Estadística avanzada para ciencia de datos

Estadística bayesiana - Proyecto de Contrastes, Regresión e Inferencia Bayesiana

Jakub Maciążek

Project description

Introduction

We will explore a dataset of volumetric brain measurements of a number of individuals, which fall into two groups (according to the variable CLASS): the value HEALTHY indicates a healthy individual, while the value AD indicates Alzheimer's disease (AD stands for Alzheimer's Disease).

These data come from calculating around 230 morphometric estimates (volume and thickness of anatomical brain regions) of some 260 individuals, belonging to the OASIS project.

Objective

The purpose of this project is to use the advanced statistical techniques seen in class to:

- Establish the existing relationship between volumetric variables and the age and sex of each individual.
- Discover markers among volumetric measures of cognitive status (healthy, Alzheimer's)
- Generate a classifier model of cognitive status to predict, from the volumetric variables, whether an individual has Alzheimer's or is healthy.
- Conduct a small Bayesian analysis of the epidemiology/incidence of Alzheimer's disease.

Dataset analysis

Analysis was started with overview of the dataset. Alzheimer.csv contains 262 records of individual patients. Each record has specified 228 properties. Two of them are of categorical type:

- SEX (MALE, FEMALE): gender of a patient.
- CLASS (HEALTHY, AD): health status of a patient, meaning healthy or with Alzheimer's disease.

Remaining 226 variables are of type double and all except one (AGE - age of patient) describe different brain parameters.

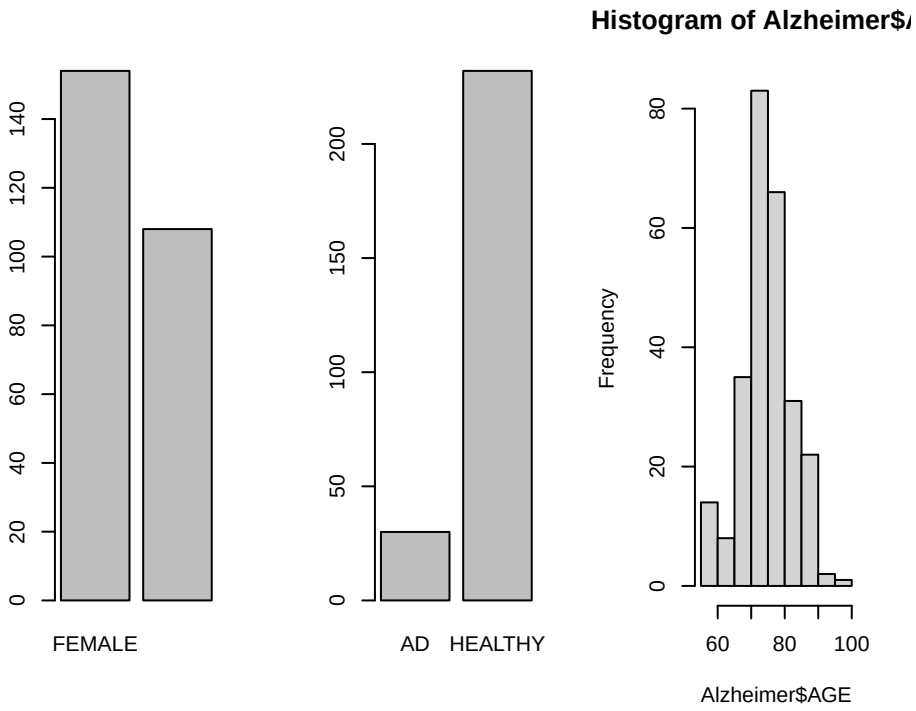
```
Alzheimer <- read_csv("Alzheimer.csv")
## Rows: 262 Columns: 228
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): SEX, CLASS
## dbl (226): AGE, BRAIN_VOLUME, GM_VOLUME, WM_VOLUME, CSF_VOLUME, GM_BRAIN_QUO...
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
head(Alzheimer)
## # A tibble: 6 x 228
##     AGE SEX    BRAIN_VOLUME GM_VOLUME WM_VOLUME CSF_VOLUME GM_BRAIN_QUOTIENT
##   <dbl> <chr>         <dbl>     <dbl>     <dbl>      <dbl>             <dbl>
## 1  68   MALE          1411.      599.      433.       379.             0.425
## 2  82.9 FEMALE        1368.      552.      446.       370.             0.403
## 3  72.8 FEMALE        1154.      444.      393.       317.             0.385
## 4  73   FEMALE        1547.      665.      492.       390.             0.430
## 5  71.1 FEMALE        1326.      559.      433.       334.             0.422
## 6  78.1 MALE          1400.      526.      441.       433.             0.376
## # i 221 more variables: WM_BRAIN_QUOTIENT <dbl>, CSF_BRAIN_QUOTIENT <dbl>,
## #   GM_WM_QUOTIENT <dbl>, PRECENTRAL_L_VOLUME <dbl>, PRECENTRAL_R_VOLUME <dbl>,
## #   FRONTAL_SUP_L_VOLUME <dbl>, FRONTAL_SUP_R_VOLUME <dbl>,
## #   FRONTAL_SUP_ORB_L_VOLUME <dbl>, FRONTAL_SUP_ORB_R_VOLUME <dbl>,
## #   FRONTAL_MID_L_VOLUME <dbl>, FRONTAL_MID_R_VOLUME <dbl>,
## #   FRONTAL_MID_ORB_L_VOLUME <dbl>, FRONTAL_MID_ORB_R_VOLUME <dbl>,
## #   FRONTAL_INF_OPER_L_VOLUME <dbl>, FRONTAL_INF_OPER_R_VOLUME <dbl>, ...
```

Brief descriptive statistic of dataset population

Dataset contains information for people over 50 years old. They are not equally represented by age. There is especially small representation for people over 90 years old. Both genders are fairly equally represented. Most of the data corresponds to healthy individuals, with only ~35 records corresponding to people with Alzheimer's disease.

```
par(mfrow=c(1,3))
barplot(table(Alzheimer$SEX))
barplot(table(Alzheimer$CLASS))
hist(Alzheimer$AGE)
```



Hypothesis testing

The aim is to establish markers or indicators of an individual's condition. Basically, we want to test whether brain atrophy is specific to Alzheimer's disease, compared to healthy subjects.

Those tests will be conducted for general brain volume, as well as white and gray matter. Also, independence of health status from gender will be studied.

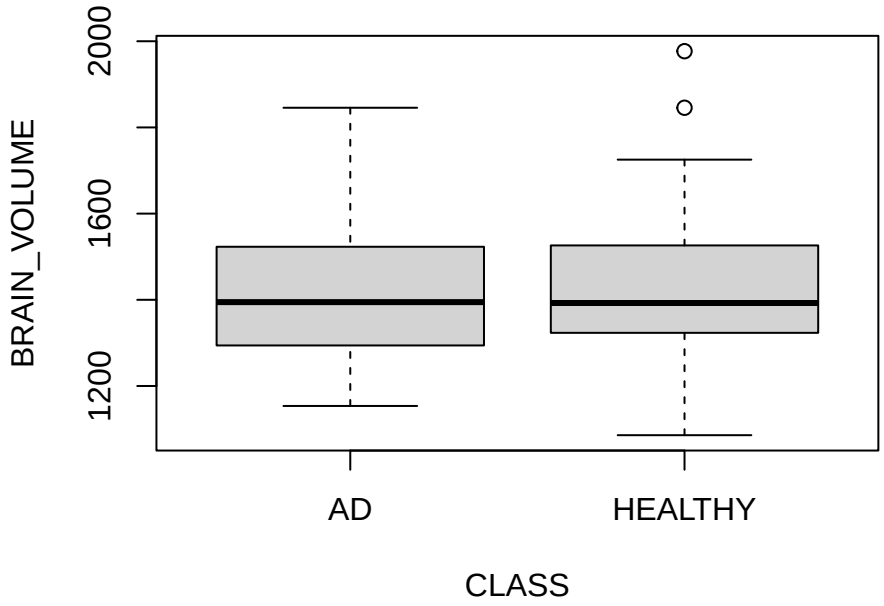
Test I: BRAIN_VOLUME & CLASS

In the first test it was examined, whether brain atrophy (loss of neurons/capacity/volume) is specific to Alzheimer's disease. In order to do that, it was tested whether mean brain volume is significantly lower for sick individuals. Independent Welsh T-test was conducted, as samples are independent and variances unknown.

First test results with *p-value* of 0.445. It means, that in general mean of brain volume for sick individuals does not significantly differ from healthy ones. It can be also observed on the boxplots comparison, where can be seen that IQR ranges and means are almost the same

```
t.test(Alzheimer$BRAIN_VOLUME ~ Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Alzheimer$BRAIN_VOLUME by Alzheimer$CLASS
## t = -0.13922, df = 35.082, p-value = 0.445
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
```

```
##      -Inf 48.50606
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              1417.723              1422.079

boxplot(BRAIN_VOLUME~CLASS, data=Alzheimer)
```

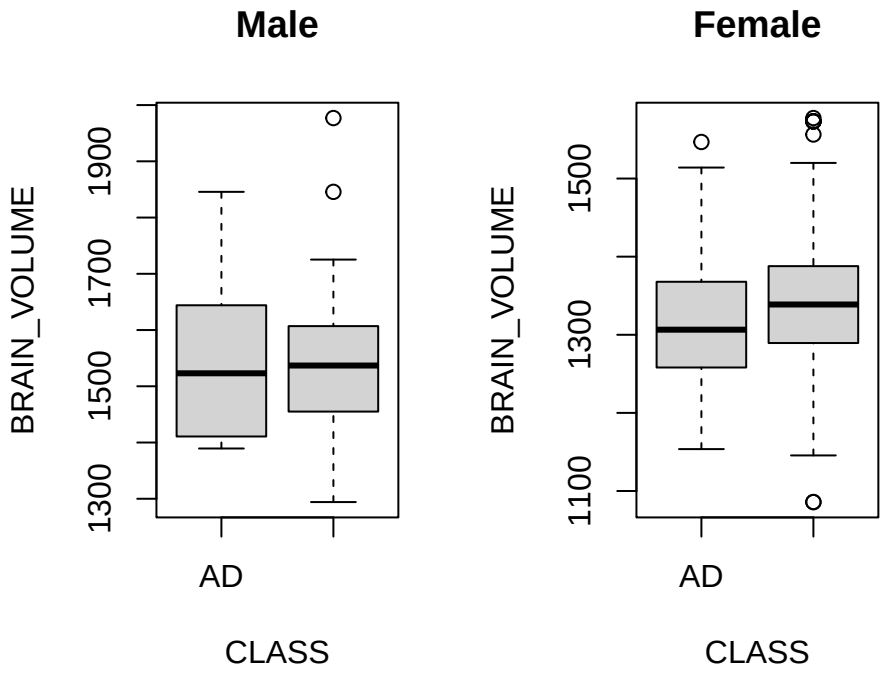


Result does not differ for groups distincted by gender, and there is no significant difference neither for men nor woman. Again, also on the boxplots it can be seen that means are almost equal.

Therefore, brain volume does not indicate individual's condition.

```
Male_Alzheimer <- filter(Alzheimer, SEX == "MALE")
#head(Male_Alzheimer)
Female_Alzheimer <- filter(Alzheimer, SEX == "FEMALE")
#head(Female_Alzheimer)

t.test(Male_Alzheimer$BRAIN_VOLUME ~ Male_Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Male_Alzheimer$BRAIN_VOLUME by Male_Alzheimer$CLASS
## t = 0.023782, df = 14.19, p-value = 0.5093
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##      -Inf 74.43254
## sample estimates:
##      mean in group AD mean in group HEALTHY
```

```
##              1542.332              1541.340
t.test(Female_Alzheimer$BRAIN_VOLUME ~ Female_Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Female_Alzheimer$BRAIN_VOLUME by Female_Alzheimer$CLASS
## t = -0.64776, df = 19.635, p-value = 0.2623
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##      -Inf 28.21545
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              1322.434              1339.380

par(mfrow=c(1,2))
boxplot(BRAIN_VOLUME~CLASS, data=Male_Alzheimer, main="Male")
boxplot(BRAIN_VOLUME~CLASS, data=Female_Alzheimer, main="Female")
```



T-tests basic assumptions

In order to conduct the tests, compared samples need to come from normal distribution. This was checked with Shapiro-Wilk test. Equality of variances was not tested, therefore Welsh t-test was used.

In the test, null hypothesis states that *"sample distribution is normal"*. Therefore, if *p-value* of the test is greater than 0.05, sample distribution does not significantly differ from normal distribution.

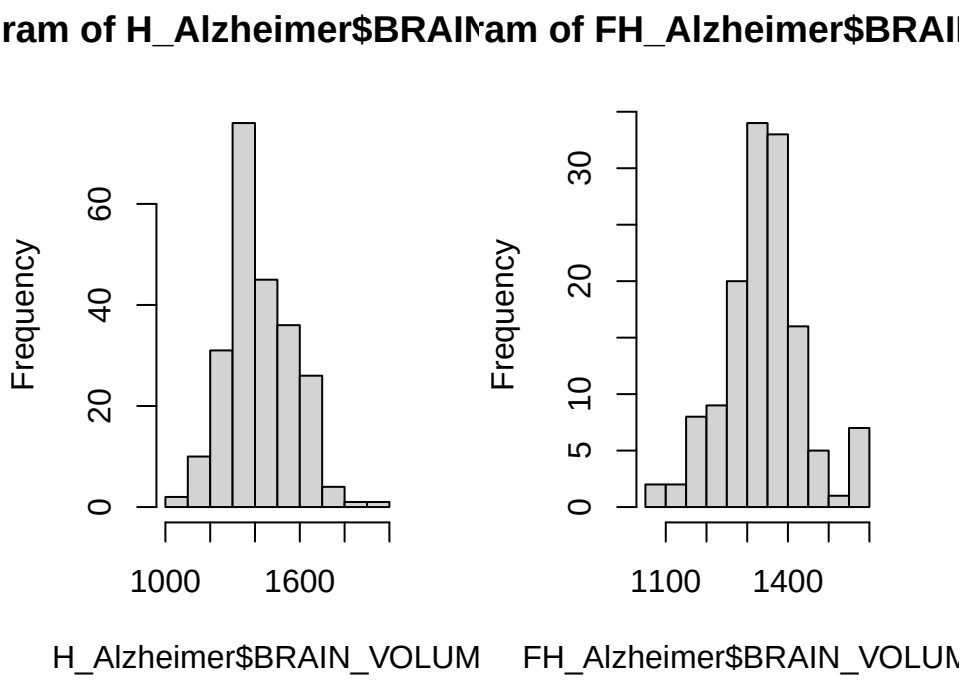
From below results, it can be seen that samples pass tests in all cases, exceplt healthy samples for general and woman population.

```r
# Normal distibution of brain volume by health status.

AD_Alzheimer <- filter(Alzheimer, CLASS == "AD")
H_Alzheimer <- filter(Alzheimer, CLASS == "HEALTHY")

shapiro.test(AD_Alzheimer$BRAIN_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  AD_Alzheimer$BRAIN_VOLUME
## W = 0.95001, p-value = 0.1692
shapiro.test(H_Alzheimer$BRAIN_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  H_Alzheimer$BRAIN_VOLUME
## W = 0.97826, p-value = 0.001233

# Normal distibution of brain volume by health status among different genders.

MAD_Alzheimer <- filter(Male_Alzheimer, CLASS == "AD")
MH_Alzheimer <- filter(Male_Alzheimer, CLASS == "HEALTHY")

shapiro.test(MAD_Alzheimer$BRAIN_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  MAD_Alzheimer$BRAIN_VOLUME
## W = 0.89677, p-value = 0.1208
shapiro.test(MH_Alzheimer$BRAIN_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  MH_Alzheimer$BRAIN_VOLUME
## W = 0.97412, p-value = 0.0565

FAD_Alzheimer <- filter(Female_Alzheimer, CLASS == "AD")
FH_Alzheimer <- filter(Female_Alzheimer, CLASS == "HEALTHY")

shapiro.test(FAD_Alzheimer$BRAIN_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  FAD_Alzheimer$BRAIN_VOLUME
## W = 0.93909, p-value = 0.3076
shapiro.test(FH_Alzheimer$BRAIN_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  FH_Alzheimer$BRAIN_VOLUME
## W = 0.97237, p-value = 0.00695
```

From below plots it can be seen that the problem is related to samples that have more than 50 observations,

and data seem to more or less follow the distribution. Therefore it was assumed that lack of normality will not affect test result significantly. (Alternatively, two-samples Wilcoxon rank test could be used, as it is suitable for data from non-normal distribution.)

```
par(mfrow=c(1,2))
hist(H_Alzheimer$BRAIN_VOLUME)
hist(FH_Alzheimer$BRAIN_VOLUME)
```

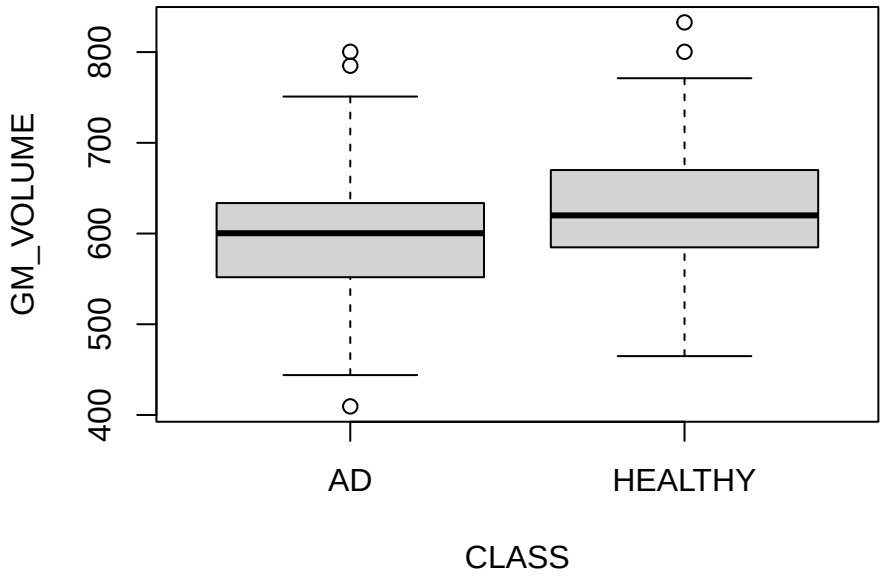


Test II: GM_VOLUME & CLASS

Tests for Grey matter were conducted with alike assumptions. For general population *p-value* of 0.08233 was received. It is close to usual significance level of 0.05, but still suggests that mean grey matter volume is not significantly lower for sick individuals. At the following boxplots it can be confirmed, that means and general ranges are very similar, with only IQR ranges being slightly different.

```
t.test(Alzheimer$GM_VOLUME ~ Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Alzheimer$GM_VOLUME by Alzheimer$CLASS
## t = -1.4209, df = 33.214, p-value = 0.08233
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##      -Inf 4.386372
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              603.4151              626.4035
```

```
boxplot(GM_VOLUME~CLASS, data=Alzheimer)
```



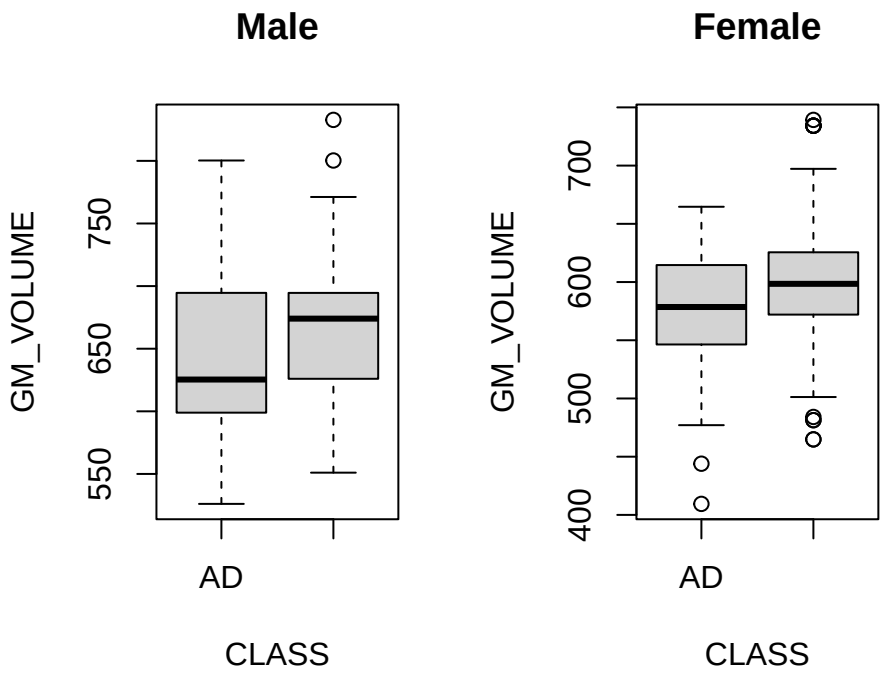
For man subgroup, test *p-value* of 0.2781 means results alike to general population. However, female subgroup with test "p-value" of 0.04279 (lower than significance level of 0.05) means that mean brain grey matter volume for sick woman is significantly lower than for healthy ones.

Therefore grey matter volume can be used as disease indicator for woman.

However, in men's boxplot huge difference in means is observed, which together with close p-value for general population can mean that gray matter may be a correct indicator for general population.

```
t.test(Male_Alzheimer$GM_VOLUME ~ Male_Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Male_Alzheimer$GM_VOLUME by Male_Alzheimer$CLASS
## t = -0.60339, df = 13.525, p-value = 0.2781
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##      -Inf 27.67754
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              649.8218              664.1903
t.test(Female_Alzheimer$GM_VOLUME ~ Female_Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Female_Alzheimer$GM_VOLUME by Female_Alzheimer$CLASS
```

```
## t = -1.8168, df = 18.405, p-value = 0.04279
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##       -Inf -1.505607
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              567.9276              600.2010

par(mfrow=c(1,2))
boxplot(GM_VOLUME~CLASS, data=Male_Alzheimer, main="Male")
boxplot(GM_VOLUME~CLASS, data=Female_Alzheimer, main="Female")
```



T-tests basic assumptions

Again, in al cases except health females, p-value is greater than *0.05* indicating that sample distribution does not significantly differ from normal one.

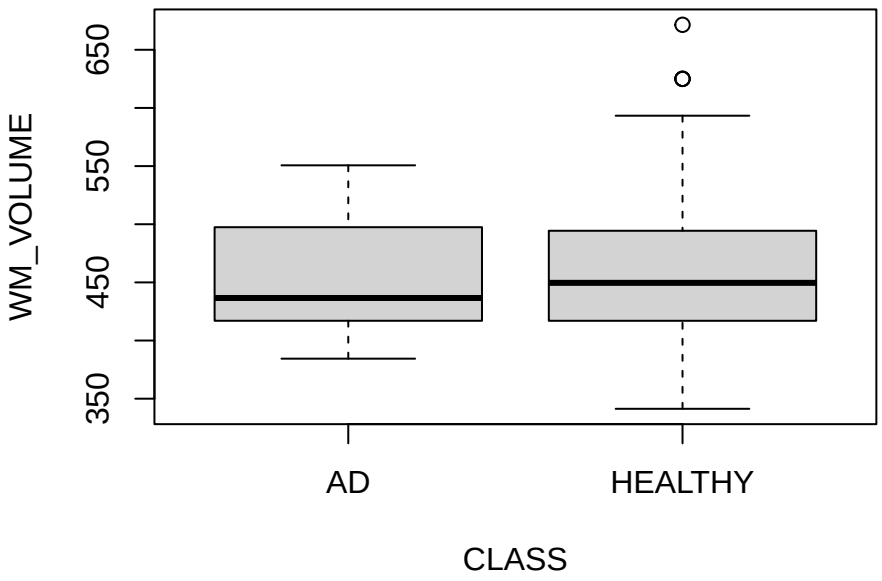
```
# Normal distibution of brain gray matter volume by health status.
shapiro.test(AD_Alzheimer$GM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  AD_Alzheimer$GM_VOLUME
## W = 0.95265, p-value = 0.1989
shapiro.test(H_Alzheimer$GM_VOLUME)
##
##  Shapiro-Wilk normality test
##
```

```
## data:  H_Alzheimer$GM_VOLUME
## W = 0.99256, p-value = 0.2938

# Normal distibution of brain gray matter volume by health status among different genders.
shapiro.test(MAD_Alzheimer$GM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  MAD_Alzheimer$GM_VOLUME
## W = 0.90104, p-value = 0.1382
shapiro.test(MH_Alzheimer$GM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  MH_Alzheimer$GM_VOLUME
## W = 0.97969, p-value = 0.147

shapiro.test(FAD_Alzheimer$GM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  FAD_Alzheimer$GM_VOLUME
## W = 0.92441, p-value = 0.1753
shapiro.test(FH_Alzheimer$GM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  FH_Alzheimer$GM_VOLUME
## W = 0.97345, p-value = 0.008894
```

Test III: WM_VOLUME & CLASS

Neither for whole population, nor for those samples divided by gender, mean brain white matter volume is significantly smaller for sick individuals.

```
t.test(Alzheimer$WM_VOLUME ~ Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Alzheimer$WM_VOLUME by Alzheimer$CLASS
## t = -0.5758, df = 39.563, p-value = 0.284
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##      -Inf 10.82296
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              452.6765              458.2984

boxplot(WM_VOLUME~CLASS, data=Alzheimer)
```

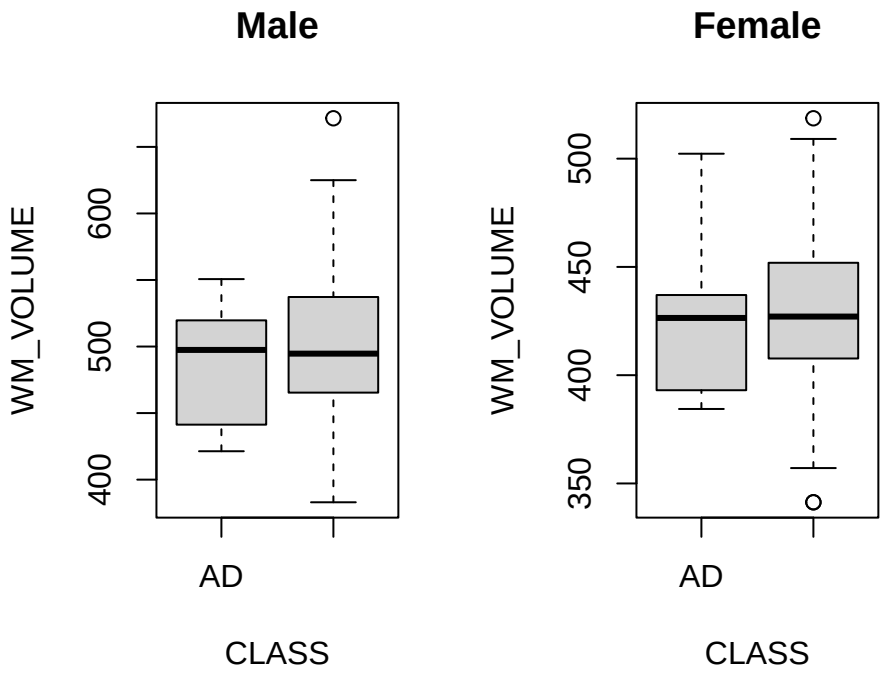


Moreover, even though on boxplots can be seen IQR ranges differ for health groups when divided by gender, their mean value is very similar.

White matter volume can not be used as health indicator.

```
t.test(Male_Alzheimer$WM_VOLUME ~ Male_Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Male_Alzheimer$WM_VOLUME by Male_Alzheimer$CLASS
## t = -1.0052, df = 17.727, p-value = 0.1642
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##      -Inf 9.656863
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              488.1153              501.4072
t.test(Female_Alzheimer$WM_VOLUME ~ Female_Alzheimer$CLASS, alternative="less")
##
##  Welch Two Sample t-test
##
## data:  Female_Alzheimer$WM_VOLUME by Female_Alzheimer$CLASS
## t = -0.31405, df = 20.071, p-value = 0.3784
## alternative hypothesis: true difference in means between group AD and group HEALTHY is less than 0
## 95 percent confidence interval:
##     -Inf 12.7055
## sample estimates:
##      mean in group AD mean in group HEALTHY
##              425.5762              428.4054
```

```
par(mfrow=c(1,2))
boxplot(WM_VOLUME~CLASS, data=Male_Alzheimer, main="Male")
boxplot(WM_VOLUME~CLASS, data=Female_Alzheimer, main="Female")
```



T-tests basic assumptions

Samples divided by genders follow normal distribution, however general ones differ from it, especially sample of healthy individuals differ very much with p-value of *5.47e-06*.

```
# Normal distibution of brain gray matter volume by health status.
shapiro.test(AD_Alzheimer$WM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  AD_Alzheimer$WM_VOLUME
## W = 0.92437, p-value = 0.03488
shapiro.test(H_Alzheimer$WM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  H_Alzheimer$WM_VOLUME
## W = 0.96075, p-value = 5.47e-06

# Normal distibution of brain gray matter volume by health status among different genders.
shapiro.test(MAD_Alzheimer$WM_VOLUME)
##
##  Shapiro-Wilk normality test
##
```

```
## data:  MAD_Alzheimer$WM_VOLUME
## W = 0.91526, p-value = 0.2163
shapiro.test(MH_Alzheimer$WM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  MH_Alzheimer$WM_VOLUME
## W = 0.9826, p-value = 0.2398

shapiro.test(FAD_Alzheimer$WM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  FAD_Alzheimer$WM_VOLUME
## W = 0.89533, p-value = 0.05678
shapiro.test(FH_Alzheimer$WM_VOLUME)
##
##  Shapiro-Wilk normality test
##
## data:  FH_Alzheimer$WM_VOLUME
## W = 0.98718, p-value = 0.2332
```

Test IV: SEX & CLASS independence

Based on a test *p-value* greater than significance level of 0.05, we cannot reject null hypothesis, which states that variables are independent and not correlated.

```
chisq.test(Alzheimer$SEX, Alzheimer$CLASS)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Alzheimer$SEX and Alzheimer$CLASS
## X-squared = 0.0027726, df = 1, p-value = 0.958
```

Regression

We are considering establishing a pattern of annual (i.e. age-dependent) atrophy in the brain. How does the volume of the brain vary with age? Does it depend on sex? To answer those questions 3 regression models predicting different brain volumes were constructed.

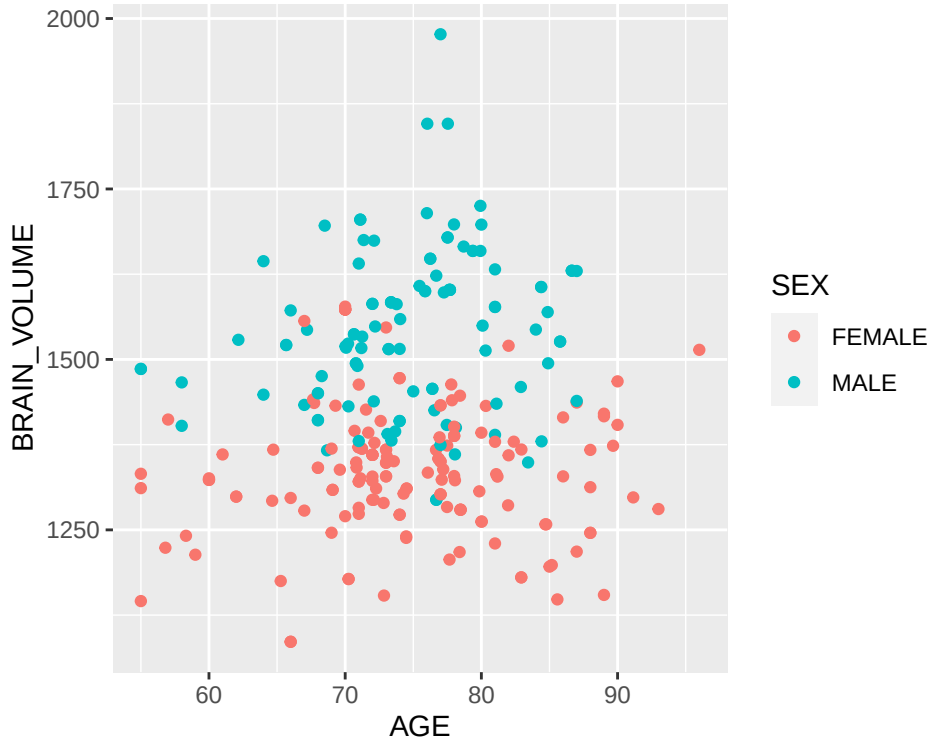
On the other hand, we want to know whether the volumetric parameters of the brain are good indicators of age. To solve this question, next 3 reversal models where constructed.

Brain volume ~ age & sex regression model

> We want to establish a regression model that uses as predictors the variables sex and age to estimate the value of brain volume (variable BRAIN_VOLUME). Use simple linear and polynomial regression and compare the models using anova(), detecting if all variables are significant and commenting and interpreting the known goodness-of-fit and error measures.

Starting with the visual analysis, it can be observed that on average man have higher brain volume. However, no more patterns are detected, therefore regression will be most likely inaccurate.

```
ggplot(data = Alzheimer, aes(x=AGE, y=BRAIN_VOLUME)) + geom_point(aes(colour=SEX))
```



Simple linear model

Based on the *Pr(>|t|)* value, it can be observed that AGE is an insignificant factor for any of the below models.

```
m1 <- lm(BRAIN_VOLUME ~ AGE, data = Alzheimer)
summary(m1)
##
## Call:
## lm(formula = BRAIN_VOLUME ~ AGE, data = Alzheimer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -325.48 -100.01  -32.25  109.35  552.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1333.203     89.820  14.843   <2e-16 ***
## AGE            1.184      1.197   0.989    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.1 on 260 degrees of freedom
```

```
## Multiple R-squared:  0.003748,   Adjusted R-squared:  -8.421e-05
## F-statistic: 0.978 on 1 and 260 DF,  p-value: 0.3236
```

```
m2 <- lm(BRAIN_VOLUME ~ SEX, data = Alzheimer)
summary(m2)
```

```
##
## Call:
## lm(formula = BRAIN_VOLUME ~ SEX, data = Alzheimer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -251.67  -64.98   -5.53   60.06  435.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1337.510      8.552  156.40   <2e-16 ***
## SEXMALE      203.949     13.320   15.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106.1 on 260 degrees of freedom
## Multiple R-squared:  0.4742, Adjusted R-squared:  0.4721
## F-statistic: 234.4 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
m3 <- lm(BRAIN_VOLUME ~ SEX + AGE, data = Alzheimer)
summary(m3)
```

```
##
## Call:
## lm(formula = BRAIN_VOLUME ~ SEX + AGE, data = Alzheimer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.84  -62.59    0.31   56.81  432.56
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1246.1563    65.3781  19.061   <2e-16 ***
## SEXMALE      204.0049    13.2947  15.345   <2e-16 ***
## AGE            1.2231     0.8678   1.409     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105.9 on 259 degrees of freedom
## Multiple R-squared:  0.4782, Adjusted R-squared:  0.4741
## F-statistic: 118.7 on 2 and 259 DF,  p-value: < 2.2e-16
```

As AGE is insignificant, for later analysis *m2* model will be used with following equation:

$ BRAIN_VOLUME = 1337.510 + 203.949*SEXMALE $

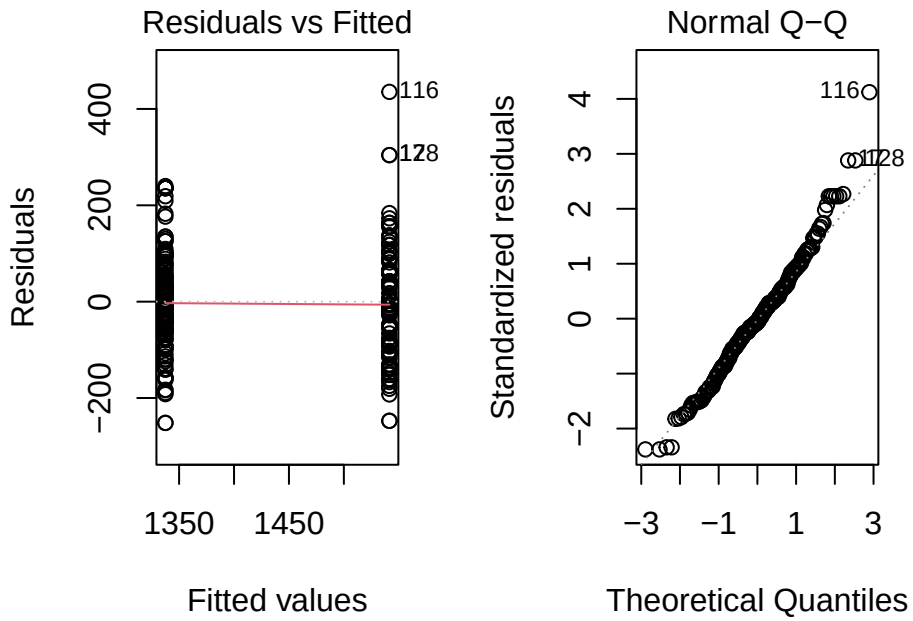
This means that on average, woman have brain volume of 1337.510, and man 203.949 higher.

Above model explains about ~47% of variance of predicted volume and has ~7% error rate.

```
rse2 <- sigma(m2)/mean(Alzheimer$BRAIN_VOLUME)
rse2
## [1] 0.07465346
```

Though residuals median does not equal zero but -5.53, which means that model is slightly lowers the score, residuals do not grow for higher values and are fairly normally distributed, which satisfies regression assumptions.

```
par(mfrow=c(1,2))
plot(m2, which=1:2)
```



Polynomial model

For test of different values of polynomials, none of them returned a model with significant value of *AGE*. Best one was below one.

```
m4 <- lm(BRAIN_VOLUME ~ SEX + I(AGE^2), data = Alzheimer)
summary(m4)
##
## Call:
## lm(formula = BRAIN_VOLUME ~ SEX + I(AGE^2), data = Alzheimer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.28  -61.73    0.00   57.46  433.15
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 1.296e+03  3.406e+01  38.059   <2e-16 ***
## SEXMALE     2.042e+02  1.331e+01  15.343   <2e-16 ***
## I(AGE^2)    7.320e-03  5.841e-03   1.253    0.211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106 on 259 degrees of freedom
## Multiple R-squared:  0.4773, Adjusted R-squared:  0.4733
## F-statistic: 118.3 on 2 and 259 DF,  p-value: < 2.2e-16
```

Models comparison

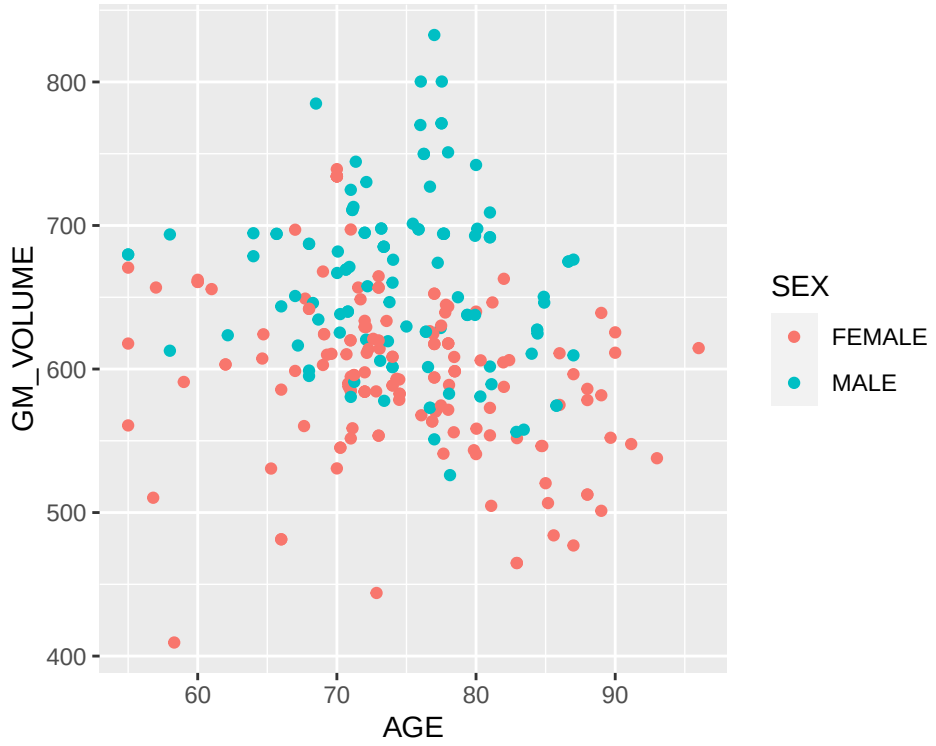
Anova test returned value of *0.2113* which is higher than significance level of 0.05, confirming that more complex model is not better enough.

```
anova(m2, m4)
## Analysis of Variance Table
##
## Model 1: BRAIN_VOLUME ~ SEX
## Model 2: BRAIN_VOLUME ~ SEX + I(AGE^2)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    260 2928304
## 2    259 2910654  1     17650 1.5706 0.2113
```

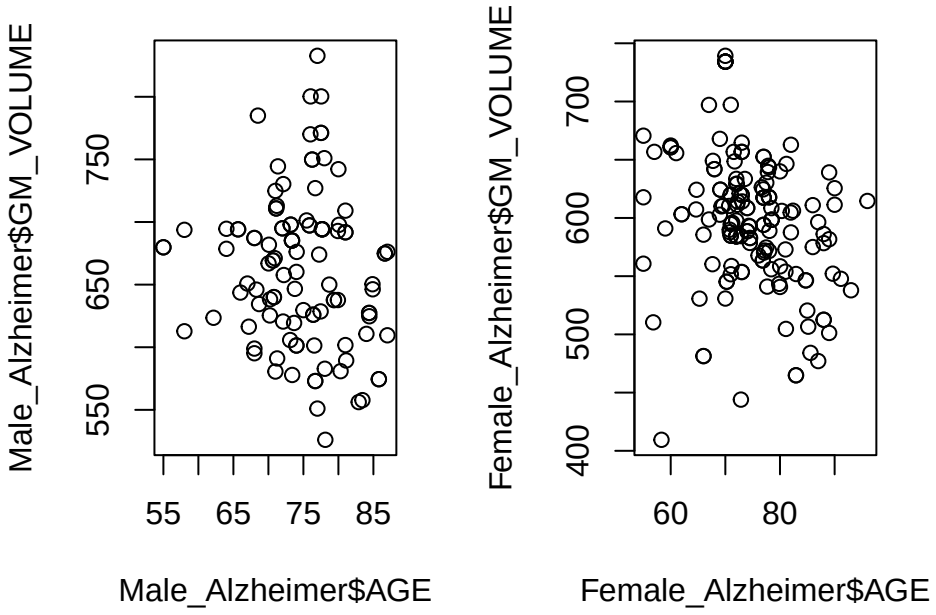
Gray matter ~ age & sex regression model

Initial visual analysis raises similar conclusions to those from the previous point.

```
ggplot(data = Alzheimer, aes(x=AGE, y=GM_VOLUME)) + geom_point(aes(colour=SEX))
```




```
par(mfrow=c(1,2))
plot(Male_Alzheimer$AGE, Male_Alzheimer$GM_VOLUME)
plot(Female_Alzheimer$AGE, Female_Alzheimer$GM_VOLUME)
```



Simple linear model

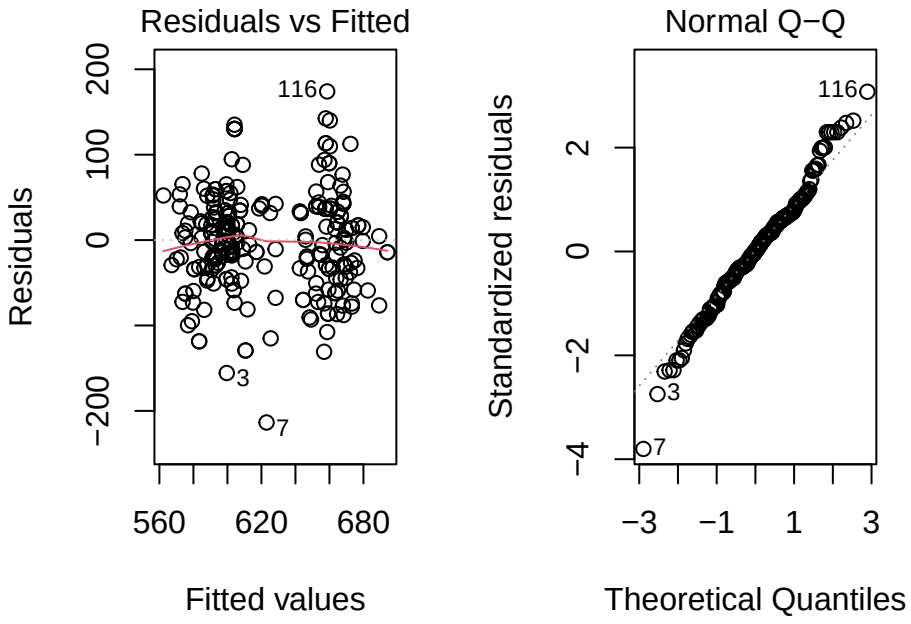
Despite similar assumptions based on visual analysis, in contrast to the previous analysis, in this case both variables are significant. Therefore the best one is model m3.

```
m1 <- lm(GM_VOLUME ~ AGE, data = Alzheimer)
summary(m1)
##
## Call:
## lm(formula = GM_VOLUME ~ AGE, data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -240.859  -41.151   -5.349   37.043  212.741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 744.8892    40.2073  18.526  < 2e-16 ***
## AGE          -1.6220     0.5357  -3.028  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.39 on 260 degrees of freedom
## Multiple R-squared:  0.03406,    Adjusted R-squared:  0.03034
## F-statistic: 9.167 on 1 and 260 DF,  p-value: 0.002712

m2 <- lm(GM_VOLUME ~ SEX, data = Alzheimer)
summary(m2)
##
## Call:
## lm(formula = GM_VOLUME ~ SEX, data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -187.188  -36.484    1.929   31.751  170.275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  596.638      4.677 127.566   <2e-16 ***
## SEXMALE       65.822      7.285   9.036   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.04 on 260 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.236
## F-statistic: 81.64 on 1 and 260 DF,  p-value: < 2.2e-16

m3 <- lm(GM_VOLUME ~ SEX + AGE, data = Alzheimer)
summary(m3)
##
## Call:
## lm(formula = GM_VOLUME ~ SEX + AGE, data = Alzheimer)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -213.548  -32.070    0.414   34.294  174.065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 716.8348    35.0933  20.427  < 2e-16 ***
## SEXMALE      65.7492     7.1363   9.213  < 2e-16 ***
## AGE          -1.6093     0.4658  -3.455 0.000644 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.86 on 259 degrees of freedom
## Multiple R-squared:  0.2725, Adjusted R-squared:  0.2669
## F-statistic: 48.51 on 2 and 259 DF,  p-value: < 2.2e-16
```

Model m3 explains ~27% of variance in the predicted value and has ~9% error rate. Moreover, it seems to satisfy regression assumptions, as error is mostly constant and close to 0 for all values, and residuals seem to follow normal distribution.

```
rse3 <- sigma(m3)/mean(Alzheimer$GM_VOLUME)
rse3
## [1] 0.09115205

par(mfrow=c(1,2))
plot(m3, which=1:2)
```



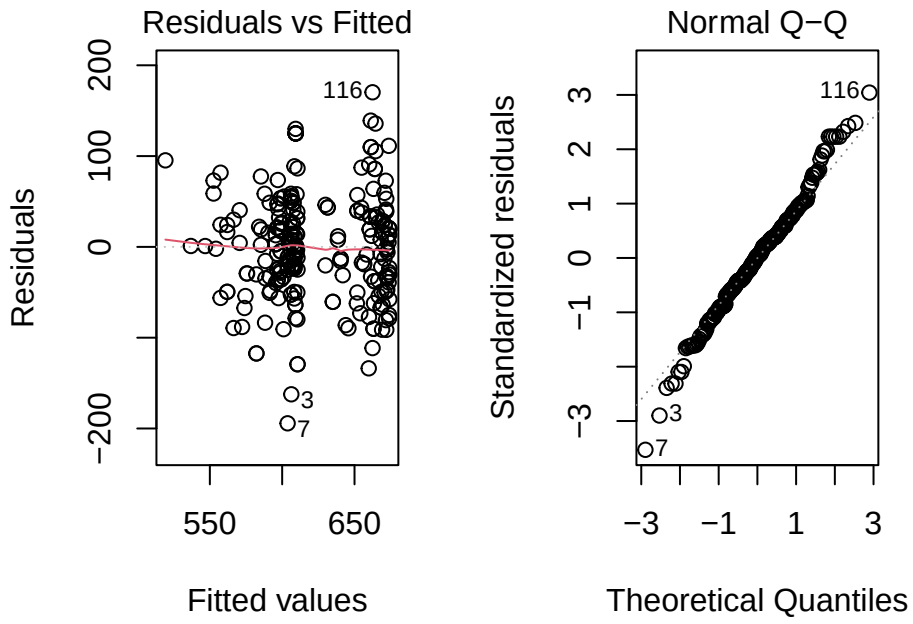
Polynomial model

When constructing a polynomial model, additional polynomials of age were added. For polynomials over 2, AGE variable was loosing significance. Therefore below model was choosen as the best.

```
m4 <- lm(GM_VOLUME ~ SEX + AGE + I(AGE^2), data = Alzheimer)
summary(m4)
##
## Call:
## lm(formula = GM_VOLUME ~ SEX + AGE + I(AGE^2), data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -194.297  -32.849    1.286   32.477  170.276
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 149.60912  223.41050   0.670   0.5037
## SEXMALE      63.50071    7.11428   8.926   <2e-16 ***
## AGE          13.88024    6.04425   2.296   0.0225 *
## I(AGE^2)     -0.10447    0.04065  -2.570   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.25 on 258 degrees of freedom
## Multiple R-squared:  0.2907, Adjusted R-squared:  0.2824
## F-statistic: 35.24 on 3 and 258 DF,  p-value: < 2.2e-16
```

Model m4 explains ~28% of variance of explained variable, and has ~9% error rate. It also ssems to satisfy regression assumptions.

```
rse4 <- sigma(m4)/mean(Alzheimer$GM_VOLUME)
rse4
## [1] 0.09018134

par(mfrow=c(1,2))
plot(m4, which=1:2)
```



Models comparison

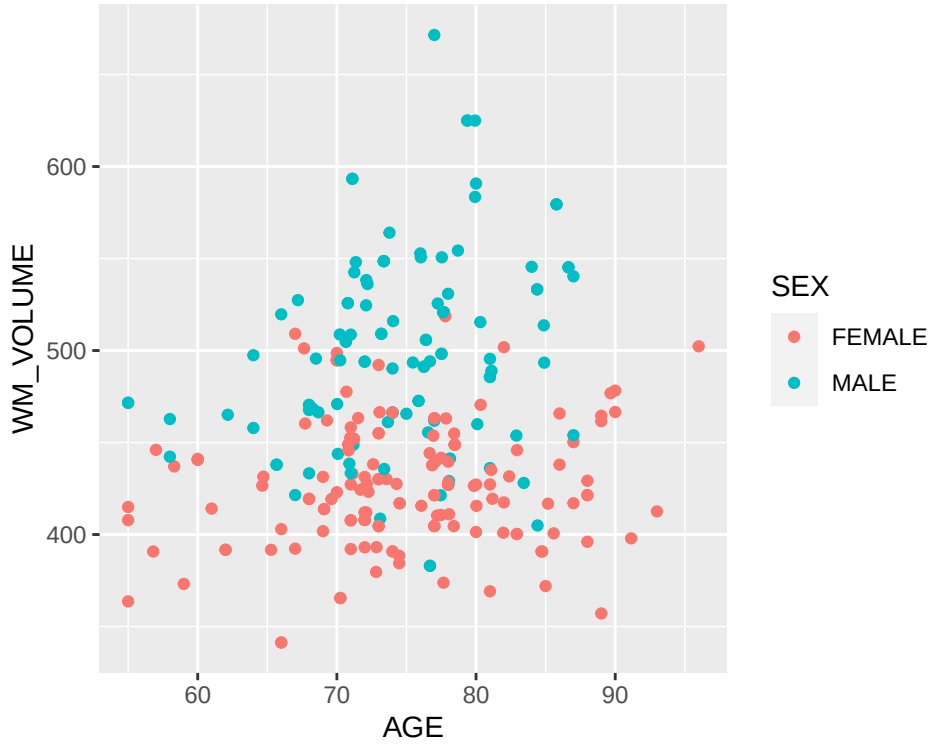
Value of *Pr(>F)* for anova test is equal to 0.01073, which is lower than significance level. Therefore it can be said that polynomial model m4 is significantly better than model m3.

```
anova(m3, m4)
## Analysis of Variance Table
##
## Model 1: GM_VOLUME ~ SEX + AGE
## Model 2: GM_VOLUME ~ SEX + AGE + I(AGE^2)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    259 837304
## 2    258 816401  1     20903 6.6057 0.01073 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

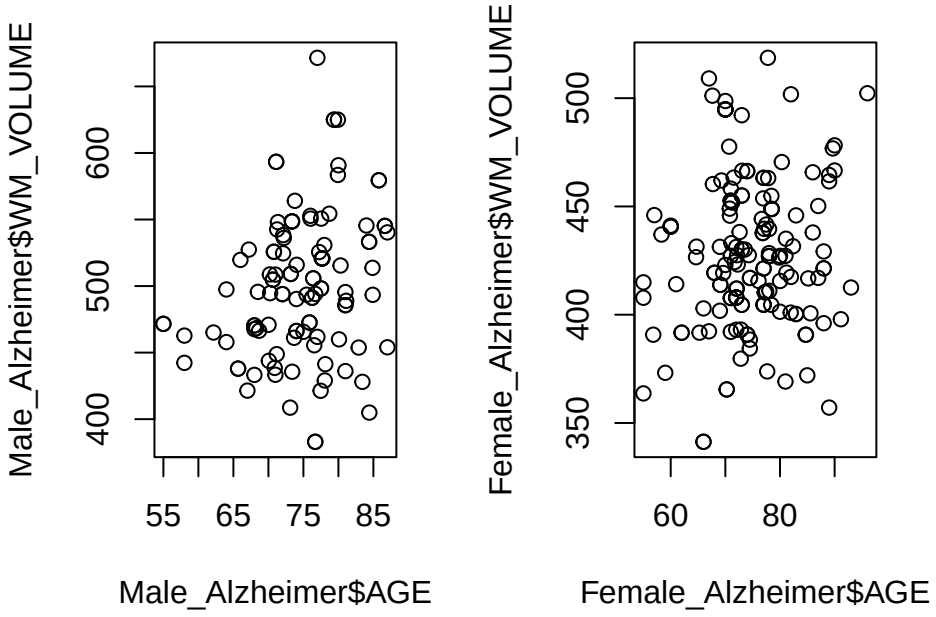
White matter ~ age & sex regression model

Initial visual analysis raise similar conclusions to those from previous points.

```
ggplot(data = Alzheimer, aes(x=AGE, y=WM_VOLUME)) + geom_point(aes(colour=SEX))
```




```
par(mfrow=c(1,2))
plot(Male_Alzheimer$AGE, Male_Alzheimer$WM_VOLUME)
plot(Female_Alzheimer$AGE, Female_Alzheimer$WM_VOLUME)
```



Simple linear model

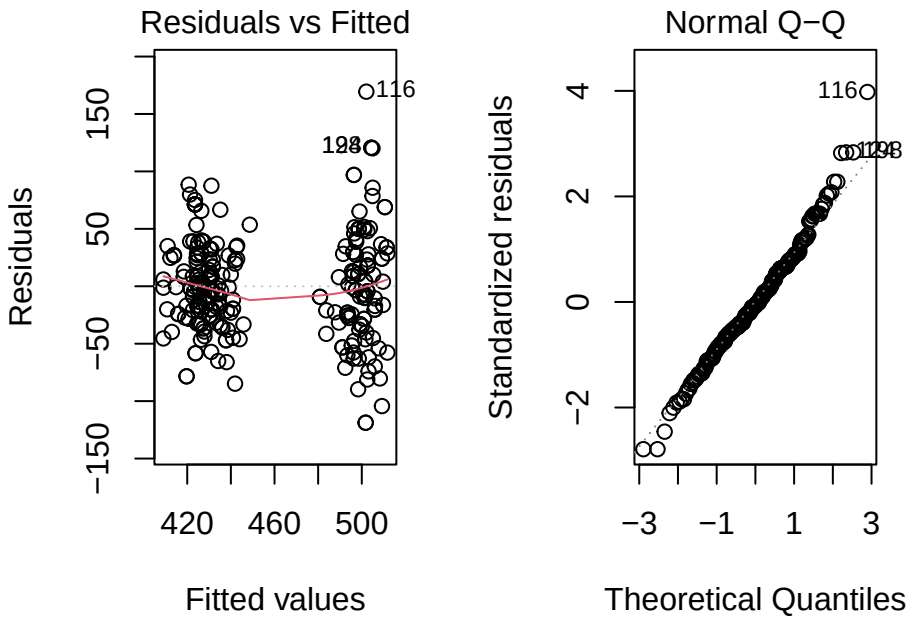
Like in previous analysis, both variables are significant, making model m3 the best one.

```
m1 <- lm(WM_VOLUME ~ AGE, data = Alzheimer)
summary(m1)
##
## Call:
## lm(formula = WM_VOLUME ~ AGE, data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -114.190  -40.527   -6.553   37.683  211.626
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 386.5803    34.1295  11.327   <2e-16 ***
## AGE           0.9518     0.4547   2.093   0.0373 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.51 on 260 degrees of freedom
## Multiple R-squared:  0.01657,    Adjusted R-squared:  0.01279
## F-statistic: 4.381 on 1 and 260 DF,  p-value: 0.03731

m2 <- lm(WM_VOLUME ~ SEX, data = Alzheimer)
summary(m2)
##
## Call:
## lm(formula = WM_VOLUME ~ SEX, data = Alzheimer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -116.82  -28.11   -2.01   25.72  171.69
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  428.093      3.491  122.62   <2e-16 ***
## SEXMALE       71.714      5.438   13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.33 on 260 degrees of freedom
## Multiple R-squared:  0.4008, Adjusted R-squared:  0.3985
## F-statistic: 173.9 on 1 and 260 DF,  p-value: < 2.2e-16

m3 <- lm(WM_VOLUME ~ SEX + AGE, data = Alzheimer)
summary(m3)
##
## Call:
## lm(formula = WM_VOLUME ~ SEX + AGE, data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -118.795  -25.695   -2.713   27.067  169.415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 355.9619    26.4085  13.479  < 2e-16 ***
## SEXMALE      71.7581     5.3702  13.362  < 2e-16 ***
## AGE           0.9657     0.3505   2.755  0.00629 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.79 on 259 degrees of freedom
## Multiple R-squared:  0.4179, Adjusted R-squared:  0.4134
## F-statistic: 92.96 on 2 and 259 DF,  p-value: < 2.2e-16
```

Obtained model m3 explains ~41% of variance of predicted value and has ~9% error rate. Moreover, model seems to satisfy regression assumptions of normality of residuals distribution and homoscedasticity, despite residual mean not equal to 0.

```
rse3 <- sigma(m3)/mean(Alzheimer$WM_VOLUME)
rse3
## [1] 0.09349173

par(mfrow=c(1,2))
plot(m3, which=1:2)
```



Polynomial model

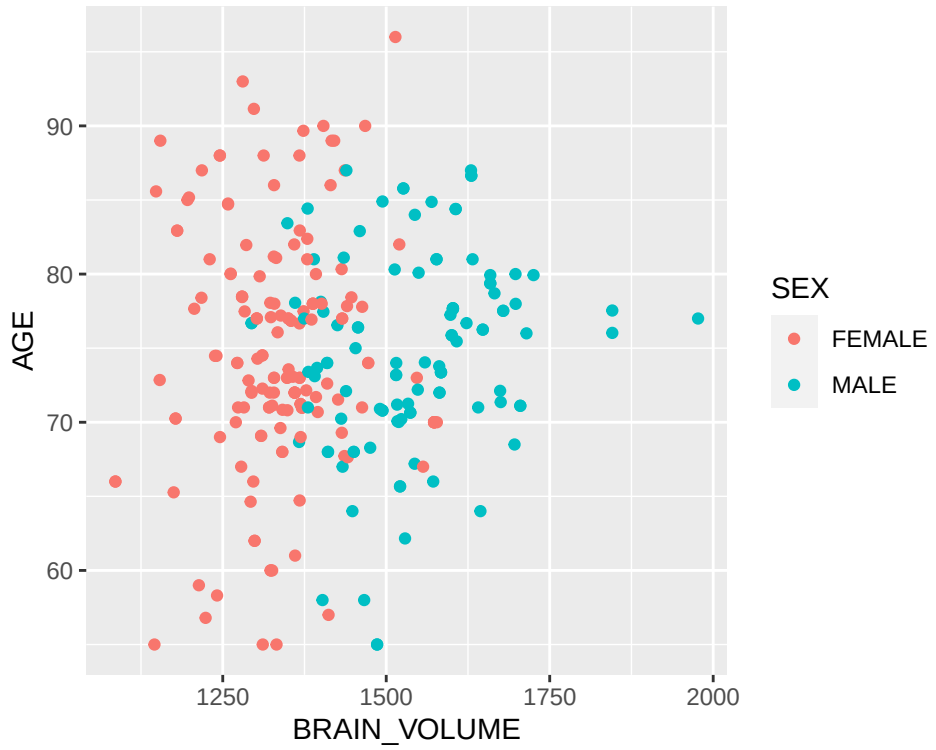
Addition of a polynomial to the previous model makes AGE loose significance and gain only little in respect of R-square quality. Moreover, with anova test *Pr(>F)* greater than significance level, it can be said that m3 model performs better than m4.

```
m4 <- lm(WM_VOLUME ~ SEX + AGE + I(AGE^2), data = Alzheimer)
summary(m4)
##
## Call:
## lm(formula = WM_VOLUME ~ SEX + AGE + I(AGE^2), data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.023  -25.879   -3.192   26.484  168.241
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.22036  169.89937   1.061    0.290
## SEXMALE      71.06148    5.41028  13.135   <2e-16 ***
## AGE           5.76479    4.59654   1.254    0.211
## I(AGE^2)     -0.03237    0.03091  -1.047    0.296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.78 on 258 degrees of freedom
## Multiple R-squared:  0.4203, Adjusted R-squared:  0.4136
## F-statistic: 62.36 on 3 and 258 DF,  p-value: < 2.2e-16

anova(m3, m4)
## Analysis of Variance Table
##
## Model 1: WM_VOLUME ~ SEX + AGE
## Model 2: WM_VOLUME ~ SEX + AGE + I(AGE^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    259 474157
## 2    258 472150  1    2006.5 1.0964  0.296
```

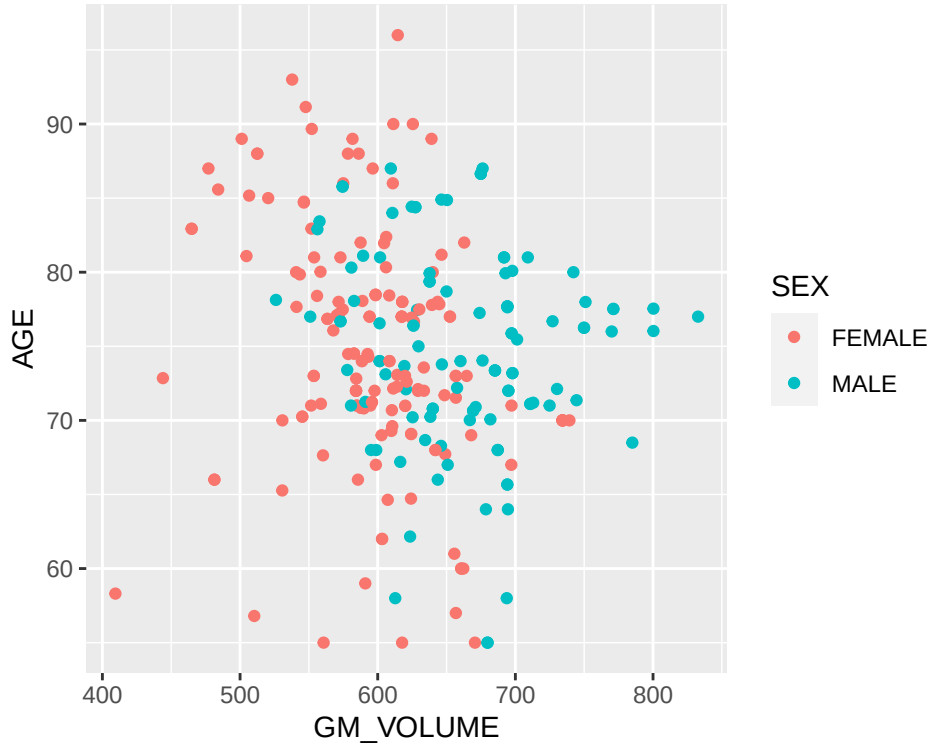
Age ~ brain volume & gray/white matter & sex regression model

In order to create model for Age prediction, visual analysis was conducted again. However, no clear patterns were observed (besides average difference by sex).

```
ggplot(data = Alzheimer, aes(x=BRAIN_VOLUME, y=AGE)) + geom_point(aes(colour=SEX))
```




```
ggplot(data = Alzheimer, aes(x=GM_VOLUME, y=AGE)) + geom_point(aes(colour=SEX))
```

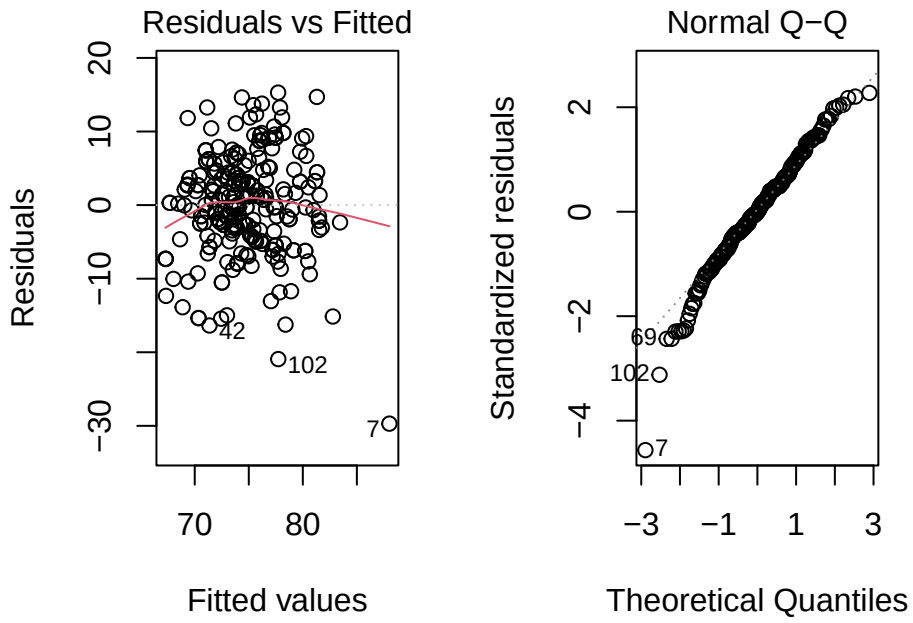


```
ggplot(data = Alzheimer, aes(x=WM_VOLUME, y=AGE)) + geom_point(aes(colour=SEX))
```

Due to the lack of obvious relationship, 1st model was constructed containing all discussed variables. It explained ~20% of variance in predicted variable and had ~9% error rate. Model satisfied regression assumptions, though it had a little pattern of growing residuals for extreme values.

```
m1 <- lm(AGE ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME + SEX, data = Alzheimer)
summary(m1)
##
## Call:
## lm(formula = AGE ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME + SEX,
##     data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.7007  -3.5316  -0.0114   4.1219  15.2854
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.45207    5.31866  12.870  < 2e-16 ***
## BRAIN_VOLUME  0.05437    0.01074   5.061 7.95e-07 ***
## GM_VOLUME    -0.09987    0.01316  -7.587 6.00e-13 ***
## WM_VOLUME    -0.01611    0.01836  -0.878  0.38103
## SEXMALE      -3.40524    1.19794  -2.843  0.00483 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.76 on 257 degrees of freedom
```

```
## Multiple R-squared:  0.2117, Adjusted R-squared:  0.1994
## F-statistic: 17.25 on 4 and 257 DF,  p-value: 1.507e-12

rse1 <- sigma(m1)/mean(Alzheimer$AGE)
rse1
## [1] 0.09053209

par(mfrow=c(1,2))
plot(m1, which=1:2)
```

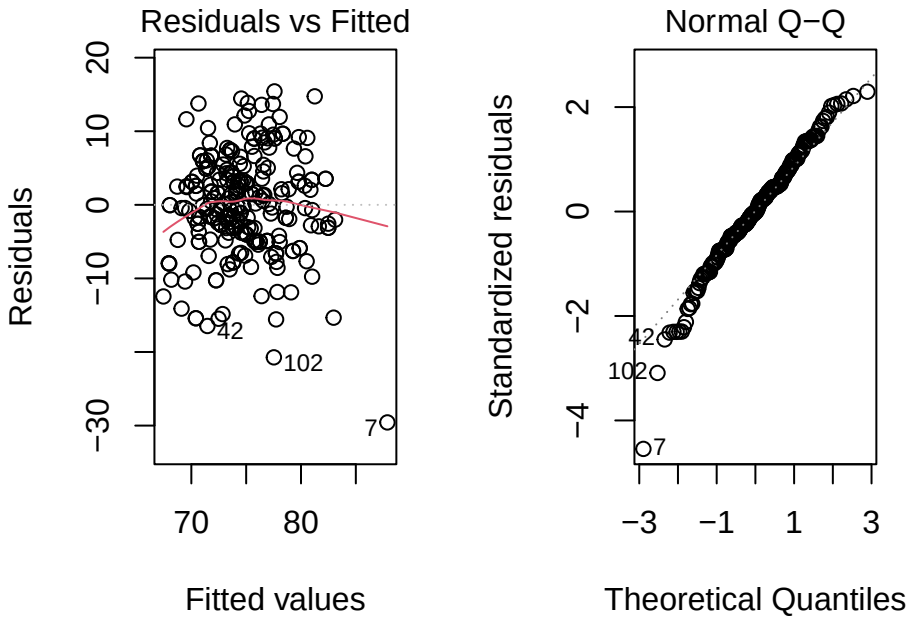


Based on *Pr(>|t|)* value in coefficients table of model m1, it can be seen that WM_VOLUME is insignificant. Therefore next constructed model m2 was constructed without it.

Newly created model explained ~20% of variance in predicted variable and had ~9% error rate. Though it showed similar to m1 problems in distribution of residuals, it seems to satisfy regression assumptions.

```
m2 <- lm(AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX, data = Alzheimer)
summary(m2)
##
## Call:
## lm(formula = AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX, data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.5696  -3.7643  -0.0512   3.9302  15.4192
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  68.502190   5.315978  12.886  < 2e-16 ***
## BRAIN_VOLUME  0.046780   0.006371   7.343 2.74e-12 ***
## GM_VOLUME    -0.094497   0.011649  -8.112 2.03e-14 ***
## SEXMALE      -3.366258   1.196580  -2.813  0.00528 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.757 on 258 degrees of freedom
## Multiple R-squared:  0.2093, Adjusted R-squared:  0.2001
## F-statistic: 22.76 on 3 and 258 DF,  p-value: 4.164e-13

rse2 <- sigma(m2)/mean(Alzheimer$AGE)
rse2
## [1] 0.09049173

par(mfrow=c(1,2))
plot(m2, which=1:2)
```



As in previous model, WM_VOLUME variable was removed, it was added again to 3rd model to check if its polynomial value would prove to be more significant.

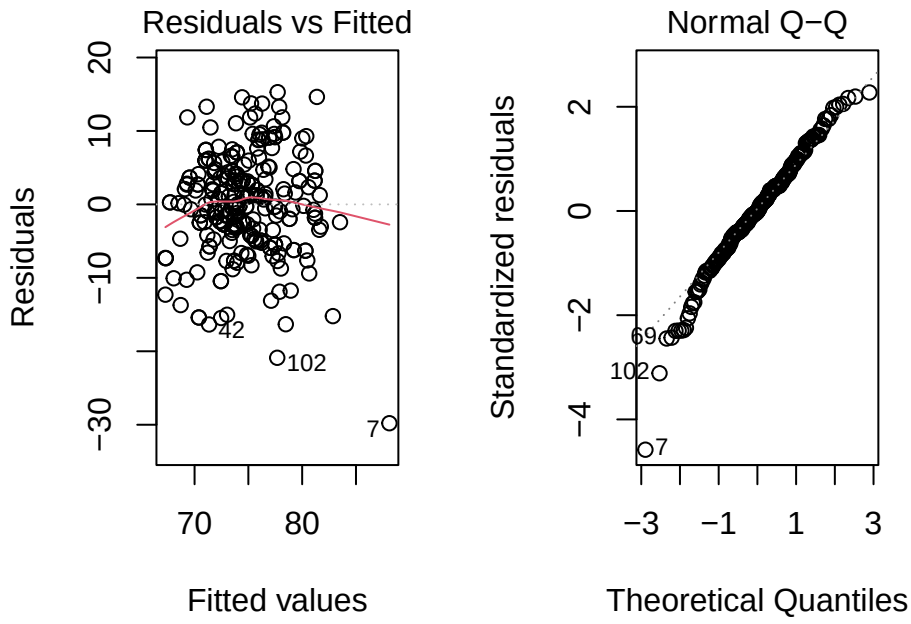
Newly obtained model m3 explained ~20% of variance and had ~9% error rate. However, based on coefficients table and *Pr(>|t|)* value, it can be see that I(WM_VOLUME^2) once again is insignificant.

```
m3 <- lm(AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX + I(WM_VOLUME^2), data = Alzheimer)
summary(m3)
##
## Call:
```

```
## lm(formula = AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX + I(WM_VOLUME^2),
##     data = Alzheimer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.7940  -3.4926  -0.0284   4.1219  15.2786
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.468e+01  6.707e+00   9.643  < 2e-16 ***
## BRAIN_VOLUME    5.459e-02  1.050e-02   5.197 4.13e-07 ***
## GM_VOLUME      -1.003e-01  1.320e-02  -7.597 5.65e-13 ***
## SEXMALE        -3.410e+00  1.198e+00  -2.847  0.00477 **
## I(WM_VOLUME^2) -1.711e-05  1.829e-05  -0.935  0.35045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.759 on 257 degrees of freedom
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.1997
## F-statistic: 17.28 on 4 and 257 DF,  p-value: 1.432e-12
```

```
rse3 <- sigma(m3)/mean(Alzheimer$AGE)
rse3
## [1] 0.09051366
```

```
par(mfrow=c(1,2))
plot(m3, which=1:2)
```



Models comparison

Model m2 proofs to be the best one, as it contains only significant variables. It is confirmed to be better by anova tests as well as in both cases *Pr(>F)* value was greater than 0.05.

```
anova(m2, m1)
## Analysis of Variance Table
##
## Model 1: AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX
## Model 2: AGE ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME + SEX
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1    258 11780
## 2    257 11745  1     35.19 0.77  0.381
anova(m2, m3)
## Analysis of Variance Table
##
## Model 1: AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX
## Model 2: AGE ~ BRAIN_VOLUME + GM_VOLUME + SEX + I(WM_VOLUME^2)
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1    258 11780
## 2    257 11740  1    39.972 0.875 0.3504
```

Even though m2 is the best model, using all significant variable from given set, it still explains only 20% of variance in the predicted age. On the other hand, its predictions have only ~9% error rate.

Therefore, it can be said that those volumetric parameters and sex, are poor indicators of age.

Logistic regression and classification

We want to be able to decide, on the basis of the brain volumetry data we are looking at, whether a new subject may have Alzheimer's disease. To do so, we will build logistic regression and classification models to solve this problem.

We begin with the split of the dataset into a training set with 80% of the data, randomly selected, and the remaining 20% to validate the models.

```
# Adding a binary variable indicating whether or not individual is healthy
Alzheimer$IS_HEALTHY <- Alzheimer$CLASS == "HEALTHY"
head(Alzheimer)
## # A tibble: 6 x 229
##     AGE SEX    BRAIN_VOLUME GM_VOLUME WM_VOLUME CSF_VOLUME GM_BRAIN_QUOTIENT
##   <dbl> <chr>         <dbl>     <dbl>     <dbl>      <dbl>              <dbl>
## 1  68   MALE          1411.      599.      433.       379.              0.425
## 2  82.9 FEMALE        1368.      552.      446.       370.              0.403
## 3  72.8 FEMALE        1154.      444.      393.       317.              0.385
## 4  73   FEMALE        1547.      665.      492.       390.              0.430
## 5  71.1 FEMALE        1326.      559.      433.       334.              0.422
## 6  78.1 MALE          1400.      526.      441.       433.              0.376
## # i 222 more variables: WM_BRAIN_QUOTIENT <dbl>, CSF_BRAIN_QUOTIENT <dbl>,
## #   GM_WM_QUOTIENT <dbl>, PRECENTRAL_L_VOLUME <dbl>, PRECENTRAL_R_VOLUME <dbl>,
## #   FRONTAL_SUP_L_VOLUME <dbl>, FRONTAL_SUP_R_VOLUME <dbl>,
## #   FRONTAL_SUP_ORB_L_VOLUME <dbl>, FRONTAL_SUP_ORB_R_VOLUME <dbl>,
## #   FRONTAL_MID_L_VOLUME <dbl>, FRONTAL_MID_R_VOLUME <dbl>,
## #   FRONTAL_MID_ORB_L_VOLUME <dbl>, FRONTAL_MID_ORB_R_VOLUME <dbl>,
```

```
## #   FRONTAL_INF_OPER_L_VOLUME <dbl>, FRONTAL_INF_OPER_R_VOLUME <dbl>, ...

# Data split (80% training, 20% validation)
training_sample_size <- floor(0.8*nrow(Alzheimer))
set.seed(12345)
selected <- sample(seq_len(nrow(Alzheimer)), size = training_sample_size)

Training_Alzheimer <- Alzheimer[selected, ]
Testing_Alzheimer <- Alzheimer[-selected, ]
```

Model generation

Having data split, next step is to create a logistic regression model based on significant variables from training dataset. Based on *Pr(>|z|)* value from coefficients table, it can be seen that SEX variable is insignificant. Therefore new model that excludes it will be trained.

```
m1 <- glm(formula = IS_HEALTHY ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME + SEX + AGE, family = binomial, da
#m1
summary(m1)
##
## Call:
## glm(formula = IS_HEALTHY ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME +
##     SEX + AGE, family = binomial, data = Training_Alzheimer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7668   0.2672   0.3565   0.4573   1.1951
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.863393   4.052774  -0.707 0.479861
## BRAIN_VOLUME -0.020512   0.007096  -2.891 0.003844 **
## GM_VOLUME     0.030329   0.008534   3.554 0.000379 ***
## WM_VOLUME     0.021748   0.011734   1.853 0.063814 .
## SEXMALE       0.692621   0.683558   1.013 0.310937
## AGE           0.071754   0.036023   1.992 0.046385 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 136.32  on 208  degrees of freedom
## Residual deviance: 120.01  on 203  degrees of freedom
## AIC: 132.01
##
## Number of Fisher Scoring iterations: 5
```

In the next generated model, value of *Pr(>|z|)* for variables WM_VOLUME and AGE is again over, significance level of 0.05, therefore next model will wxclude them.

```
m2 <- glm(formula = IS_HEALTHY ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME + AGE, family = binomial, data = 

summary(m2)
```

```
##
## Call:
## glm(formula = IS_HEALTHY ~ BRAIN_VOLUME + GM_VOLUME + WM_VOLUME +
##     AGE, family = binomial, data = Training_Alzheimer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7587   0.2599   0.3635   0.4661   1.2299
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.491676   3.704185  -1.213 0.225285
## BRAIN_VOLUME -0.017767   0.006437  -2.760 0.005777 **
## GM_VOLUME     0.028544   0.008280   3.447 0.000566 ***
## WM_VOLUME     0.020538   0.011315   1.815 0.069496 .
## AGE           0.067244   0.035308   1.904 0.056848 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 136.32  on 208  degrees of freedom
## Residual deviance: 121.06  on 204  degrees of freedom
## AIC: 131.06
##
## Number of Fisher Scoring iterations: 5
```

Last obtained model contains only significant variables.

```
m3 <- glm(formula = IS_HEALTHY ~ BRAIN_VOLUME + GM_VOLUME, family = binomial, data = Training_Alzheimer)

summary(m3)
##
## Call:
## glm(formula = IS_HEALTHY ~ BRAIN_VOLUME + GM_VOLUME, family = binomial,
##     data = Training_Alzheimer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5310   0.3201   0.4008   0.4958   0.8285
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.143574   2.383889   0.060   0.9520
## BRAIN_VOLUME -0.005687   0.002793  -2.036   0.0417 *
## GM_VOLUME     0.016490   0.006002   2.748   0.0060 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 136.32  on 208  degrees of freedom
## Residual deviance: 128.52  on 206  degrees of freedom
```

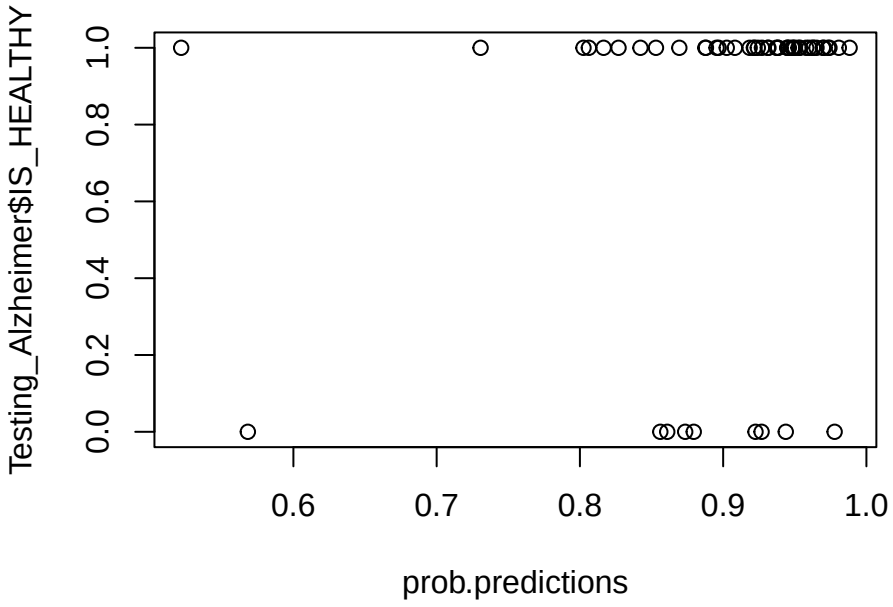
```
## AIC: 134.52
##
## Number of Fisher Scoring iterations: 5
```

Model validation

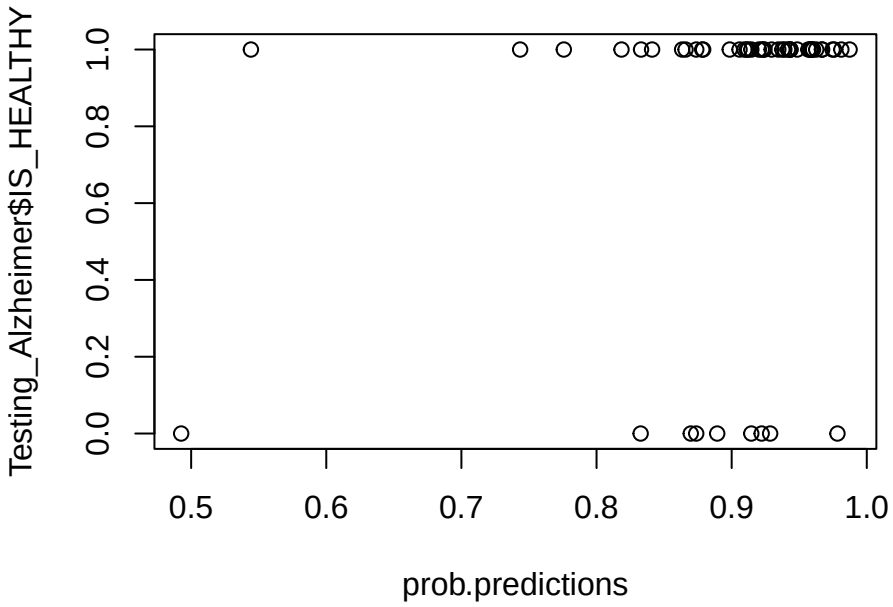
How accurate is this model if we run it (using predict()) on the validation set?

After compariosn of 3 models, it can be seen that the second one, even though contains variables with significance level over 0.05, has the highest accuracy of ~85%, with ~15% error rate.

```
#m1

#logit predictions
logit.predictions <- predict(object = m1, newdata = Testing_Alzheimer)
#inverse logit to transform to probabilities
prob.predictions <- 1 / (1 + exp(-logit.predictions))

Testing_Alzheimer$PREDICTION_IS_HEALTHY <- prob.predictions >= 0.5

accuracy <- mean(Testing_Alzheimer$PREDICTION_IS_HEALTHY == Testing_Alzheimer$IS_HEALTHY)
error_rate <- 1-accuracy

accuracy
## [1] 0.8301887
error_rate
## [1] 0.1698113

plot(prob.predictions, Testing_Alzheimer$IS_HEALTHY)
```

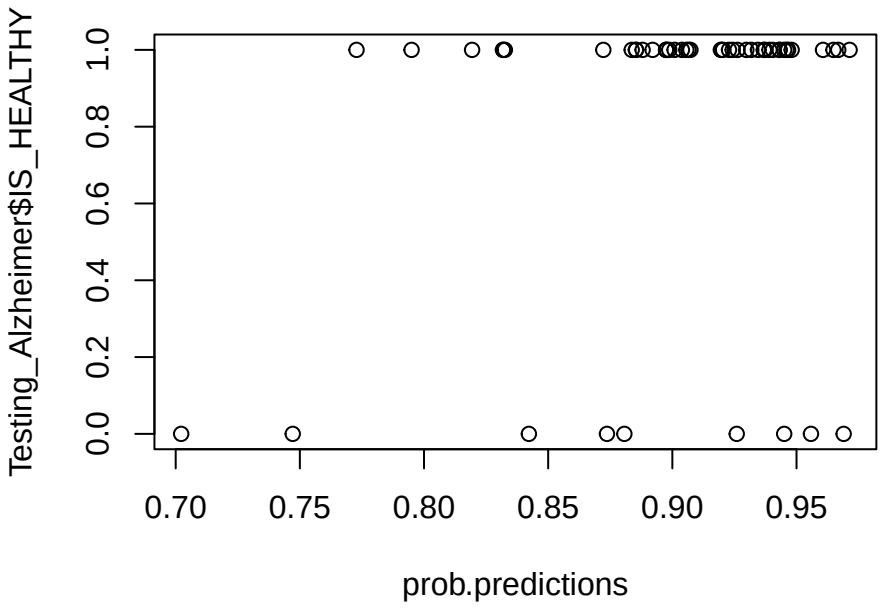


```r
#m2

#logit predictions
logit.predictions <- predict(object = m2, newdata = Testing_Alzheimer)
#inverse logit to transform to probabilities
prob.predictions <- 1 / (1 + exp(-logit.predictions))

Testing_Alzheimer$prob.predictions <- prob.predictions
Testing_Alzheimer$PREDICTION_IS_HEALTHY <- prob.predictions >= 0.5

accuracy <- mean(Testing_Alzheimer$PREDICTION_IS_HEALTHY == Testing_Alzheimer$IS_HEALTHY)
error_rate <- 1-accuracy

accuracy
## [1] 0.8490566
error_rate
## [1] 0.1509434

plot(prob.predictions, Testing_Alzheimer$IS_HEALTHY)
```




```r
#m3

#logit predictions
logit.predictions <- predict(object = m3, newdata = Testing_Alzheimer)
#inverse logit to transform to probabilities
prob.predictions <- 1 / (1 + exp(-logit.predictions))
```

```r
Testing_Alzheimer$PREDICTION_IS_HEALTHY <- prob.predictions >= 0.5

accuracy <- mean(Testing_Alzheimer$PREDICTION_IS_HEALTHY == Testing_Alzheimer$IS_HEALTHY)
error_rate <- 1-accuracy

accuracy
## [1] 0.8301887
error_rate
## [1] 0.1698113

plot(prob.predictions, Testing_Alzheimer$IS_HEALTHY)
```

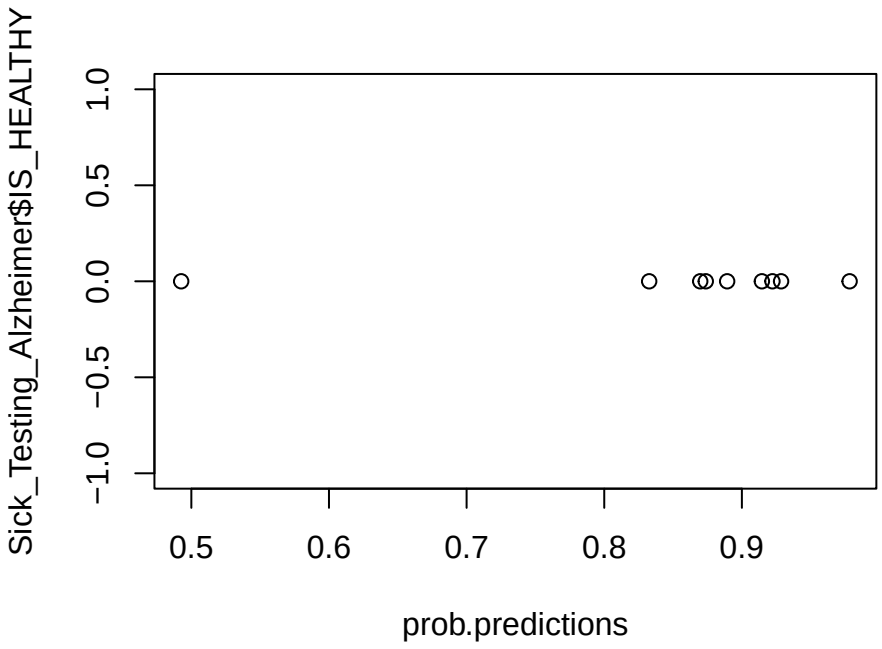


However, looking at the plots of predictions vs real health status, it can be seen that in most cases sick individuals are also classifies as healthy ones. When calculating fals positive error rate for m2 model, it can be seen that it is only ~11% accurate and has ~89% false positive error rate.

Moreover, it is impossible to improve predictions by adjusting probability threshold, because in most cases their predictions overlap in value with healthy ones.

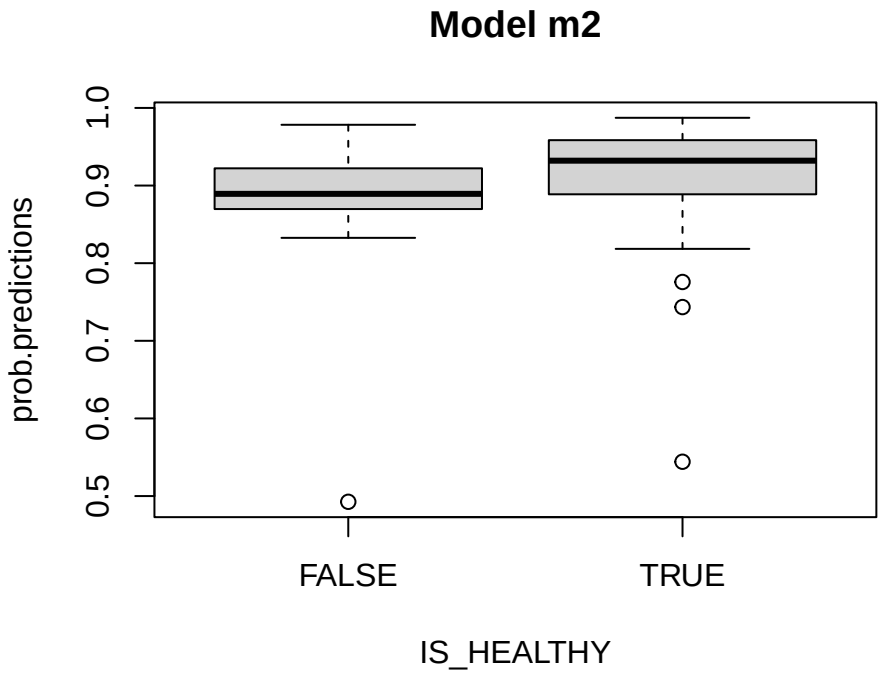
```r
Sick_Testing_Alzheimer <- filter(Testing_Alzheimer, CLASS == "AD")

logit.predictions <- predict(object = m2, newdata = Sick_Testing_Alzheimer)
#inverse logit to transform to probabilities
prob.predictions <- 1 / (1 + exp(-logit.predictions))

Sick_Testing_Alzheimer$PREDICTION_IS_HEALTHY <- prob.predictions >= 0.5

accuracy <- mean(Sick_Testing_Alzheimer$PREDICTION_IS_HEALTHY == Sick_Testing_Alzheimer$IS_HEALTHY)
```

```r
error_rate <- 1-accuracy

accuracy
## [1] 0.1111111
error_rate
## [1] 0.8888889

plot(prob.predictions, Sick_Testing_Alzheimer$IS_HEALTHY)
```



However, if correct predicition of sick status is more important, probability threshold can be adjusted for some value between the means of probabilities for health and sick characters. Such a model would prove over 50% accuracy for both cases.

```r
#m2

#logit predictions
logit.predictions <- predict(object = m2, newdata = Testing_Alzheimer)
#inverse logit to transform to probabilities
Testing_Alzheimer$prob.predictions <- 1 / (1 + exp(-logit.predictions))

box_plot <- boxplot(prob.predictions~IS_HEALTHY, data=Testing_Alzheimer, main="Model m2")
```



From below summary and the plot, it can be seen that first quantile for healthy individuals is similar to sick ones median value of probability. At the same time, third quantile of probabilities for sick characters is little below healthy ones median.

This means, that if the threshold was adjusted to be sick ones median, accuracy in detecting sick patients would rise to 50%, with false negative ration for healthy patients around 25%. To increase accuracy of detecting sick individuals to 75%, probability threshold set at 3rd quantile of sick ones probabilities, would also result in ~50% error rate for predicting healthy individuals.

```
Summary<-box_plot$stats
colnames(Summary)<-c("Sick","Healthy")
rownames(Summary)<-c("Min","First Quartile","Median","Third Quartile","Maximum")
Summary
##                      Sick   Healthy
## Min             0.8326413 0.8184616
## First Quartile 0.8697292 0.8887461
## Median          0.8893121 0.9320412
## Third Quartile 0.9221978 0.9584820
## Maximum         0.9782783 0.9872911
```

To achieve better results, possible use of other volumetric data from the file, ones that we do not consider, might be useful.

Inferencia Bayesiana

> *ToDo: We want to make a study of the incidence of Alzheimer's disease in the population, since it is estimated that, a priori, the incidence of Alzheimer's disease in the population over 60 years of age is between 7% and 9%.*

- Bayesian inference will be based on studied dataset
- Estimate the a posteriori distribution of the theta proportion of Alzheimer's cases twice, first starting from a non-informative prior (uniform in [0,1]), and secondly starting from a trapezoid-shaped prior,
- In both cases, give the maximum a posteriori estimator (using both the mode and the mean) and the 95% maximum posterior density credible interval for theta.

Data preparation

At first, dataset was reduced to the population over 60 years old as given by the task, and initial statistics were calculated.

```
Alzheimer_60Plus_Data <- filter(Alzheimer, AGE >= "60")
Alzheimer_60Plus <- Alzheimer_60Plus_Data$CLASS

N <- length(Alzheimer_60Plus)
N
## [1] 251

nHealthy <- sum(Alzheimer_60Plus == 'HEALTHY')
nHealthy
## [1] 222

nSick <- sum(Alzheimer_60Plus == 'AD')
nSick
## [1] 29
```

Theta definition

Our goal is to learn about the incidence of Alzheimer's disease in the given population. As it is said to be between 7% and 9%, following probabilities could be considered:

```
theta <- seq(from = 0.07,
             to = 0.09,
             length = 100)
theta
##   [1] 0.07000000 0.07020202 0.07040404 0.07060606 0.07080808 0.07101010
##   [7] 0.07121212 0.07141414 0.07161616 0.07181818 0.07202020 0.07222222
##  [13] 0.07242424 0.07262626 0.07282828 0.07303030 0.07323232 0.07343434
##  [19] 0.07363636 0.07383838 0.07404040 0.07424242 0.07444444 0.07464646
##  [25] 0.07484848 0.07505051 0.07525253 0.07545455 0.07565657 0.07585859
##  [31] 0.07606061 0.07626263 0.07646465 0.07666667 0.07686869 0.07707071
##  [37] 0.07727273 0.07747475 0.07767677 0.07787879 0.07808081 0.07828283
##  [43] 0.07848485 0.07868687 0.07888889 0.07909091 0.07929293 0.07949495
##  [49] 0.07969697 0.07989899 0.08010101 0.08030303 0.08050505 0.08070707
##  [55] 0.08090909 0.08111111 0.08131313 0.08151515 0.08171717 0.08191919
##  [61] 0.08212121 0.08232323 0.08252525 0.08272727 0.08292929 0.08313131
##  [67] 0.08333333 0.08353535 0.08373737 0.08393939 0.08414141 0.08434343
##  [73] 0.08454545 0.08474747 0.08494949 0.08515152 0.08535354 0.08555556
##  [79] 0.08575758 0.08595960 0.08616162 0.08636364 0.08656566 0.08676768
##  [85] 0.08696970 0.08717172 0.08737374 0.08757576 0.08777778 0.08797980
```

```
##  [91] 0.08818182 0.08838384 0.08858586 0.08878788 0.08898990 0.08919192
##  [97] 0.08939394 0.08959596 0.08979798 0.09000000
```

However, as we would like to compare results fro two priors, from given range of distribution from 0 to 1, following theta set will be considered.

```
theta <- seq(from = 0,
             to = 1,
             length = 100)
theta
##   [1] 0.00000000 0.01010101 0.02020202 0.03030303 0.04040404 0.05050505
##   [7] 0.06060606 0.07070707 0.08080808 0.09090909 0.10101010 0.11111111
##  [13] 0.12121212 0.13131313 0.14141414 0.15151515 0.16161616 0.17171717
##  [19] 0.18181818 0.19191919 0.20202020 0.21212121 0.22222222 0.23232323
##  [25] 0.24242424 0.25252525 0.26262626 0.27272727 0.28282828 0.29292929
##  [31] 0.30303030 0.31313131 0.32323232 0.33333333 0.34343434 0.35353535
##  [37] 0.36363636 0.37373737 0.38383838 0.39393939 0.40404040 0.41414141
##  [43] 0.42424242 0.43434343 0.44444444 0.45454545 0.46464646 0.47474747
##  [49] 0.48484848 0.49494949 0.50505051 0.51515152 0.52525253 0.53535354
##  [55] 0.54545455 0.55555556 0.56565657 0.57575758 0.58585859 0.59595960
##  [61] 0.60606061 0.61616162 0.62626263 0.63636364 0.64646465 0.65656566
##  [67] 0.66666667 0.67676768 0.68686869 0.69696970 0.70707071 0.71717172
##  [73] 0.72727273 0.73737374 0.74747475 0.75757576 0.76767677 0.77777778
##  [79] 0.78787879 0.79797980 0.80808081 0.81818182 0.82828283 0.83838384
##  [85] 0.84848485 0.85858586 0.86868687 0.87878788 0.88888889 0.89898990
##  [91] 0.90909091 0.91919192 0.92929293 0.93939394 0.94949495 0.95959596
##  [97] 0.96969697 0.97979798 0.98989899 1.00000000
```

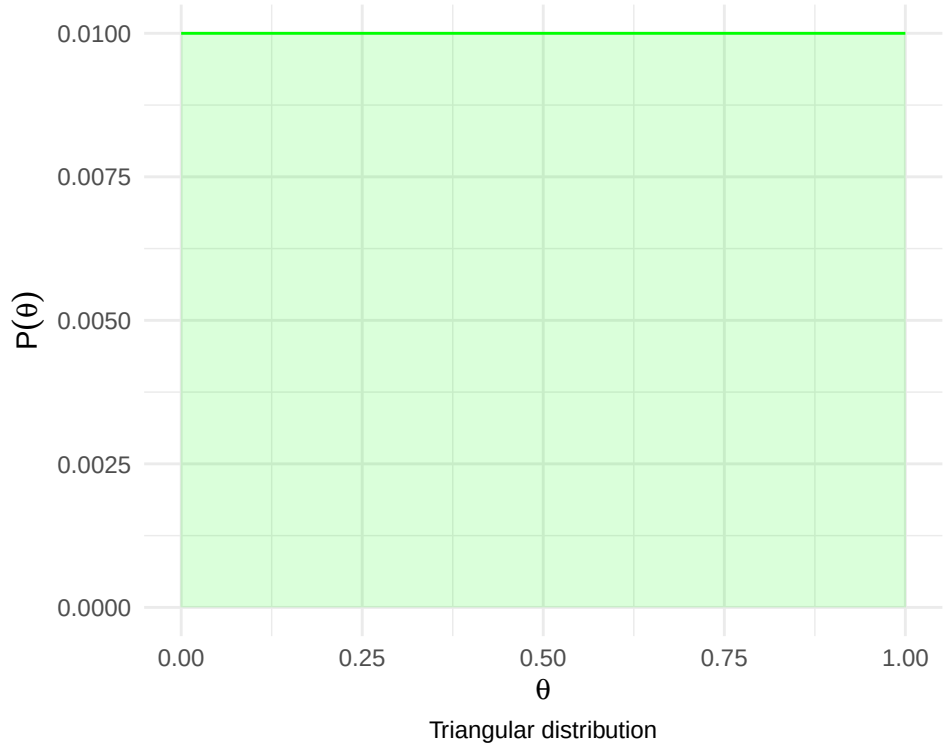
Analysis with non-informative prior

In this case, knowledge about the probability being between 7% and 9% is not used.

Priori

Non-informative prior (uniform in [0,1]) is used.

```
pTheta <- dunif(theta, min = 0, max = 1)
pTheta <- pTheta / sum(pTheta)

#plot(theta, pTheta)
#lines(theta, pTheta)
```

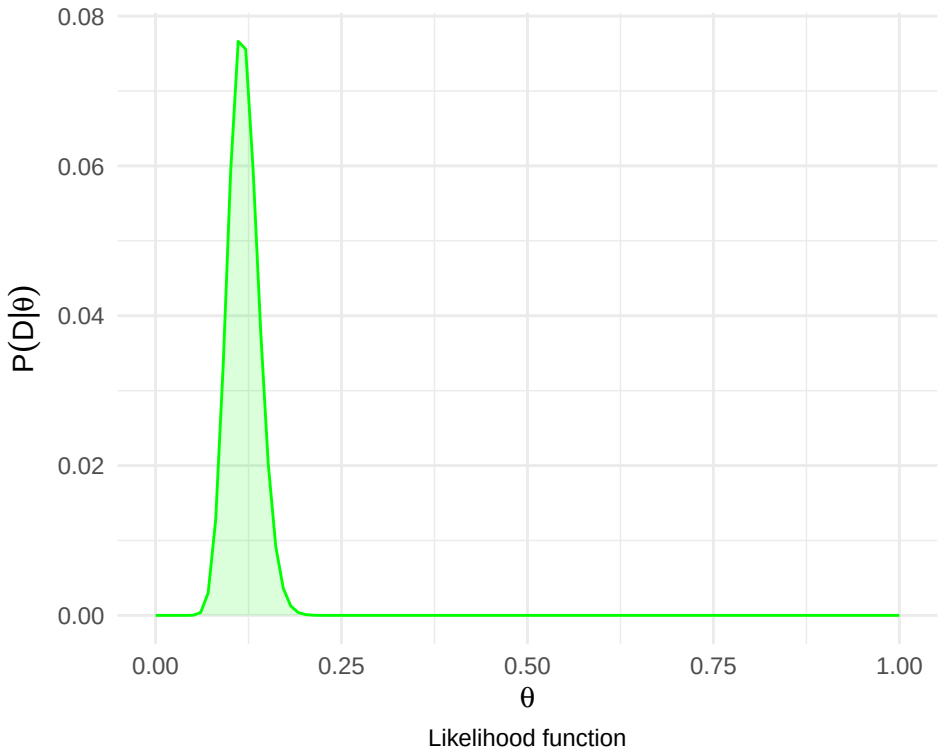


Likelyhood function

In this case, likelihood function is given by binomial distribution, which models how many patients from given population have Alzheimer's disease.

```
pDataGivenTheta <- choose(N, nSick) * theta^nSick * (1 - theta)^nHealthy
```

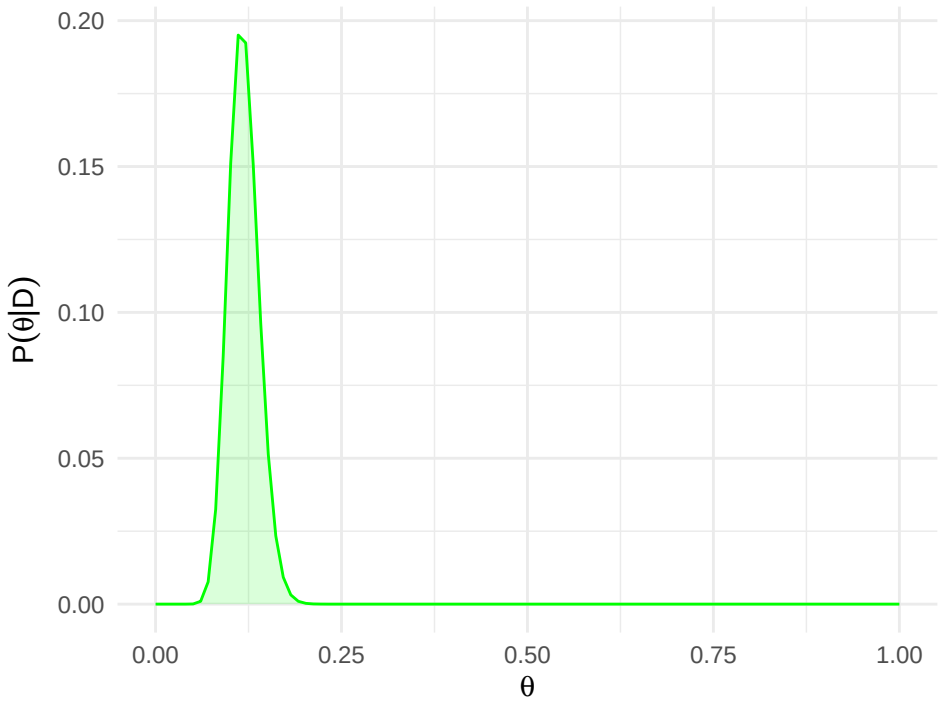
```
df <- data.frame(theta = theta, prior = pDataGivenTheta)
ggplot(df, aes(x = theta, y = prior)) +
  geom_line(color = "green") +
  geom_area(fill = "green", alpha = 0.15) +
  xlab(TeX("$\\theta$")) + ylab(TeX("$P(D | \\theta)$")) +
  labs(caption = "Likelihood function") +
  theme_minimal()  +
  theme(plot.caption = element_text(hjust = 0.5))
```



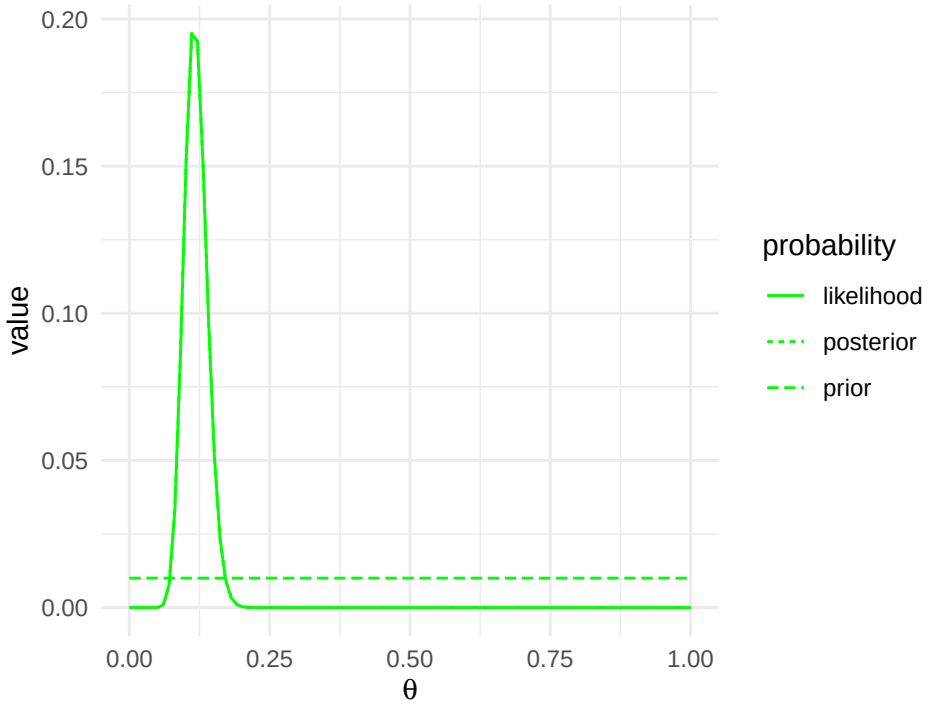
Posteriori distribution

```
# Marginal probability of data
pData <- sum(pDataGivenTheta * pTheta)
# Bayes Theorem
pThetaGivenData <- pDataGivenTheta * pTheta / pData
```

```
posterior_uniform <- pThetaGivenData
```



A posteriori distribution of θ, using as prior distribution the uniform distribution



Prior, likelihood and posterior comparison

```
df1 <- data.frame(theta = theta,
                  prior = pTheta,
                  likelihood = pDataGivenTheta,
```

```
                  posterior = pThetaGivenData)

# id stores the row with the maximum "posterior".
id <- which.max(df1$posterior)
kable(df1[(id - 3):(id + 3), ], row.names = FALSE)
```

theta	prior	likelihood	posterior
0.0808081	0.01	0.0127547	0.0324666
0.0909091	0.01	0.0333997	0.0850174
0.1010101	0.01	0.0593546	0.1510844
0.1111111	0.01	0.0766326	0.1950649
0.1212121	0.01	0.0755793	0.1923837
0.1313131	0.01	0.0591406	0.1505398
0.1414141	0.01	0.0378083	0.0962393

Maximum a posteriori estimator

Investigated **maximum a posteriori estimatori** is equal to **0.(1)** when computed by mode, and **0.12** when computed by mean.

```
# Using mode:
theta_estimated_mode <- theta[which.max(pThetaGivenData)]
theta_estimated_mode
## [1] 0.1111111

# Using the mean or expectation:
theta_estimated_mean <- sum(theta * pThetaGivenData)
theta_estimated_mean
## [1] 0.1185771
```

Interval estimation

```
density <- data.frame(x = theta,
                      y = pThetaGivenData)
class(density) <- "density"

# By default, it considers 95% intervals
interval <- hdi(density, credMass = 0.95)
interval
##      lower      upper
## 0.08080808 0.15151515
## attr(,"credMass")
## [1] 0.95
## attr(,"height")
## [1] 0.03246659
```

```
lower_interval <- interval["lower"]
upper_interval <- interval["upper"]
```

```r
p <- sum(pThetaGivenData[(theta > lower_interval) & (theta < upper_interval)])
p
## [1] 0.8703295
```

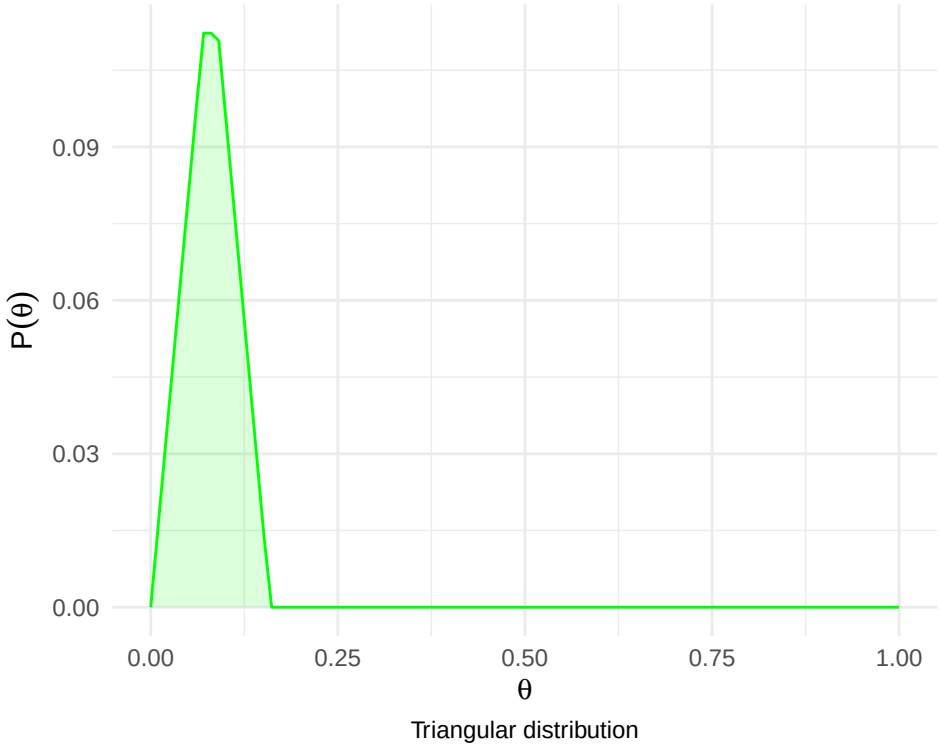
Analysis with given priori

In this case, knowledge about the probability being between 7% and 9% is used.

Priori

Given trapezoid-shaped prior is used.

```r
pTheta <- pmax(0, pmin(1, pmin((1 / 0.07)*theta , -(1/0.07)*theta + (1/0.07)*0.09+1 )))
pTheta <- pTheta / sum(pTheta)
```



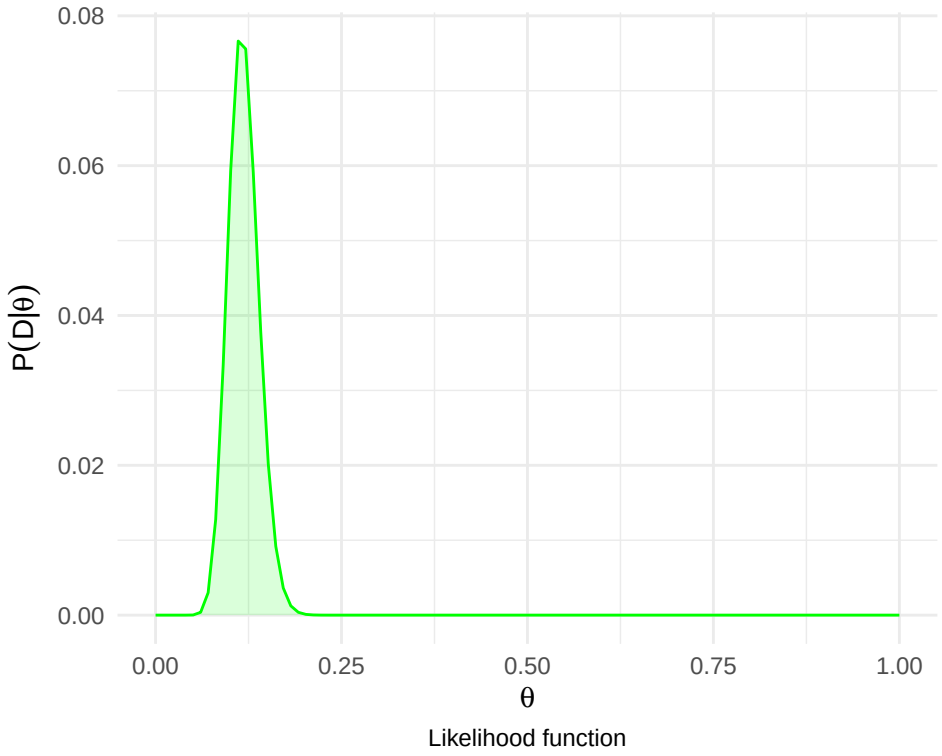
Likelyhood function

In this case, likelihood function is again given by binomial distribution, which models how many patients from given population have Alzheimer's disease.

```r
pDataGivenTheta <- choose(N, nSick) * theta^nSick * (1 - theta)^nHealthy
```

```r
df <- data.frame(theta = theta, prior = pDataGivenTheta)
ggplot(df, aes(x = theta, y = prior)) +
  geom_line(color = "green") +
```

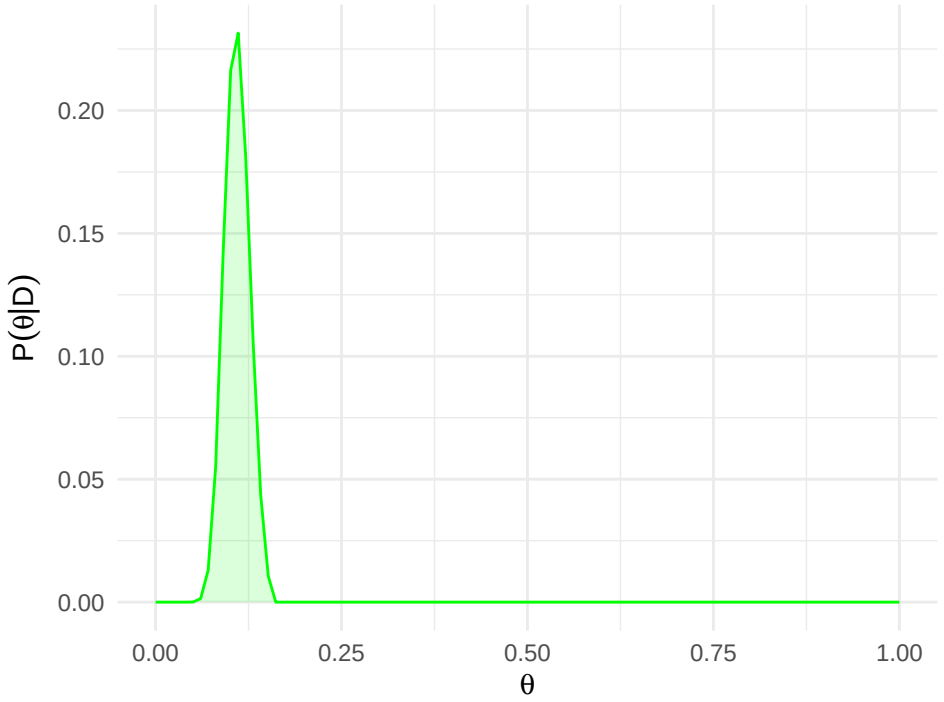
```
  geom_area(fill = "green", alpha = 0.15) +
  xlab(TeX("$\\theta$")) + ylab(TeX("$P(D | \\theta)$")) +
  labs(caption = "Likelihood function") +
  theme_minimal()  +
  theme(plot.caption = element_text(hjust = 0.5))
```



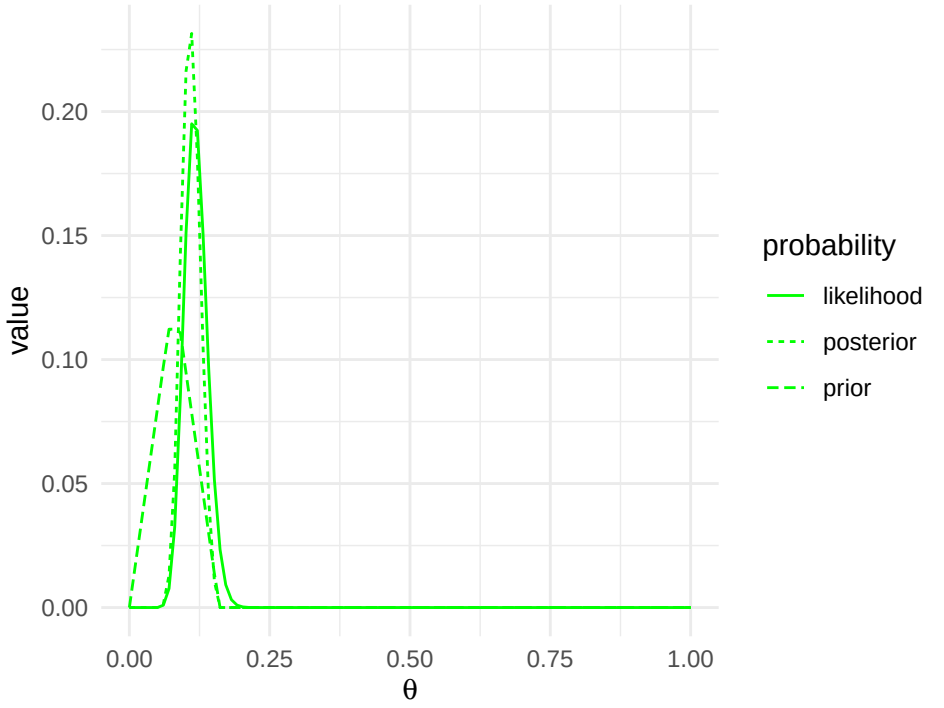
Posteriori distribution

```
# Marginal probability of data
pData <- sum(pDataGivenTheta * pTheta)
# Bayes Theorem
pThetaGivenData <- pDataGivenTheta * pTheta / pData
```

```
posterior_trapezoidal <- pThetaGivenData
```



A posteriori distribution of θ, using as prior distribution the trapezoidal distributior



Prior, likelihood and posterior comparison

```
df1 <- data.frame(theta = theta,
                  prior = pTheta,
                  likelihood = pDataGivenTheta,
```

```
                  posterior = pThetaGivenData)

# id stores the row with the maximum "posterior".
id <- which.max(df1$posterior)
kable(df1[(id - 3):(id + 3), ], row.names = FALSE)
```

theta	prior	likelihood	posterior
0.0808081	0.1122449	0.0127547	0.0551649
0.0909091	0.1107872	0.0333997	0.1425793
0.1010101	0.0945902	0.0593546	0.2163340
0.1111111	0.0783933	0.0766326	0.2314819
0.1212121	0.0621963	0.0755793	0.1811307
0.1313131	0.0459994	0.0591406	0.1048244
0.1414141	0.0298024	0.0378083	0.0434173

Maximum a posteriori estimator

Investigated **maximum a posteriori estimator** is equal to **0.(1)** when computed by mode, and ~**0.11** when computed by mean.

```
# Using mode:
theta_estimated_mode <- theta[which.max(pThetaGivenData)]
theta_estimated_mode
## [1] 0.1111111

# Using the mean or expectation:
theta_estimated_mean <- sum(theta * pThetaGivenData)
theta_estimated_mean
## [1] 0.1094644
```

Interval estimation

```
density <- data.frame(x = theta,
                      y = pThetaGivenData)
class(density) <- "density"

# By default, it considers 95% intervals
interval <- hdi(density, credMass = 0.95)
interval
##      lower      upper
## 0.08080808 0.14141414
## attr(,"credMass")
## [1] 0.95
## attr(,"height")
## [1] 0.04341731
```

```
lower_interval <- interval["lower"]
upper_interval <- interval["upper"]
```

```r
p <- sum(pThetaGivenData[(theta > lower_interval) & (theta < upper_interval)])
p
## [1] 0.8763503
```

Summary

Received results:

Non-informative apriori:

- maximum a posteriori estimator:
  - for mode: 0.(1)
  - for mean: 0.1185771
- Interval estimation:
  - lower: 0.08080808
  - upper: 0.15151515
  - p: 0.8703295

Informative apriori:

- maximum a posteriori estimator:
  - for mode: 0.(1)
  - for mean: 0.1094644
- Interval estimation:
  - lower: 0.08080808
  - upper: 0.14141414
  - p: 0.8763503

Analysis that was based on prior knowledge, concluded with more concise mode and mean estimators, as well as more compact interval with higher probability.