Cardiovascular

Naomi Zubeldia and Drew Millane

2023-03-14

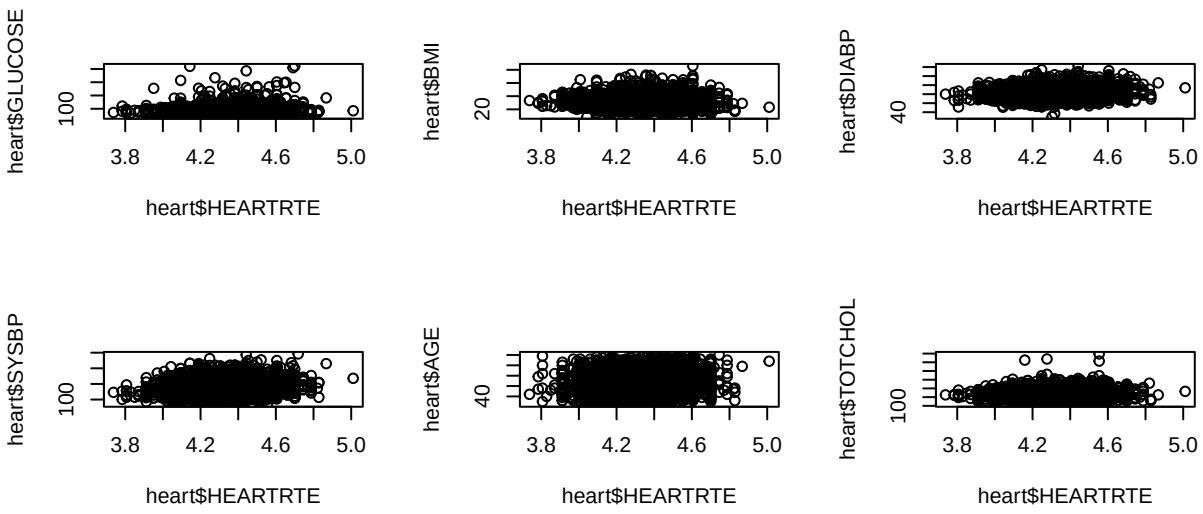
```
#reading in libraries
library(tidyverse)
library(nlme)
library(grid)
library(gridExtra)
```

1. Explanatory Data Analysis

```
#reading in data and changing things to factors
heart<-read.table('Tachycardia.txt',header=T)
heart$HEARTRTE<-log(heart$HEARTRTE)
heart$SEX<-as.factor(heart$SEX)
heart$CURSMOKE<-as.factor(heart$CURSMOKE)
heart$DIABETES<-as.factor(heart$DIABETES)
heart$BPMEDS<-as.factor(heart$BPMEDS)
```

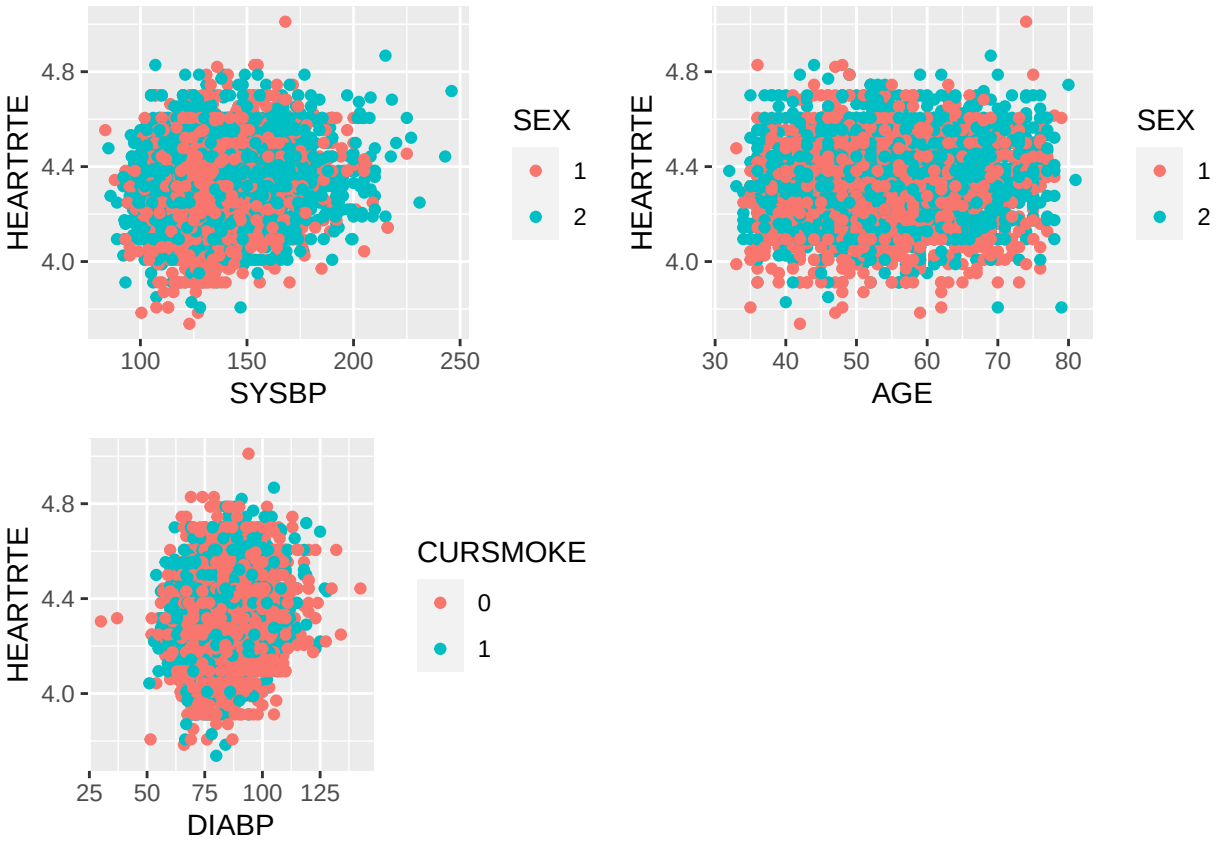
To observe how the log transformed heart rate variable interacts with the explanatory variables, we created the plots below. There does not seem to be any linear relationships between the variables. It is interesting that alot of the observations seem to be spread along the bottom of most of the plots, with an exception of age that is spread all over.

```
par(mfrow=c(3,3))
#plots of explanatory variables
plot(heart$HEARTRTE,heart$GLUCOSE)
plot(heart$HEARTRTE,heart$BMI)
plot(heart$HEARTRTE,heart$DIABP)
plot(heart$HEARTRTE,heart$SYSBP)
plot(heart$HEARTRTE,heart$AGE)
plot(heart$HEARTRTE,heart$TOTCHOL)
```



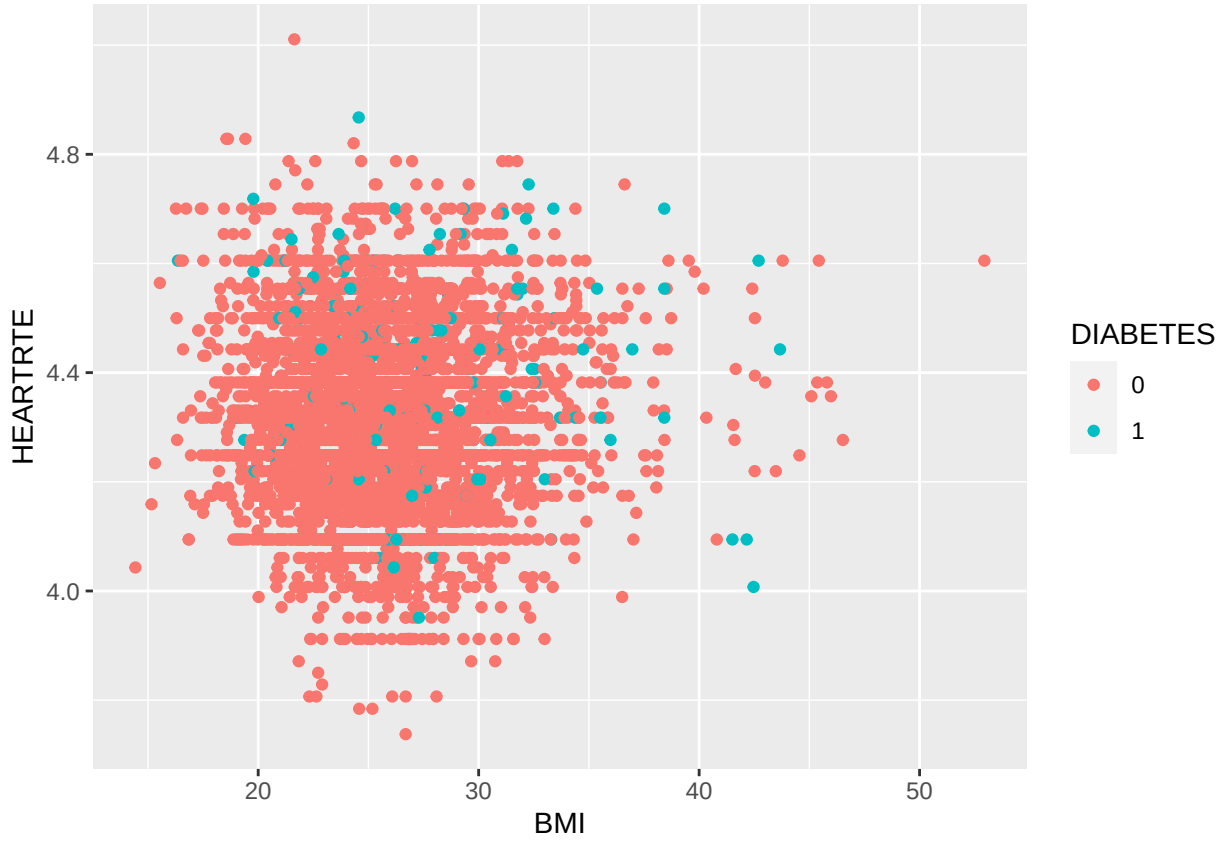
We wanted to see the if gender or smoking was a dominate influence on these observations. It looks like from the plots below that it was pretty even.

```
#observing three variables from data
p1<-ggplot(mapping=aes(x=SYSBP,y=HEARTRTE,color=SEX),data=heart)+geom_point()

p2<-ggplot(mapping=aes(x=AGE,y=HEARTRTE,color=SEX),data=heart)+geom_point()

p3<-ggplot(mapping=aes(x=DIABP,y=HEARTRTE,color=CURSMOKE),data=heart)+geom_point()


grid.arrange(p1,p2,p3, ncol=2)
```



Since people who are more overweight tend to have diabetes, we wanted to observe that in the data in the plot below. It looks like most of the people in this study do not have diabetes and those who do have a pretty even spread.

```
#looking at how diabetes influences bmi
ggplot(mapping=aes(x=BMI,y=HEARTRTE,color=DIABETES),data=heart)+geom_point()
```



3. Basic Linear Model

To learn what model is best for this data, the intuitive approach is to fit a basic linear model. We then checked if the residuals are correlated to get a quick answer if this is a good model for the data.

```
#setting up basic linear model
base<- lm(HEARTRTE~.-RANDID-PERIOD,data=heart)

#correlation matrix of base
round(cor(matrix(resid(base),1734,3,byrow=T)),2)
```

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.46 0.39
## [2,] 0.46 1.00 0.50
## [3,] 0.39 0.50 1.00
```

From the correlation matrix above, it appears that there is quite a bit of correlation in this basic model, which means it may not be the best fit.

3. Testing Correlation Structure

Since the basic linear model did not fit the data, we tested a couple different gls models with varying correlation structures to get the best fit. We tested if the AR1, MA1 or general symmetric correlations were the best.

```
#gls models with AR1, MA1, and symmetric correlations
ar<-gls(model=HEARTRTE~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOSE, data=heart, co

ma<-gls(model=HEARTRTE~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOSE, data=heart, co

sy<-gls(model=HEARTRTE~SEX+TOTCHOL+AGE+SYSBP+DIABP+CURSMOKE+BMI+DIABETES+BPMEDS+GLUCOSE, data=heart, co
```

After fitting the different models using the maximum likelihood method, we compared the different AIC values of each model, which we extracted below. The model with the smallest AIC was the general symmetrically correlated model, which makes it the best model for the data.

```
#finding the AIC of the gls models above
AIC(ar)
```

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

```
AIC(ma)
```

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

```
AIC(sy)
```

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

##4. Logintudinal Model Notation

The longitudinal model is explained below:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{B})$$

$\boldsymbol{y}$: A nx1 vector of our response variable. In this case, this is the heart rate of the patient. Each observation in the data set has a row in this vector with the corresponding heart rate.

$\boldsymbol{X}$: A model matrix of explanatory variables (covariates). In this case, we first have a column of 1's for the intercept and an additional column for each explanatory variable. We have n rows in this matrix, where each row corresponds to a different observation.

$\boldsymbol{\beta}$: A vector of coefficients for each explanatory variable. In our case, these are the coefficients representing the intercept and the effect of the explanatory factors.

σ^2: This represents the variability of $\boldsymbol{y}$ about the regression line.

$\boldsymbol{B}$: This is a matrix that represents the correlation between measurements within each patient. It has 0's on the off-diagonal and sub matrices on the diagonal. These matrices which we will call R represent the correlation of measurements within each patient block. Since the model that fits the data the best is a general symmetric model, each R matrix has 1's on the diagonal and correlations between measurements within subjects on the off diagonal. Overall, matrix B will have 1's on the diagonal due to each R matrix having 1's along their diagonals. ## 5. Fitting the Model

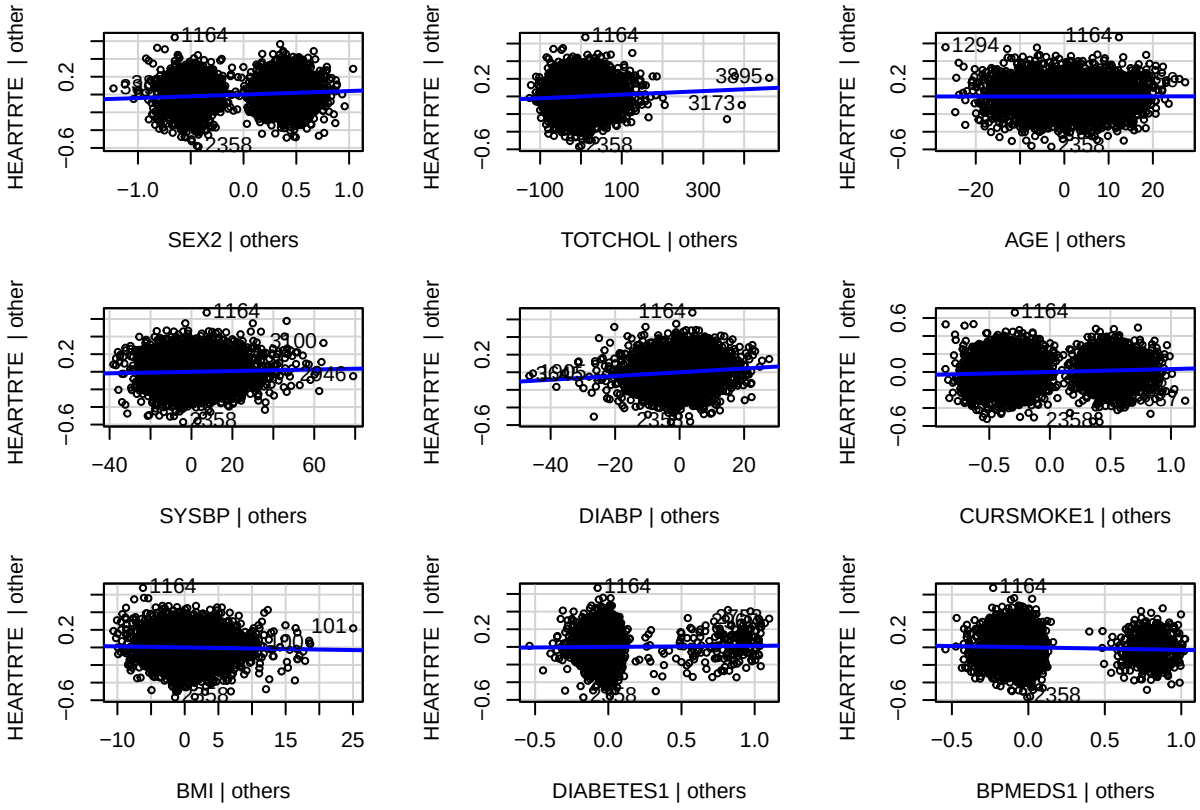
```
#gls model with general symmetric correlation
gls <- gls(model=HEARTRTE~.-RANDID-PERIOD, data=heart, correlation=corSymm(form=~1:3|RANDID), method="M
```

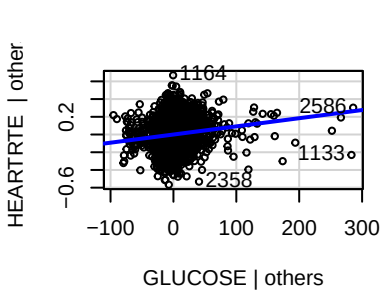
After fitting the longitudinal model with the best correlation parameters, we need to check the LINE assumptions to make sure that our model is valid.

```r
#Testing line assumptions

library(car)
```

```
## Warning: package 'car' was built under R version 4.2.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.2

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```r
source("../STAT469/glstools-master/glstools-master/stdres.gls.R")
#residuals
sres <- stdres.gls(gls)

#linearity assumption
avPlots(base)
```


HEARTRTE | other
SEX2 | others
TOTCHOL | others
AGE | others
SYSBP | others
DIABP | others
CURSMOKE1 | others
BMI | others
DIABETES1 | others
BPMEDS1 | others

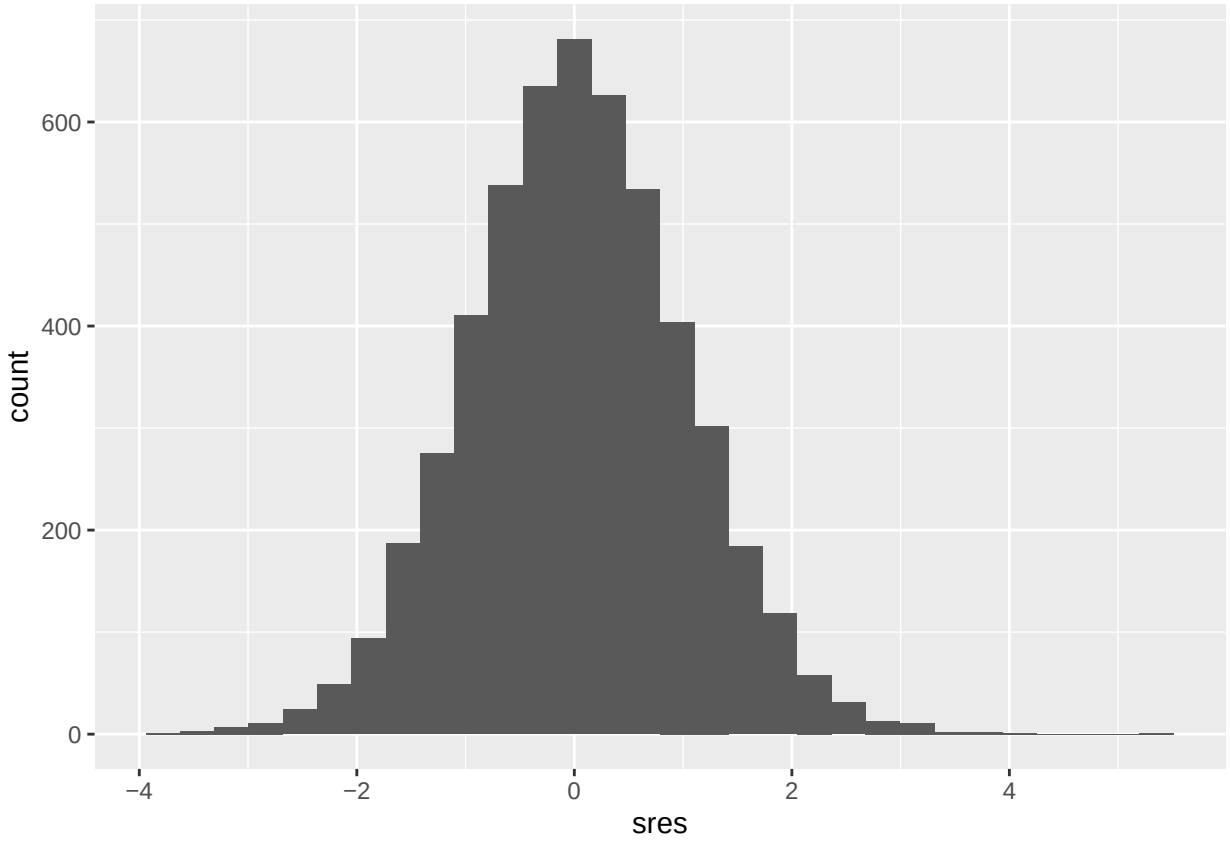



```
#Independence
resid <- matrix(sres, nrow = 1734, ncol = 3,byrow=T)
cor(resid)
```

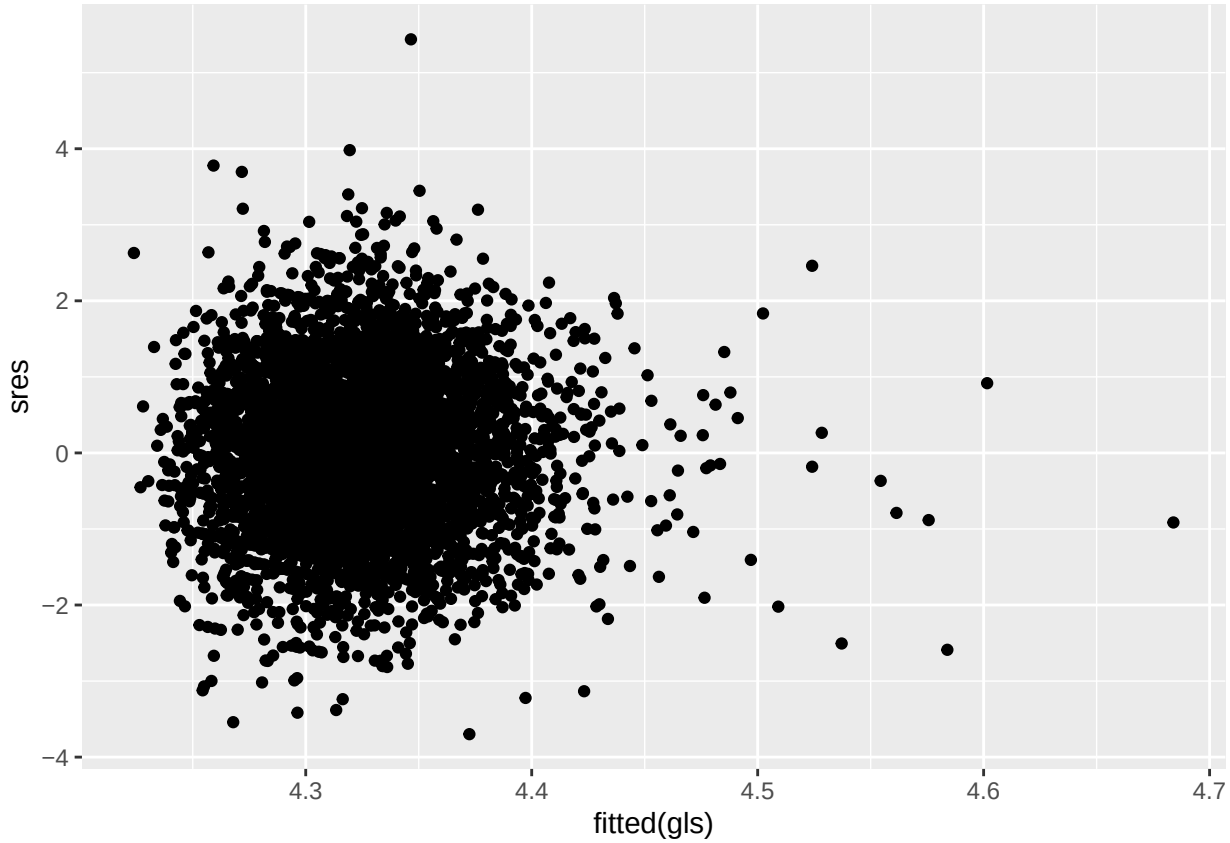
```
##              [,1]         [,2]         [,3]
## [1,] 1.000000000  0.010689327  0.005603399
## [2,] 0.010689327  1.000000000 -0.004182295
## [3,] 0.005603399 -0.004182295  1.000000000
```

```
#normality assumption
ggplot(data = heart)+geom_histogram(mapping=aes(x=sres))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Equal variance
ggplot(data = heart) +
  geom_point(mapping = aes(x = fitted(gls), y = sres))
```



In the added variable plot above, we can see that the variables have a positive relationship, which validates the normality assumption. The correlation matrix of standardized residuals shows that there is little correlation, which makes the observations independent meeting the next assumption. We then plotted a histogram of the standardized residuals, which is distributed normally meeting normality. Lastly, we plotted the fitted values versus the standardized residuals and the spread is pretty normal, which means the equal variance assumption is met. ## 6. Is Diabetes a risk factor?

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.2.2
```

```
#performing t test on diabetes
summary(gls)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: HEARTRTE ~ . - RANDID - PERIOD
##   Data: heart
##         AIC       BIC   logLik
##   -5939.495 -5841.143 2984.748
##
## Correlation Structure: General
##  Formula: ~1:3 | RANDID
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.473
```

```
## 3 0.397 0.510
##
## Coefficients:
##                  Value   Std.Error   t-value p-value
## (Intercept)  3.964716 0.028251412 140.33691  0.0000
## SEX2         0.038283 0.005971589   6.41092  0.0000
## TOTCHOL      0.000158 0.000051645   3.05077  0.0023
## AGE          0.000957 0.000288883   3.31190  0.0009
## SYSBP        0.000490 0.000163155   3.00576  0.0027
## DIABP        0.001293 0.000282446   4.57831  0.0000
## CURSMOKE1    0.029674 0.005052659   5.87287  0.0000
## BMI          0.000084 0.000732499   0.11529  0.9082
## DIABETES1    0.010842 0.012130455   0.89380  0.3715
## BPMEDS1     -0.021195 0.007430880  -2.85233  0.0044
## GLUCOSE      0.000745 0.000092794   8.03360  0.0000
##
##  Correlation:
##           (Intr) SEX2   TOTCHO AGE    SYSBP  DIABP  CURSMO BMI    DIABET BPMEDS
## SEX2      -0.192
## TOTCHOL   -0.234 -0.122
## AGE       -0.421  0.047 -0.138
## SYSBP      0.051 -0.076  0.038 -0.427
## DIABP     -0.327  0.088 -0.135  0.324 -0.687
## CURSMOKE1 -0.302  0.122 -0.030  0.217  0.003  0.027
## BMI       -0.484  0.122 -0.073 -0.004 -0.017 -0.174  0.118
## DIABETES1  0.112  0.030  0.006 -0.066 -0.054  0.034 -0.010  0.013
## BPMEDS1    0.131 -0.052 -0.006 -0.128 -0.109  0.034  0.001 -0.029 -0.029
## GLUCOSE   -0.223  0.002  0.059 -0.080 -0.067  0.069  0.010 -0.051 -0.331  0.012
##
## Standardized residuals:
##          Min           Q1          Med           Q3          Max
## -3.667883599 -0.646135632 -0.009491491  0.654628952  4.409208848
##
## Residual standard error: 0.150611
## Degrees of freedom: 5202 total; 5191 residual
```

```
intervals(gls,0.95)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                     lower          est.         upper
## (Intercept)  3.909331e+00  3.964716e+00  4.0201004764
## SEX2         2.657655e-02  3.828338e-02  0.0499902093
## TOTCHOL      5.631156e-05  1.575581e-04  0.0002588047
## AGE          3.904189e-04  9.567510e-04  0.0015230831
## SYSBP        1.705522e-04  4.904051e-04  0.0008102579
## DIABP        7.394116e-04  1.293125e-03  0.0018468380
## CURSMOKE1    1.976825e-02  2.967359e-02  0.0395789316
## BMI         -1.351559e-03  8.444733e-05  0.0015204537
## DIABETES1   -1.293862e-02  1.084218e-02  0.0346229781
## BPMEDS1     -3.576300e-02 -2.119535e-02 -0.0066276941
## GLUCOSE      5.635549e-04  7.454705e-04  0.0009273861
##
```

```
##  Correlation structure:
##              lower      est.     upper
## cor(1,2) 0.4351377 0.4727194 0.5086567
## cor(1,3) 0.3555596 0.3971656 0.4372006
## cor(2,3) 0.4748022 0.5103756 0.5442841
##
##  Residual standard error:
##     lower      est.     upper
## 0.1471653 0.1506110 0.1541374
```

To determine whether diabetes is a risk factor for Tachycardia we performed a hypothesis test. With a p-value of 0.3715, we conclude that diabetes does not have a significant effect on heart rate. We also calculated a 95% confidence interval for diabetes. With 95% confidence, we conclude that as the diabetes unit increases by 1 the heart rate will change between -0.01 and 0.03. 0 is contained in this interval so we conclude that it does not have a significant effect on heart rate.

7. Difference in smoking

```
#vectors of each patient
b <- t(c(1,1,0,35,0,0,1,0,0,0,0))
c <- t(c(1,1,0,45,0,0,0,0,0,0,0))

#difference
d <- b-c

#Testing the difference
my.test <- glht(gls, linfct=d, alternative="two.sided")
summary(my.test)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = HEARTRTE ~ . - RANDID - PERIOD, data = heart, correlation = corSymm(form = ~1:3 |
##     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)
```

```
#finding the confidence interval
confint(my.test, .95)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: gls(model = HEARTRTE ~ . - RANDID - PERIOD, data = heart, correlation = corSymm(form = ~1:3 |
##     RANDID), method = "ML")
##
```

```
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate lwr      upr
## 1 == 0 0.020106 0.009819 0.030393
```

With 95% confidence, we expect the difference between two women (one who smokes and one who does not) to be between 0.009 and 0.03. Smoking has an effect on increasing heart rate and possibly causing heart problems in individuals.