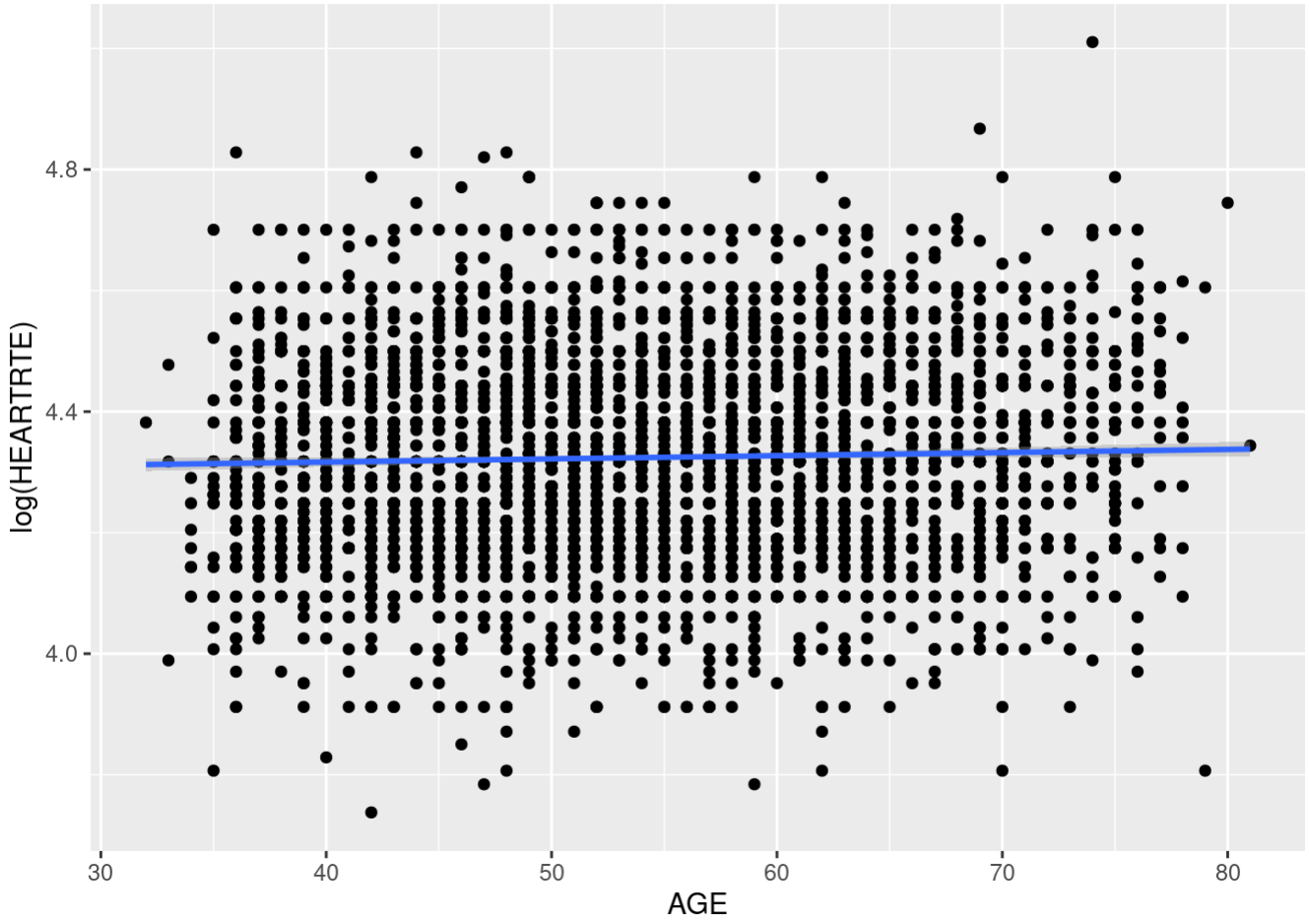
```
ggplot(data=t, mapping = aes(x=TOTCHOL, y=log(HEARTRTE)))+
 geom_point()+geom_smooth(method="lm",formula = y ~ x)
```



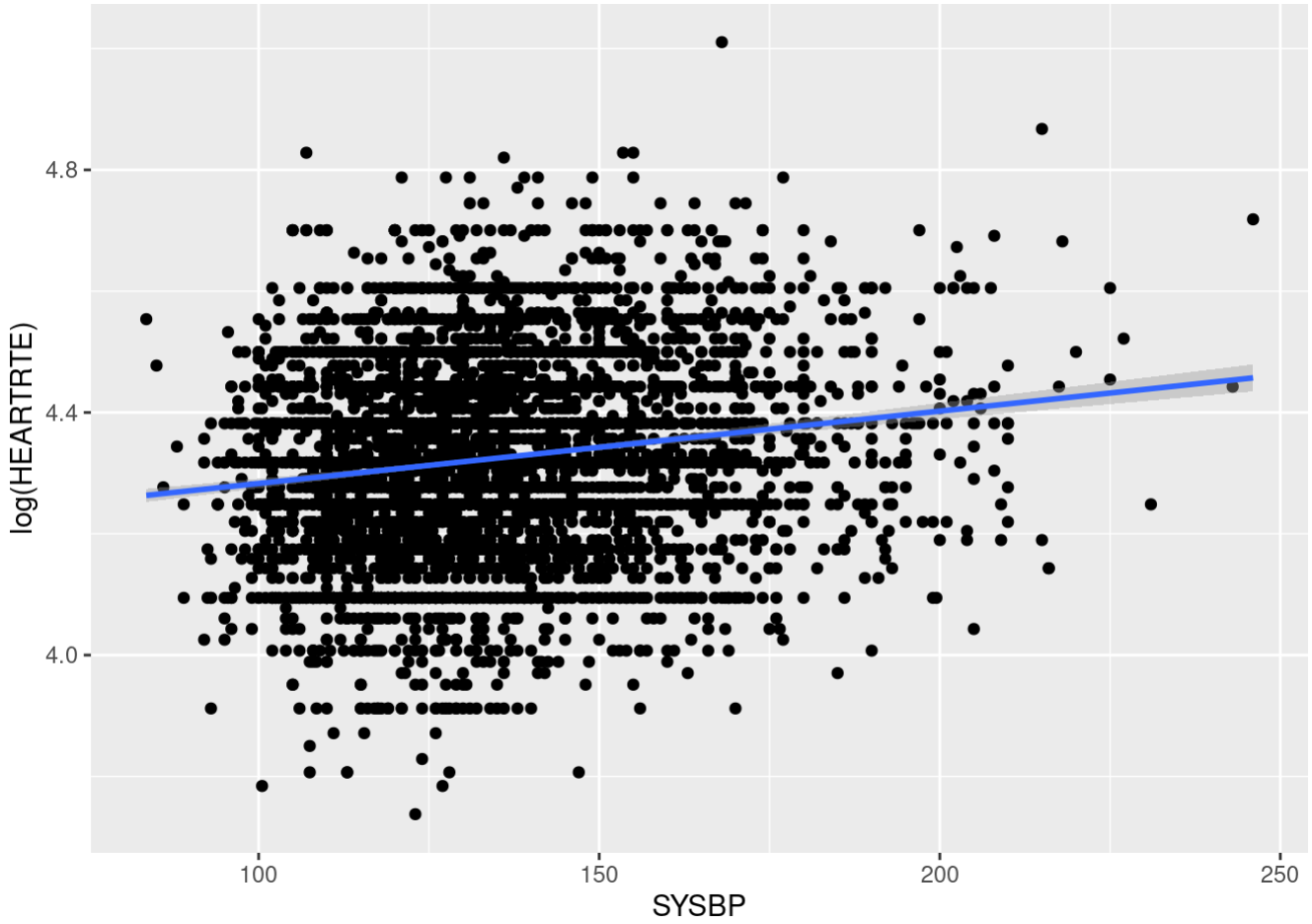

```
ggplot(data=t, mapping = aes(x=SEX, y=log(HEARTRTE)))+
 geom_boxplot()
```



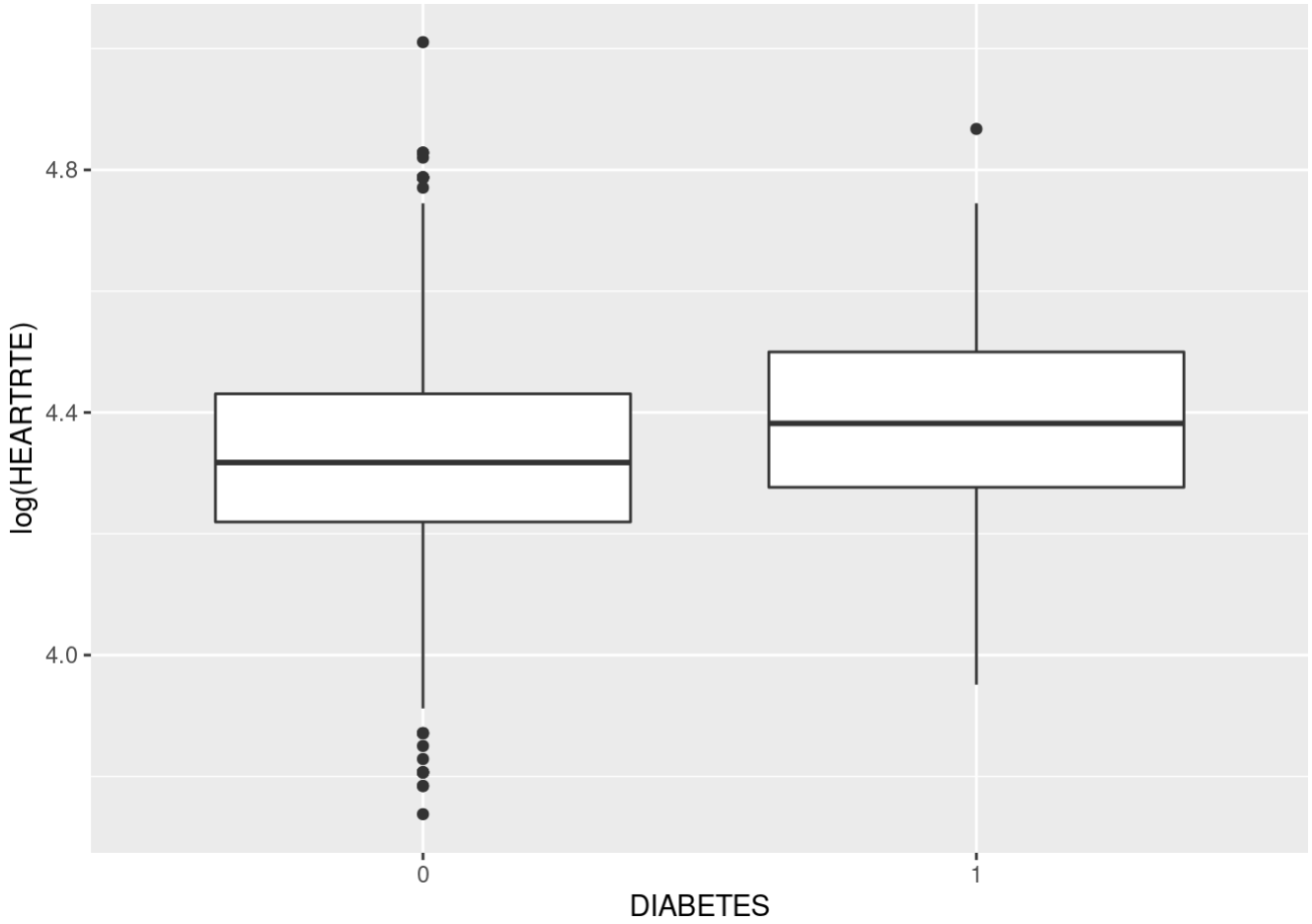

```
ggplot(data=t, mapping = aes(x=AGE, y=log(HEARTRTE)))+
 geom_point()+geom_smooth(method="lm",formula = y ~ x)
```



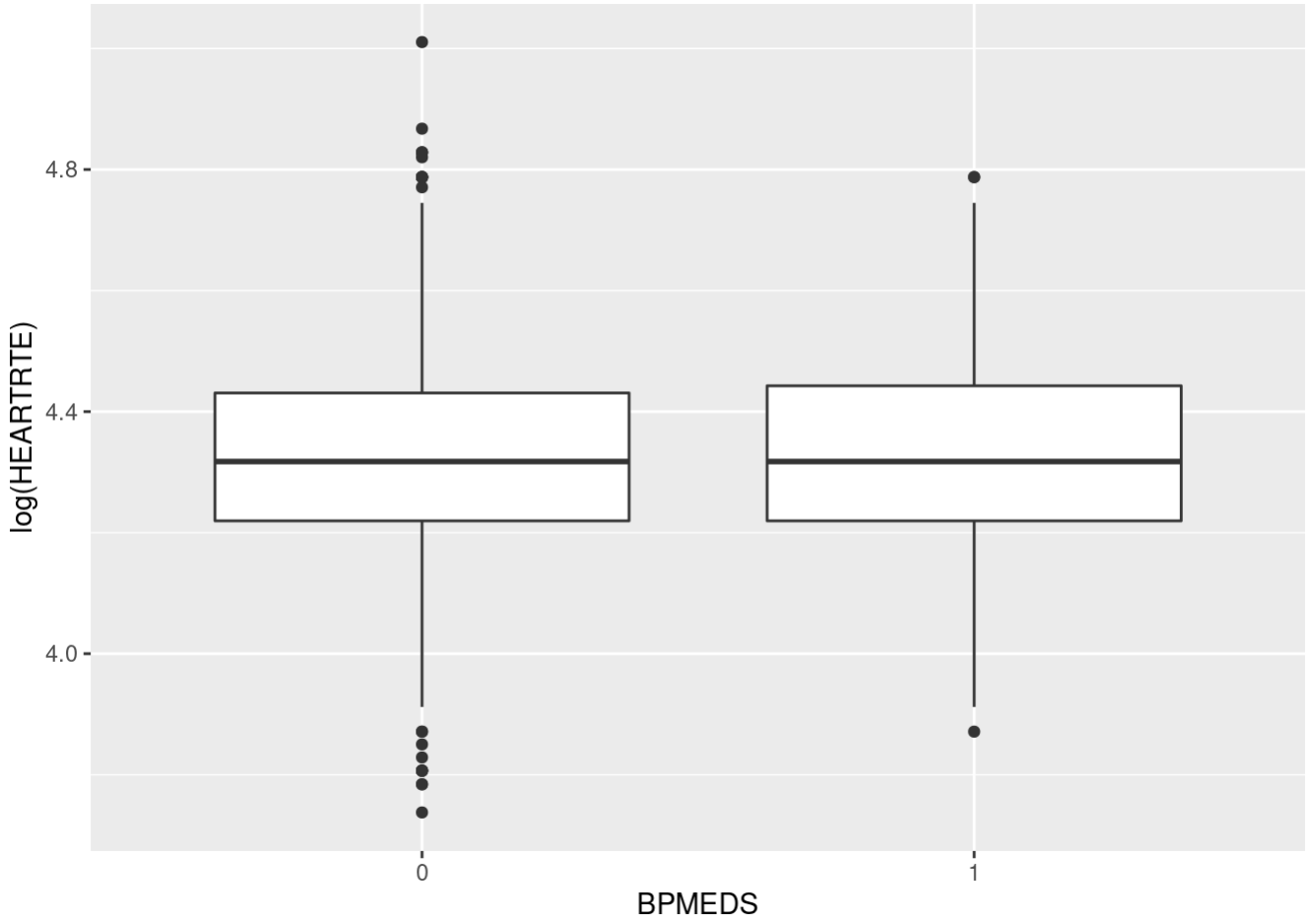

```
ggplot(data=t, mapping = aes(x=SYSBP, y=log(HEARTRTE)))+
 geom_point()+geom_smooth(method="lm",formula = y ~ x)
```



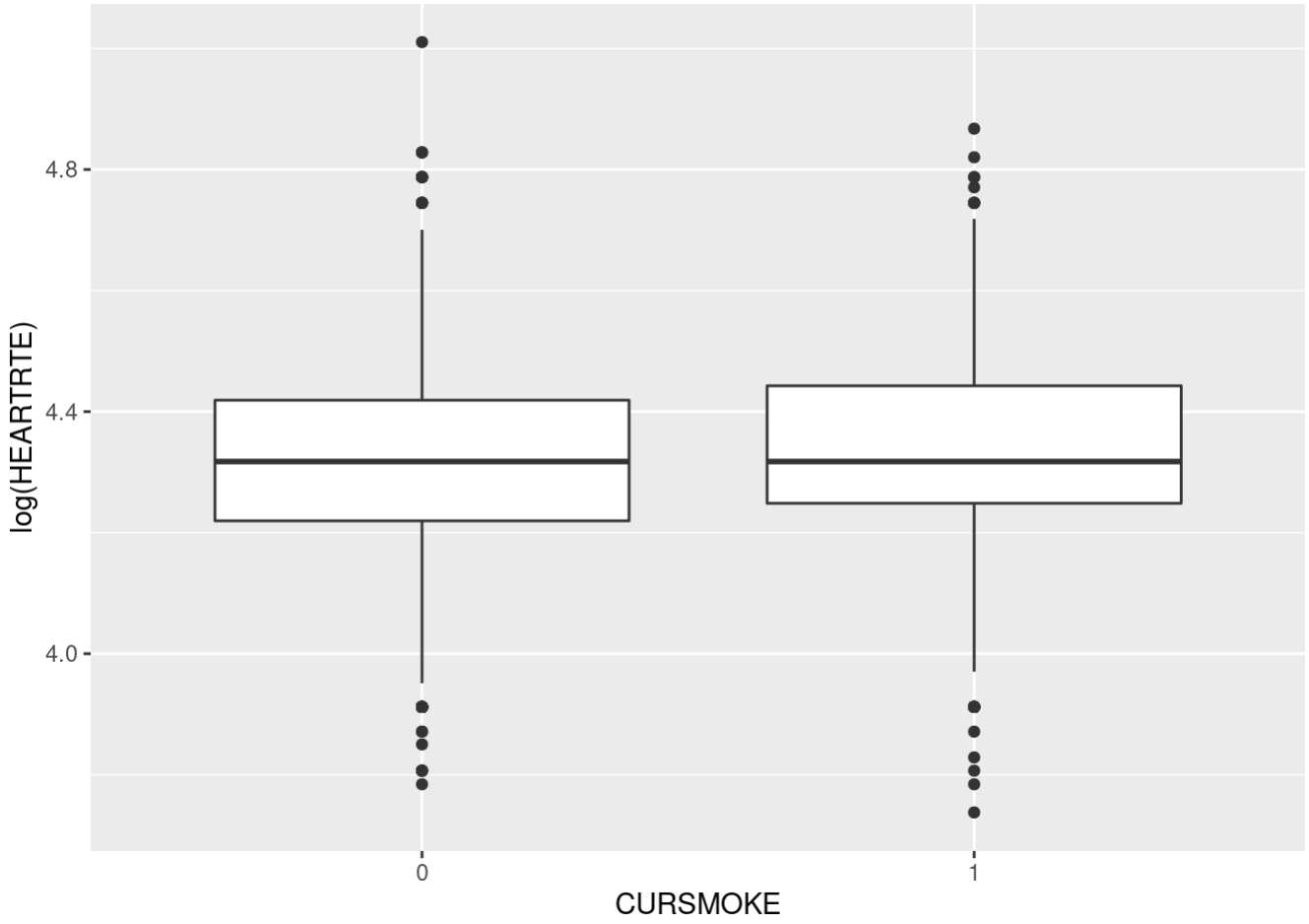

```
ggplot(data=t, mapping = aes(x=DIABETES, y=log(HEARTRTE)))+
 geom_boxplot()
```




```
ggplot(data=t, mapping = aes(x=BPMEDS, y=log(HEARTRTE)))+
 geom_boxplot()
```

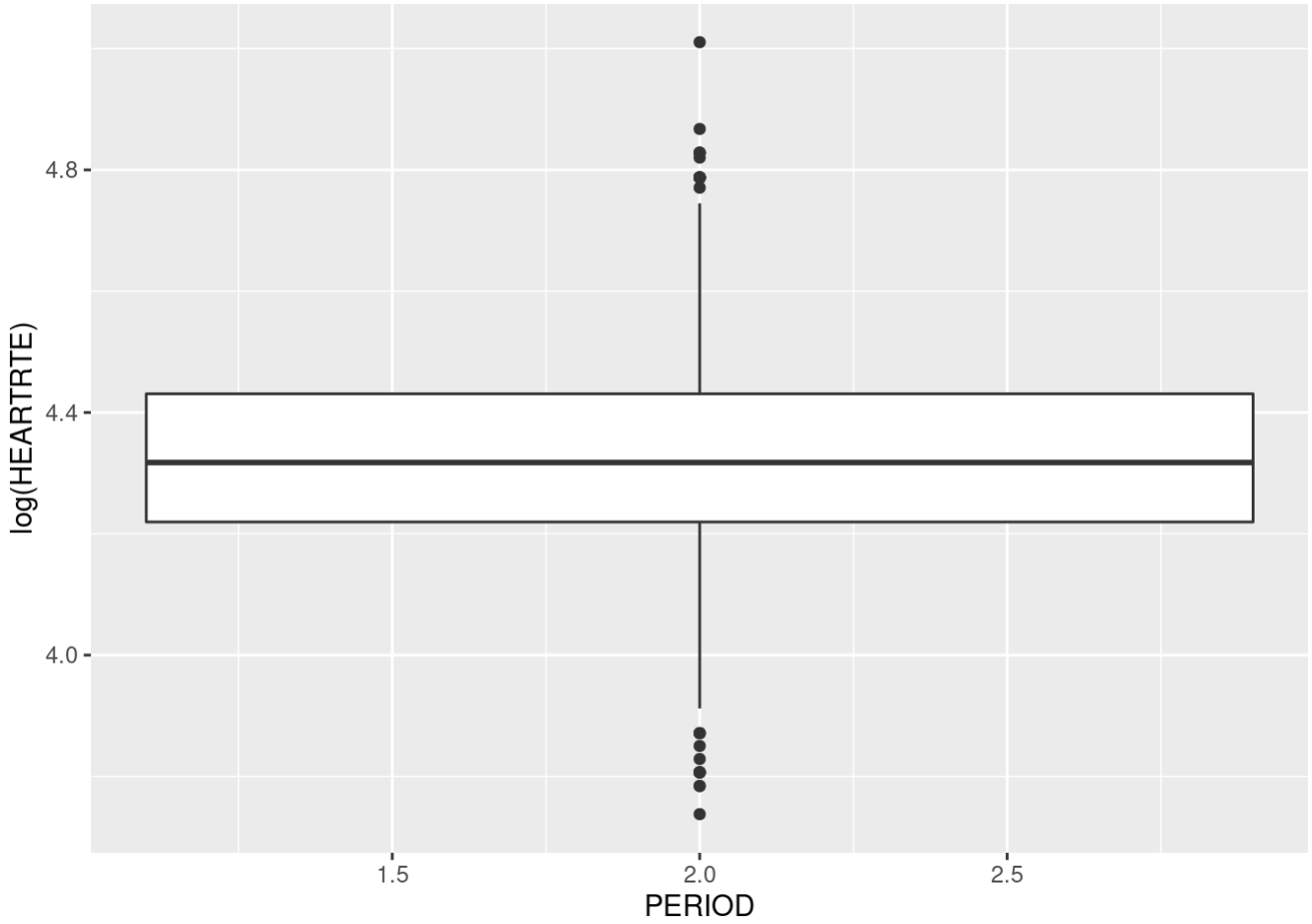



```
ggplot(data=t, mapping = aes(x=CURSMOKE, y=log(HEARTRTE)))+
 geom_boxplot()
```

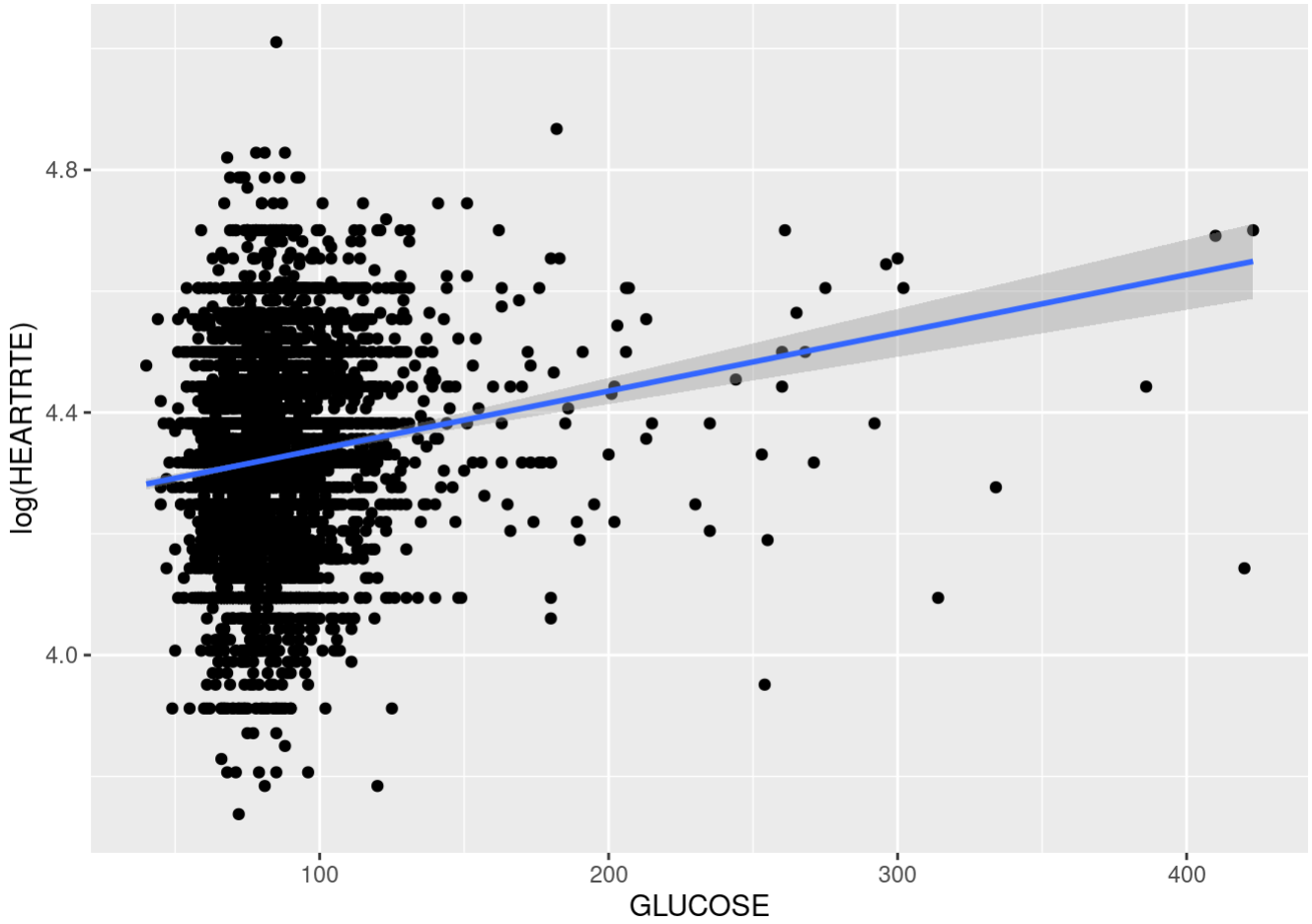



```
ggplot(data=t, mapping = aes(x=PERIOD, y=log(HEARTRTE)))+
 geom_boxplot()
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```




```
ggplot(data=t, mapping = aes(x=GLUCOSE, y=log(HEARTRTE)))+
 geom_point()+geom_smooth(method="lm",formula = y ~ x)
```



We are not seeing any noticeable correlations between the x-variables and the log(heart rate).

2. Fit an independent MLR model with a linear effect of all variables except RANDID and PERIOD. Explore the residuals to see if there is evidence of correlation within a patients from period to period (visit to visit).

```
t.lm <- lm(formula=log(HEARTRTE)~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOS
E,data=t)
t.lm
```

```
##
## Call:
## lm(formula = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = t)
##
## Coefficients:
## (Intercept)         SEX2      TOTCHOL          AGE        SYSBP        DIABP
##   3.959e+00    3.948e-02    2.028e-04    2.974e-05    4.370e-04    2.126e-03
##   CURSMOKE1          BMI    DIABETES1      BPMEDS1      GLUCOSE
##   3.254e-02   -1.191e-03    1.235e-02   -2.928e-02    9.214e-04
```

```
t.maxres <- matrix(resid(t.lm),byrow = TRUE, ncol = 3)
cor(t.maxres)
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.4600650 0.3914696
## [2,] 0.4600650 1.0000000 0.5010486
## [3,] 0.3914696 0.5010486 1.0000000
```

Based on looking at this matrix, we can say that there is an evidence of correlation within a patient from period to period.

3. To determine an appropriate correlation structure to use, fit a longitudinal MLR model with an AR1, MA1 and general symmetric correlation matrix within each patient but independent across patients. Compare the model fits using AIC (which can be extracted from a gls() object using AIC()).

```
x <- gls(model=log(HEARTRTE)~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOSE, d
ata=t, correlation=corSymm(form=~1|RANDID), method="ML")

AIC(x)
```

```
## [1] -5939.495
```

```
y <- gls(model=log(HEARTRTE)~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOSE, d
ata=t, correlation=corAR1(form=~PERIOD | RANDID), method="ML")

AIC(y)
```

```
## [1] -5868.104
```

```
z <- gls(model=log(HEARTRTE)~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOSE, d
ata=t, correlation=corARMA(form=~PERIOD | RANDID, q = 1), method="ML")

AIC(z)
```

```
## [1] -5647.617
```

We will use the general symmetric correlation structure based on the fact that it gave us the lowest AIC value which means that it is the best model for model fit in comparison to the other correlation structures that we tested.

```
t.model <- x
```

4. Write out your model for analyzing the Tachycardia data in terms of parameters. Explain and interpret any parameters associated with the model.

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 B)$$

$$B = \text{diag}(R, \ldots, R)$$

-y is a vector of the responses aka log(heart rate).

-X is the matrix of all the explanatory variables in the model such as age, sex, glucose level, total blood cholesterol, etc.

-Beta is the coefficients of the explanatory variables in the model.

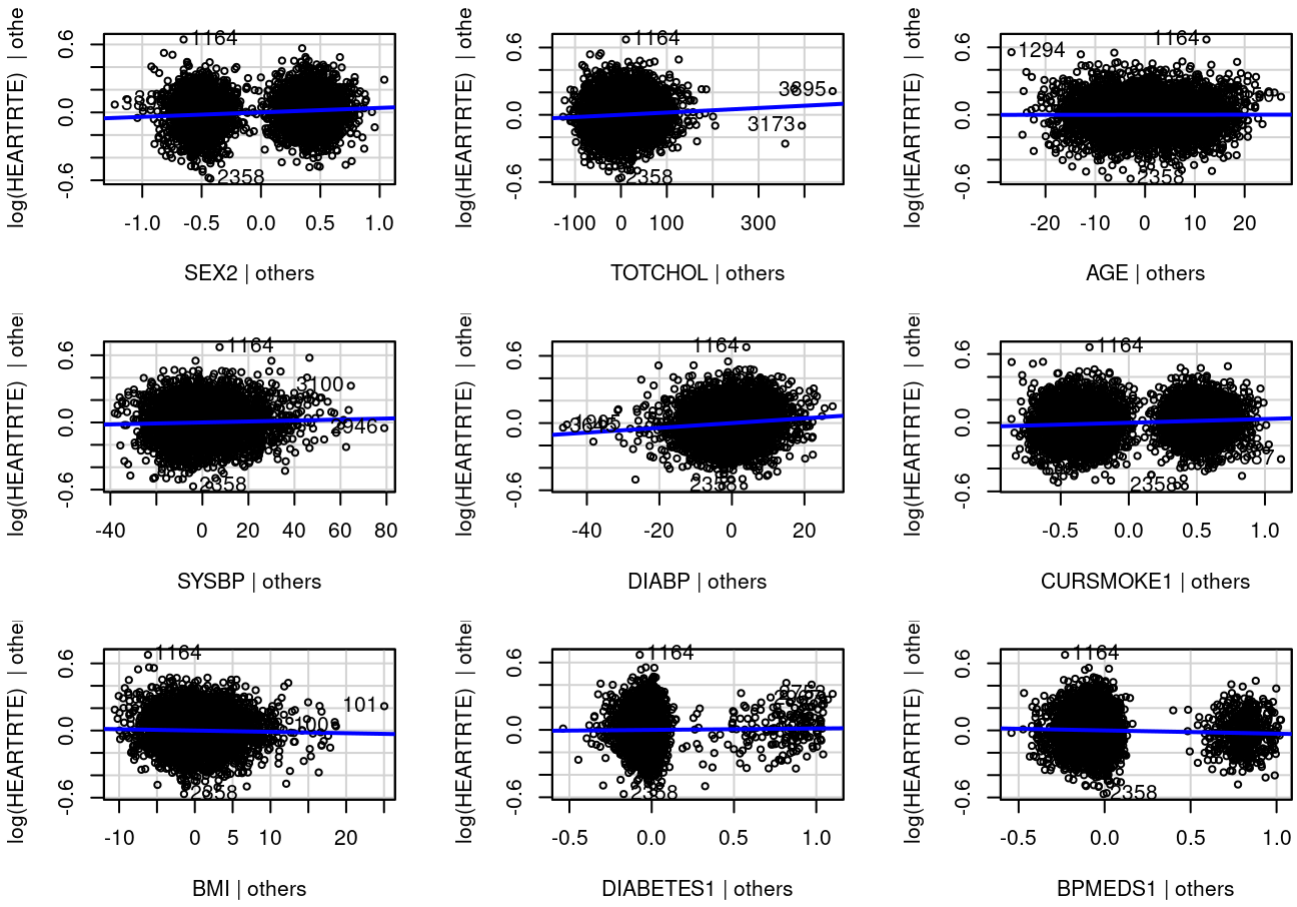
-Sigma squared is the the constant that contributes to the variance.

-e is the error ~ N(0,sigma-squared*B).

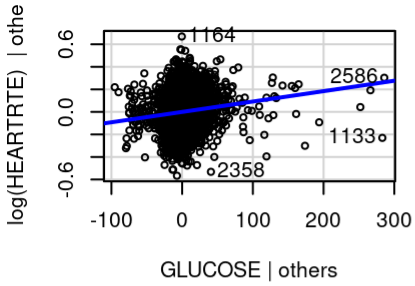
-R is a 3 by 3 matrix of the general symmetric correlation structure.

5. Fit your longitudinal model and validate any assumptions you made to fit the model.

```
#Linearity
avPlots(t.lm,ask=FALSE)
```



Added-Variable Plots



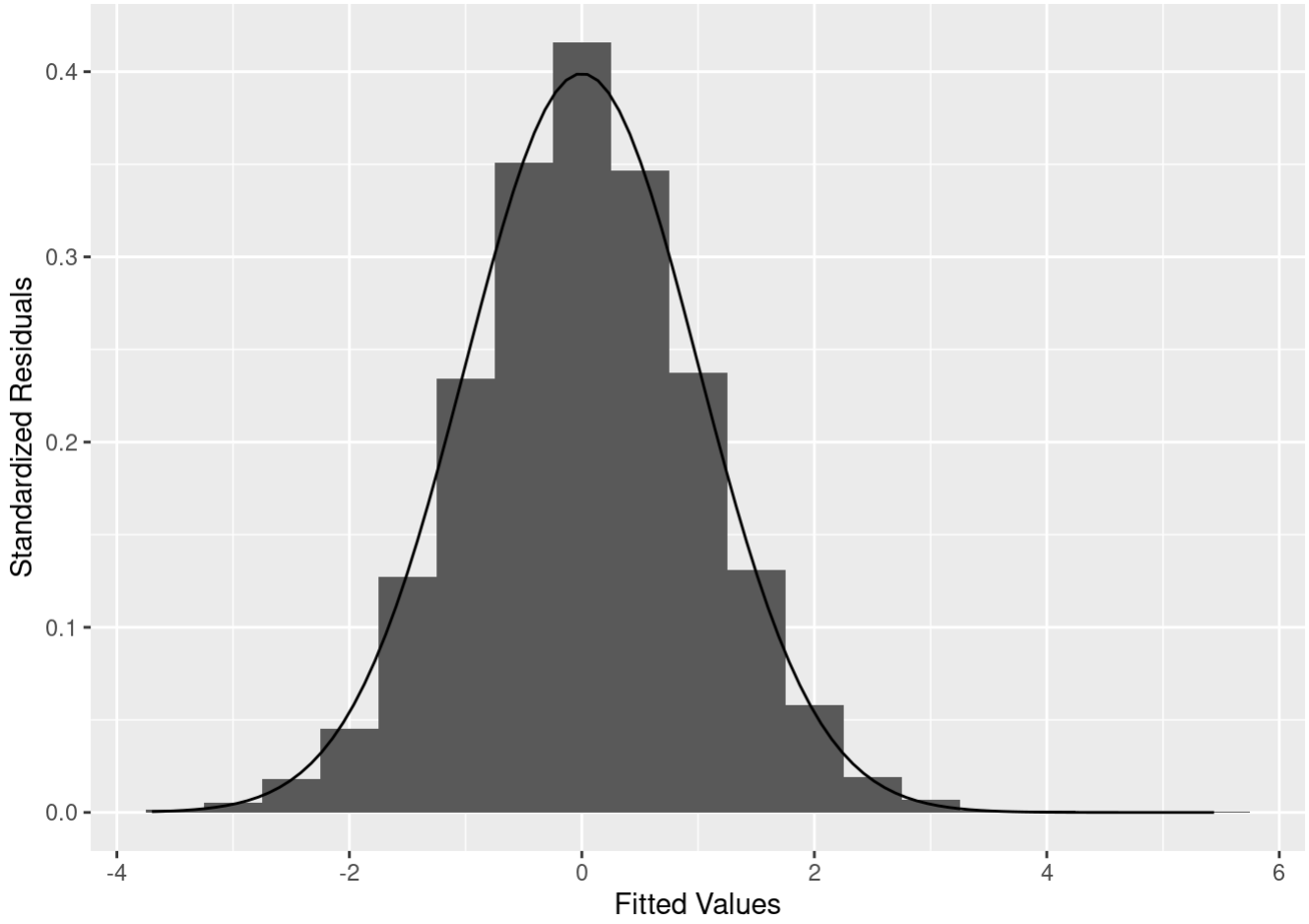
The av plots of the numeric vs numeric variables appear to be linear so it meets the linearity assumption.

```
#independence
source("stdres.R")
sres1 <- stdres.gls(t.model)

t.maxres1 <- matrix(sres1,byrow = TRUE, ncol = 3)

cor(t.maxres1)
```

```
##             [,1]         [,2]         [,3]
## [1,] 1.000000000  0.010689330  0.005603363
## [2,] 0.010689330  1.000000000 -0.004182272
## [3,] 0.005603363 -0.004182272  1.000000000
```

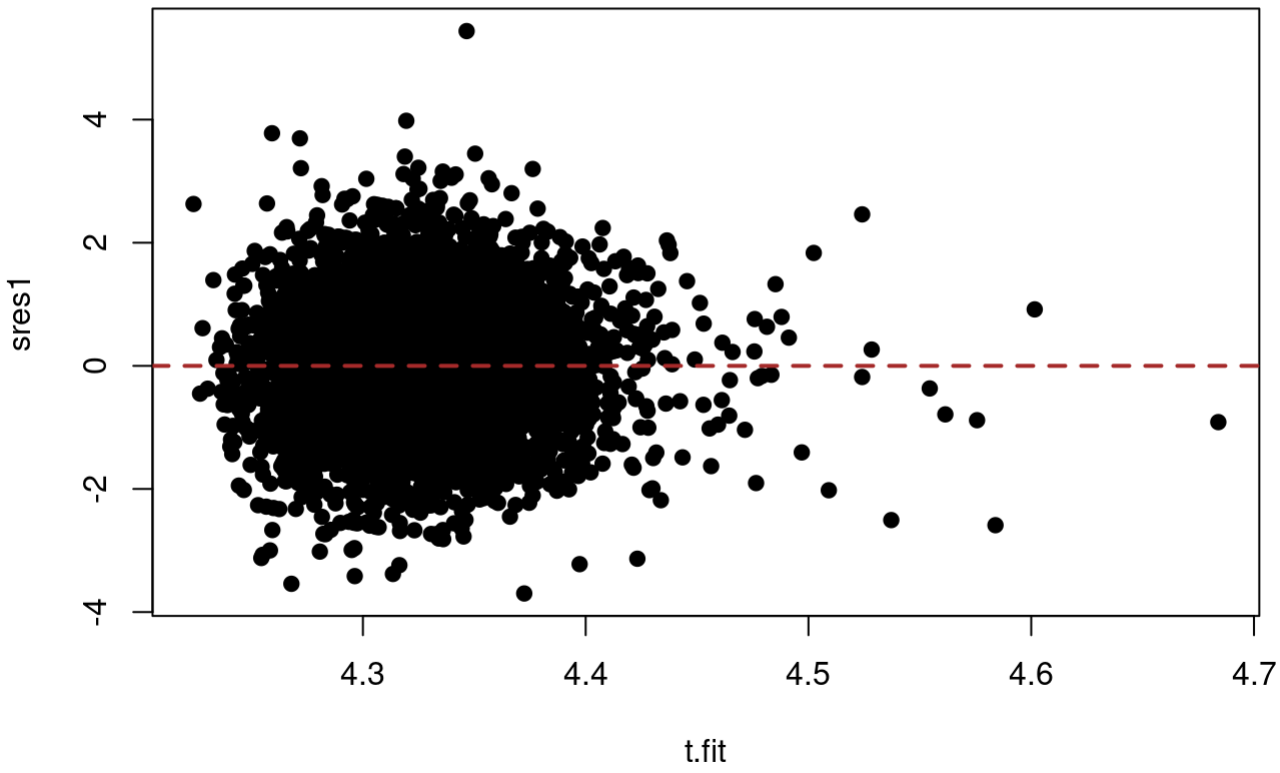
There isn't a strong correlation in the residuals based on looking at this matrix so we can assume independence here.

```
ggplot(t, aes(x = sres1)) +
  geom_histogram(aes(y = ..density..), binwidth = .5) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = sd(sres1))) + # add a standard normal cu
rve
  labs(y= "Standardized Residuals", x = "Fitted Values")
```

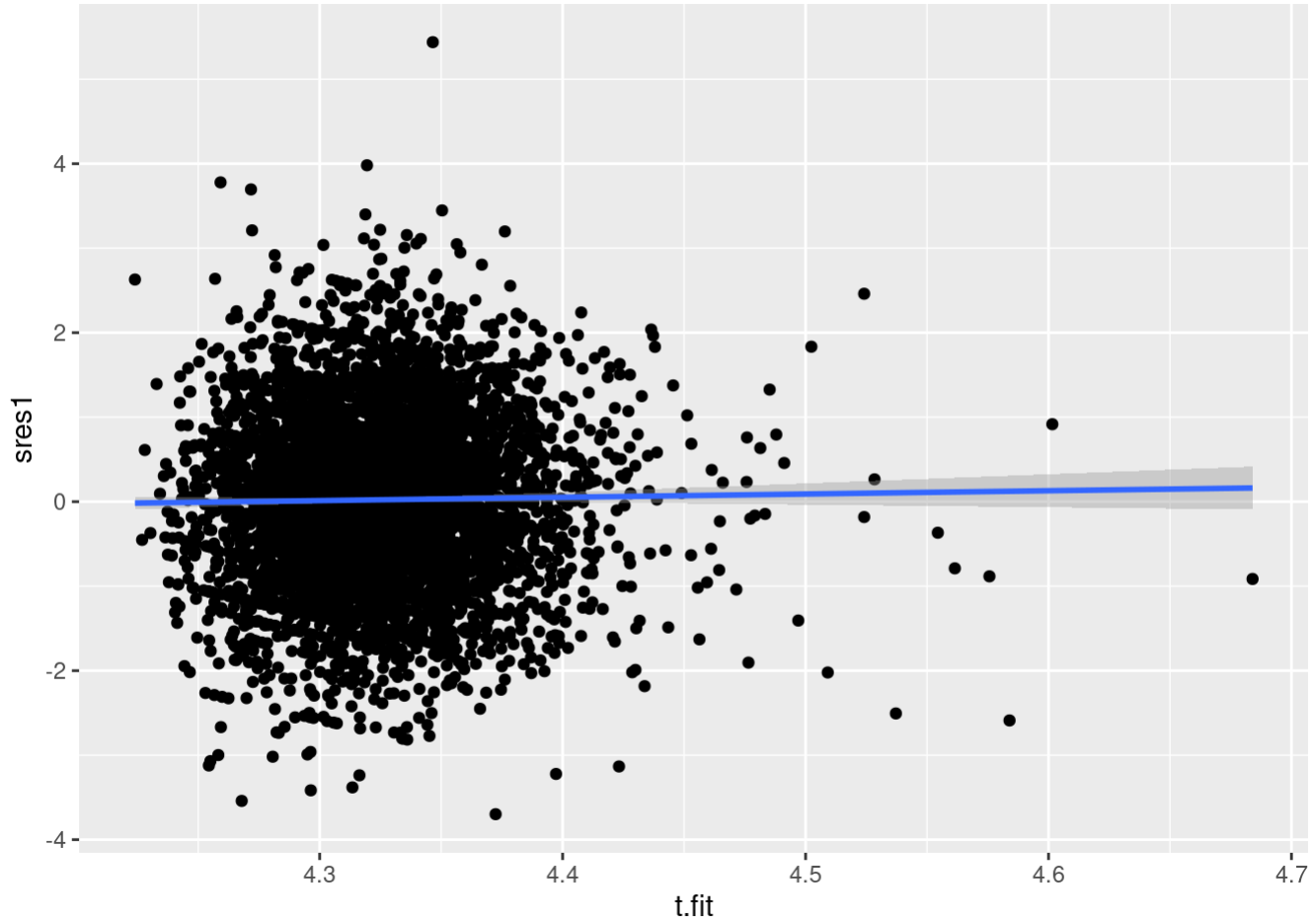


The histogram of the standardized residuals looks normal with a Gaussian curve so we can assume normality here as well.

```
#equal variance
t.fit <- fitted(t.model)

plot(t.fit,sres1,pch=19)
abline(0,0,col = "brown",lwd = 2,lty = 2)
```




```
ggplot(data=t, mapping = aes(x=t.fit, y=sres1))+
 geom_point()+geom_smooth(method="lm",formula = y ~ x)
```



The scatter plot of the fitted values vs residuals looks good because even though there is a big concentration of data points on the left side of the graph, they are evenly spread apart from the y axis of 0 so it meets the assumption of equal variance.

6. Is DIABETES a risk factor for Tachycardia? Justify your answer and explain any effect of DIABETES on heart rate (include uncertainty in your conclusions).

```
coef(t.model)
```

```
##   (Intercept)          SEX2       TOTCHOL           AGE         SYSBP
##  3.964716e+00  3.828338e-02  1.575581e-04  9.567510e-04  4.904051e-04
##         DIABP     CURSMOKE1           BMI     DIABETES1       BPMEDS1
##  1.293125e-03  2.967359e-02  8.444734e-05  1.084218e-02 -2.119535e-02
##       GLUCOSE
##  7.454705e-04
```

```
diabetes <- matrix(c(1,0,0,0,0,0,0,0,1,0,0), nrow=1)
no_diabetes <- matrix(c(1,0,0,0,0,0,0,0,0,0,0), nrow=1)
diff_d <- diabetes - no_diabetes
a <- glht(t.model, linfct = diff_d, alternative = "two.sided")
summary(a)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = t, correlation = corSymm(form = ~1 |
##     RANDID), method = "ML")
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(>|z|)
## 1 == 0  0.01084    0.01213   0.894    0.371
## (Adjusted p values reported -- single-step method)
```

```
confint(a)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = t, correlation = corSymm(form = ~1 |
##     RANDID), method = "ML")
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate lwr      upr
## 1 == 0  0.01084 -0.01293  0.03462
```

Ho: Diabetes is not a risk factor for Tachycardia.

Ha: Diabetes is a risk factor for Tachycardia.

Based on the p-value of 0.371, we fail to reject the null hypothesis that it is a risk factor and conclude that diabetes is not a risk factor for Tachycardia.

We also found the confidence interval to be (-0.01293,0.03462) which includes 0 in the interval. This means that there is not a significance and we conclude that diabetes is not a risk factor for Tachycardia.

7. What is the expected difference in heart rate for a female patient with at age 35 who is a smoker vs. an older female of 45 but not a smoker (assume the other characteristics are the same)? What does this say about the effect of smoking?

```
no_smoke <- matrix(c(1,0,0,45,0,0,0,0,0,0,0), nrow=1)
smoke <- matrix(c(1,0,0,35,0,0,1,0,0,0,0), nrow=1)
diff_d1 <- smoke - no_smoke
b <- glht(t.model, linfct = diff_d1, alternative = "two.sided")
summary(b)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = t, correlation = corSymm(form = ~1 |
##     RANDID), method = "ML")
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(>|z|)
## 1 == 0 0.020106   0.005249   3.831 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(b)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: gls(model = log(HEARTRTE) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
##     CURSMOKE + BMI + DIABETES + BPMEDS + GLUCOSE, data = t, correlation = corSymm(form = ~1 |
##     RANDID), method = "ML")
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate lwr      upr
## 1 == 0 0.020106 0.009819 0.030393
```

```
exp(0.020106)
```

```
## [1] 1.020309
```

```
c <- confint(b)
exp(c$confint[c(1,2,3)])
```

```
## [1] 1.020310 1.009867 1.030860
```

Ho: The difference of the log heart rates between a 35 year old female patient that smokes and a 45 year old female patient that does not smoke is 0.

Ha: The difference of the log heart rates between a 35 year old female patient that smokes and a 45 year old female patient that does not smoke is not 0.

Based on the p-value of 0.000128, we reject the null hypothesis and say that the difference of the log heart rates between a 35 year old female patient that smokes and a 45 year old female patient that does not smoke is not 0. This conclusion can also be supported by the confidence interval that we calculated which is (0.009819,0.030393). Since 0 is not included in the confidence interval, we can conclude the same and reject the null.