Statistical Modelling and Forecasting Case Study Report, MA7007.

Date-may15th,2023.

Student id - 21025221

SAMRAT RAI

WORD COUNT- 3,587

Table of Contents

Introduction:

This assignment is conducted in the software called R Studio and the language implemented to carry this project is R. R is a famous programming language used for analysing data, for calculation, as a statistical software and used often used for statistics and machine learning (1). R is also flexible enough to use in operating systems that are used quite often like Microsoft Windows, Mac OS, and Linux. R studio is an integrated platform where various data analysis is done like transforming, importing, exploring, modelling etc. data's (2). In this assignment project we are analysing three different data sets. Each data set is broken down into three different sections. Each dataset has its own task and its sets of problems to solve. The first dataset contains the Body mass index taken from the study called fourth Dutch growth. The body mass index of Dutch boys ages from 10 to 20 has been given in the first dataset. The second dataset contains the handgrip strength of English schoolchildren. The data is stored inside R studio under the package named gamlss dot grip which we will be using to solve the second dataset. the goal her in the second dataset is to create centile curves for grip given age. The third dataset is a dataset of student's choice where we must find a dataset which is approved by the professor. The dataset should contain one target variable with some explanatory variables and the dataset should contain numerical values.

1. First data set (fitting distributions to the data):

1.1 Comment on the different distributions you are using.

Here in the first dataset, six different distribution methods have been chosen to carry out data set analyzation. the distribution methods are Box-Cox Power Exponential (BCPE), Box-Cox Cole and Green (BCCG), Box-Cox-t(BCT), exponential (EXP), generalized beta type 2(GB2) and log-Normal Box-Cox (LNO).

1.2 Distribution chosen for the 1st dataset.

The distribution chosen for the first dataset are Box-Cox Power Exponential (BCPE) and generalized beta type 2(Gb2). The age chosen to find a suitable distribution of BMI is from 15-16 years old.

1.3 Reason for choosing the distribution methods:

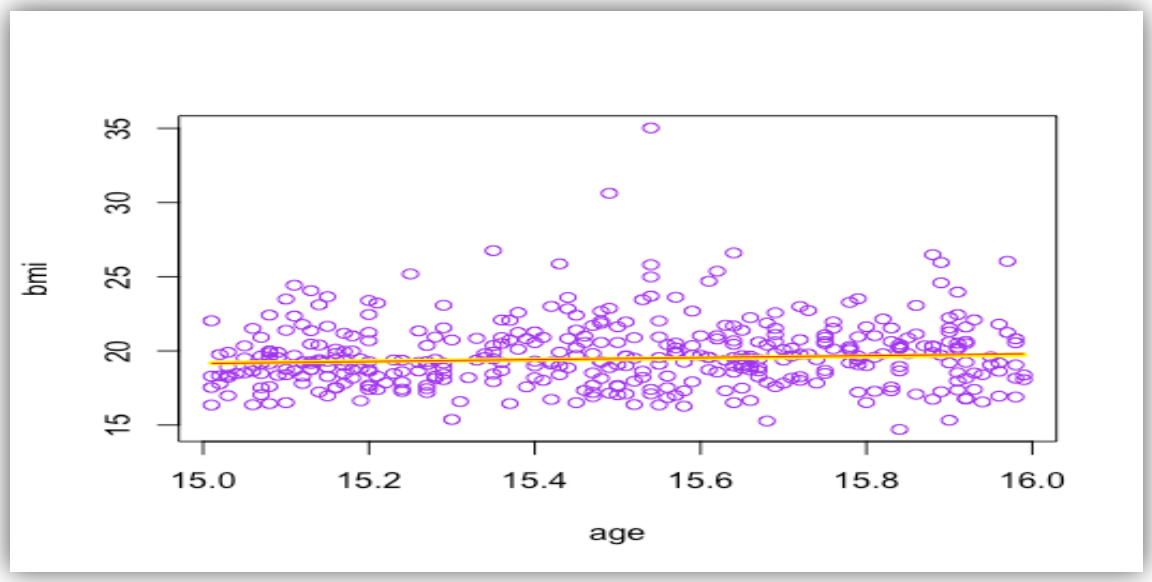
BCPE is a four-parameter distribution used to transform non-normal data into normal distribution (3). BCPE can be used to explain random generation, distribution function, quantile function and density. BCPE and GB2 have been chosen out of the six distribution methods chosen to calculate the BMI of 15–16-year-old Dutch boys because the lower global deviance level the better the fit and the model will also result is being better. With BCPE having the lowest global deviance level of Global Deviance of 1718.241 and BCT with a global deviance of 1729.437. whereas other distribution methods have higher global deviance or Akaike information criterion (AIC) which means the model with have lower best fit chance with them. Also, the models with lowest AIC levels tend to have higher degrees of freedom. Hence is the reason of choosing BCPE and BCT distributions as they both have the lowest AIC level and highest degrees of freedom which is 5.

DISTRIBUTION METHODS.	DF	AIC
AIR2(BCPE)	5	1728.241
AIR3(BCT)	5	1729.437
AIR5(GB2)	5	1731.005
AIR1 (BCCG)	4	1731.392
AIR6 (LNO)	3	1757.605
AIR4 (EXP)	2	3214.335

```
     df       AIC
air2  5 1728.241
air3  5 1729.437
air5  5 1731.005
air1  4 1731.392
air6  3 1757.605
air4  2 3214.335
```

DF and AIC levels 1

1.3 Plotting the fitted distribution methods.



BCPE&BCT lines 1

The above figure shows the result of BCPE and BCT line. Red being the BCPE line and yellow being the BCT line. As you can see both distribution methods go fit along the same line because both have almost identical AIC levels and the same degrees of freedom level with 5. The scatter plot above shows that boys ranging from 15 to 16 years tend to have BMI of mostly 17 as the data points are densely condensed along that line. Which the data points above BMI above 17 and below 17 comprises mostly of outlier data points which are less likely to occur. So, in majority boys ranging from 15 to 16 years tend to have a BMI of around 17.28-17.30 according to the fitted distribution models of BCPE and BCT.

1.4 The fitted parameter values of the final chosen models.

```
> air2 <- gamlss(bmi~age,data=dt,family=BCPE)
GAMLSS-RS iteration 1: Global Deviance = 1731.6
GAMLSS-RS iteration 2: Global Deviance = 1718.491
GAMLSS-RS iteration 3: Global Deviance = 1718.249
GAMLSS-RS iteration 4: Global Deviance = 1718.241
GAMLSS-RS iteration 5: Global Deviance = 1718.241
```

BCPE global deviances 1

There are five iterations of global deviances in which the lowest deviance level is chose which is also the Akaike information criterion level (AIC) of the parameter chosen for the first dataset which is BCPE. In terms of parameters for BCPE the AIC level is 1718.241 and degrees of freedom is 5. BCT also has global deviances iterated five times, the lowest deviance is also the AIC level. The parameters for BCT are AIC level is 1719.437 and degrees of freedom is 5.

```
> air3 <- gamlss(bmi~age,data=dt,family=BCT)
GAMLSS-RS iteration 1: Global Deviance = 1720.843
GAMLSS-RS iteration 2: Global Deviance = 1719.463
GAMLSS-RS iteration 3: Global Deviance = 1719.439
GAMLSS-RS iteration 4: Global Deviance = 1719.438
GAMLSS-RS iteration 5: Global Deviance = 1719.437
```

BCT global deviances 1

Second data set (centile estimation)

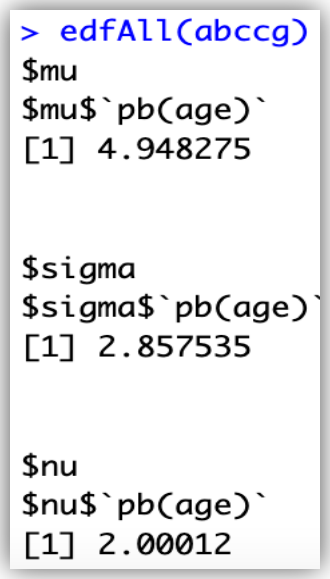
Comment on the different models you are using.

The distribution models used for dataset two are Box-Cox Cole and Green (BCCG), Box-Cox Power Exponential (BCPE) and Box-Cox-t(BCT). Here in dataset two, we are using 1000 sample out of 3766 English boys to analyse the handgrip strength. We are using two variables grip and age. We must plot grip against age, we are using the LMS method to fit the data which uses BCCG distribution as the response variable. Here we are trying to fit BCCG to grip variable. We are using LMS model to fit BCT and BCPE to the data as well.

Answer the explicit questions in section 1.2.

The second explicit question asks How many degrees of freedom were used for smoothing in the model?

The effective degrees of freedom fitted for the parameter BCCG is 9.80593, for the parameter BCPE is 11.76359 and for the parameter BCT is 11.87961. degrees of freedom are an estimation of the total number of independent variables of a statistical analysis (4).



smoothing results for age 1

For BCCG there were three types of parameters used for smoothing the variable age. The function edFall was used to calculate the degrees of freedom using the three parameters mu, sigma, and nu for smoothing the model. To estimate the smoothens of age, one of the best smoothing tool called p-splines is used along with the function pb() for fitting. There was a total of three parameters used for degrees of freedom calculation in BCCG using edFall and the results for it are for the parameter mu is 4.948275, for the parameter sigma is 2.857535 and for the parameter nu is 2.00012. The degree of freedom for BCCG in total is 9.80593 which is a result of a sum when all three parameters MU, SIGMA and NU is added.

```
> edfAll(abct)
$mu
$mu$`pb(age)`
[1] 4.971954


$sigma
$sigma$`pb(age)`
[1] 2.907516


$nu
$nu$`pb(age)`
[1] 2.000122


$tau
$tau$`pb(age)`
[1] 2.000015
```

smoothing results for age 2

For the distribution function BCT four parameters were used to smooth the variable age. The result for parameter MU is 4.971954, for parameter SIGMA is 2.907516, for parameter NU is 2.00122 and for parameter TAU is 2.000015. in terms of degrees of freedom, the higher the DF is the better the data fits the model so the parameter MU has the best DF in this case for BCPE. So, when smoothing the variable age, the best parameter would be MU. When all the four parameters MU, SIGMA, NU, and TAU are added the result is 11.87961 which is the degree of freedom of the distribution method BCT.

```
> edfAll(abcpe)
$mu
$mu$`pb(age)`
[1] 4.959257


$sigma
$sigma$`pb(age)`
[1] 2.803945


$nu
$nu$`pb(age)`
[1] 2.000108


$tau
$tau$`pb(age)`
[1] 2.00028
```

smoothing results for age 3

For the distribution method BCPE, four parameters MU, SIGMA, NU, and TAU were used to calculate the degrees of freedom for the smoothness of the model. The result for the degrees of freedom for MU is 4.959257, the degrees of freedom for SIGMA is 2.803945, the degrees of freedom for NU is 2.00108 and the degrees of freedom for TAU is 2.00028. The total result for the degreed of freedom for BCPE is 11.76359.
So, to answer the question a total of 11 parameters were used to calculate the degrees of freedom for the distribution methods BCCG, BCPE AND BCT.

The second explicit question is What are the effective degrees of freedom fitted for the parameters? Try to interpret the effective degrees of freedom.
In terms of the effectiveness of degrees of freedom, the higher the degrees of freedom the better the fit for the model is indicated. So, for the parameters used in BCCG, the parameter MU has a degree of freedom of 4.948275 which is higher than the parameters SIGMA and NU which means the parameter MU proves to be the most effective degree of freedom for BCCG. In terms of BCT out of the four

parameter MU, SIGMA, NU, and TAU. MU has the highest degree of freedom which is 4.971954 which means MU has the most effective degrees of freedom for BCT. Lastly for the distribution model BCPE, the parameter MU again has the highest degrees of freedom with the value of 4.959257. so, to answer the question the smoother parameter for fitting the model would be MU for all the three distribution methods BCCG, BCT and BCPE. Hence in terms of the effective degrees of freedom, the parameter MU would be the most effective fit to smooth the variable age and the model.

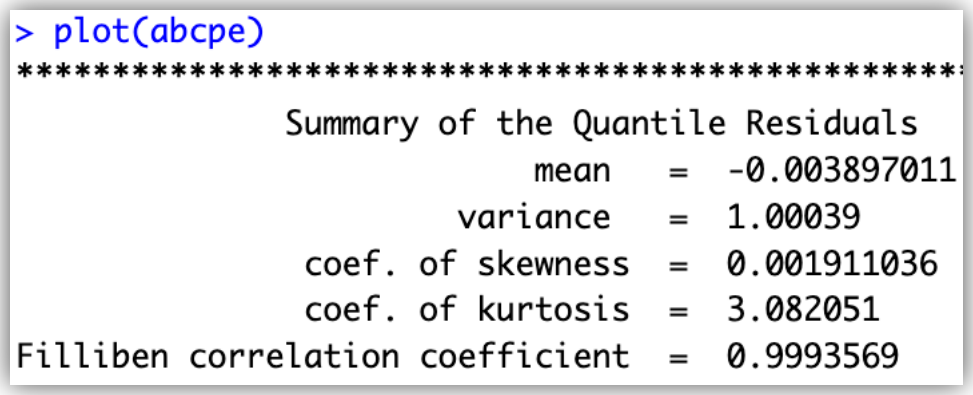
Use residual diagnostics for checking the model.

```
> plot(abccg)
**************************************************
              Summary of the Quantile Residuals
                           mean   =  0.0002727109
                       variance   =  1.001002
                   coef. of skewness  =  0.002864923
                   coef. of kurtosis  =  3.376945
Filliben correlation coefficient  =  0.9986508
```

residual summary BCCG 1

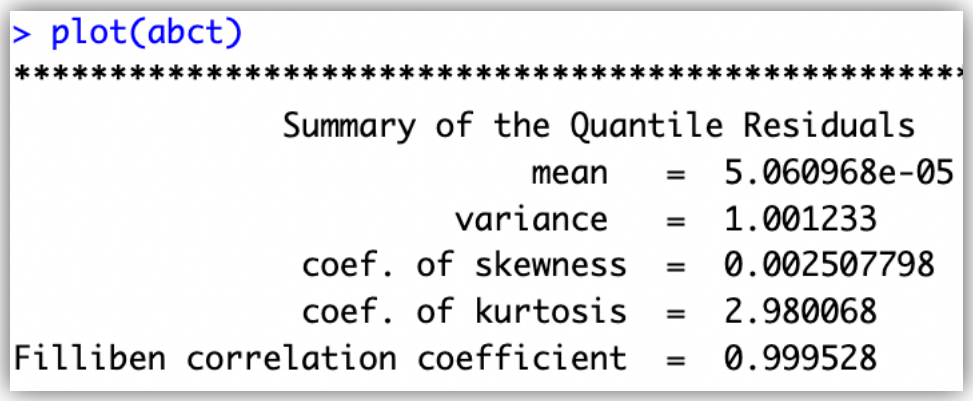
The mean of the model is 0.0002727109 which is positive, so the model turns out to be fit good. The variance of the model is 1.001002 which means that the datapoints or data is more likely to be scattered from the mean and from each other (5). variance value lower than 1 is considered low variance and higher than 1 is considered high variance (5). If the variance is low, it will be easier to make predictions and if it is high, it is harder to make predictions so low variance level tend to be preferred. The skewness id 0.002864923 that means we have a symmetrical skewness where the data has a normal distribution without too much negative or positive outliers (6). A good range of kurtosis for symmetry has a to be within the value of -2 to +2 (7), a value higher than the 3 in kurtosis is known to have a positive kurtosis where the datapoints or numbers are in the top of the positive kurtosis curve rather than being around the mean of the curve. This also points out that there is a

higher probability of age being a factor when it comes to grip strength. A Filliben correlation coefficient of 0.99 indicates that we have a standard uniform quantile which will plot a uniform Q-Q plot (7).



residual summary BCPE 1

the mean of the model is -0.003897011 which points out to the model being is poorly fit. The variance of the model is 1.00039 which is high and means that the datapoints are dispersed from the mean and each other. A skewness of 0.001911036 means that the skewness of the graph is symmetrical, and data has very little negative and positive outlying data. the kurtosis has a value of 3.082051 which is close to 3 which means that the kurtosis is mesokurtic which means it has a normal distribution where there is a 50-50 probability of the age being a factor when it comes to grip strength. A Filliben correlation coefficient of 0.99 indicates that we have a standard uniform quantile which will plot a uniform Q-Q plot.



residual summary BCT 1

the mean of the plot BCT here is also negative which means the data is poorly fitted. The variance is 1.001233 which means that the data has chances of being more scattered although the variance is not that high, it is still considered a little high as is it above 1. The skewness here is 0.002507798 which means it has a skewness of 0 which gives the plot a symmetrical skewness. The kurtosis here is below 3 which means it has a normal mesokurtic curve with a 50-50 probability of the age being a factor when it comes to grip strength. A Filliben correlation coefficient of 0.99 indicates that we have a standard uniform quantile which will plot a uniform Q-Q plot. So, in conclusion the first plot of BCCG is better in comparison to the other two plots of BCPE and BCT as BCCG has a positive mean which indicates a good fit whereas the other two plots have negative mean. The variance of all the three plots is the same with variance level of 1. The skewness is also the same with all three plots having a symmetrical skewness of 0. In terms of kurtosis BCPE and BCT has a normal mesokurtic kurtosis where probability chances are 50-50 which means age has a fifty percent chance of playing a factor in grip strength which makes it hard to predict the outcome while BCCG has a high kurtosis which is called leptokurtic that indicated that there is more probability of age playing a factor in grip strength which makes it easier to predict an outcome. The Filliben correlation coefficient has the same standard uniform line for all three plots. Hence, the residual diagnostics for the plot/model BCCG is better in comparison to the other two plot/models.

Comment on how you selected your final model.

For the Selection of final model, I analysed the worm plot, Q-stats plot and analysed the degrees of freedom for the fit. In terms of the worm plot the BCCG model had the most dispersed curve in the plot with outliers' data points, BCPE also had a curve but with less data outliers and BCT had the best worm plot with condensed data points in a uniform line.

```
> Q.stats(abccg)
                          Z1          Z2          Z3             Z4 AgostinoK2    N
   0.5 to  100.5  -1.4600056  0.16347489  0.141766745  2.2615945378  5.1349077  100
 100.5 to  200.5  -0.4936753  1.07002971  0.155046618  1.0912209971  1.2148027  100
 200.5 to  300.5  -0.2246487 -0.66159623 -1.082128268  0.0006675172  1.1710020  100
 300.5 to  400.5   1.1333984 -0.62881877  0.844650506  0.6068474666  1.0816983  100
 400.5 to  500.5   1.5391864  0.78128627 -0.004245411  1.1971420657  1.4331671  100
 500.5 to  600.5  -0.4492139 -0.99418540 -1.334066369  1.2093584461  3.2422809  100
 600.5 to  700.5   1.1437637 -2.03309327  0.089731974 -1.0128422253  1.0339012  100
 700.5 to  800.5  -1.4975295 -0.65154059  1.240987635  1.1374776212  2.8339056  100
 800.5 to  900.5   0.1712928  2.30839919 -0.112690646  0.9249698534  0.8682684  100
 900.5 to 1000.5   0.1647028 -0.22372484  0.208735245 -1.1130734285  1.2825029  100
TOTAL Q stats      9.8885368 13.54035610  5.312696152 13.9837407866 19.2964369 1000
df for Q stats     5.0517246  8.07123256  7.999880333 10.0000000000 17.9998803    0
p-val for Q stats  0.0807219  0.09742154  0.723680384  0.1737345337  0.3737790    0
```

```
> Q.stats(abcpe)
                          Z1          Z2          Z3          Z4 AgostinoK2    N
   0.5 to  100.5  -1.49943835 -0.02517537  0.23974281  1.76102538  3.1586870  100
 100.5 to  200.5  -0.53210515  0.95926206  0.09624036  0.54335381  0.3044956  100
 200.5 to  300.5  -0.22754414 -0.60468566 -0.90290403 -0.57372688  1.1443982  100
 300.5 to  400.5   1.08331906 -0.52548214  0.66502201  0.01703635  0.4425445  100
 400.5 to  500.5   1.48131710  0.75573780 -0.02109060  0.80831525  0.6538184  100
 500.5 to  600.5  -0.47195183 -1.00178911 -1.16547321  0.65023624  1.7811350  100
 600.5 to  700.5   1.10873927 -1.85721402 -0.01401154 -1.41272932  1.9960005  100
 700.5 to  800.5  -1.55463949 -0.66812721  1.24479861  0.67669034  2.0074334  100
 800.5 to  900.5   0.12646220  2.21459751 -0.26762055  0.34422435  0.1901112  100
 900.5 to 1000.5   0.09614022 -0.08036365  0.22422593 -1.56188768  2.4897704  100
TOTAL Q stats      9.84528898 11.94385309  4.35461933  9.81377473 14.1683941 1000
df for Q stats     5.04074288  8.09802751  7.99989205  7.99972028 15.9996123    0
p-val for Q stats  0.08154452  0.15939444  0.82378766  0.27831713  0.5861426    0
```

```
> Q.stats(abct)
                          Z1           Z2           Z3           Z4  AgostinoK2    N
   0.5 to  100.5  -1.47106931 -0.108332431  0.310263108  1.584579e+00  2.60715373  100
 100.5 to  200.5  -0.50474228  0.946561262  0.085955479  3.523574e-01  0.13154410  100
 200.5 to  300.5  -0.18275150 -0.561620021 -0.836284344 -6.854740e-01  1.16924617  100
 300.5 to  400.5   1.11847762 -0.505108889  0.505113702 -3.618387e-01  0.38606709  100
 400.5 to  500.5   1.52766212  0.722282791 -0.057420889  6.198225e-01  0.38747706  100
 500.5 to  600.5  -0.42529539 -0.977472378 -1.055080999  4.832500e-01  1.34672653  100
 600.5 to  700.5   1.15746539 -1.768258564  0.004120461 -1.401822e+00  1.96512280  100
 700.5 to  800.5  -1.53094814 -0.647693242  1.230302392  5.752517e-01  1.84455848  100
 800.5 to  900.5   0.16658295  2.161689230 -0.294588388 -7.239589e-05  0.08678232  100
 900.5 to 1000.5   0.14967951 -0.007970211  0.252004139 -1.593706e+00  2.60340389  100
TOTAL Q stats      9.95150990 11.174619594  3.838605329  8.689477e+00 12.52808217 1000
df for Q stats     5.02804591  8.046241948  7.999878087  7.999985e+00 15.99986283    0
p-val for Q stats  0.07782145  0.195099256  0.871376035  3.691631e-01  0.70688971    0
```

In terms of Q-stats plot p-val for BCCG looks better than BCT and BCPE although the p-value for all the models is above 0.05 which indicates that results are not that

significant for all the models in terms of p-value having said that BCCG has the lowest p-val compared to the other two models. In terms of degrees of freedom for Q-stats the model BCCG has the best df. Lastly in terms of Total Q stats BCCG has the highest Q-stats making it the slightly better than the other two models.

```
> GAIC(abccg,abct,abcpe)
             df      AIC
abct  11.87961 6339.679
abcpe 11.76359 6341.504
abccg  9.80593 6341.626
```

GAIC summary of 3 models 1

Lastly in terms of degrees of freedom which signifies the best fitted model, BCT has the highest degrees of freedom making It the best fit. In conclusion the final model chosen would be BCT as the worm plot, degrees of freedom and quantile residuals summary suggest BCT to be the optimal best fit for modelling.

Discussion of the final centile plots.

```
> centiles(abccg)
% of cases below  0.4 centile is  0.9
% of cases below  2 centile is  2.5
% of cases below  10 centile is  9.6
% of cases below  25 centile is  24.6
% of cases below  50 centile is  49.3
% of cases below  75 centile is  76.5
% of cases below  90 centile is  90.6
% of cases below  98 centile is  97.7
% of cases below  99.6 centile is  99.3
> centiles(abcpe)
% of cases below  0.4 centile is  0.8
% of cases below  2 centile is  2.4
% of cases below  10 centile is  9.7
% of cases below  25 centile is  25.4
% of cases below  50 centile is  49.7
% of cases below  75 centile is  75.4
% of cases below  90 centile is  90.5
% of cases below  98 centile is  97.7
% of cases below  99.6 centile is  99.4
> centiles(abct)
% of cases below  0.4 centile is  0.5
% of cases below  2 centile is  2.3
% of cases below  10 centile is  9.7
% of cases below  25 centile is  25.4
% of cases below  50 centile is  49.3
% of cases below  75 centile is  75.2
% of cases below  90 centile is  90.4
% of cases below  98 centile is  97.7
% of cases below  99.6 centile is  99.4
```

centiles % levels for each model 1

for BCCG model, In the 0.4th centile age plays a 0.9 percent role in being a factor for grip strength.

Whereas in the 99.6th centile age plays a 99.3 percent role in being a factor for grip strength.

for BCPE model, In the 0.4th centile age plays a 0.8 percent role in being a factor for grip strength.

Whereas in the 99.6th centile age plays a 99.4 percent role in being a factor for grip strength.

for BCT model, In the 0.4th centile age plays a 0.5 percent role in being a factor for grip strength.

Whereas in the 99.6th centile age plays a 99.4 percent role in being a factor for grip strength.
For all centiles plots there is a gradual increase in each centile curve line from 0.4th to 99.6th which indicates a normal growth range. Having a normal growth range in all three centiles plots of BCCG, BCPE and BCT indicates that age is a factor for grip strength and that age plays a gradual role in terms of grip strength which means that the older a child gets the stronger their grip strength becomes in a gradual manner from year to year.

Third dataset-students selected data:

Explanation of the collected data:

The third dataset is of supermarket sales in the growing most populated cities of Myanmar. It contains 3 months of data where future prediction is made easier with the data's containing within the dataset. the dataset contains 17 columns in total out of which eight variables in specific has been chosen to explain the modelling. I have chosen variables like branch, city, customer type, gender, product line, unit price, quantity, and payment. The main goal here was to choose one target variable but I have experimented with 3 variables as target variables to examine which target variable would have the best fit and outcome. All the variables are in numerical values but there were some categorical values which were not used.

Source of dataset:

https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

preliminary analysis on the collected data and discussion on the reliability of the data.

```
> dim(superstorenew)
[1] 809   8
> head(superstorenew,5)
  Branch City Customer.type Gender Product.line Unit.price Quantity Payment
1      1    3             1      1            4      74.69        7       3
3      1    3             2      2            5      46.33        7       2
4      1    3             1      2            4      58.22        8       3
5      1    3             2      2            6      86.31        7       3
6      3    2             2      2            1      85.39        7       3
```

snapshot of supermarket sales dataset 1

as you can see the supermarket sales dataset has 809 rows with 8 columns. This is the dataset which will be used for the modelling process. This dataset is sourced from Kaggle which has good reliable sources of data's which are mostly used by data scientists (9). This dataset is used for the purpose of predictive data analysis hence making it a perfect dataset for modelling and analysing.

Models used to fit the data:

The distribution models used to fit the data are EXP (exponential), GB2 (generalized beta type 2) and BCT (Boc-Cox-t).

```
> plot(r1)
**************************************************
             Summary of the Quantile Residuals
                          mean   =  0.2388393
                      variance   =  0.2935645
               coef. of skewness  =  -0.4065595
               coef. of kurtosis  =  1.941177
Filliben correlation coefficient  =  0.9523902
**************************************************
```

EXP quantile results 1

```
> plot(r2)
******************************************************
                Summary of the Quantile Residuals
                             mean   =  0.02358032
                         variance   =  1.000684
                  coef. of skewness =  -0.07694077
                  coef. of kurtosis =  1.976783
Filliben correlation coefficient    =  0.975414
******************************************************
```

GB2 quantile results 1

```
> plot(r3)
******************************************************
                Summary of the Quantile Residuals
                             mean   =  -0.04647599
                         variance   =  1.038344
                  coef. of skewness =  0.003738751
                  coef. of kurtosis =  1.875919
Filliben correlation coefficient    =  0.9813967
```

BCT quantile results 1

As you can see from the figures above are the results of using the distributions to fit the model. Plot r1 is the plot for EXP, plot r2 is the plot for GB2 and plot r3 is the plot for BCT.

Final model selection diagnosis:

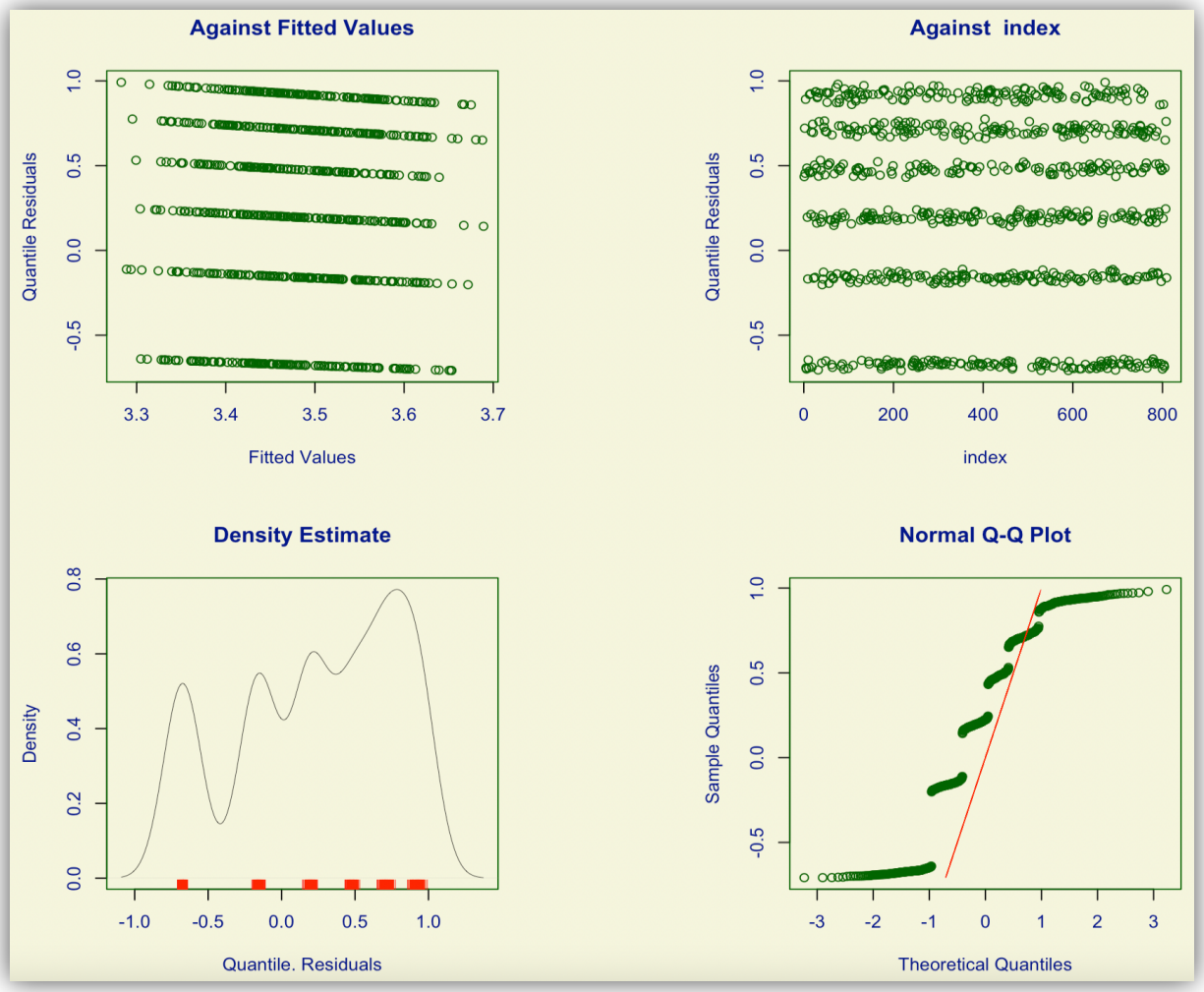
```
> GAIC(r1,r2,r3)
   df      AIC
r1  5 3646.233
r2  6 3910.215
r3  6 7547.427
```

GAIC of all three models 1

The final model has been selected by taking in the best values for Akaike information criterion and degrees of freedom which has been calculated by using GAIC. As the plot r1 which represents the distribution method called exponential (EXP) has the lowest AIC score of 3646.233 which means that EXP is the model with the fewest parameter and will better fit the model than other distribution methods used which are GB2 and BCT.

Model used for prediction and plotting the distribution.

The model used for prediction and plotting is the EXP model as it had the best outcome when comparing the Akaike information criterion values. The skewness, mean and variance are also better of the EXP model when compared to GB2 and BCT models. EXP has a lower level of variance of 0.2935 which means that data points are close to each other and not spread out while the other two models GB2 and BCT have a high variance above 1 where the data points are spread out and further away from the mean. EXP model also has a Filliben correlation coefficient of 0.95 which means it that the model has 95 percent chance of being true and is very reliable.



final distribution model plotted 1

this the plots of the final distribution model EXP which is chosen as it has the best results and best fits the model.

Peer Review:

The peer review selected is of my fellow peer named Kunik nayaar with a seed number of 1114. My peer has used the distribution models BCCG, EXP and GB2 with unit price as a target variable. The best result formulated out of the three plots was from the distribution model BCCG. The Filliben correlation coefficient of 0.97 was achieved through the BCCG model which is extremely good and proves to be better than my model which achieved Filliben correlation coefficient of 0.95. the choice of unit price as the target variable is also good as it fits in well when

comparing it with other independent variables such as quantity, payment, product line, city etc.

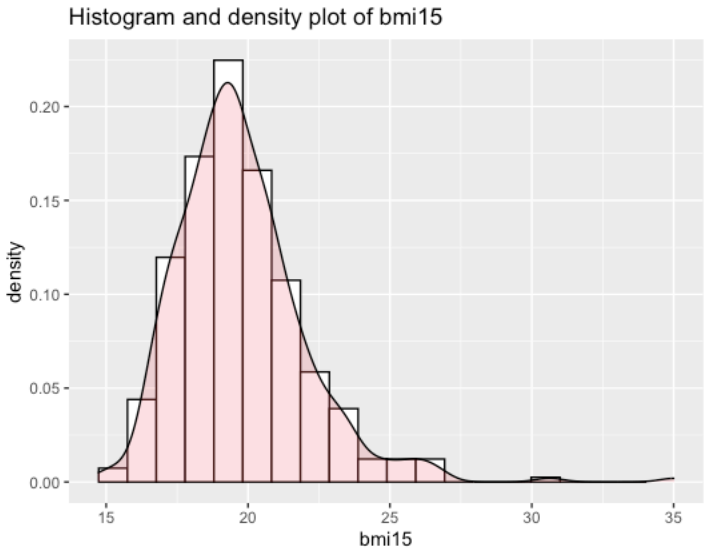
Conclusion:

This case study report has given me knowledge which I haven't had before. It has made me familiar with the programming language of R and given me the statistical knowledge that is needed for data analysis which was unknown to me before. It has given me the knowledge to code and plot various machine learning models. using Gamlss library, I know understand many parameters like MU, NU, TAU, SIGMA etc., different distribution families like GB2, BCCG, EXP, BCT, BCPE, LNP etc. through R and gamlss the calculations of Akaike information criterion, the degrees of freedom, global deviance, skewness & kurtosis plot, mean etc. were easy to be carried. This project has given me an idea and confidence to carry out my post graduate thesis project.

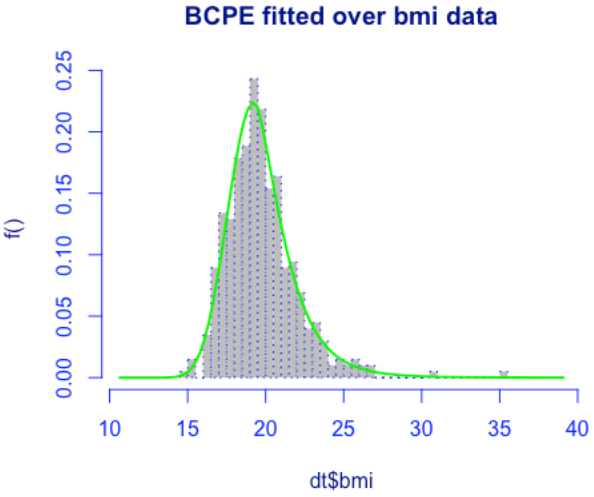
Appendix:

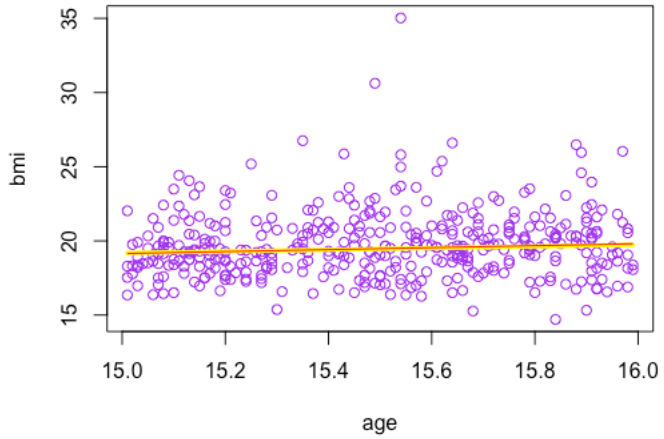
DATASET 1 APPENDIX:

```
install.packages("gamlss")
library(gamlss)
data(dbbmi)
year <- 15
dt<-with(dbbmi,subset(dbbmi,age>year & age<year+1))
bmi15 <- dt$bmi
install.packages("gamlss.ggplots")
library(gamlss,ggplots)
gamlss.ggplots:::y_hist(bmi15)
air1 <- gamlss(bmi~age,data=dt,family=BCCG)
air2 <- gamlss(bmi~age,data=dt,family=BCPE)
air3 <- gamlss(bmi~age,data=dt,family=BCT)
air4 <- gamlss(bmi~age,data=dt,family=EXP)
air5 <- gamlss(bmi~age,data=dt,family=GB2)
air6 <- gamlss(bmi~age,data=dt,family=LNO)
GAIC(air1,air2,air3,air4,air5,air6)
plot(bmi~age,data=dt,col='purple')
lines(fitted(air2)~dt$age,col='red',lwd=1)
lines(fitted(air5)~dt$age,col='yellow',lwd=3)
hist<-histDist(dt$bmi,family=BCPE,nbins=30,line.col='green',line.wd='2',main='BCPE fitted over bmi data')
hist<-histDist(dt$bmi,family=GB2,nbins=20,line.col='maroon',line.wd='2',main='GB2 fitted over bmi data')
```

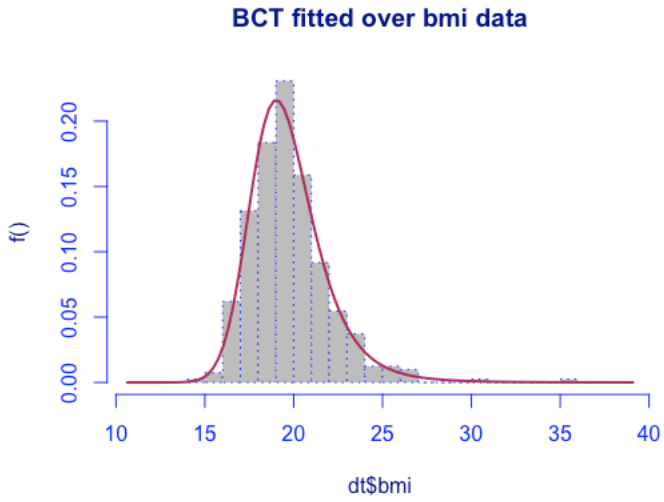



```
> air2 <- gamlss(bmi~age,data=dt,family=BCPE)
GAMLSS-RS iteration 1: Global Deviance = 1731.6
GAMLSS-RS iteration 2: Global Deviance = 1718.491
GAMLSS-RS iteration 3: Global Deviance = 1718.249
GAMLSS-RS iteration 4: Global Deviance = 1718.241
GAMLSS-RS iteration 5: Global Deviance = 1718.241
```

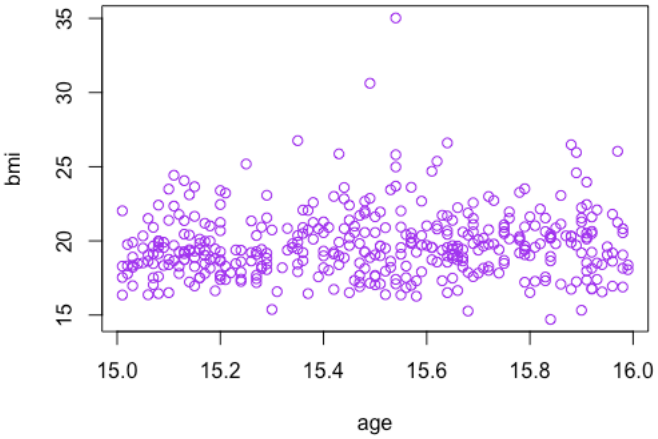




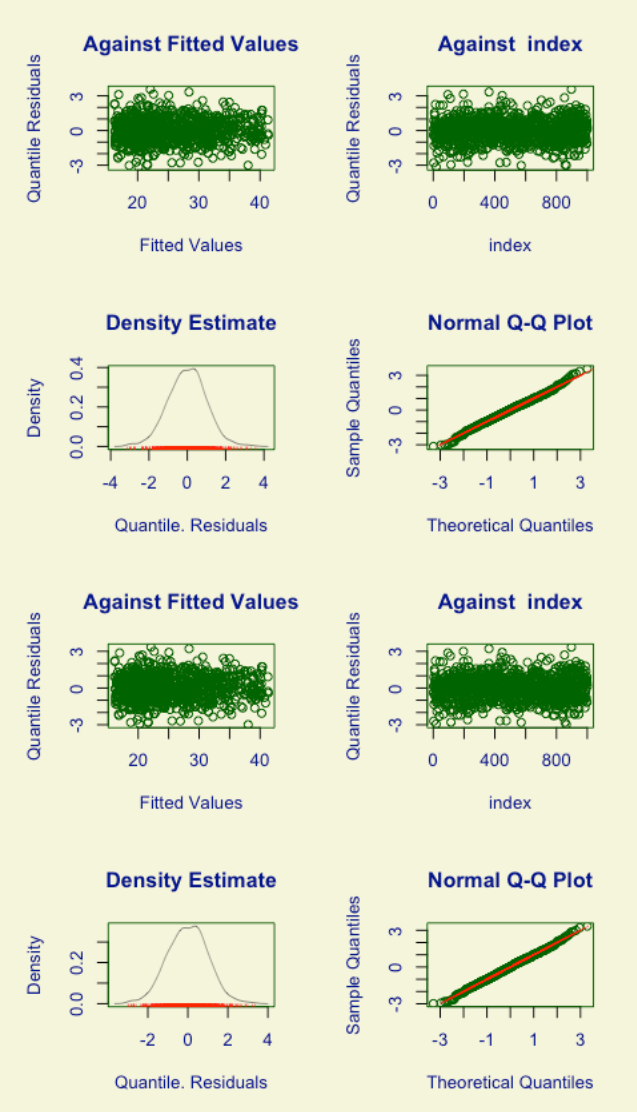

```
> air3 <- gamlss(bmi~age,data=dt,family=BCT)
GAMLSS-RS iteration 1: Global Deviance = 1720.843
GAMLSS-RS iteration 2: Global Deviance = 1719.463
GAMLSS-RS iteration 3: Global Deviance = 1719.439
GAMLSS-RS iteration 4: Global Deviance = 1719.438
GAMLSS-RS iteration 5: Global Deviance = 1719.437
```

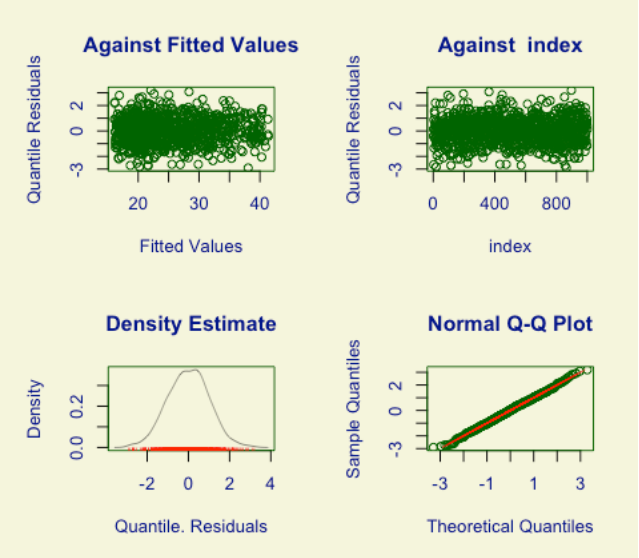


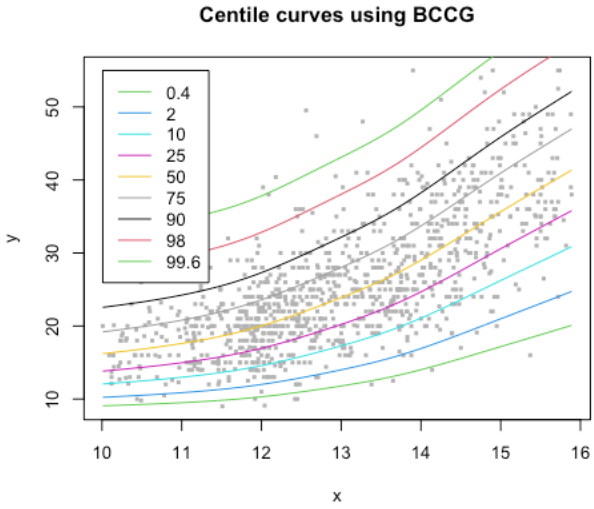

```
     df      AIC
air2  5 1728.241
air3  5 1729.437
air5  5 1731.005
air1  4 1731.392
air6  3 1757.605
air4  2 3214.335
```

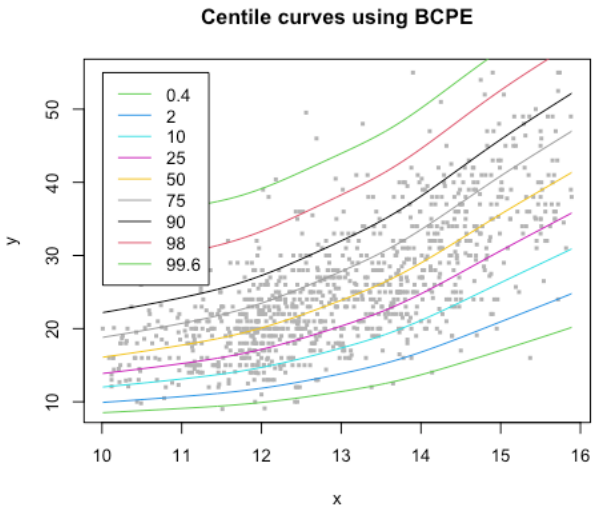


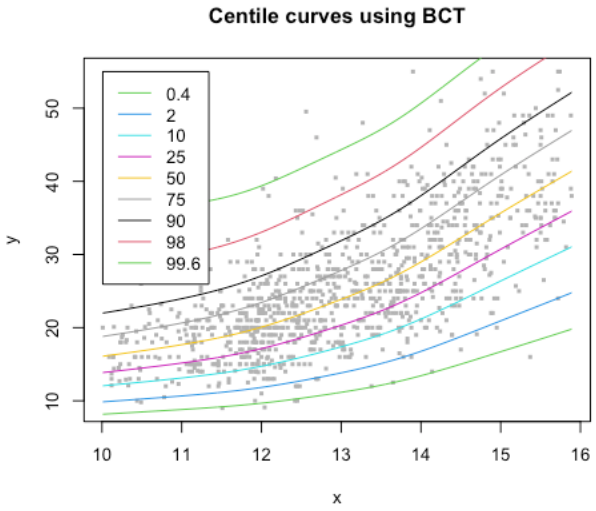
DATSET 2 APPENDIX:




Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

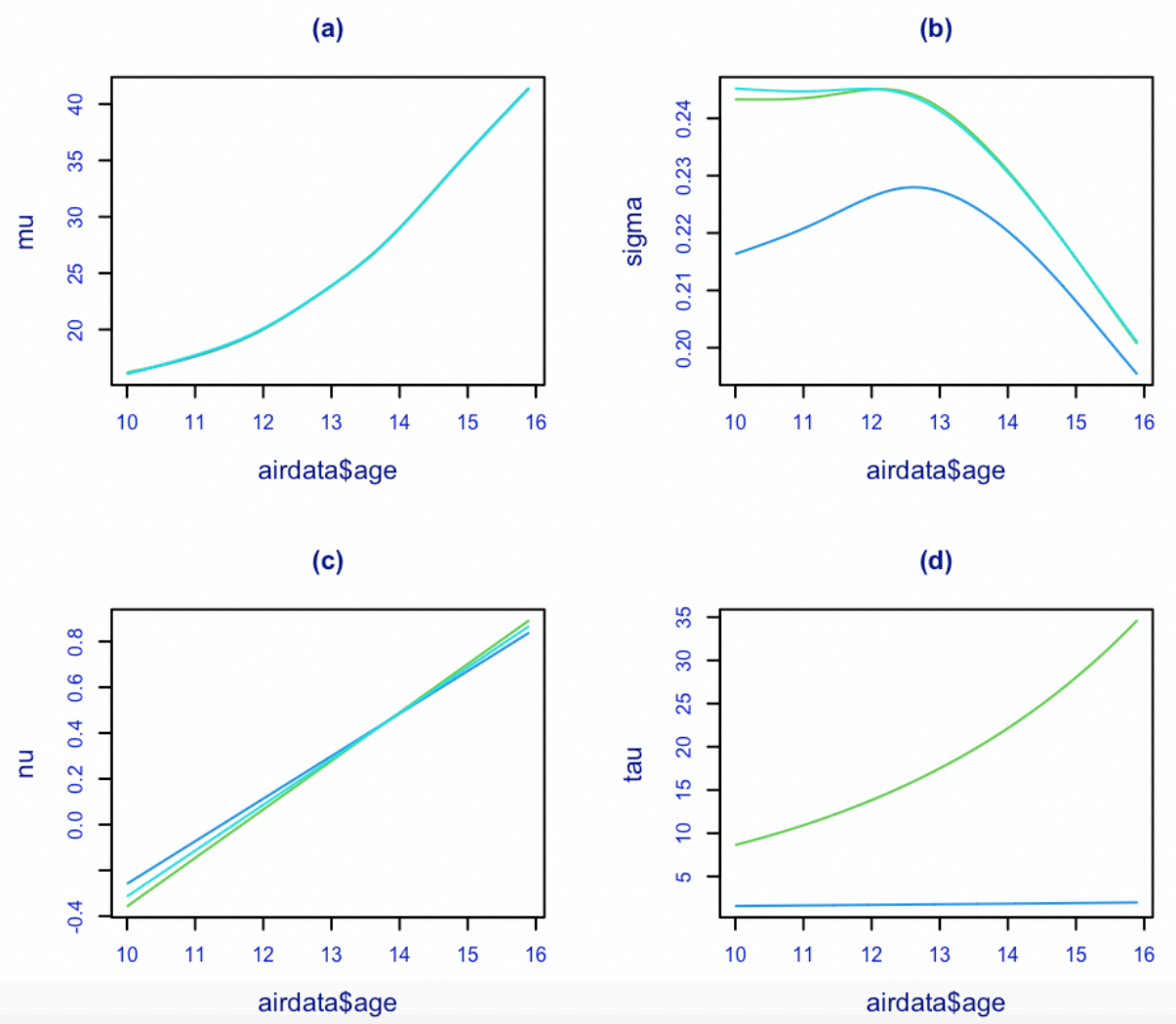

Centile curves using BCCG
0.4
2
10
25
50
75
90
98
99.6
y
x


Centile curves using BCPE
0.4
2
10
25
50
75
90
98
99.6
y
x


Centile curves using BCT
0.4
2
10
25
50
75
90
98
99.6
y
x
10
20
30
40
50
10
11
12
13
14
15
16

```
> centiles(abccg)
% of cases below  0.4 centile is  0.9
% of cases below  2 centile is  2.5
% of cases below  10 centile is  9.6
% of cases below  25 centile is  24.6
% of cases below  50 centile is  49.3
% of cases below  75 centile is  76.5
% of cases below  90 centile is  90.6
% of cases below  98 centile is  97.7
% of cases below  99.6 centile is  99.3
> centiles(abcpe)
% of cases below  0.4 centile is  0.8
% of cases below  2 centile is  2.4
% of cases below  10 centile is  9.7
% of cases below  25 centile is  25.4
% of cases below  50 centile is  49.7
% of cases below  75 centile is  75.4
% of cases below  90 centile is  90.5
% of cases below  98 centile is  97.7
% of cases below  99.6 centile is  99.4
> centiles(abct)
% of cases below  0.4 centile is  0.5
% of cases below  2 centile is  2.3
% of cases below  10 centile is  9.7
% of cases below  25 centile is  25.4
% of cases below  50 centile is  49.3
% of cases below  75 centile is  75.2
% of cases below  90 centile is  90.4
% of cases below  98 centile is  97.7
% of cases below  99.6 centile is  99.4
```

```
> centiles.split(abccg)
     10.005 to 11.965 11.965 to 12.905 12.905 to 14.055 14.055 to 15.895
0.4         0.7843137         1.214575              0.8        0.8064516
2           1.9607843         4.048583              1.6        2.4193548
10          8.6274510        10.526316              8.8       10.4838710
25         27.0588235        22.672065             26.8       21.7741935
50         46.6666667        48.178138             54.8       47.5806452
75         78.0392157        77.327935             78.0       72.5806452
90         92.5490196        91.497976             88.0       90.3225806
98         98.4313725        96.761134             97.6       97.9838710
99.6       99.6078431        98.380567             99.2      100.0000000
> centiles.split(abcpe)
     10.005 to 11.965 11.965 to 12.905 12.905 to 14.055 14.055 to 15.895
0.4         0.3921569         1.214575              0.8        0.8064516
2           1.9607843         4.048583              1.2        2.4193548
10          8.6274510        10.526316              9.2       10.4838710
25         27.8431373        23.886640             27.2       22.5806452
50         48.2352941        48.582996             54.4       47.5806452
75         77.2549020        74.898785             76.8       72.5806452
90         92.1568627        91.497976             88.0       90.3225806
98         98.4313725        96.761134             97.6       97.9838710
99.6       99.6078431        98.785425             99.2      100.0000000
> centiles.split(abct)
     10.005 to 11.965 11.965 to 12.905 12.905 to 14.055 14.055 to 15.895
0.4         0.3921569         1.214575              0.0        0.4032258
2           1.5686275         4.048583              1.2        2.4193548
10          8.6274510        10.526316              9.2       10.4838710
25         27.8431373        23.886640             27.2       22.5806452
50         46.6666667        48.178138             54.0       48.3870968
75         76.4705882        74.898785             76.8       72.5806452
90         92.1568627        91.093117             88.0       90.3225806
98         98.4313725        96.761134             97.6       97.9838710
99.6       99.6078431        98.785425             99.2      100.0000000
```

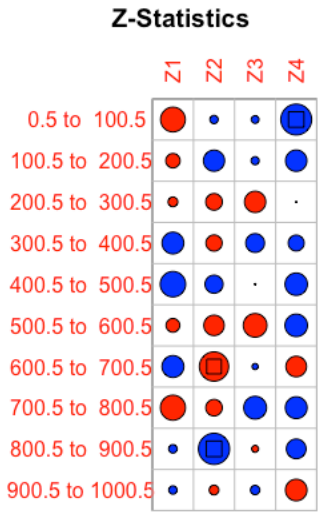


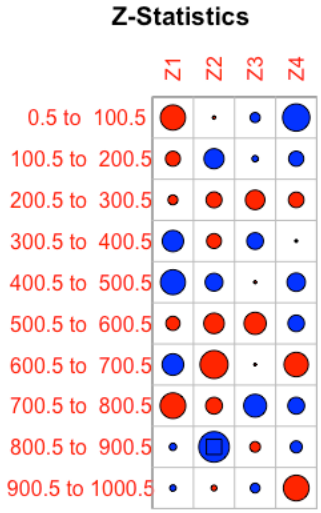

```
> GAIC(abccg,abct,abcpe)
             df      AIC
abct  11.87961 6339.679
abcpe 11.76359 6341.504
abccg  9.80593 6341.626

> GAIC(abccg,abct,abcpe)
             df      AIC
abct  11.87961 6339.679
abcpe 11.76359 6341.504
abccg  9.80593 6341.626
```

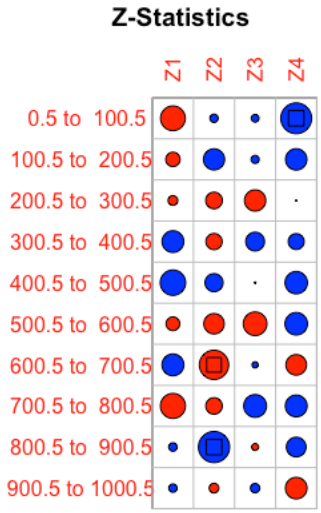
```
> edfAll(abccg)
$mu
$mu$`pb(age)`
[1] 4.948275


$sigma
$sigma$`pb(age)`
[1] 2.857535


$nu
$nu$`pb(age)`
[1] 2.00012
```

```
> edfAll(abct)
$mu
$mu$`pb(age)`
[1] 4.971954


$sigma
$sigma$`pb(age)`
[1] 2.907516


$nu
$nu$`pb(age)`
[1] 2.000122


$tau
$tau$`pb(age)`
[1] 2.000015
```

```
> edfAll(abcpe)
$mu
$mu$`pb(age)`
[1] 4.959257


$sigma
$sigma$`pb(age)`
[1] 2.803945


$nu
$nu$`pb(age)`
[1] 2.000108


$tau
$tau$`pb(age)`
[1] 2.00028
```





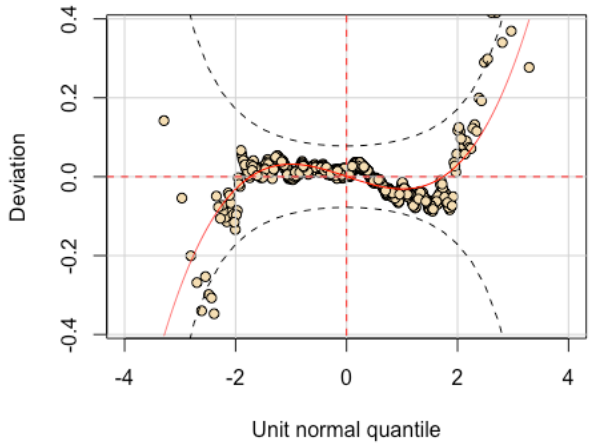


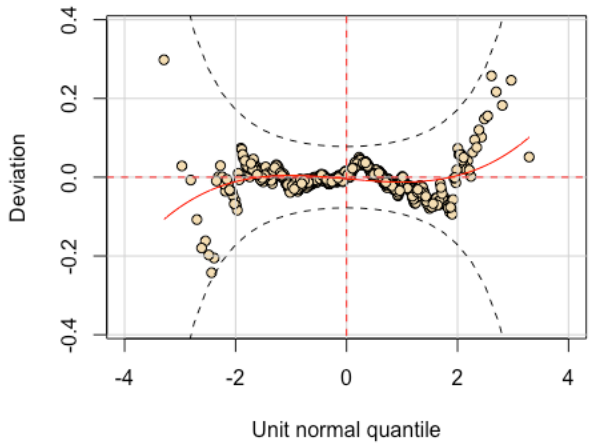

```
> Q.stats(abcpe)
                          Z1          Z2          Z3          Z4 AgostinoK2    N
   0.5 to  100.5  -1.49943835 -0.02517537  0.23974281  1.76102538  3.1586870  100
 100.5 to  200.5  -0.53210515  0.95926206  0.09624036  0.54335381  0.3044956  100
 200.5 to  300.5  -0.22754414 -0.60468566 -0.90290403 -0.57372688  1.1443982  100
 300.5 to  400.5   1.08331906 -0.52548214  0.66502201  0.01703635  0.4425445  100
 400.5 to  500.5   1.48131710  0.75573780 -0.02109060  0.80831525  0.6538184  100
 500.5 to  600.5  -0.47195183 -1.00178911 -1.16547321  0.65023624  1.7811350  100
 600.5 to  700.5   1.10873927 -1.85721402 -0.01401154 -1.41272932  1.9960005  100
 700.5 to  800.5  -1.55463949 -0.66812721  1.24479861  0.67669034  2.0074334  100
 800.5 to  900.5   0.12646220  2.21459751 -0.26762055  0.34422435  0.1901112  100
 900.5 to 1000.5   0.09614022 -0.08036365  0.22422593 -1.56188768  2.4897704  100
TOTAL Q stats      9.84528898 11.94385309  4.35461933  9.81377473 14.1683941 1000
df for Q stats     5.04074288  8.09802751  7.99989205  7.99972028 15.9996123    0
p-val for Q stats  0.08154452  0.15939444  0.82378766  0.27831713  0.5861426    0
```

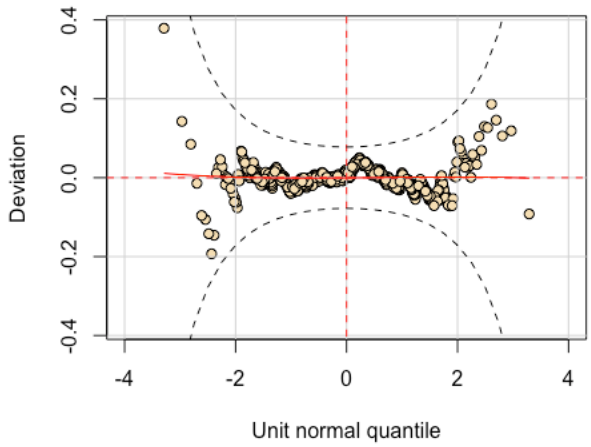
```
> Q.stats(abct)
                          Z1           Z2           Z3           Z4  AgostinoK2    N
   0.5 to  100.5  -1.47106931 -0.108332431  0.310263108  1.584579e+00  2.60715373  100
 100.5 to  200.5  -0.50474228  0.946561262  0.085955479  3.523574e-01  0.13154410  100
 200.5 to  300.5  -0.18275150 -0.561620021 -0.836284344 -6.854740e-01  1.16924617  100
 300.5 to  400.5   1.11847762 -0.505108889  0.505113702 -3.618387e-01  0.38606709  100
 400.5 to  500.5   1.52766212  0.722282791 -0.057420889  6.198225e-01  0.38747706  100
 500.5 to  600.5  -0.42529539 -0.977472378 -1.055080999  4.832500e-01  1.34672653  100
 600.5 to  700.5   1.15746539 -1.768258564  0.004120461 -1.401822e+00  1.96512280  100
 700.5 to  800.5  -1.53094814 -0.647693242  1.230302392  5.752517e-01  1.84455848  100
 800.5 to  900.5   0.16658295  2.161689230 -0.294588388 -7.239589e-05  0.08678232  100
 900.5 to 1000.5   0.14967951 -0.007970211  0.252004139 -1.593706e+00  2.60340389  100
TOTAL Q stats      9.95150990 11.174619594  3.838605329  8.689477e+00 12.52808217 1000
df for Q stats     5.02804591  8.046241948  7.999878087  7.999985e+00 15.99986283    0
p-val for Q stats  0.07782145  0.195099256  0.871376035  3.691631e-01  0.70688971    0

> Q.stats(abct)
                          Z1           Z2           Z3           Z4  AgostinoK2    N
   0.5 to  100.5  -1.47106931 -0.108332431  0.310263108  1.584579e+00  2.60715373  100
 100.5 to  200.5  -0.50474228  0.946561262  0.085955479  3.523574e-01  0.13154410  100
 200.5 to  300.5  -0.18275150 -0.561620021 -0.836284344 -6.854740e-01  1.16924617  100
 300.5 to  400.5   1.11847762 -0.505108889  0.505113702 -3.618387e-01  0.38606709  100
 400.5 to  500.5   1.52766212  0.722282791 -0.057420889  6.198225e-01  0.38747706  100
 500.5 to  600.5  -0.42529539 -0.977472378 -1.055080999  4.832500e-01  1.34672653  100
 600.5 to  700.5   1.15746539 -1.768258564  0.004120461 -1.401822e+00  1.96512280  100
 700.5 to  800.5  -1.53094814 -0.647693242  1.230302392  5.752517e-01  1.84455848  100
 800.5 to  900.5   0.16658295  2.161689230 -0.294588388 -7.239589e-05  0.08678232  100
 900.5 to 1000.5   0.14967951 -0.007970211  0.252004139 -1.593706e+00  2.60340389  100
TOTAL Q stats      9.95150990 11.174619594  3.838605329  8.689477e+00 12.52808217 1000
df for Q stats     5.02804591  8.046241948  7.999878087  7.999985e+00 15.99986283    0
p-val for Q stats  0.07782145  0.195099256  0.871376035  3.691631e-01  0.70688971    0
```

```
library(gamlss)
data(grip)
set.seed(1143)
index<-sample(3766,1000)
airdata<-grip[index,]
dim(airdata)
plot(grip~age,data=mydata,col='maroon')
abccg<-gamlss(grip~pb(age),sigma.fo=~pb(age),nu.fo=~pb(age),data=airdata,family=BCCG)
edfAll(abccg)
abct<-gamlss(grip~pb(age),sigma.fo=~pb(age),nu.fo=~pb(age),tau.fo=~pb(age),data=airdata,family=BCT,start.from=abccg)
edfAll(abct)
abcpe<-gamlss(grip~pb(age),sigma.fo=~pb(age),nu.fo=~pb(age),tau.fo=~pb(age),data=airdata,family=BCPE,start.from=abccg)
edfAll(abcpe)
GAIC(abccg,abct,abcpe)
fittedPlot(abccg,abct,abcpe,x=airdata$age)
centiles(abccg)
centiles(abcpe)
centiles(abct)
centiles.split(abct)
plot(abccg)
plot(abct)
plot(abcpe)
wp(abccg)
wp(abct)
wp(abcpe)

Q.stats(abccg)
```

```
> plot(abccg)
*******************************************************
               Summary of the Quantile Residuals
                           mean   =  0.0002727109
                       variance   =  1.001002
                   coef. of skewness  =  0.002864923
                   coef. of kurtosis  =  3.376945
Filliben correlation coefficient  =  0.9986508
> plot(abcpe)
*******************************************************
               Summary of the Quantile Residuals
                           mean   =  -0.003897011
                       variance   =  1.00039
                   coef. of skewness  =  0.001911036
                   coef. of kurtosis  =  3.082051
Filliben correlation coefficient  =  0.9993569
> plot(abct)
*******************************************************
               Summary of the Quantile Residuals
                           mean   =  5.060968e-05
                       variance   =  1.001233
                   coef. of skewness  =  0.002507798
                   coef. of kurtosis  =  2.980068
Filliben correlation coefficient  =  0.999528
```

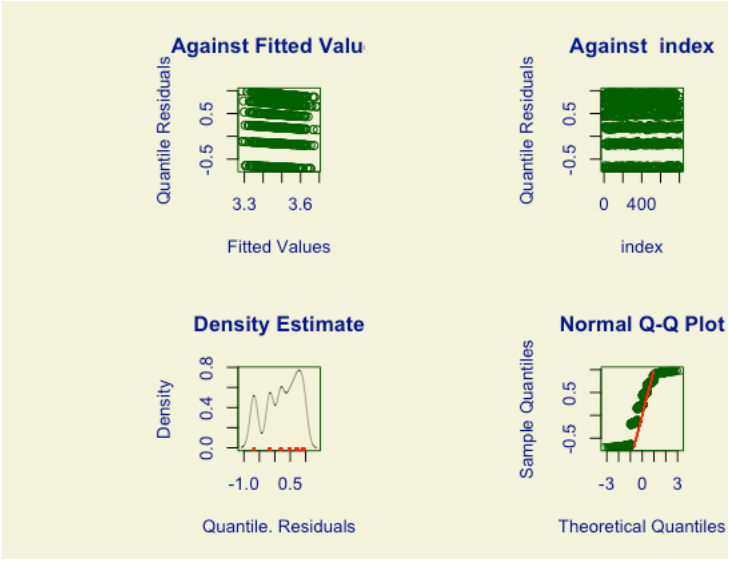

Deviation
0.4
0.2
0.0
-0.2
-0.4
-4
-2
0
2
4
Unit normal quantile

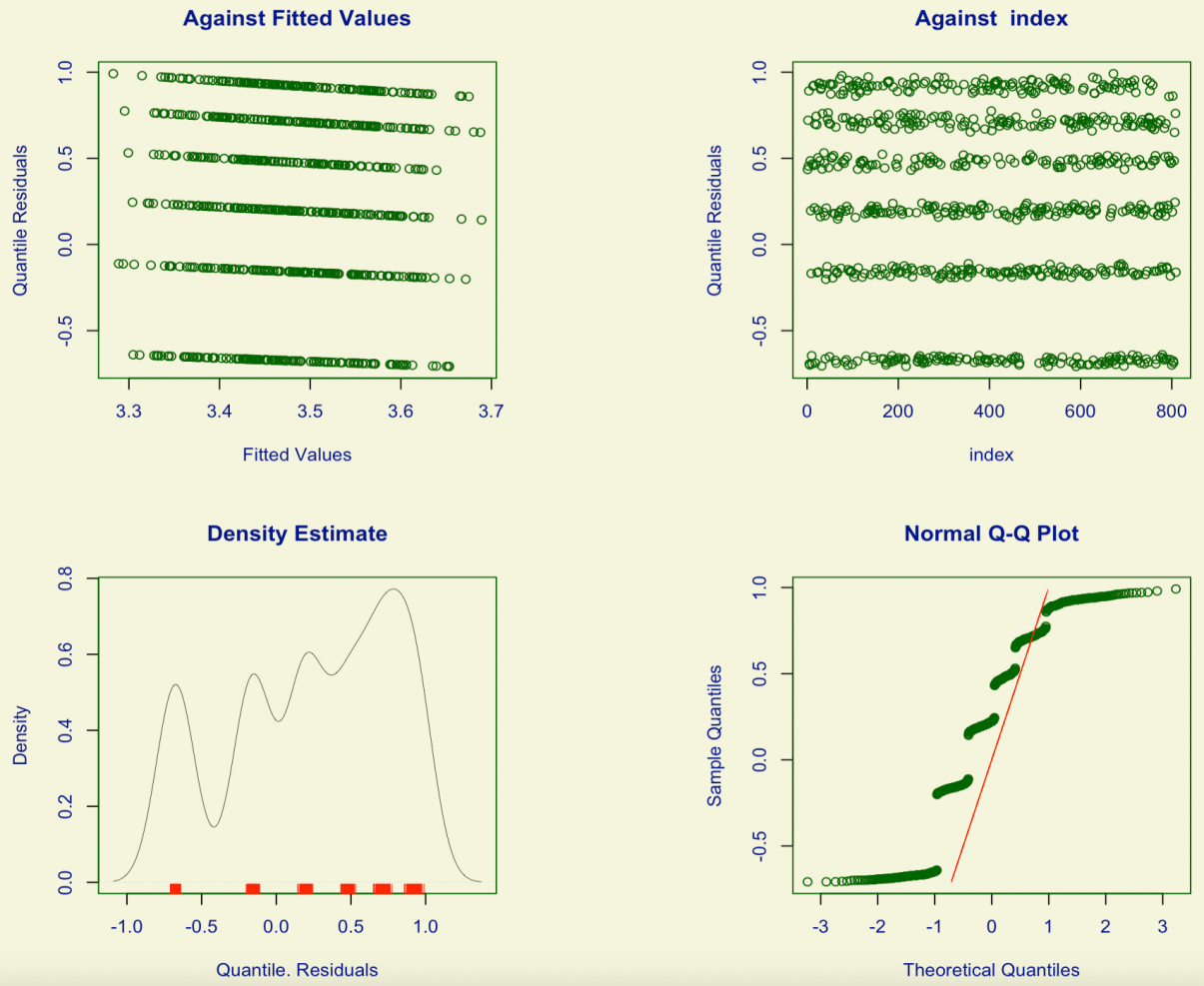

Deviation
0.4
0.2
0.0
-0.2
-0.4
-4
-2
0
2
4
Unit normal quantile


Deviation
0.4
0.2
0.0
-0.2
-0.4
-4
-2
0
2
4
Unit normal quantile

DATASET 3 APPENDIX:

```
> plot(r3)
**************************************************
                Summary of the Quantile Residuals
                              mean   =  -0.04647599
                          variance   =  1.038344
                  coef. of skewness  =  0.003738751
                  coef. of kurtosis  =  1.875919
Filliben correlation coefficient   =  0.9813967
```

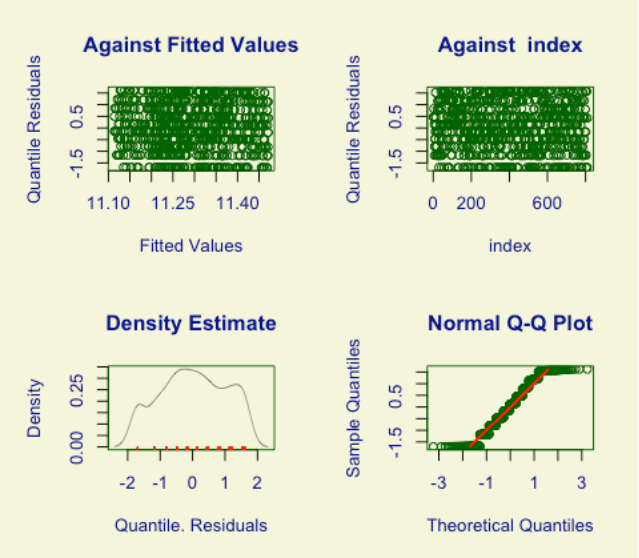


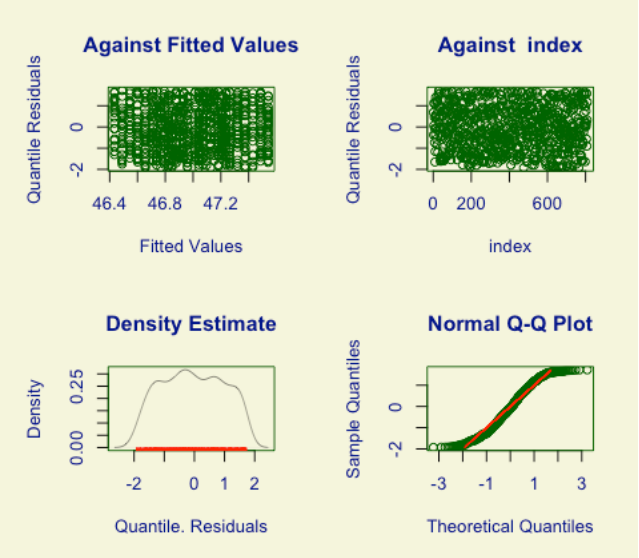



```
> plot(r1)
**************************************************
                Summary of the Quantile Residuals
                              mean   =  0.2388393
                          variance   =  0.2935645
                  coef. of skewness  =  -0.4065595
                  coef. of kurtosis  =  1.941177
Filliben correlation coefficient     =  0.9523902
**************************************************
```

```
> plot(r1)
**************************************************
                Summary of the Quantile Residuals
                             mean   =  0.2388393
                         variance   =  0.2935645
                   coef. of skewness  = -0.4065595
                   coef. of kurtosis  =  1.941177
Filliben correlation coefficient  =  0.9523902
```

```
> GAIC(r1,r2,r3)
   df      AIC
r1  5 3646.233
r2  6 3910.215
r3  6 7547.427
```






```
> dim(superstorenew)
[1] 809   8
> head(superstorenew,5)
  Branch City Customer.type Gender Product.line Unit.price Quantity Payment
1      1    3             1      1            4      74.69        7       3
3      1    3             2      2            5      46.33        7       2
4      1    3             1      2            4      58.22        8       3
5      1    3             2      2            6      86.31        7       3
6      3    2             2      2            1      85.39        7       3
```