# STATISTICAL MODELLING AND FORECASTING

# MA7007

# SPRING SEMESTER 2022-2023

UNEGBU, OLISA UDOCHUKWU 22020843

COURSEWORK

LONDON METROPOLITAN UNIVERSITY

## 1. Introduction

The rationale for this study is to analyse three (3) different data sets. The first section will focus on analysing the subset of Body Mass index (BMI) data to find a suitable Generalized Addictive Models for Location, Scale and Shape (GAMLSS) distribution of the BMI of Dutch boys between 19 to 20 years old. The second section will focus on analysing a sample of the handgrip (HG) strength in relation to gender and age of English school children to create centile curves for the grip given age. The last section will focus on analysing a sample of the different factors affecting the quality of red wine to find an appropriate statistical model for the wine quality.

The Generalized Addictive Model for Location, Scale and Shape (GAMLSS) is a general class of statistical models for a univariate response variable. The distribution for the response variable in GAMLSS can be selected from a general family of distributions which includes highly skew and/or kurtotic continuous and discrete distributions, Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019).

The first data set is the Body Mass Index (BMI) of Dutch Boys. The data is from the Fourth Dutch Growth Study, Fredriks et al. (2000a, 2000b), which is a cross-sectional data that measures growth and development of the Dutch population between the ages 0 and 21 years. The study measured , among other variables, height, weight, head circumference, all as bmi, then age for 7482 males and 7018 females. The BMI of Dutch boys only is what been analyse in this study as a subset of Body Mass Index (BMI) data obtained from the Fourth Dutch Growth Study. The BMI of Dutch Boys is stored in gamlss.data package as 'dbbmi'. The format of the data is a data frame with 7294 observations of 2 variables. The variables names are age and bmi. The age variable is a numeric vector as well as the bmi.

The second data set is subset (only boys) from the hand grip strength data analysed by Cohen et al. (2010) in relation to gender and age in English school children. A sample of 1000 of the 3766 observations of the boys was analysed in this study to create a centile curve for the grip of the boys using my unique set seed (1163). The hand grip strength data is stored in gamlss.data package as 'grip'. The format of the data is a data frame with 1000 observations of 2 variables namely, age and grip. The age represents the age of the participant while the grip represents the handgrip strength of the participant. The age variable is numeric vector as well as the grip.

The third data set is related to red vinho verde wine samples, from the north of Portugal analysed by Cortez et al. (2009). The data was gotten from http://www3.dsi.uminho.pt/pcortez/wine/. The data is based on physiochemical tests to model the wine quality. The data set can be viewed as a regression task. The classes are ordered and not balanced due to many more normal wines than just the excellent or poor wines used for the physiochemical tests. The reason I chose to analyse the wine quality data is to analyse the factors that made what was once considered a luxury good is now increasingly consumed across the world. An 80% subset of the data was analysed using my unique set seed (1163). The format of the data is a data frame with 1278 observations of the 1599 observations of the original data with 12 variables. The data has 11 explanatory variables namely, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and a target variable quantity with score

between 0 and 10. All the 11 independent variables are numeric vectors while the dependent variable is an integer.

## 1. Fitting Distributions for BMI of Dutch Boys data set

The data set is a subset of the Body Mass Index (BMI) of Dutch population between the ages 0 and 21 years old. Ages between 19 to 20 years were used for this coursework. The is.na() function was first used to check for null values, and none was found. The with() and subset() functions were used to extract the needed ages by adding 1 to the initial age 19. The new data frame of the selected ages has 269 observations of 2 variables. The y_hist() function in the gamlss.ggplots package in R library was used to plot a histogram chart for the Dutch boys between the ages of 19 and 20 years old as seen in figure 2.1 below.

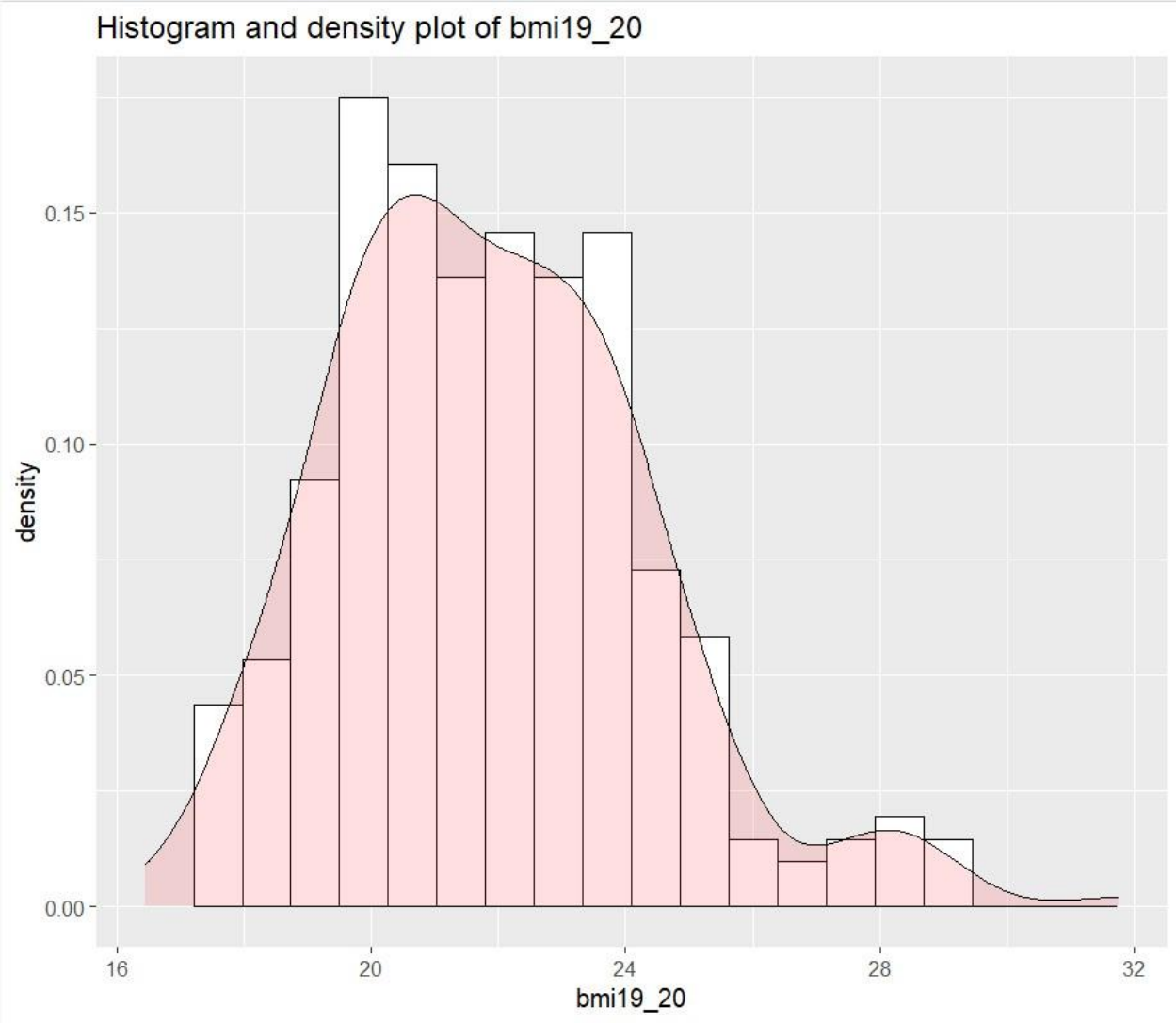



*Figure 2.1: Histogram showing the Dutch boys between the ages of 19 and 20 years.*

The histogram plot in figure 2.1 above shows that the data distribution for the Dutch boys between the ages of 19 and 20 years has a heavier tail on the right which makes it skewed to the right (positive skewness). This means that it has higher values to the right than to the left resulting in an overall asymmetry in the distribution. It also shows a bit of leptokurtic property of thicker tail than the normal distribution. The shape of the distribution shows where the tails of the distribution are thicker (that is, have more extreme values) than the tails of a normal distribution, and the peak of the distribution is higher and sharper than the peak of a normal distribution. Therefore, to fit the response variable using the GAMLSS distribution, the fitDist() function from the gamlss package was used to select all the gamlss family that fits the data using 'realline' type as shown in figure 2.2 below.

```
      SN1       SN2    exGAUS       JSU      JSUo       ST5      EGB2      SEP4       ST1       ST2
1243.802  1244.226  1244.376  1244.513  1244.513  1244.598  1244.655  1244.765  1245.689  1245.723
     SEP1      SEP2       ST3       SST     SHASH      SEP3        RG   SHASHo2    SHASHo       ST4
1245.801  1245.801  1245.815  1245.815  1246.084  1246.096  1246.128  1246.748  1246.748  1249.579
       GT        LO        TF       TF2        NO       PE2        PE       NET        GU
1253.944  1256.608  1257.297  1257.297  1259.295  1259.973  1259.973  1265.158  1361.656
```

*Figure 2.2: showing the different gamlss family distribution that fits the data.*

The fitDist() function fits a set of distributions to the response variable and chooses the one with the smallest GAIC (Generalized Akaike Information Criterion), Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019). The SN1 gamlss family with the smallest GAIC of 1243.802 as shown in figure 2.2 above the best model suitable for the BMI of Dutch Boys data between the ages of 19 to 20 years old. The SN1 family in gamlss is a member of the family of distributions known as the skew-normal family. It features a location parameter which determines the centre of the distribution, a scale parameter which determines the spread of the distribution, and a skewness parameter which determines the direction and degree of skewness.

Figure 2.3 below shows the plot of the quantile residuals using the plot() function.

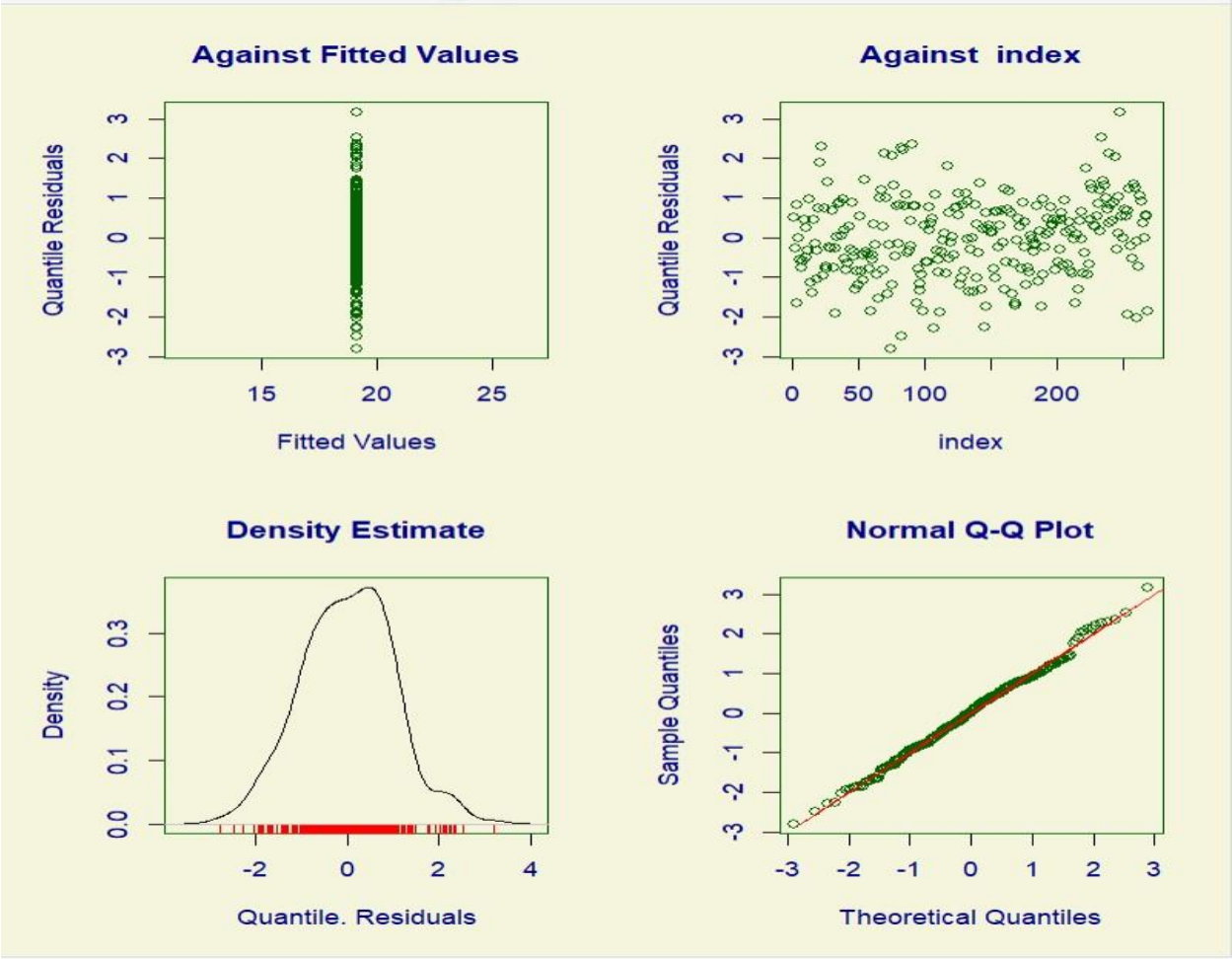



*Figure 2.3: showing the summary quantile residuals plots.*

Figure 2.4 below shows the result of the summary of the quantile residual of the SN1 family statistical model.

```
**********************************************************************
            Summary of the Quantile Residuals
                        mean   =  0.0005081966
                    variance   =  1.001017
             coef. of skewness  =  0.03366683
             coef. of kurtosis  =  3.03161
Filliben correlation coefficient  =  0.9977691
**********************************************************************
```

*Figure 2.4: showing the summary of the quantile residuals.*

The summary of the quantile residuals as shown in figure 2.4 above shows a mean of 0.0005081966 which is very close to zero indicates the model is unbiased on average. The variance of the quantile residuals 1.001017, which is close to 1 indicates that the residuals are approximately normally distributed with a unit variance. The coefficient of kurtosis is 3.03161 which is greater than 3 indicates that the distribution of the residuals has more extreme vails in the tails. The Filliben correlation coefficient of 0.9977691 measures the correlation between the quantile residuals and their expected values.

Figure 2.5 below shows a histogram plot of the fitted SN1 distribution of the BMI of Dutch Boys between the ages of 19 to 20 years data set using the hisDist() function with the nbins = 3.

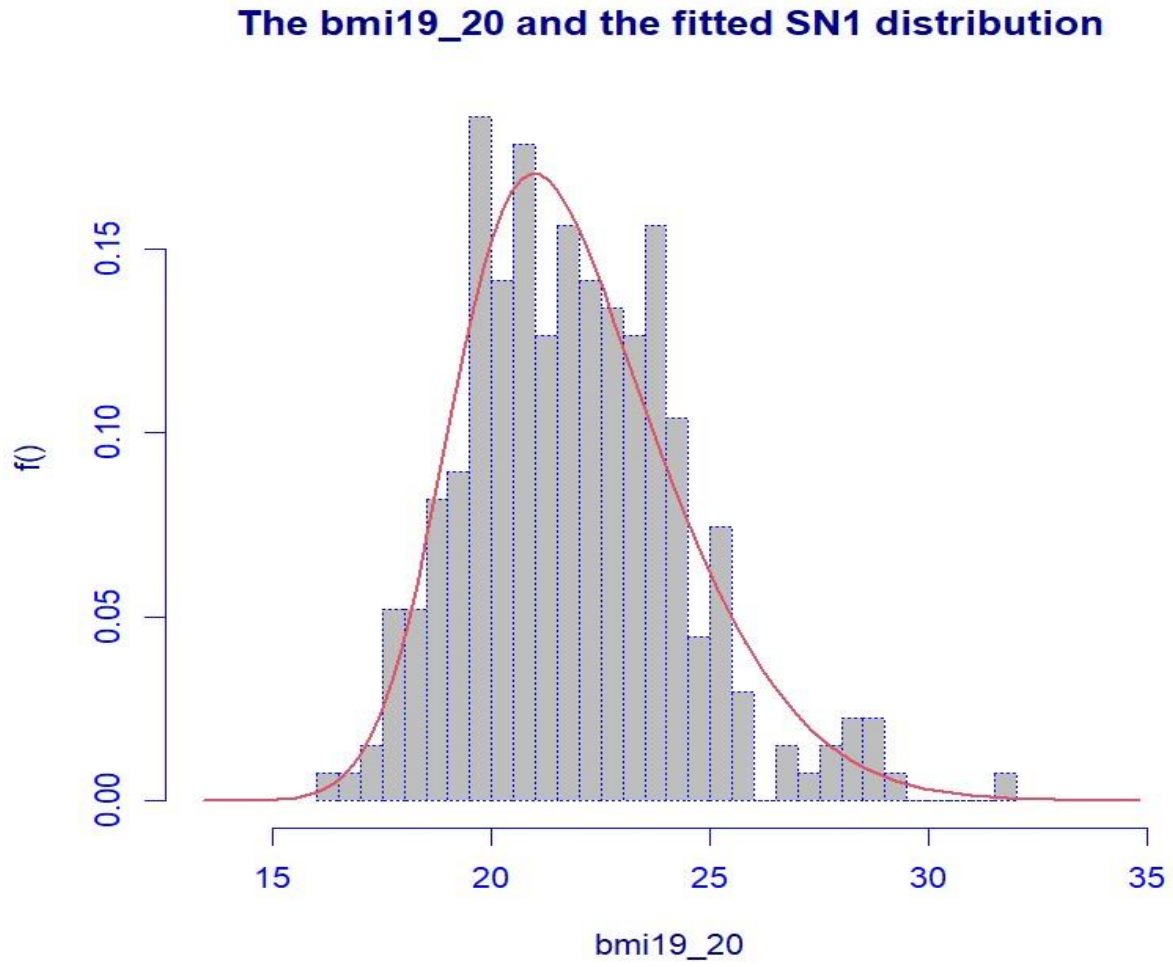



*Figure 2.5: showing the histogram plot of the fitted SN1 distribution.*

Figure 2.6 below shows the summary statistics for the SN1 distribution for the BMI of Dutch Boys between the ages of 19 to 20 years data set using the summary() function.

```
Family:  c("SN1", "Skew normal type 1 (Azzalini type 1)")

Call:  gamlssML(formula = y, family = DIST[i])

Fitting method: "nlminb"


Coefficient(s):
           Estimate  Std. Error  t value   Pr(>|t|)
eta.mu    19.1084045   0.2839335 67.29888 < 2.22e-16 ***
eta.sigma  1.3201019   0.0711042 18.56575 < 2.22e-16 ***
eta.nu     2.6244108   0.6349979  4.13294 3.5815e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Degrees of Freedom for the fit: 3 Residual Deg. of Freedom   266
Global Deviance:     1237.8
            AIC:     1243.8
            SBC:     1254.59
```

*Figure 2.6: showing the summary statistics of the SN1 distribution.*

The summary statistics of SN1 distribution also known as the Azzalini type 1 distribution for the BMI of Dutch Boys between the ages of 19 to 20 years data set as shown in figure 2.6 above used a fitting method ‘nlminb’ to optimize the algorithm used to fit the model. The mu shows the estimated coefficient of 19.11 for the location parameter of the distribution which determines the centre of the distribution. The sigma shows the estimated coefficient of 1.32 for the scale parameter of the distribution which determines the spread of the distribution. The nu shows the estimated coefficient2.62 for the shape parameter of the distribution which determines the degree and direction of skewness in the distribution. The degrees of freedom that fits the model is 3 for the SN1 family and 266 residual degrees of freedom. The global deviance of 1237.8 measures how well the model fits the data. The AIC of 1243.8 shows the Akaike information Criterion (AIC) value for the model which measures the model’s goodness of fit and complexity. The SBC of 1254.59 shows the Schwarz Bayesian Criterion (SBC) value for the model which also measures the goodness of fit and complexity.

2. Centile estimation for the hand grip strength data

The data is a subset (only boys) from the data analysed by Cohen et al. (2010) in relation to gender and age in English school children. The data consist of 3766 observations with 2 variables, age and grip. A sample of 1000 observations was analysed for the purpose of this study using a unique set seed (1163). The sample was extracted from the grip data suing the sample() function. The age and grip variables are both numeric vectors. The grip was plotted against the age using the plot() function as shown in figure 3.1 below.

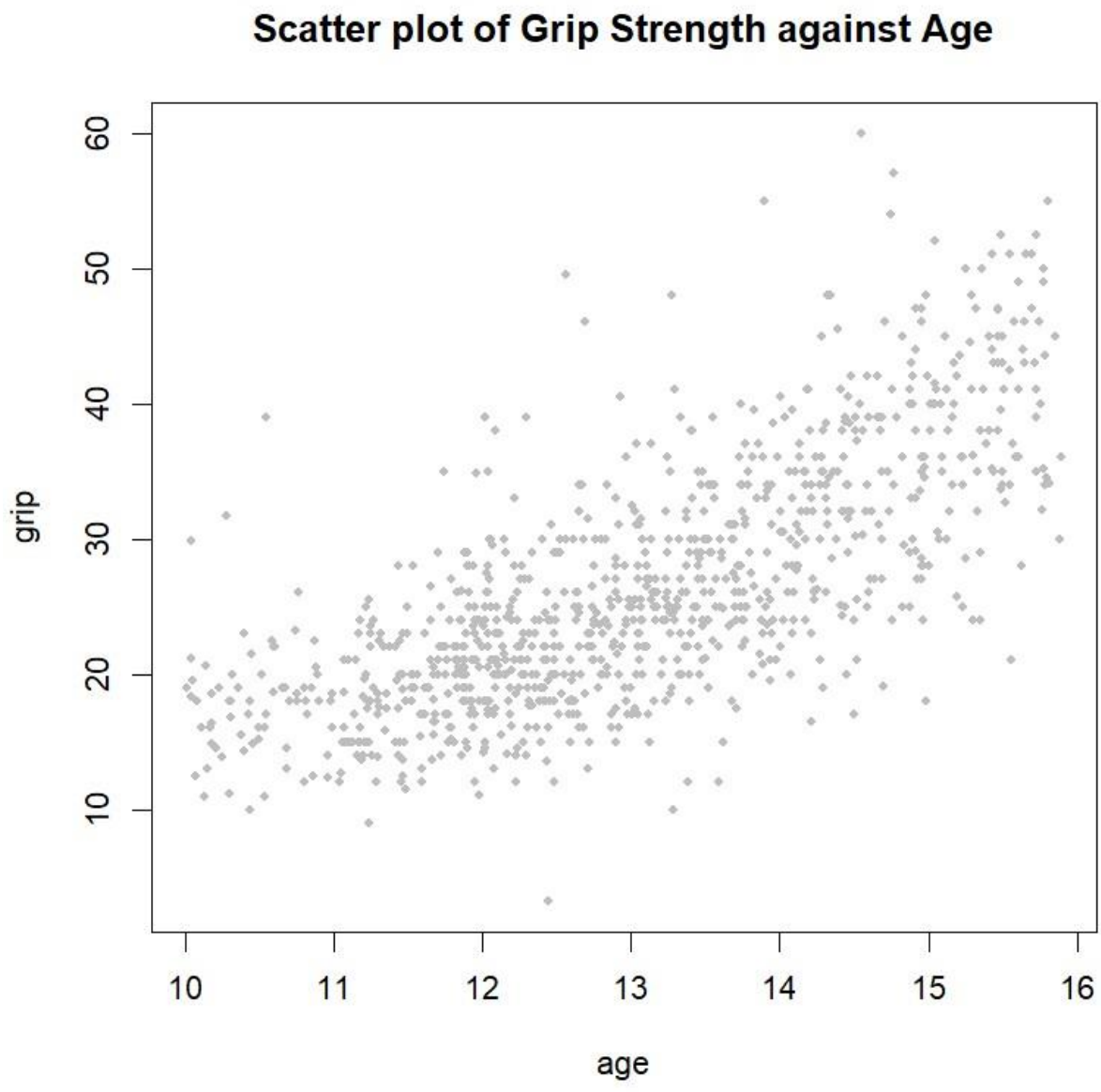



*Figure 3.1: showing plot of grip~age.*

The power transformation was not used for this data set because the data set is already normally distributed as can be seen in figure 3.1 above. Power transformations are usually used in data analysis to

correct for non-normality, heteroscedasticity, or other problems that violate assumptions of statistical models or tests.

For this study, the BCCG, BCT, BCPE distributions were used to fit the hand grip data set. The Box-Cox Cole and Green (BCCG) also known as the LMS method in centile estimation uses three parameters BCCG(μ,σ,ν). The Box-Cox t (BCT) uses four parameters BCT(μ,σ,ν,τ) and BCTo(μ,σ,ν,τ) for skewness and leptokurtosis. The Box-Cox power exponential (BCPE) uses four parameters as well BCPE(μ,σ,ν,τ) and BCPEo(μ,σ,ν,τ) for skewness and platy-lepto Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019).

The BCCG distribution was used first to fit the data and the result is shown in figure 3.2 below.

```
GAMLSS-RS iteration 1: Global Deviance = 6308.841
GAMLSS-RS iteration 2: Global Deviance = 6301.183
GAMLSS-RS iteration 3: Global Deviance = 6301.154
GAMLSS-RS iteration 4: Global Deviance = 6301.156
GAMLSS-RS iteration 5: Global Deviance = 6301.156
> edfAll(gbccg)
$mu
$mu$`pb(age)`
[1] 5.09299


$sigma
$sigma$`pb(age)`
[1] 2.002745


$nu
$nu$`pb(age)`
[1] 2.000126
```

*Figure 3.2: BCCG distribution result.*

The BCCG distribution result is shown in figure 3.2 above indicates that mu parameter represents the estimated location (mean) parameter with an estimated coefficient of 5.09299 for the smooth function of age. The sigma parameter represents the estimated scale (standard deviation) parameter with an estimated coefficient of 2.002745 for the smooth function of age. The nu parameter represents the estimated shape (degrees of freedom) with an estimated coefficient of 2.000126 for the smooth function of age.

The BCT distribution was used next to fit the data using the BCCG distribution as starting values and the result is shown in figure 3.3 below.

```
GAMLSS-RS iteration 1: Global Deviance = 6268.815
GAMLSS-RS iteration 2: Global Deviance = 6267.938
GAMLSS-RS iteration 3: Global Deviance = 6267.902
GAMLSS-RS iteration 4: Global Deviance = 6267.905
GAMLSS-RS iteration 5: Global Deviance = 6267.912
GAMLSS-RS iteration 6: Global Deviance = 6267.915
GAMLSS-RS iteration 7: Global Deviance = 6267.917
GAMLSS-RS iteration 8: Global Deviance = 6267.918
> edfAll(gbct)
$mu
$mu$`pb(age)`
[1] 5.257096


$sigma
$sigma$`pb(age)`
[1] 2.002237


$nu
$nu$`pb(age)`
[1] 2.000097


$tau
$tau$`pb(age)`
[1] 2.829418
```

*Figure 3.3: BCT distribution result.*

The BCT distribution result is shown in figure 3.3 above indicates that mu parameter represents the estimated location (mean) parameter with an estimated coefficient of 5.257096 for the smooth function of age. The sigma parameter represents the estimated scale (standard deviation) parameter with an estimated coefficient of 2.002237 for the smooth function of age. The nu parameter represents the estimated shape (degrees of freedom) with an estimated coefficient of 2.829418 for the smooth function of age. The tau parameter represents an additional parameter with an estimated coefficient of 2.763941 indicating that the tails of the distribution are heavier than that of a normal distribution.

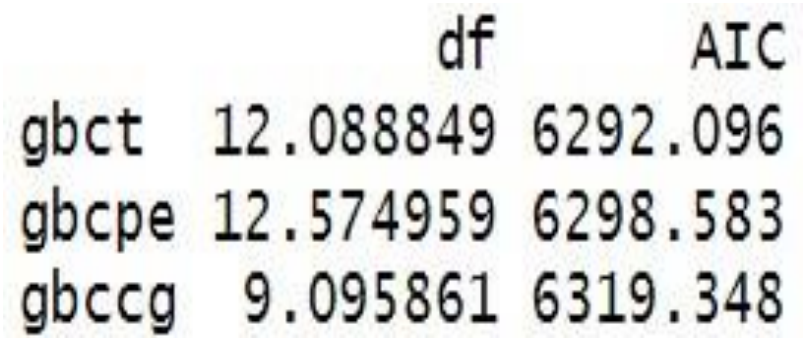
The BCPE distribution was finally used to fit the data using the BCCG distribution as starting values and the result is shown in figure 3.4 below.

```
GAMLSS-RS iteration 1: Global Deviance = 6275.833
GAMLSS-RS iteration 2: Global Deviance = 6273.677
GAMLSS-RS iteration 3: Global Deviance = 6273.493
GAMLSS-RS iteration 4: Global Deviance = 6273.45
GAMLSS-RS iteration 5: Global Deviance = 6273.438
GAMLSS-RS iteration 6: Global Deviance = 6273.434
GAMLSS-RS iteration 7: Global Deviance = 6273.433
GAMLSS-RS iteration 8: Global Deviance = 6273.433
> edfAll(gbcpe)
$mu
$mu$`pb(age)`
[1] 5.311763


$sigma
$sigma$`pb(age)`
[1] 2.49915


$nu
$nu$`pb(age)`
[1] 2.000105


$tau
$tau$`pb(age)`
[1] 2.763941
```

*Figure 3.4: BCPE distribution result.*

The BCPE distribution result is shown in figure 3.4 above indicates that mu parameter represents the estimated location (mean) parameter with an estimated coefficient of 5.311763 for the smooth function of age. The sigma parameter represents the estimated scale (standard deviation) parameter with an estimated coefficient of 2.49915 for the smooth function of age. The nu parameter represents the estimated shape (degrees of freedom) with an estimated coefficient of 2.000105 for the smooth function of age. The tau parameter represents an additional parameter with an estimated coefficient of 2.763941 indicating that the tails of the distribution are heavier than that of a normal distribution.

The Generalized Akaike Information Criterion (GAIC) was used to compare the BCCG, BCT, and BCPE and the result is shown in figure 3.5 below.



| | df | AIC |
|---|---|---|
| gbct | 12.088849 | 6292.096 |
| gbcpe | 12.574959 | 6298.583 |
| gbccg | 9.095861 | 6319.348 |

*Figure 3.5: GAIC result for BCCG, BCT, and BCPE models.*

The results shown in figure 3.5 above shows the degree of freedom and the Akaike Information Criterion (AIC) the Box-Cox Cole and Green (BCCG), Box-Cox t (BCT), and Box-Cox power exponential (BCPE) models fitted to a sample of 1000 observations of the hand grip strength data set. The Akaike Information Criterion (AIC) measures the relative quality of statistical models for a dataset which is calculated based on the likelihood function and the number of parameters in the model, Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019). The lower the AIC value, the more suitable the model. The results show that the BCT model has the highest degrees of freedom (df) of 12.088849 and the lowest AIC value of 6292.096. Followed by the BCPE model which has a moderate degrees of freedom of 12.574959 with a slightly higher AIC value of 6298.583. The BCCG has the lowest degrees of freedom of 9.095861 and the highest AIC value of 6319.348. The result indicates that the BCT model is the more suitable model and fits better than the BCCG and BCPE.

The fittedplot() function in the gamlss.tools package in R programming language was used to create plots of the fitted values for the BCCG, BCT, and BCPE GAMLSS models as shown in figure 3.6 below.

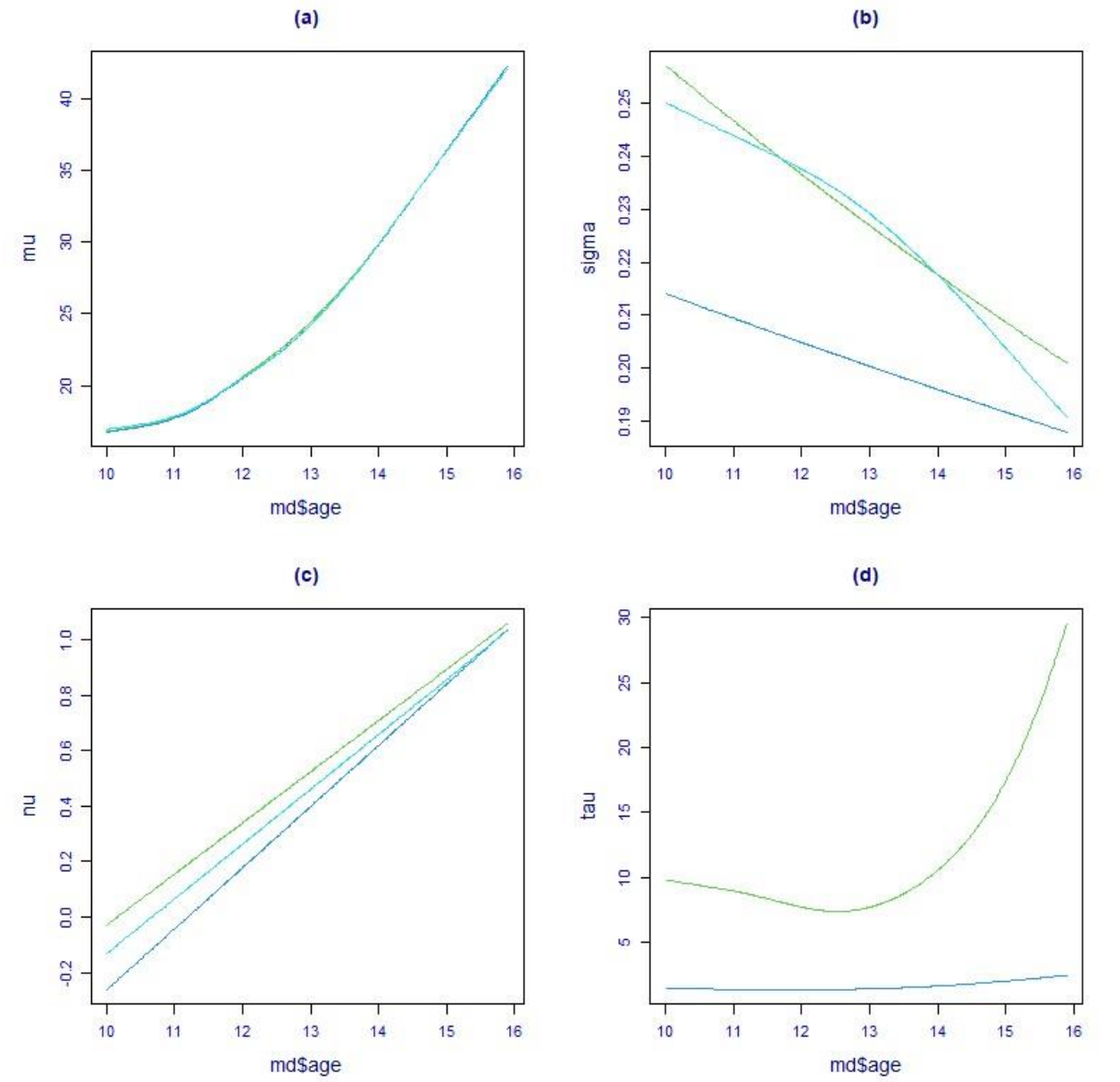



*Figure 3.5: Fitted plot for BCCG, BCT, and BCPE models.*

The centile plot for the fitted BCCG model was obtained using the centile() function as shown in figure 3.6 below.

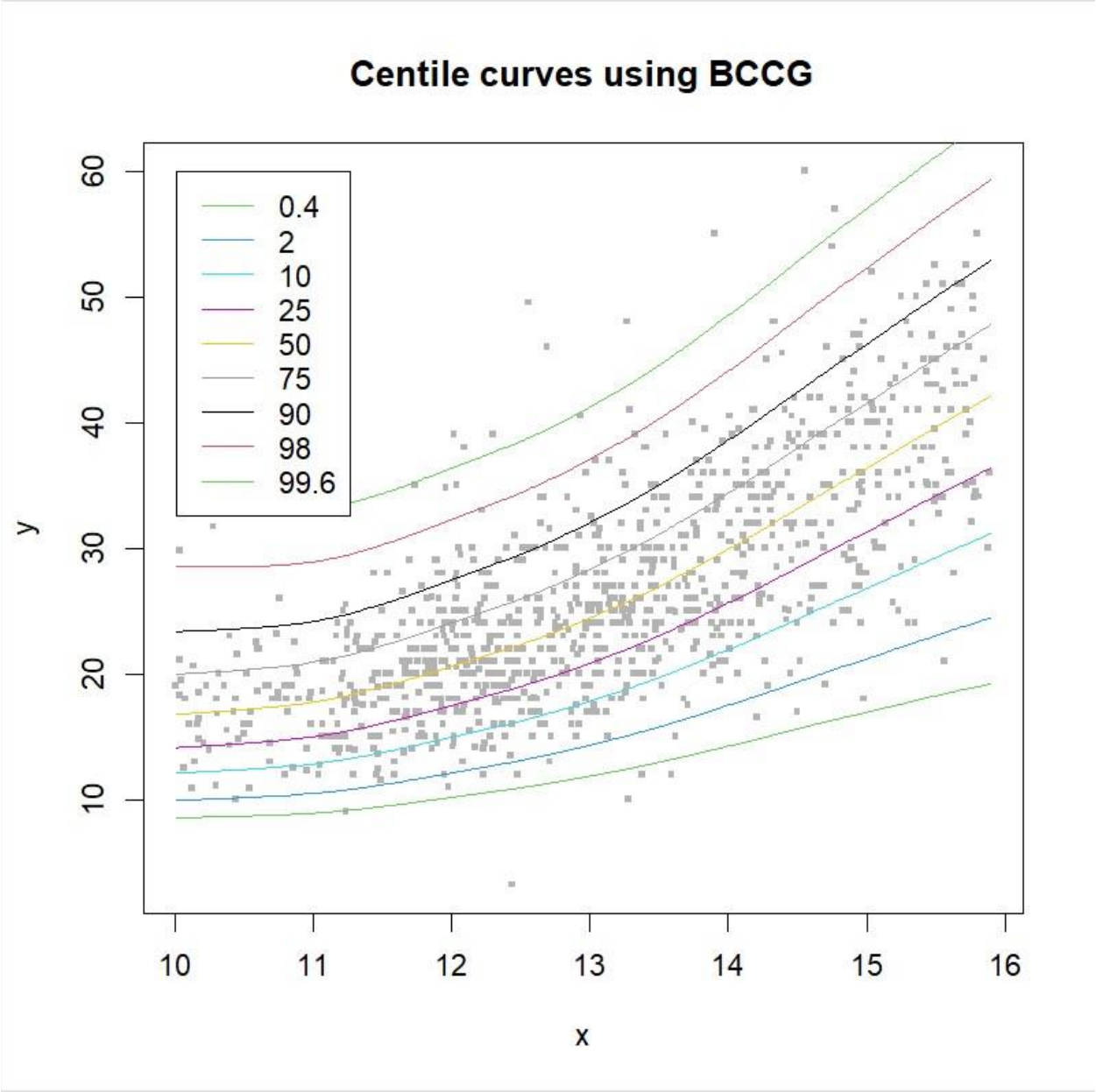



*Figure 3.6: Centile plot for the BCCG model.*

The centile plot for the fitted

BCT model was obtained using the centile() function as shown in figure 3.7 below.

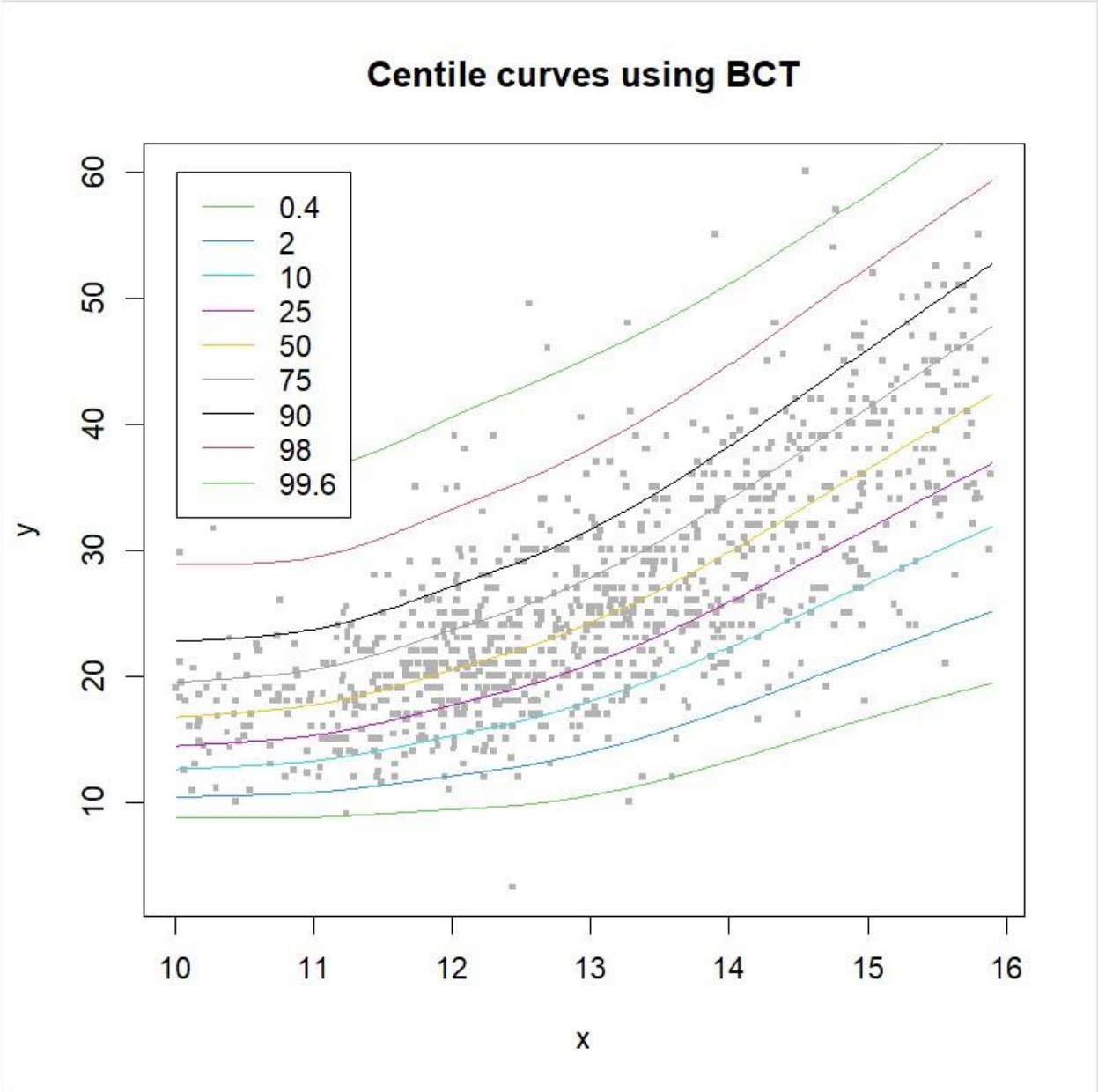



*Figure 3.7: Centile Plot for the BCT model.*

The centile plot for the fitted

BCPE model was obtained using the centile() function as shown in figure 3.8 below.

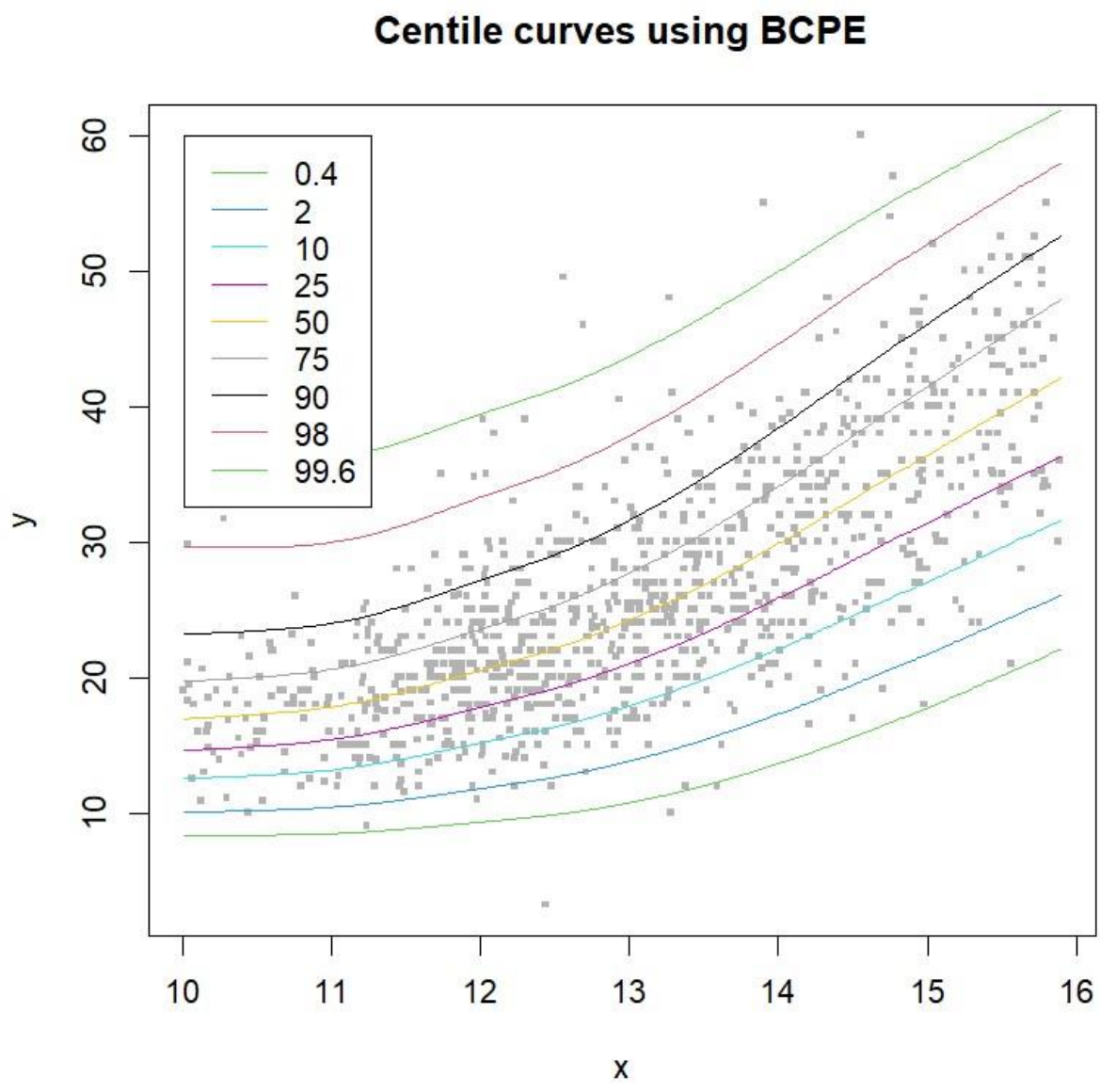



*Figure 3.8: Centile Plot for the BCPE model.*

The centile plot for the fitted

The centile() function also calculates the percentiles of the data at different cut-off points of the x-axis (age) as shown in figure 3.9 below.

```
> centiles(gbccg, xvar=md$age)
% of cases below  0.4 centile is  0.5
% of cases below  2 centile is  1.7
% of cases below  10 centile is  9.1
% of cases below  25 centile is  25.3
% of cases below  50 centile is  50.7
% of cases below  75 centile is  76.9
% of cases below  90 centile is  91.1
% of cases below  98 centile is  98
% of cases below  99.6 centile is  99
```

```
> centiles(gbct, xvar=md$age)
% of cases below  0.4 centile is  0.2
% of cases below  2 centile is  1.6
% of cases below  10 centile is  10.5
% of cases below  25 centile is  25.9
% of cases below  50 centile is  50
% of cases below  75 centile is  74.5
% of cases below  90 centile is  90.3
% of cases below  98 centile is  98
% of cases below  99.6 centile is  99.3
```

```
> centiles(gbcpe, xvar=md$age)
% of cases below  0.4 centile is  0.3
% of cases below  2 centile is  1.6
% of cases below  10 centile is  9.7
% of cases below  25 centile is  26.3
% of cases below  50 centile is  50.3
% of cases below  75 centile is  74.6
% of cases below  90 centile is  90.4
% of cases below  98 centile is  98
% of cases below  99.6 centile is  99.3
```

*Figure 3.9: showing the percentile values for the BCCG, BCT, and BCPE models.*

The results shown in figure 3.9 above show the percentage of cases that fall below the specified cut-off points for each model. For instance, 25.3% of cases are below the 25$^{th}$ percentile, while 76.9% of cases are below the 75$^{th}$ percentile for the BCCG model. Likewise, for the BCT model, 25.9% of cases are below the 25$^{th}$ percentile, while 74.5% of cases are below the 75$^{th}$ percentile. Lastly for the BCPE model, 26.3% of cases are below the 25$^{th}$ percentile, while 74.6% of cases are below the 75$^{th}$ percentile. The overall results suggest that three (3) of the models, BCCG, BCT, and BCPE are similar in the sense of their capability to predict percentiles of the hand of strength data, with few slight differences in their performance at certain cut-off points.

The residuals from the fitted models of BCCG, BCT, and BCPE were investigated using the plot(), wp(), and Q-stats() functions. The residual plot, worm plot, Q-statistics plot for the BCT model is shown in figures 3.10, 3.11, and 3.12 respectively.

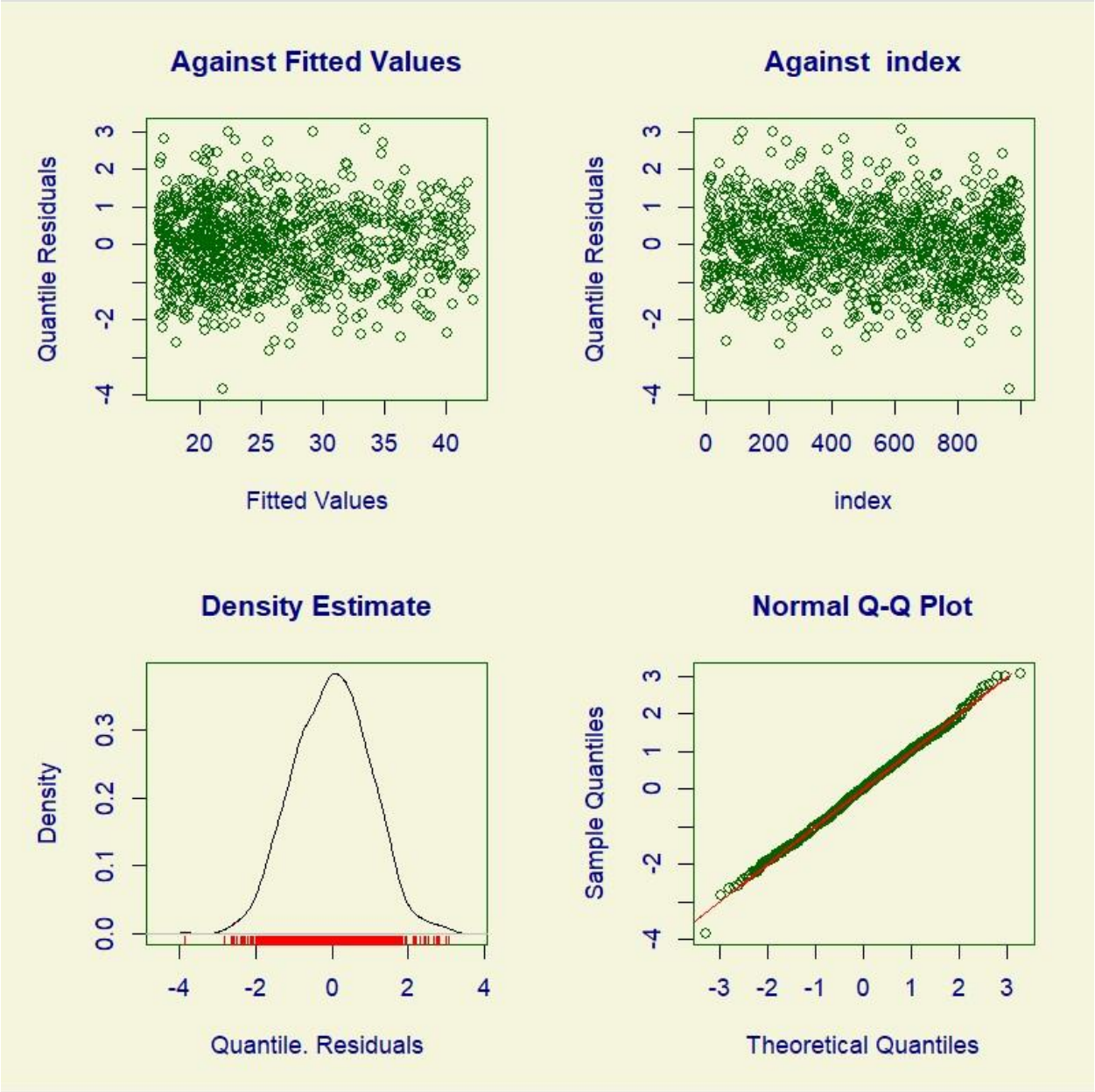



*Figure 3.10: Quartile Residual Plot for BCT model.*

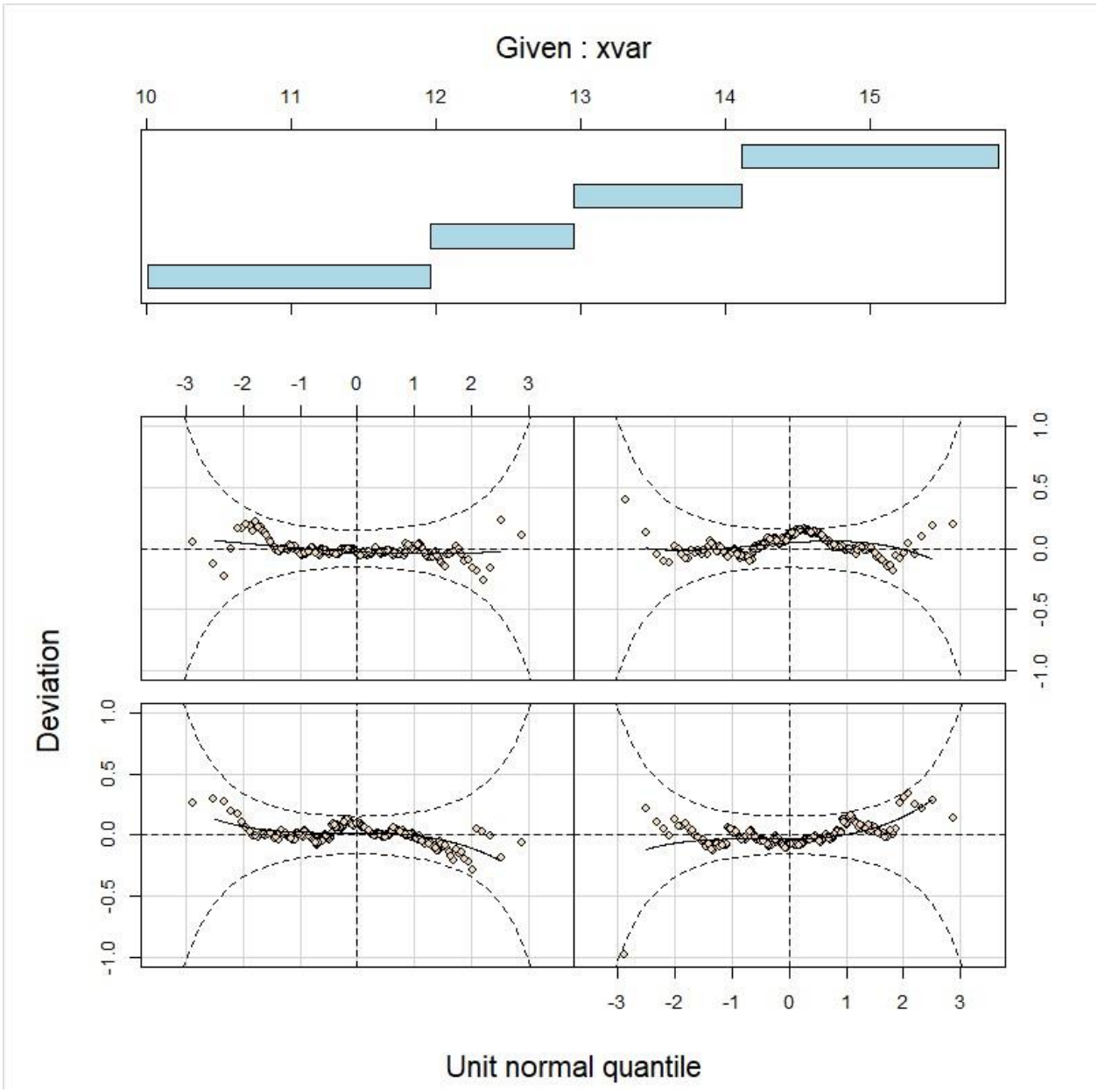



*Figure 3.11:Worm plot for BCT model.*

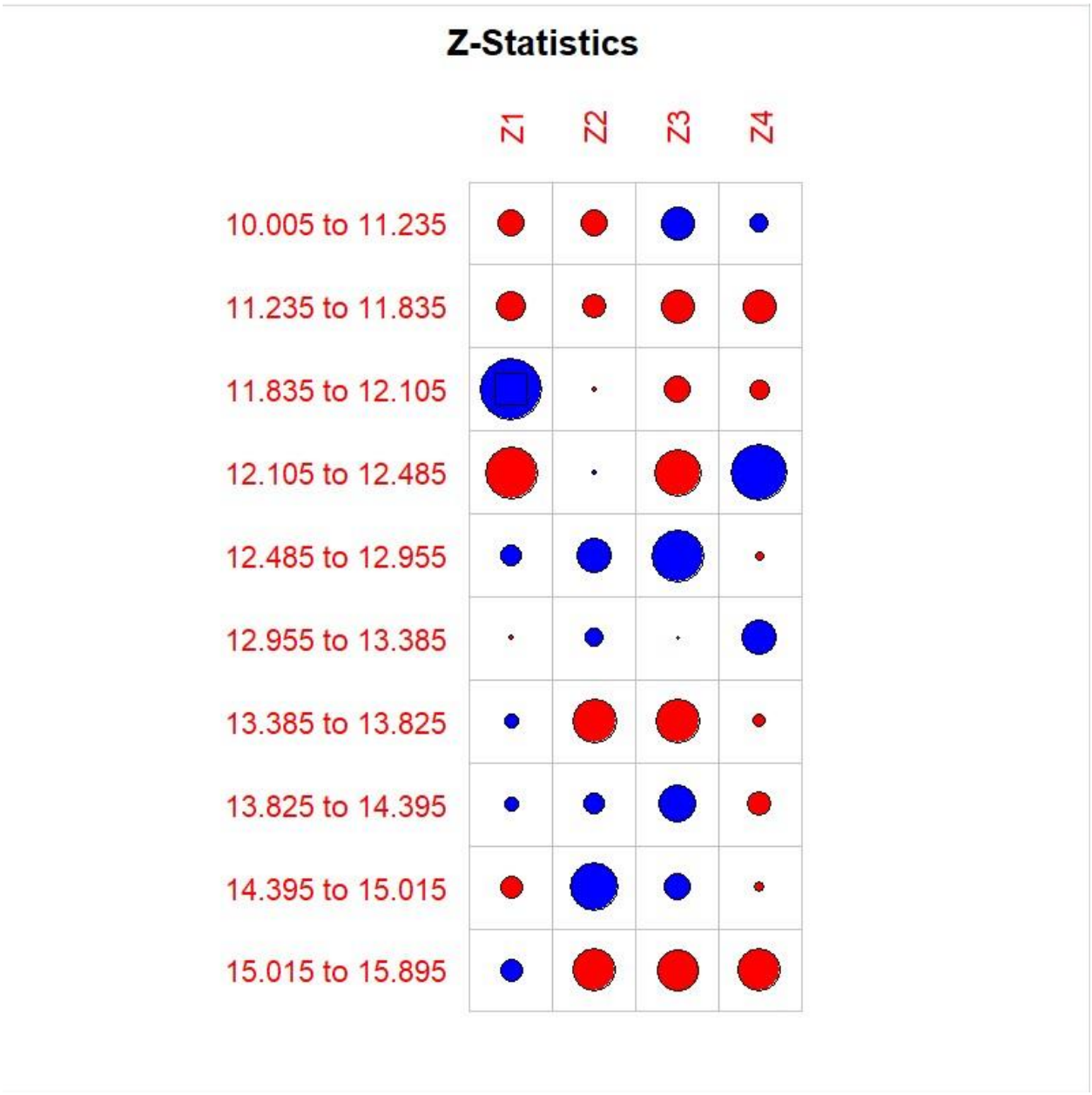



*Figure 3.12: Q-Statistics for BCT model.*

The checkMomentSK() function was used to check the moment skewness and kurtosis of the fitted BCT model as shown in figure 3.13 below.

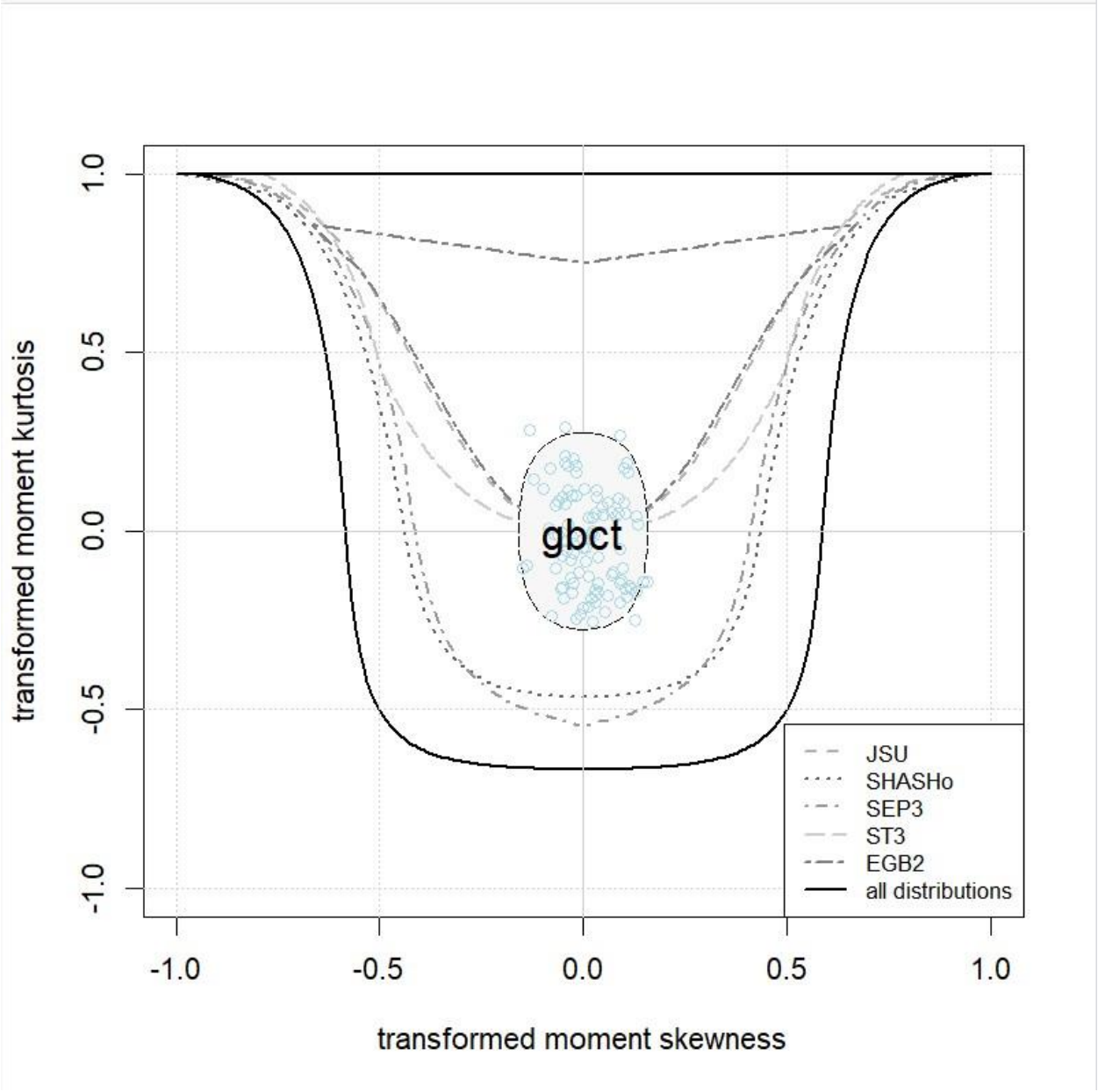



*Figure 3.13: Moment Plot for BCT model.*

The Box-Cox t (BCT) model was chosen as the best model that fit a subset of 1000 observations of the hand grip strength data because has the lowest Akaike Information Criterion (AIC) value of 6292.096 as compared to the BCPE model with 6298.583 and BCCG with 6319.348. Also as shown in figure 3.13, the BCT model data points fit more at the centre than others even though it has a few data points outside the centre box.

## 3. Analysis of Wine Quality in North Portugal

This section focuses on analysing factors affecting wine quality like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol using GAMLSS package. Wine was once viewed as a luxury good, but not so anymore as wider range of consumers increasingly enjoy wine now. Portugal is ranked 10$^{th}$ among wine producing countries in the world with an average production of 6.5million hectolitres according to world population review, 2021. Mankind has been fermenting grapes into wine for thousands of years which has led to the increasing consumption of wine across the globe. This among several factors led to analyse different factors affecting the quality of wine using vinho verde samples from the Minho region of Portugal as case study.

### 3.1. Dataset information and source

The wine quality dataset was collected from http://www3.dsi.uminho.pt/pcortez/wine/ . The data was used to test (ordinal) regression method where the classes are ordered and not balanced due to many more normal wines than just the excellent or poor wines used for the physiochemical tests. The data is based on physiochemical tests to model the wine quality.

The descriptions of the variables of the wine quality dataset is shown in figure 4.1.1 below.

**Table 4.1.1 Descriptions of the wine quality dataset variables**

| Variables | Descriptions |
|---|---|
| Fixed acidity | total amount of acids present in the wine, mainly tartaric acid |
| Volatile acidity | presence of volatile acids, such as acetic acid, in the wine |
| Citric acid | naturally occurring acid found in wine, which can contribute to its overall acidity and freshness |
| Residual sugar | amount of sugar that remains in the wine after fermentation |
| Chlorides | salts present in wine, primarily derived from the grape must or winemaking process |
| Free sulfur dioxide | amount of sulfur dioxide that remains in the wine in its unbound form |
| Total sulfur dioxide | sum of both free and bound forms of sulfur dioxide (SO2) present in wine |
| Density | mass of wine per unit volume and is often an indicator of the wine's body and richness |
| pH | measure of acidity or alkalinity in the wine |
| Sulphates | prevent oxidation and inhibit the growth of unwanted microbes |
| Alcohol | impacts the wine's body, mouthfeel, and perceived sweetness |
| Quality | overall evaluation or assessment of a wine's characteristics |

*Figure 4.1.1: showing descriptions of the wine quality dataset variables.*

The dataset has 1599 observations of 12 variables which includes 11 explanatory variables and 1 response variable. The explanatory variables are numeric vectors while the response variable is integer with scores from 0 to 10.

During the data cleaning process, some of the variable names were renamed due to having long names. The fixed acidity was renamed to f_acid, volatile acidity to v_acid, citric acid to c_acid, residual sugar to r_sugar, chlorides remain same, free sulfur dioxide to fs_oxide, total sulfur dioxide to ts_dioxide, density, pH, sulphates, alcohol, and quality all remained same as they have short names. The shortening of the names helps reduces time spent writing and calling out the variables. The renamed variables are shown in figure 4.1.2 below with their first six (6) observations.

| | f_acid | v_acid | c_acid | r_sugar | chlorides | fs_dioxide | ts_dioxide | density |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11 | 34 | 0.9978 |
| 2 | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25 | 67 | 0.9968 |
| 3 | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15 | 54 | 0.9970 |
| 4 | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17 | 60 | 0.9980 |
| 5 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11 | 34 | 0.9978 |
| 6 | 7.4 | 0.66 | 0.00 | 1.8 | 0.075 | 13 | 40 | 0.9978 |

| | pH | sulphates | alcohol | quality |
|---|---|---|---|---|
| 1 | 3.51 | 0.56 | 9.4 | 5 |
| 2 | 3.20 | 0.68 | 9.8 | 5 |
| 3 | 3.26 | 0.65 | 9.8 | 5 |
| 4 | 3.16 | 0.58 | 9.8 | 6 |
| 5 | 3.51 | 0.56 | 9.4 | 5 |
| 6 | 3.51 | 0.56 | 9.4 | 5 |

*Figure 4.1.2: showing the renamed variables of the wine quality dataset.*

## 3.2. Preliminary analysis for the wine quality dataset

First, a correlation analysis was performed to ascertain the correlation between the response variable and the explanatory variables. The values obtained in the matrix as shown the correlation heatmap in figure 4.2.1 below represents the correlation coefficient, which measures the direction and strength of the linear relationship between the response variable and one explanatory variable at each time. The correlation result shows that alcohol has the highest correlation with the response variable (quality) with a value of 0.476 indicating a moderately positive relationship which suggests that higher alcohol content is related with higher wine quality. The sulphates followed next with a correlation coefficient value of 0.251 which can be seen as a weak positive relationship. This may suggest that wines with higher sulphate levels may be related to slightly higher quality. Next is c_acid with a positive correlation coefficient of 0.226 with the response variable, then f_acid with 0.124, r_sugar with 0.014 which suggests a negligible relationship indicating that residual sugar levels in the wine may not strongly influence its quality. The fs_dioxide, pH, chorides, density, ts_dioxide, and v_acid all has a negative correlation coefficient of -0.051, -0.058, -0.129, -0.175, -0.185, and -0.391 with the response variable respectively.

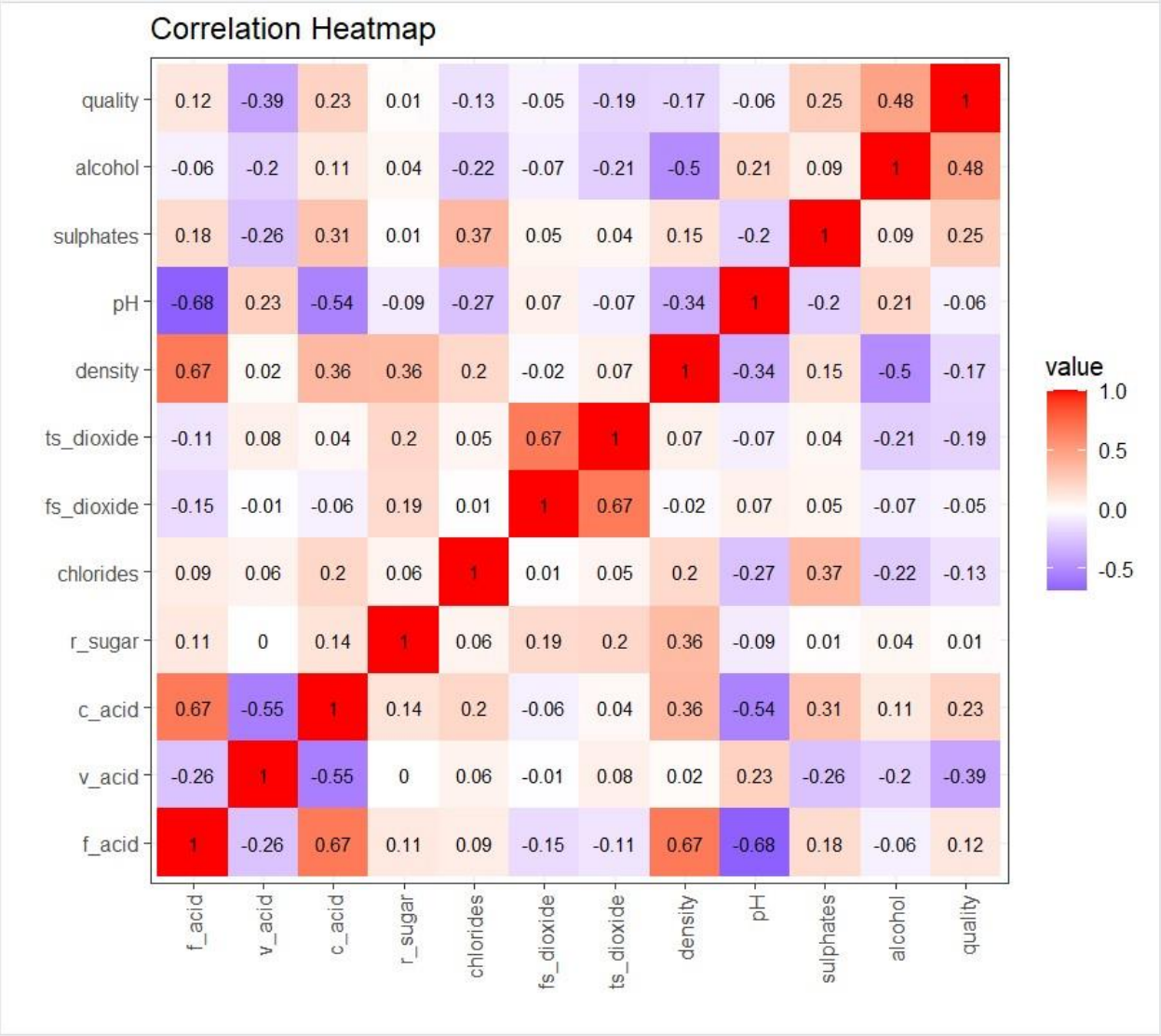



*Figure 4.2.1: showing the correlation heatmap of the wine quality dataset.*

For this study, my unique set seed number (1163) was used, and an 80% subset of the data was randomly selected. This changed the observations of the data from 1599 to 1278 of the same 12 variables. The new data variables were then changed from numeric vectors to logical vectors to enhance the readability and understandability of the code. A correlation analysis was also performed on the new dataset and shows a slight difference with the initial dataset as shown in figure 4.2.2 below.

```
     quality      alcohol    sulphates      c_acid      f_acid      r_sugar
  1.00000000   0.48340391   0.24429477   0.21505711  0.11271803  0.01902585
          pH   fs_dioxide    chlorides      density  ts_dioxide       v_acid
-0.05040876 -0.05415639 -0.14741722 -0.18532198 -0.19107482 -0.39536769
```

*Figure 4.2.2: correlation result of the subset of the wine quality dataset.*

As shown in figure 4.2.2 above, alcohol correlation coefficient with quality slightly increased from 0.476 to 0.483, sulphates decreased from 0.251 to 0.244, c_acid decreased from 0.226 to 0.215, f_acid

decreased from 0.124 to 0.112, r_sugar increased from 0.014 to 0.019, pH increased from -0.058 to 0.05, fs_dioxide decreased from -0.051 to -0.054, chlorides decreased from -0.129 to -0.147, density decreased from -0.175 to -0.185, ts_dioxide decreased from -0.185 to -0.191, and v_acid decreased from -0.391 to -0.395. The differences may have little or no influence on the analysis in 4.2.

## 3.3. Model selection and diagnostic

Model selection in GAMLSS involves finding the right distribution for the response variable by considering the explanatory variable and estimating the location (μ), scale (σ), skewness (ν), and the kurtosis(τ) of the explanatory variable based on the selected distribution, Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019). For the wine quality dataset which requires a binomial kind of distribution, several GAMLSS binomial distribution models were used to fit the data. The Binomial Family (BI) was first used to fit the model. The BI family is used when the response variable follows a binomial distribution. Then the Zero-Altered Binomial family (ZABI) which is an extension of the binomial distribution that accounts for excess zeros in the data. Next was the Zero-Inflated Binomial family(ZIBI) which is another extension of the binomial distribution. Lastly used was the Double Binomial family (DBI) which is used for modelling responses that exhibit overdispersion, when variance is greater than mean, Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019).

The DBI was chosen as the best distribution that fits the data using the Generalized Akaike Information Criterion (GAIC) because it had the lowest Akaike Information Criterion (AIC) value of 3109.684 and degrees of freedom of 2 as shown in figure 4.3.1 below.

| | df | AIC |
|---|---|---|
| mdbi | 2 | 3109.684 |
| mbi | 1 | 3881.055 |
| mzabi | 2 | 3882.443 |
| mzibi | 2 | 3883.056 |

*Figure 4.3.1: showing the GAIC results.*

The fitDist() function was alternatively used to confirm the DBI family distribution that fits the data and has the lowest AIC value of 3109.684 as shown in figure 4.3.2 below.

| DBI | BI | ZABI | BB | ZIBI | ZABB | ZIBB |
|---|---|---|---|---|---|---|
| 3109.684 | 3881.055 | 3882.442 | 3883.055 | 3883.055 | 3884.442 | 3885.055 |

*Figure 4.3.2: showing the different family distributions that fit the data.*

The gamlss() function was later used to refit the DBI distribution by exploring the relationship between the response variable and the explanatory variables and specifies only the explanatory variables without the response variable. The DBI distribution that specifies only the explanatory variables without the response variable was selected as the best DBI distribution model that fits the wine quality dataset using the GAIC to compare the refitted distribution to the initial one having the lowest AIC value of 2488.951

as shown in figure 4.3.3 below which is lower than the initial AIC value of 3109.684 as shown in figure 4.3.2 above.

| | df | AIC |
|---|---|---|
| mdbi2 | 24 | 2488.951 |
| mdbi1 | 13 | 2555.666 |
| mdbi | 2 | 3109.684 |

*Figure 4.3.3: showing the GAIC results of the refitted DBI distribution model.*

Figure 4.3.4 below shows the worm plot for the best fitted DBI distribution model for the wine quality dataset.

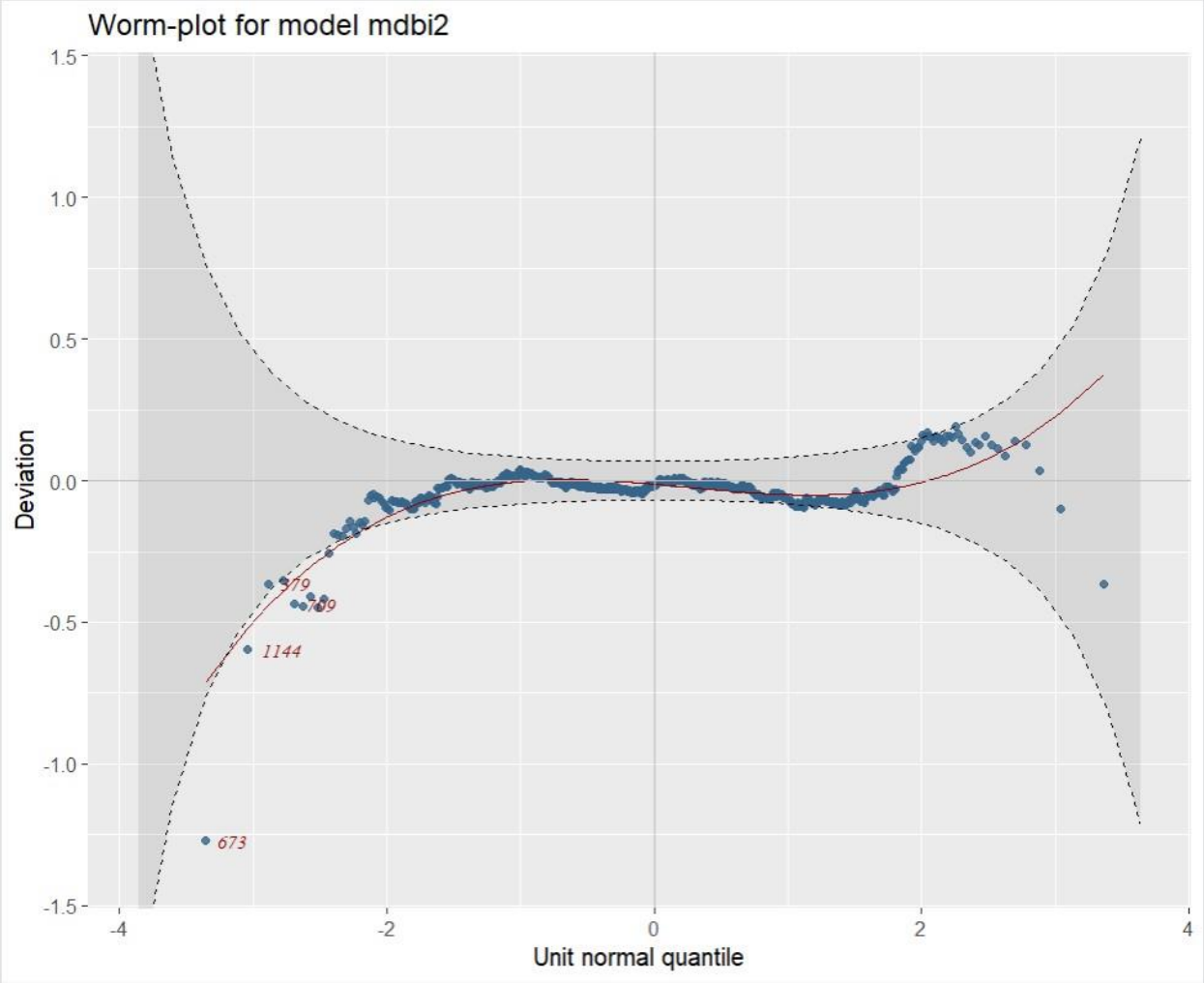



*Figure 4.3.4: worm plot for the DBI distribution model.*

The worm plot shown in figure 4.3.4 above is not perfect but good enough as there needs to be further analysis on the data set which could focus on the outliers in the data, overfitting or underfitting, or other factors that affect wine quality that is not included in the data.

Figure 4.3.5 below shows the randomized quantile residuals of the DBI distribution that fits the wine quality data.

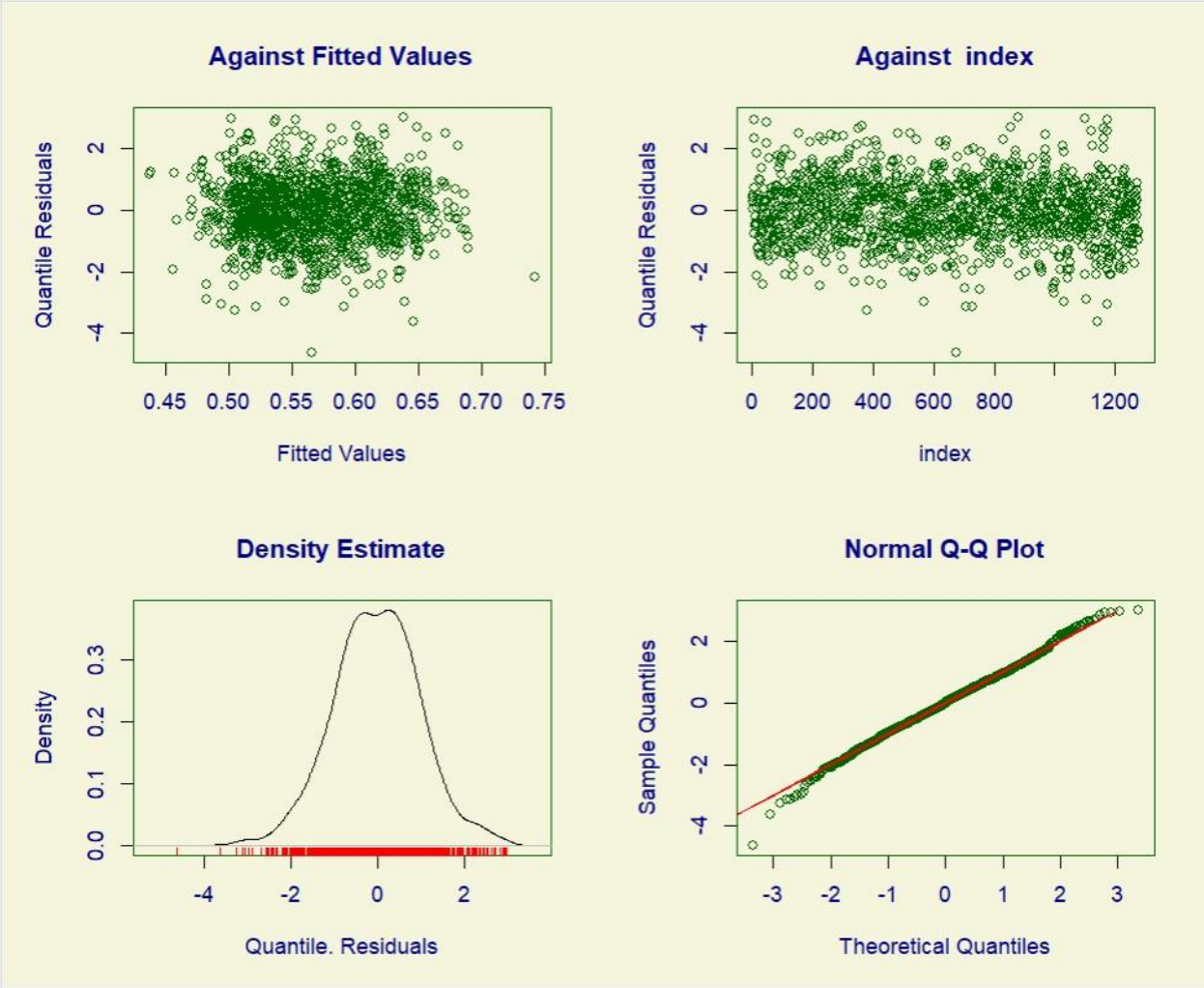



*Figure 4.3.5: showing randomized quantile residuals the DBI distribution.*

The summary of the randomized quantile residuals is shown in figure 4.3.5 below.

```
**********************************************************************
        Summary of the Randomised Quantile Residuals
                           mean   =  -0.02632755
                       variance   =  1.028388
                  coef. of skewness  =  -0.09616726
                  coef. of kurtosis  =  3.512835
Filliben correlation coefficient  =  0.9980958
**********************************************************************
```

*Figure 4.3.5: showing the summary of the randomized quantile residuals for the DBI family.*

The summary of the randomized quantile residuals from the statistical model shows a mean of the randomized quantile residuals as -0.0263 which indicates the average deviation of the actual data form the model's prediction. The variance of the randomized quantile residuals is 1.0284 which represents the

dispersion of the residuals around the mean. The coefficient of skewness is -0.0962 which is negative suggests that the distribution is slightly skewed to the left. Coefficient of kurtosis equals 3.512835 which is greater than 3 indicates heavier tails compared to a normal distribution. The Filliben Correlation Coefficient is 0.9981 which measures the correlation between the ordered residuals and their corresponding quantiles. The statistical summary provides insights into the distributional properties and the goodness of fit of the model's residuals.

The DBI family distribution model was used to make prediction on an out-of-sample observation which gives a prediction score of 0.565 suggests that 56%  of the predicted response represents the value for the first out-of-sample observation based on the DBI family distribution model that best fits the wine quality data set.

## 4. Conclusion

The coursework generally was successfully completed despite the challenges faced throughout the process. The first section of the coursework, I was able to fit and choose the appropriate parametric distribution for the BMI of Dutch Boys dataset given for the age of group 19 to 20 years. The GAMLSS distribution gave me the opportunity to estimate four (4) parameters which includes location (μ), scale (σ), skewness (ν), and the kurtosis(τ) for the SN1 GAMLSS family distribution that best fits the data. The second data for the coursework, I was able to plot the centile curves for the hand grip strength data set against age. Three (3) different GAMLSS family distributions were fitted, BCCG, BCT, ang BCPE to plot the centile curves and choose the one that fits best. The BCT was chosen as the best GAMLSS family distribution with the lowest AIC value. The last section of the coursework shows the alcohol has the highest correlation with wine quality and 56% of the predicted response represents the value for the first out-of-sample observation based on the DBI family distribution model that best fits the wine quality data set.

## References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009

http://www3.dsi.uminho.pt/pcortez/wine/ https://worldpopulationreview.com/country-rankings/wine-producing-countries

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z. & De Bastiani, F.,2020. Distributions for Modeling Location, Scale, and Shape Using GAMLSS in R. London: CRC Press Taylor & Francis Group

Stasinopoulos, M.D.et al., 2017. Flexible Regression and Smoothing Using GAMLSS in R.London CRC Press Taylor&Francis Group.