MA7007 STATISTICAL MODELLING AND FORECASTING- SPRING 2022-23

ALEKYA DAKARAPU|22023427| MSC DATA ANALYTICS

INTRODUCTION

A detail method of creating a sample data and making a real time forecasting using various statistical models and making assumptions accordingly is Statistical modelling and forecasting.

The First Dataset is given from gamlss to analyse it according to the Instructions provided. First dataset is body mass index (BMI) data which is the subset taken from the fourth Dutch growth study. The data of the body mass index consist of distinct ages in years and BMI of the boys aged between(10-22) which is in galmss.data package under the name of dbbmi.

The Second Dataset is given from gamlss to analyse handgrip strength of the English boys according to the gender and age in school children. The data is taken from the packages gamlss.data which is under the name grip. The data consist of grip (strength) and age of the children in school.

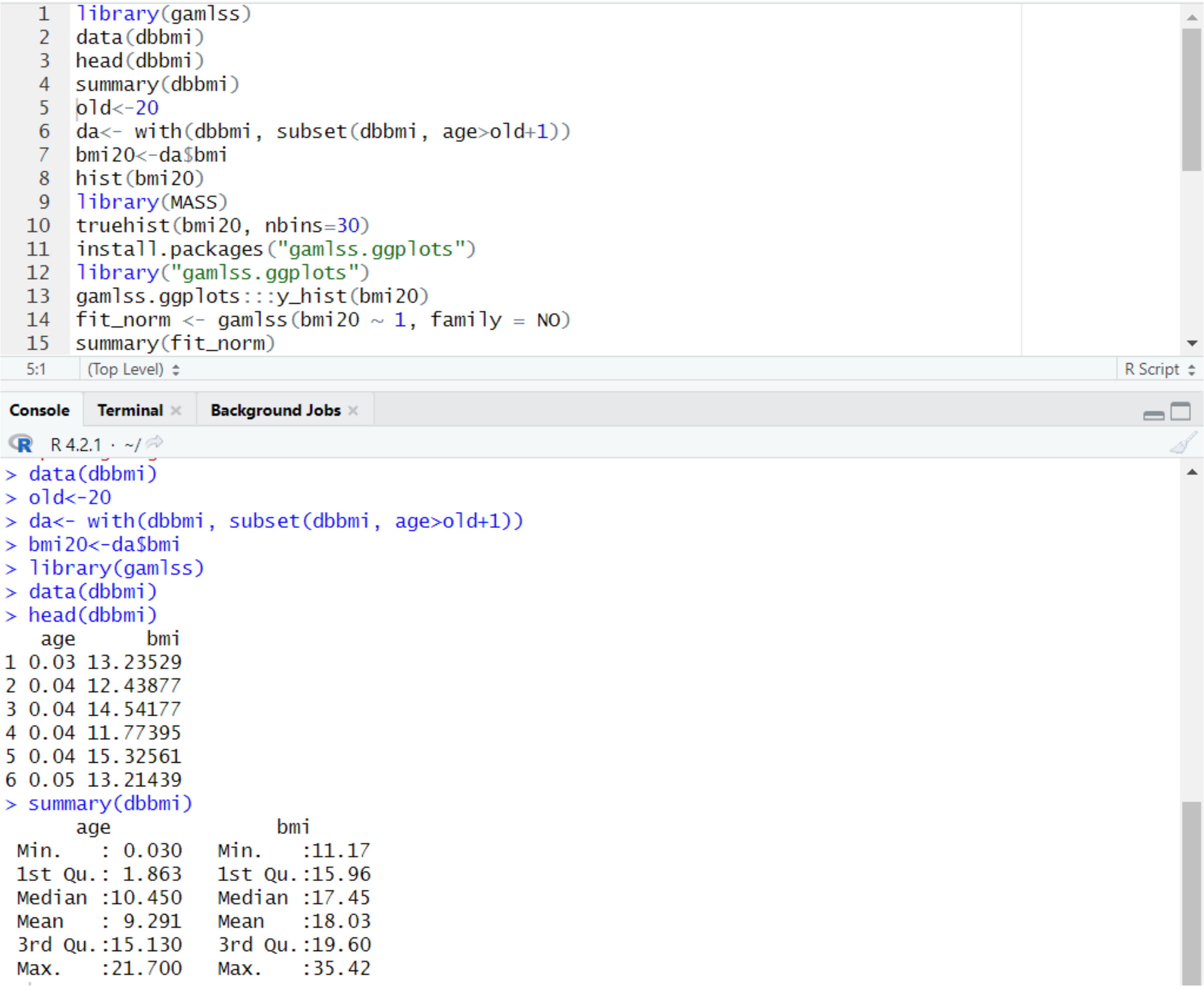
The Third dataset is chosen by group of 10 which is Price paid data taken from the gov.uk data where the data consist of the latest prices of the houses for sale in Wales and England. The data consist of Price, Property/type, old/new, Duration, Town/City, District, County, PPDCategorytype and Record Status.

2. FIRST DATA SET (FITTING DISTRIBUTIONS TO THE DATA)

According to the instruction given we have to find the suitable distribution for the BMI of a particular age. The particular age chosen from the BMI data is 20 years. I would like to find the suitable distribution for the BMI of 20 years aged boys. The data dbbmi consist of 2 variables, age and bmi with 7294 observations. Firstly install library called gamlss. Gamlss consist of dbbmi where the BMI data is located. Here we are taking the sample from the BMI data by creating a subset. The subset of the data consist of the age chosen which is 20 and the number of observations in the subset is 14 observations with 2 variables which is Age and BMI. The subset of the data is stored in the variable bmi20. Now we use bmi20 to perform the analysis and to gain more insights from the fourth Dutch growth study. Then we plot histogram of the subset taken which is bmi20. It shows the

BMI of boys who are 20 years old. To plot true hist we first install the library called mass. To plot a histogram using gamlss install packages called “gamlss.ggplots” from library “gamlss.ggplot”.

- The first step is to install the gamlss package.
- Then load the libraries gamlss which is used to perform the further analysis
- We install gamlss, ggplot2, MASS, gamlss.dist libraries initially.
- Secondly we will load the dataset from the dbbmi which can be found in the gamlss data package.
- To check the first few rows of the dataset we perform head(dbbmi) and summaries the dataset which is loaded by using summary(dbbmi) in the code and it is seen in the figure 1 as the output.



```
library(gamlss)
data(dbbmi)
head(dbbmi)
summary(dbbmi)
old<-20
da<- with(dbbmi, subset(dbbmi, age>old+1))
bmi20<-da$bmi
hist(bmi20)
library(MASS)
truehist(bmi20, nbins=30)
install.packages("gamlss.ggplots")
library("gamlss.ggplots")
gamlss.ggplots:::y_hist(bmi20)
fit_norm <- gamlss(bmi20 ~ 1, family = NO)
summary(fit_norm)
```

```
> data(dbbmi)
> old<-20
> da<- with(dbbmi, subset(dbbmi, age>old+1))
> bmi20<-da$bmi
> library(gamlss)
> data(dbbmi)
> head(dbbmi)
   age      bmi
1 0.03 13.23529
2 0.04 12.43877
3 0.04 14.54177
4 0.04 11.77395
5 0.04 15.32561
6 0.05 13.21439
> summary(dbbmi)
      age              bmi
 Min.   : 0.030   Min.   :11.17
 1st Qu.: 1.863   1st Qu.:15.96
 Median :10.450   Median :17.45
 Mean   : 9.291   Mean   :18.03
 3rd Qu.:15.130   3rd Qu.:19.60
 Max.   :21.700   Max.   :35.42
```

Figure1

Information of the packages

"gamlss": This package allows us to fit and explore generalized additive models for location, size, and form parameters. It not only enables us to estimate various distributions but also provides a versatile framework for modeling complex interactions between variables.

"ggplot2": This powerful data visualization tool in R offers a user-friendly and flexible visual syntax to create high-quality plots and charts. It empowers us to produce visually appealing and adaptable visualizations.

"MASS" (Modern Applied Statistics with S): This comprehensive program offers a wide range of statistical operations and datasets. It includes regression, classification, and multivariate analysis, and also provides practical utilities for model fitting, data manipulation, and statistical inference.

"gamlss.dist": As an extension of the gamlss package, this package enhances our modeling capabilities by offering additional distribution families. By integrating these distributions into the gamlss framework, it expands the range of modeling choices available to us.

In the initial steps of our analysis, we also check for any null values in the dataset. After running the sum(is.na(dbbmi)) code, we find that there are zero (0) null values in the dataset, which is ideal for our analysis.

We then set our age to 20. Then we execute the code where we see:

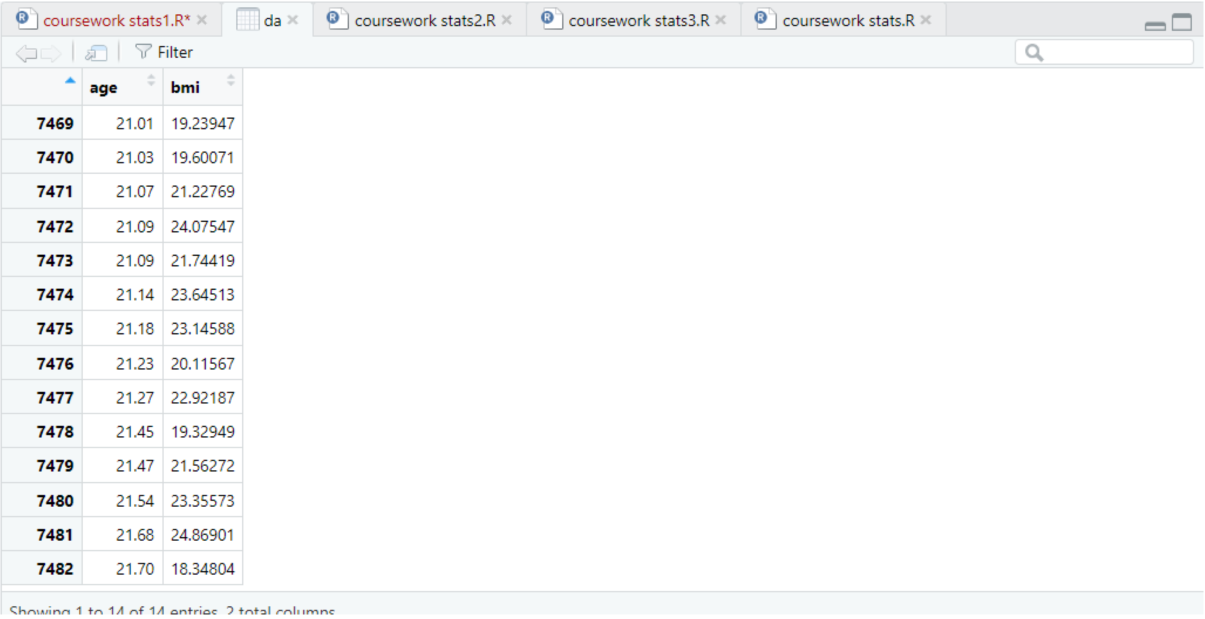
- The programme extracts a subset of data from the "dbbmi" dataset, which is much larger.

- It only selects rows from the dataset whose ages fall within a predetermined range.

- The condensed set of data is stored in a new data frame called "da." When we run the code to see the new data frame, it appears as shown in the figure.

Let's get started on our main project.

Finding the suitable value for the plot: We chose 20 as our age, but there are many people with months as the value here, for example, 20 years 5 months, 20 years 2 months, and so on. To avoid this, we take all of the values from 20 to the starting point of 21, which can be considered the value between 20 and 21. And we take those BMI values and save them, as shown on the right side of Figure 2. We can now plot a histogram to see the initial phase of our data frame and get a sense of it, which will help us choose the best-fitted distribution.



	age	bmi
7469	21.01	19.23947
7470	21.03	19.60071
7471	21.07	21.22769
7472	21.09	24.07547
7473	21.09	21.74419
7474	21.14	23.64513
7475	21.18	23.14588
7476	21.23	20.11567
7477	21.27	22.92187
7478	21.45	19.32949
7479	21.47	21.56272
7480	21.54	23.35573
7481	21.68	24.86901
7482	21.70	18.34804

Showing 1 to 14 of 14 entries, 2 total columns

Figure2

Histogram: The histogram (figure 3) is made by segmenting the BMI range into several evenly spaced intervals known as "bins" or "buckets." The x-axis represents the intervals or bins, and the y-axis shows the frequency or count of BMI values that fall within each bin. Because of the histogram's shape, it is possible to comprehend the distribution of BMI data.

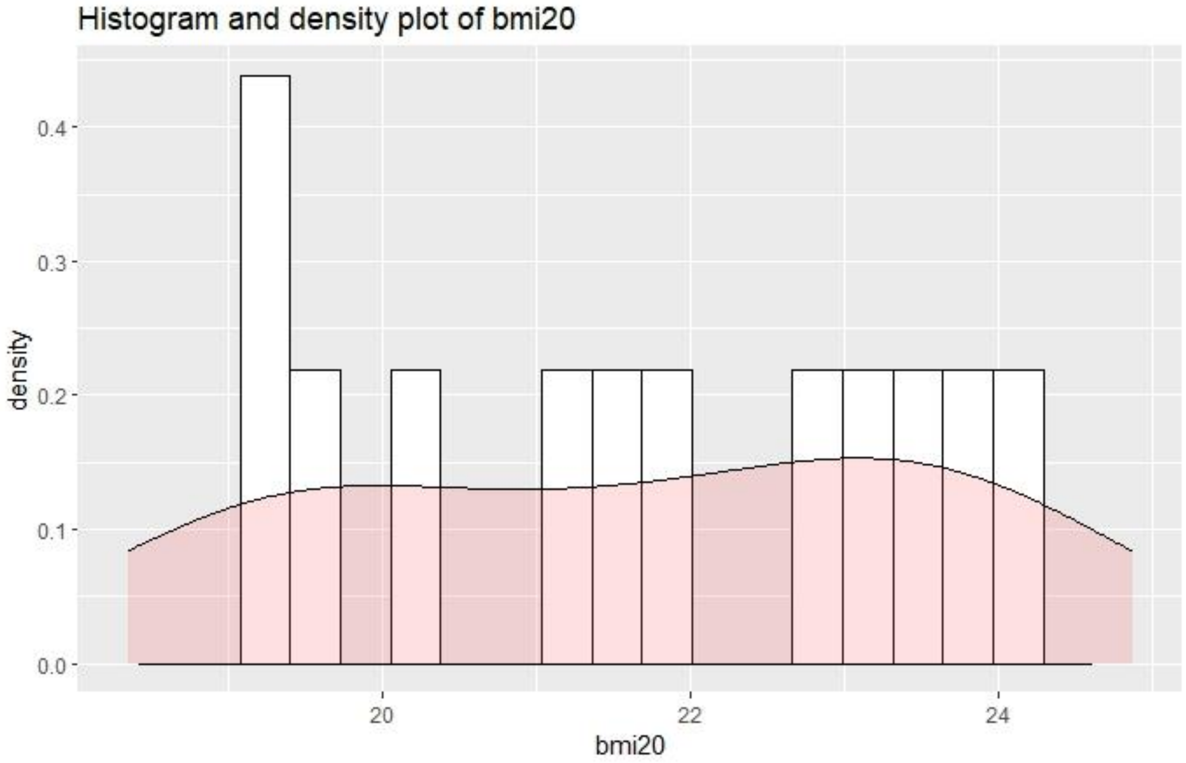


Figure3

(A) COMMENT ON THE DIFFERENT DISTRIBUTIONS YOU ARE USING.

Continuous probability distributions are those in which a random variable can take any quantity within a given range. Some examples of continuous probability distributions in R using the GAMLSS package are:

Normal Distribution: A normal distribution is a continuous probability distribution with a symmetric bell-shaped curve around the mean. It is defined by the mean and variance.

Gamma distribution: A right-skewed continuous probability distribution that is commonly used to model intervals or durations. It has shape and scale parameters.

Log-normal: The log-normal distribution is a continuous probability distribution that is frequently utilised to model positive variables with skewed distributions. It is defined by the mean and variance of the variable's logarithm.

Inverse Gaussian: The inverse Gaussian distribution is a continuous probability distribution that is frequently used to model the time needed for an event to occur. It can be defined by the mean and shape.

A continuous probability distribution that is frequently utilised to model proportions or probabilities is the beta distribution. It can be determined by the shape parameters and.

These are only a few of the numerous continuous probability distributions available in R.

The different parametric distributions that we are using are Normal distribution, Gamma distribution, Lognormal distribution, Inverse gaussian distributions. **Normal Distribution:** The normal distribution is frequently used as the default choice for modelling continuous response variables that are assumed to be normally distributed in GAMLSS (Generalised Additive Models for Location, Scale, and Shape) in R. Because it has several desirable properties, the normal distribution is a popular choice in statistical modelling. The Continuous probability of the normal distribution is symmetric at its mean. It is used to model the data which is Continuous and that are normally distributed such as people weights, heights of people, test scores and measurement errors. It models the response variables mean.

Gaussian Distribution: Gamma Distribution also comes under continuous distribution which is skewed to the right and the its peak are on left. When the skewness is towards the right then it is considered as positive skewness. It is used to model the non-negative data which has skew distributions .It models data such as time of failure, income, number of events are few examples of the data. The response variable which has the scale parameter is modelled by the Gamma distribution.

Lognormal Distribution: This distribution is a continuous probability distribution which is skewed to the right hand side and it has a long tail to the right hand side. Even in lognormal it is used to model positive continuous data that has skewed distributions. The distribution lognormal is defined on the +ve real line which has two parameters. The first parameter is location parameter which is known as Mu and the second parameter is scale parameter which is known as sigma. The response variable of the logarithm or response variable directly can be modelled by Lognormal Distribution.

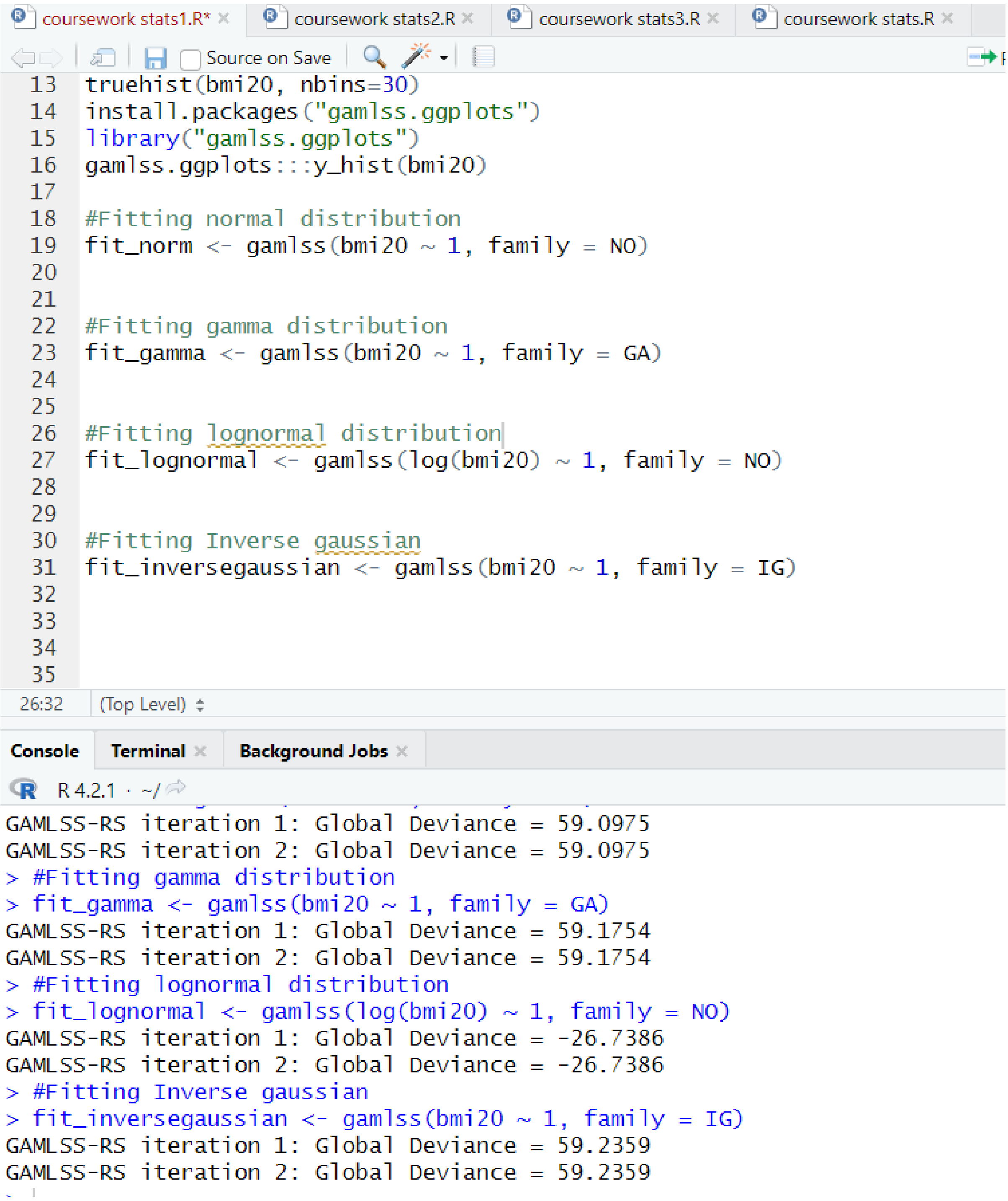
Inverse Gaussian Distribution: Even inverse Gaussian distribution is also a continuous probability distribution that is also a positive skewness which is towards right and has a long tail but it has more peak than the distribution lognormal. It is used to model +ve continuous data that has distribution which is skewed and has heavy tails. Inverse gaussian distribution has two parameters with positive real line. The first parameter is known as location parameter which is Mu and the second parameter is known as shape parameter which is lambda. The reciprocal of the response variable or the response variable is directly modelled by the inverse gaussian distribution.

These are the 4 different parameter distribution used.

(B) CHOOSE THE APPROPRIATE DISTRIBUTION:

Before selecting the best-fitted distribution, we must first examine the AIC values for the various distributions, which will assist us in determining or selecting the best fit.

AIC (Akaike Information Criterion): The Akaike Information Criterion, or AIC, is a metric for evaluating the overall quality of statistical models. In this case, AIC is used to compare several distributions fitted to the BMI data. AIC penalises models with more parameters by taking into account the model's complexity as well as its quality of fit. The lower the AIC score, the better the model fits the data while avoiding overfitting. By comparing the AIC values of several distributions, we can identify the distribution that provides the best balance of goodness of fit and model complexity. Figure 4 shows the output of the code after execution.

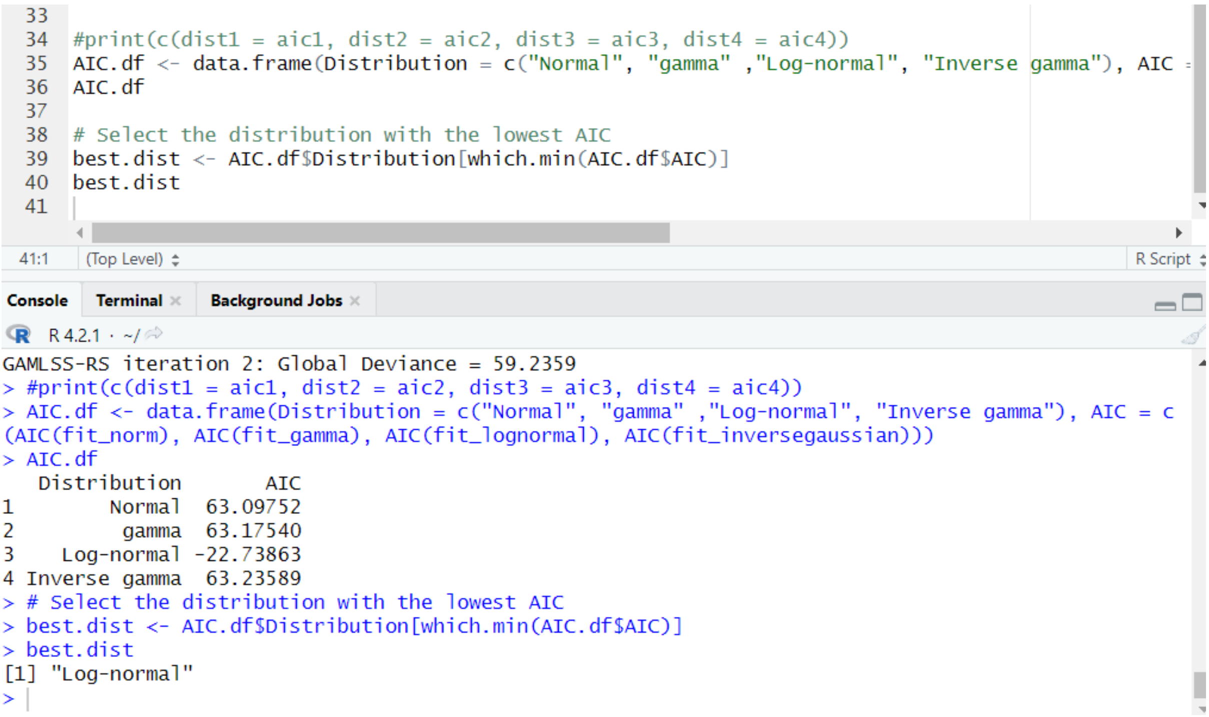


```
truehist(bmi20, nbins=30)
install.packages("gamlss.ggplots")
library("gamlss.ggplots")
gamlss.ggplots:::y_hist(bmi20)

#Fitting normal distribution
fit_norm <- gamlss(bmi20 ~ 1, family = NO)


#Fitting gamma distribution
fit_gamma <- gamlss(bmi20 ~ 1, family = GA)


#Fitting lognormal distribution
fit_lognormal <- gamlss(log(bmi20) ~ 1, family = NO)


#Fitting Inverse gaussian
fit_inversegaussian <- gamlss(bmi20 ~ 1, family = IG)
```

```
GAMLSS-RS iteration 1: Global Deviance = 59.0975
GAMLSS-RS iteration 2: Global Deviance = 59.0975
> #Fitting gamma distribution
> fit_gamma <- gamlss(bmi20 ~ 1, family = GA)
GAMLSS-RS iteration 1: Global Deviance = 59.1754
GAMLSS-RS iteration 2: Global Deviance = 59.1754
> #Fitting lognormal distribution
> fit_lognormal <- gamlss(log(bmi20) ~ 1, family = NO)
GAMLSS-RS iteration 1: Global Deviance = -26.7386
GAMLSS-RS iteration 2: Global Deviance = -26.7386
> #Fitting Inverse gaussian
> fit_inversegaussian <- gamlss(bmi20 ~ 1, family = IG)
GAMLSS-RS iteration 1: Global Deviance = 59.2359
GAMLSS-RS iteration 2: Global Deviance = 59.2359
```

Figure 4

Now, after fitting the different distributions and comparing them in the AIC, we get to see the comparing values in figure 5. There we can see the Global Deviance value for each of the distributions.



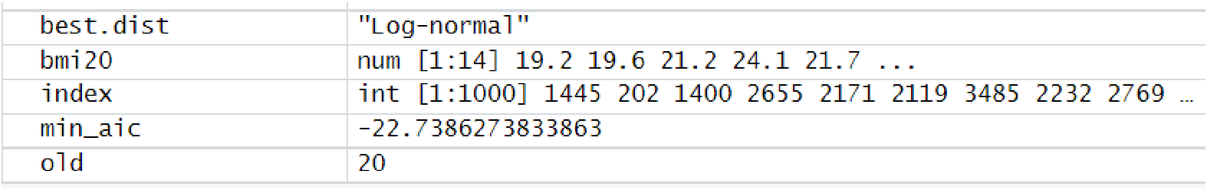
```r

#print(c(dist1 = aic1, dist2 = aic2, dist3 = aic3, dist4 = aic4))
AIC.df <- data.frame(Distribution = c("Normal", "gamma" ,"Log-normal", "Inverse gamma"), AIC =
AIC.df

# Select the distribution with the lowest AIC
best.dist <- AIC.df$Distribution[which.min(AIC.df$AIC)]
best.dist

```

```
GAMLSS-RS iteration 2: Global Deviance = 59.2359
> #print(c(dist1 = aic1, dist2 = aic2, dist3 = aic3, dist4 = aic4))
> AIC.df <- data.frame(Distribution = c("Normal", "gamma" ,"Log-normal", "Inverse gamma"), AIC = c
(AIC(fit_norm), AIC(fit_gamma), AIC(fit_lognormal), AIC(fit_inversegaussian)))
> AIC.df
   Distribution       AIC
1        Normal  63.09752
2         gamma  63.17540
3    Log-normal -22.73863
4 Inverse gamma  63.23589
> # Select the distribution with the lowest AIC
> best.dist <- AIC.df$Distribution[which.min(AIC.df$AIC)]
> best.dist
[1] "Log-normal"
> 
```

Figure5

Finally, we can say that the Log-normal distribution has the lowest AIC value, which is -22.73863, and thus we get our desired best-fitted distribution. Figure 6 illustrates the details.

Justification: The "Log-normal" distribution was chosen based on the premise of finding the distribution with the best balance of goodness of fit and model complexity. According to the lower AIC, the "Log-normal" distribution is a good choice for modelling and data analysis because it effectively represents the properties of the BMI data.



best.dist	"Log-normal"
bmi20	num [1:14] 19.2 19.6 21.2 24.1 21.7 ...
index	int [1:1000] 1445 202 1400 2655 2171 2119 3485 2232 2769 ...
min_aic	-22.7386273833863
old	20

Figure6

(C) GIVE REASONS WHY YOU CHOSE THE DISTRIBUTION IN PART(B).

The code fits several distributions (such as normal, log-normal, gamma, and Inverse gamma) to the BMI data based on the lowest AIC value. Lower values of the AIC (Akaike Information Criterion), used to quantify model fit, indicate better fit. The AIC values for each distribution are computed and saved in a data frame. By comparing the AIC values, the distribution with the lowest AIC is determined to be the best-fitting distribution. In this case, the "Log-normal distribution" is used.

(D) PLOT THE FITTED DISTRIBUTION AND COMMENT.

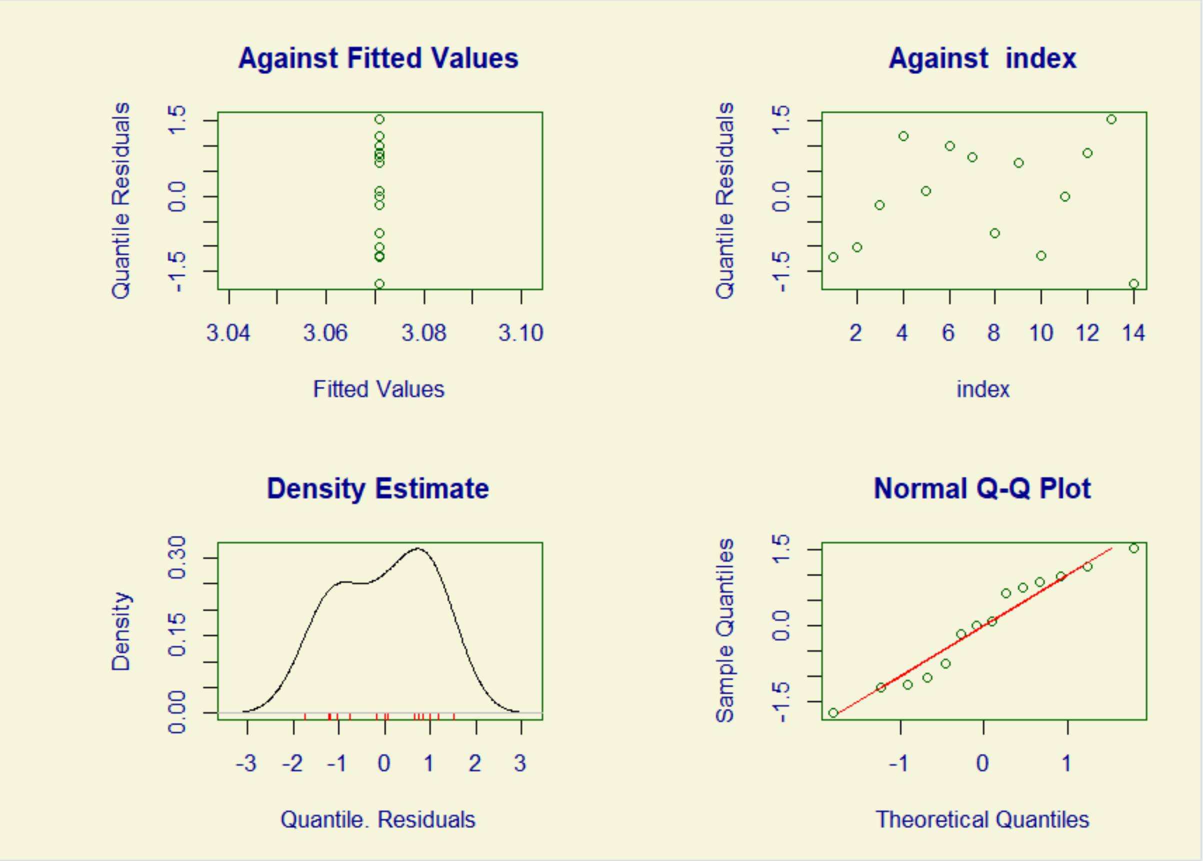


Figure7

These diagnostic plots are critical for evaluating the Log-normal model's suitability and identifying any potential issues or broken assumptions. They provide perceptions of the model's performance and, if necessary, serve as a roadmap for additional research and improvement..

For instance:

- In the residuals, for example, it aids in identifying influential data points that have a large impact on the model's fit. Points outside of a specific range may have a significant impact on the estimated coefficients.

- A Q-Q plot is used to compare the quantiles of the observed data to the quantiles of the Lognormal distribution. Deviations from a straight line represent deviations from the expected distribution.

The residuals vs. fitted plot depicts the relationship between the fitted and residual values (differences between observed and forecasted values). Patterns in the plot may indicate that the model has nonlinearity or heteroscedasticity issues.

(E) STATE THE FITTED PARAMETER VALUES OF THE FINAL CHOSEN MODEL

The summary output includes the fitted parameter values for the best-fitting distribution (Log-normal). The following sections describe the main components of the output:

- Family: The fitted distribution is known as the "Lognormal" distribution (lognormal).

- Mu link function: To calculate the location parameter (mu), use the "identity" link function. The intercept's estimated coefficient (mu) is 18.973 with a standard error of 0.127.

- Sigma link function: The "log" link function is used to compute the scale parameter (sigma). The predicted coefficient for the intercept (sigma) is -2.06938, with a standard error of 0.03635.

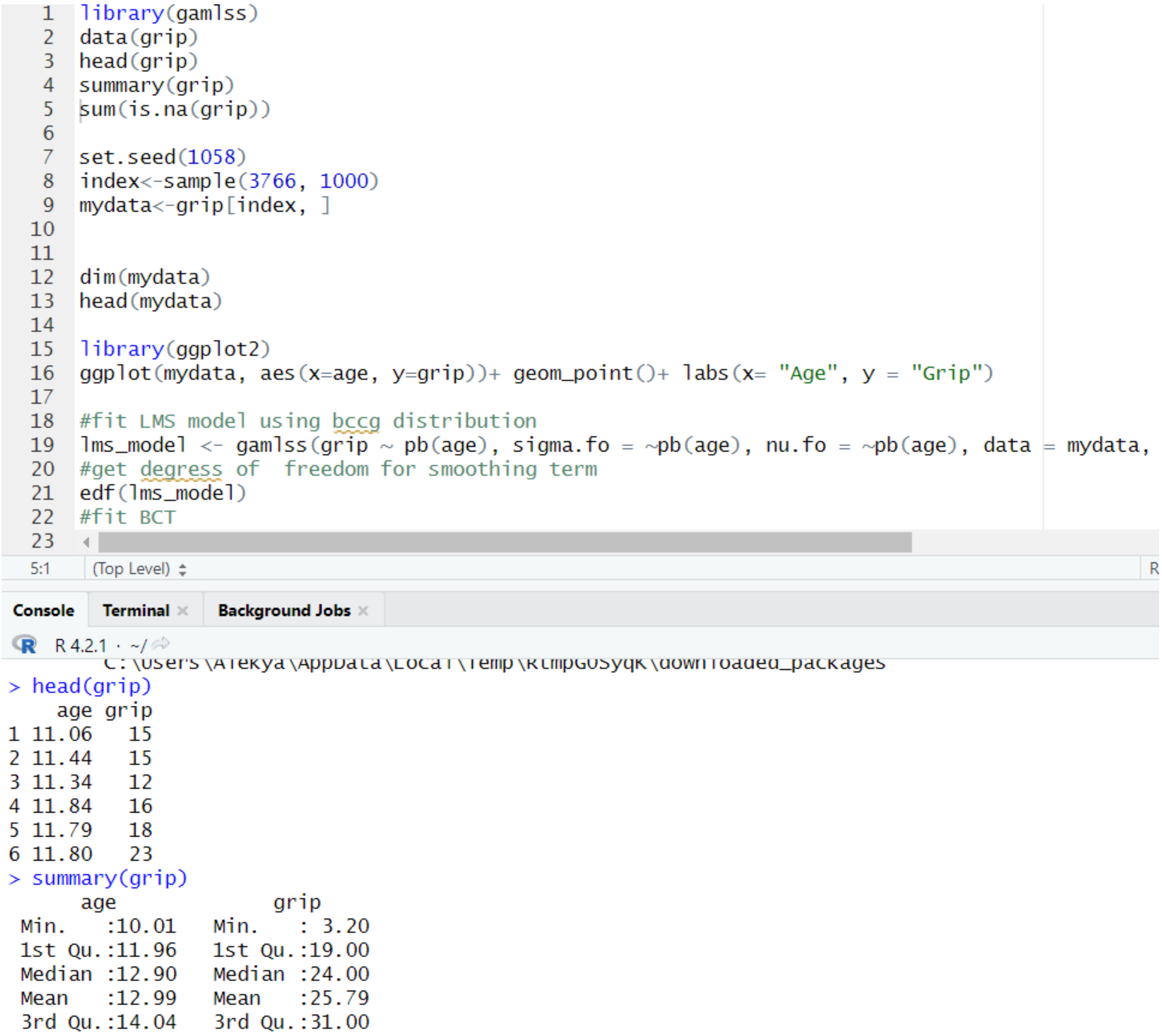
- Nu link function: The link function "identity" is used to calculate the kurtosis

- Number of observations: In total, 14 observations were used in the fitting procedure.
- Degrees of Freedom: There are three degrees of freedom in the fit, which correspond to the number of estimated parameters.
- Residual Degrees of Freedom: 0 residual degrees of freedom are obtained by subtracting the number of observations from the number of computed parameters
- Global Deviance: The global deviation is 59.0975.
- AIC: The Akaike Information Criterion (AIC) value is -22.73863. Lower numbers indicate better-fitting models, and these criteria provide model fit metrics.

The estimated parameter values for the Log-normal distribution are included in the overall summary result, allowing comprehension of the distribution's unique characteristics such as location, scale, and kurtosis.

3. SECOND DATA SET (CENTILE ESTIMATION)

Cohen et al investigated the relationship between handgrip (HG) strength, gender, and age in English schoolchildren. Each student was given a different sample of 1000 people to analyse from the initial group of 3766 English men. Grip (handgrip strength) and age are variables in the dataset "grip," which is included in the gamlss.data package. This study's main goal is to develop age-based centile curves for handgrip strength. We hope that this study will shed some important light on how handgrip strength changes with age in a community of English students. Along with the 3766 of obs, we can see two variables in the dataset taken from the gamlss. As a result, in this section, we must conduct our analysis while answering the questions that have arisen. Let's start with the data while loading it with gamlss. Following that, I was assigned a seed number of 1058. Before we can use the seed number, we must first configure our R-Studio platform with GAMLSS for further analysis and work. We can do them similarly to the first dataset. Figure 8 depicts the dataset's summary and the head. Now we must select the sample data by running the code provided in the question:



```
library(gamlss)
data(grip)
head(grip)
summary(grip)
sum(is.na(grip))

set.seed(1058)
index<-sample(3766, 1000)
mydata<-grip[index, ]


dim(mydata)
head(mydata)

library(ggplot2)
ggplot(mydata, aes(x=age, y=grip))+ geom_point()+ labs(x= "Age", y = "Grip")

#fit LMS model using bccg distribution
lms_model <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), data = mydata,
#get degress of  freedom for smoothing term
edf(lms_model)
#fit BCT
```

```
    C:\Users\Atekya\AppData\Local\Temp\RtmpGUSyqK\downloaded_packages
> head(grip)
    age grip
1 11.06   15
2 11.44   15
3 11.34   12
4 11.84   16
5 11.79   18
6 11.80   23
> summary(grip)
      age             grip      
 Min.   :10.01   Min.   : 3.20  
 1st Qu.:11.96   1st Qu.:19.00  
 Median :12.90   Median :24.00  
 Mean   :12.99   Mean   :25.79  
 3rd Qu.:14.04   3rd Qu.:31.00  
```

Figure8

Plotting Grip against Age: When we run the sample data (mydata), we see that there are 1000 values with the two variables age and grip. We discovered 0 null values in age and 0 null values in grip. Now we plot the age against grip, as shown in figure 9.

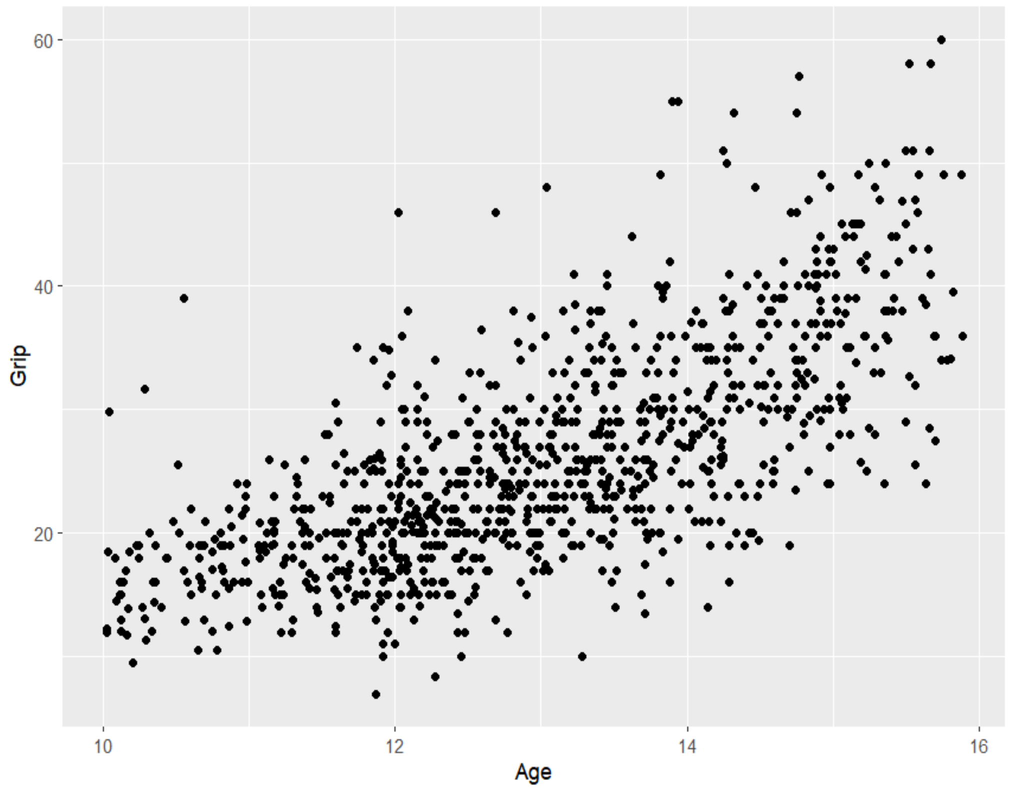


Figure9

We can see that, most of the values stay between 11-13 while the grip seems to be more from 15 to 35 around. When plotting grip versus age in the supplied data set, there is no need for the modification of the age variable.

Here is a justification: When deciding whether to convert data, the goals of the analysis and the distributional properties of the variable are typically taken into account. In this case, the predictor is the age variable, and the result of interest is the grip variable. It appears that age does not need to be power converted for any specific purpose or indication. The grip variable may already be in a usable form for analysis, and a linear relationship may suffice to represent the relationship between grip and age. It is critical to remember that the decision to change a variable may be influenced by the specifics of the data as well as the current research issue. Based on domain knowledge and statistical assumptions, I recommend carefully investigating the distributional features of variables and, if necessary, considering alternative transformations. However, in my case, it is not required.

3(A) COMMENT ON THE DIFFERENT MODELS YOU ARE USING.

Now, we'll look into the models to see which ones are the best for getting our desired centile plot for the grip where age is given. The three different distributions are introduced here. Statistical modelling, as we know, employs the Box-Cox Power Exponential (BCPE) distribution, Box-Cox Cole and Green (BCCG) transformation, and Box-Cox t transformation to handle various data features. In order to describe data with various tail behaviours, the BCPE distribution combines the BoxCox transformation with the Power Exponential distribution. The BCCG transformation, which extends the Box-Cox transformation, can deal with heteroscedasticity and skewness in variables with positive values. When applied to high-dimensional data, the Box-Cox t transformation uses a power transformation to improve symmetry and stabilise variance. These methods are critical for establishing normalcy and capturing the distinct properties of different datasets, allowing for more precise statistical modelling and analysis.

3(A.1)Model Analyze

3 Fit LMS model using BCCG distribution Let’s understand the code first:

#Fitting the LMS model by using BCCG distribution lms model

The fitted GAMLSS model is saved in a variable called gbccg. To fit GAMLSS models, the gamlss package's gamlss() method is used. The formula for the mod el is grip pb(age), where grip is the response variable and pb(age) is the P-spline function that was used to smooth the age predictor variable. The symbol represe nts the connection between the answer and the predictor. Independent P-spline f unctions specify the smoothing terms of the scale (sigma) and shape (nu) param eters (sigma.fo = pb(age) and nu.fo = pb(age)). This enables estimation of the ag e-dependent variability and shape of the BCCG distribution. The data frame dat a = da indicates the data frame (da) from which the variables (grip and age) are collected. As a result, we have the LMS model with 50 elements. If we want to know how many degrees there are, we perform another analysis and get the resu lt. The edf(lms model) code in the lms model object determines the smoothing t erm's effective degrees of freedom (df). It describes how adaptable or complex t he smoothing function used for each prediction in the model is. The output you described is specifically related to the effective df of the mu model's pb(age) ter m. Because the effective degrees of freedom for the age predictor are around 4.

700247, the association between grip strength and age can be captured with a significant degree of flexibility in this example. The effective df, which takes into account any penalties or restrictions used during the model estimating process, is a useful statistic for assessing the smoothness of the fitted model.

3.A.2 Fitting BCT,BCPE distribution using LMS model’s starting values This one is a bit larger than the previous one, as we can see there are 50 elements. Now, let’s understand them:

The fitted GAMLSS model is saved in a variable called gbct using the BCT distribution. The model's formula is grip pb(age), where grip is the response variable and pb(age) is the smoothing Pspline function used to smooth the age predictor variable. The smoothing terms of the scale (sigma), shape (nu), and skewness (tau) parameters are smoothed by the P-spline functions specified by the expressions sigma.fo = pb(age), nu.fo = pb(age), and tau.fo = pb(age), respectively. This allows for the prediction of the BCT distribution's age-dependent variability, shape, and skewness.The expression data = mydata indicates that the variables (grip and age) are derived from the data frame (mydata). The GAMLSS model, which simulates the grip variable's position, size, shape, and skewness, defines the BCT distribution as the family using the expression family = BCT. The BCT model's initial parameters should be obtained from the previously fitted LMS model, as indicated by start.from=lmsmodelsetting(lmsmodel).This aids in providing parameter estimates for the BCT model. The code edf(gbct) determines the effective degrees of freedom (df) for the smoothing term in the gbct GAMLSS model object, particularly for the mu model. The smoothing function employed for the age predictor in the mu model of the BCT The smoothing function used for the age predictor in the BCT GAMLSS model's mu model is flexible or complicated, as evidenced by the computed value of around 4.893655. A higher effective df indicates a more adaptable grip strength and age relationship, allowing for the capture of more complex patterns. A lower effective df, on the other hand, indicates a more limited or direct connection. The effective df accounts for any penalties or restrictions applied during model estimation and provides a useful way to assess how well the fitted model fits the data while taking the smoothing method into account

2. BCPE:

The smoothing terms of the scale (sigma), shape (nu), and skewness (tau) parameters are smoothed by the P-spline functions specified by the expressions sigma.fo = pb(age), nu.fo = pb(age), and tau.fo = pb(age), respectively. This allows us to forecast the variability, shape, and skewness of the BCPE distribution as a function of age.The expression data = mydata indicates that the variables (grip and age) are derived from the data frame (mydata). The GAMLSS model, which simulates the grip variable's position, size, shape, and skewness, defines the family = BCPE as the BCPE distribution. start.from = lms model tells the BCPE mod el to start with the values from the previously fitted LMS model (lms model). This helps to provide a good first approximation for the parameter estimations in the BCPE model. When you run this code, you will get a new GAMLSS model that incorporates the data and starting values from the LMS model and employs the BCPE distribution. This allows for an examination of the relationship between grip strength and age while accounting for the skewness, variability, and shape of the BCPE distribution. The code edf(gbcpe) determines the effective degrees of freedom (df) for the smoothing term in the gbcpe GAMLSS model object, particularly for the mu model. The smoothing function used for the age predictor in the BCPE GAMLSS model's mu model has a level of flexibility or complexity indicated by the obtained value of approximately 4.882. A higher effective df indicates a more flexible connection that can capture complex patterns and interactions between grip strength and age. A lower effective df, on the other hand, indicates a more limited or direct connection.

3(B) COMPARING ALL THE THREE MODELS USING GAIC

To compare the goodness-of-fit and complexity of the bccg, bct, and bcpe models, the Generalised Akaike Information Criterion (GAIC) is used. The GAIC value for the gbcpe model is 6392.523, indicating a good tradeoff between model fit and complexity. It also indicates the best fit. The gbct model also performs well, with a slightly higher GAIC score of 6392.523. The lms model, on the other hand, has the highest GAIC value (6403.770), indicating a slightly worse match. Based on the GAIC findings for the supplied data and analysis, the gbcpe model appears to be the best option among the three. It stands out for its greater fit and reduced complexity.

3(C)PLOTTING FITTED PARAMETER

We can see the plottings among gbccg, gbct, gbbcpe in figure 10

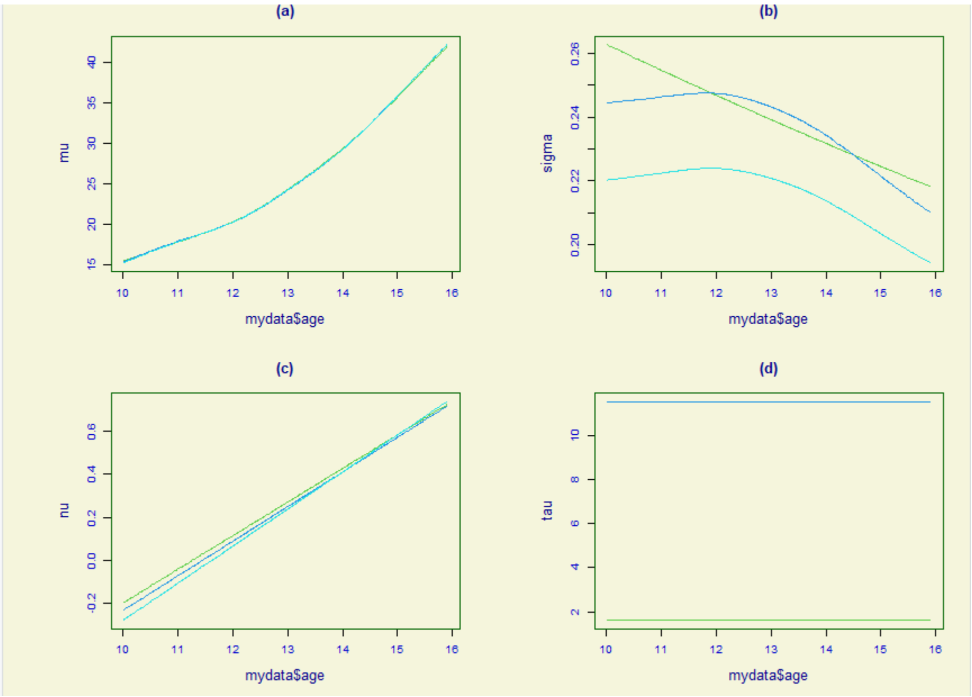


Figure10 Fitted parameters for fitted data set

3(D)OBTAINING CENTILE PLOT AND COMPARE THEM

By running certain code we got some plot views in figure 11, figure 12, and figure 13. For the LMS, In this case, we see that: 0.4 centile: 0.6% of cases are below this centile. 2 centile: 2.5% of cases are below this centile. 10 centile: 8.6% of cases are below this centile. 25 centile: 24.9% of cases are below this centile. 50 centile (Median): 50% of cases are below this centile. 75 centile: 75.5% of cases are below this centile. 90 centile: 91% of cases are below this centile. 98 centile: 97.6% of cases are below this centile. 99.6 centile: 99.4% of cases are below this centile. These centiles provide information about the distribution of the data and can be used to understand the relative position of a given value within the distribution. For the gbct, 0.4 centile: 0.4% of cases are below this centile. 2 centile: 2.3% of cases are below this centile. 10 centile: 9.3% of cases are below this centile. 25 centile: 25.5% of cases are below this centile. 50 centile (Median): 49.8% of cases are below this centile. 75 centile: 74.2% of cases are below 19 this centile. 90 centile: 90.7% of cases are below this centile. 98 centile: 97.6% of cases are below this centile. 99.6 centile: 99.5% of cases are below this centile. These centiles provide information about the distribution of the data and can be used to understand the relative position of a given value within the distribution. For the gbbcpe, 0.4 centile: 0.6% of cases are below this centile. 2 centile: 2.3% of cases are below this centile. 10 centile: 9.1% of cases are below this centile. 25 centile: 25.5% of cases are below this centile. 50 centile (Median): 50% of cases are below this centile. 75 centile: 74.4% of cases are below this centile. 90

centile: 90.9% of cases are below this centile. 98 centile: 97.6% of cases are below this centile. 99.6 centile: 99.5% of cases are below this centile

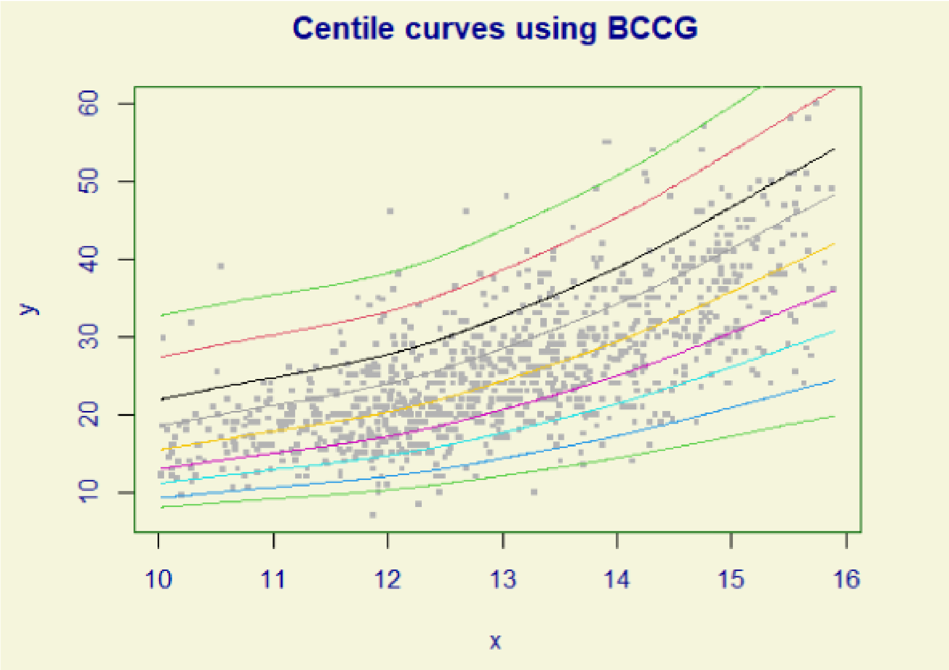


Figure 11

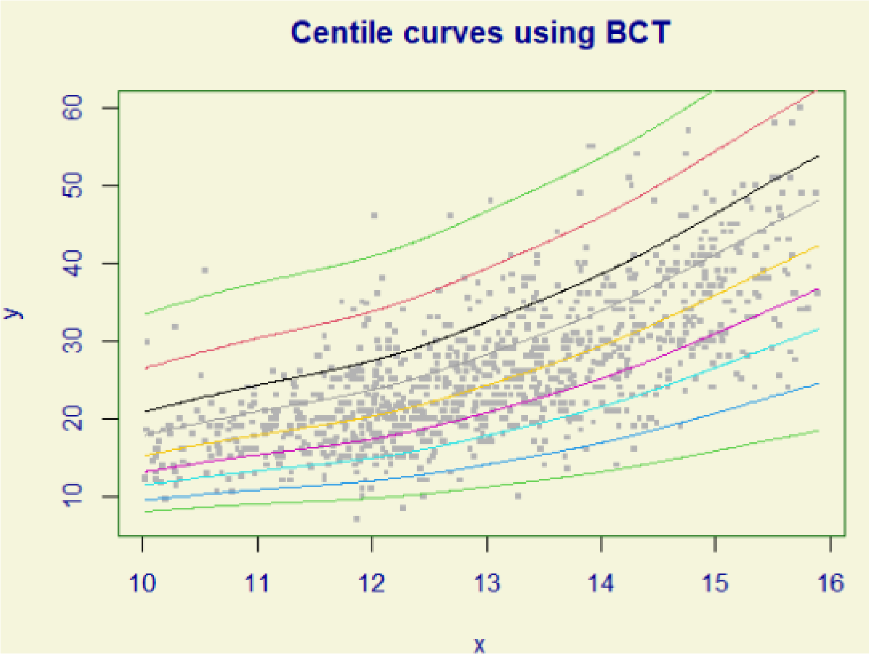


Figure12

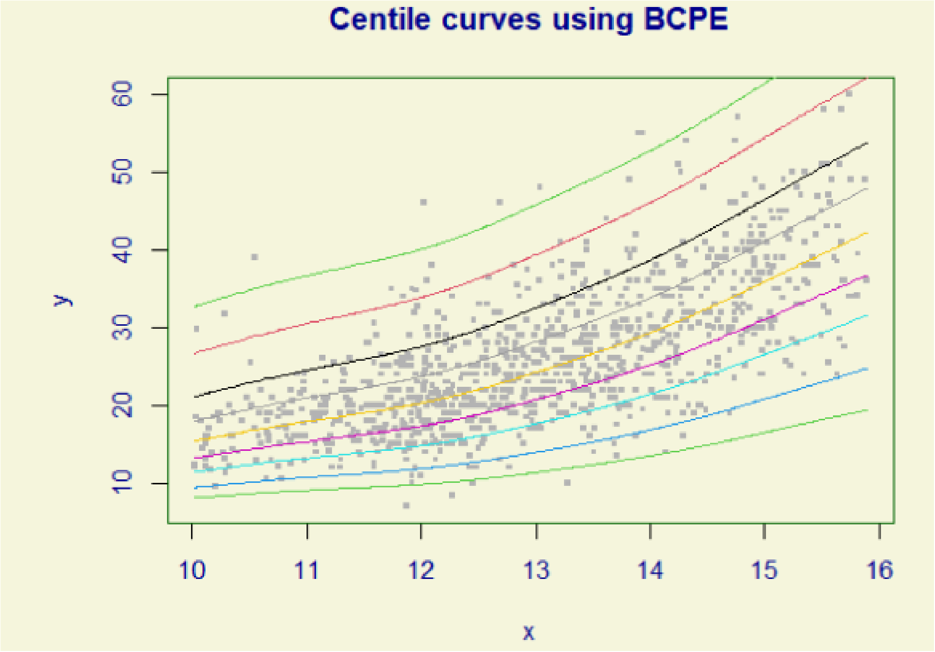


Figure13

3(E) INVESTIGATING RESIDUALS

- Residuals for LMS:

Here from figure 14, 15, 16. Summary of the Quantile Residuals, mean = 0.00 03442107, variance = 1.001002, coef. of skewness = 0.007545975, coef. of k urtosis = 3.64967, Filliben correlation coefficient = 0.997681

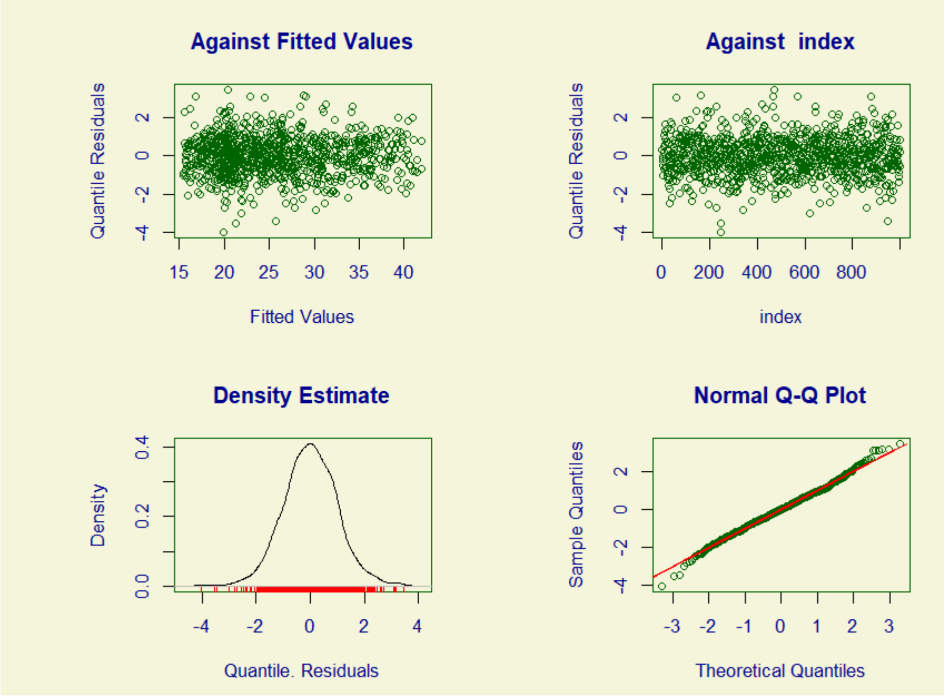


Figure14

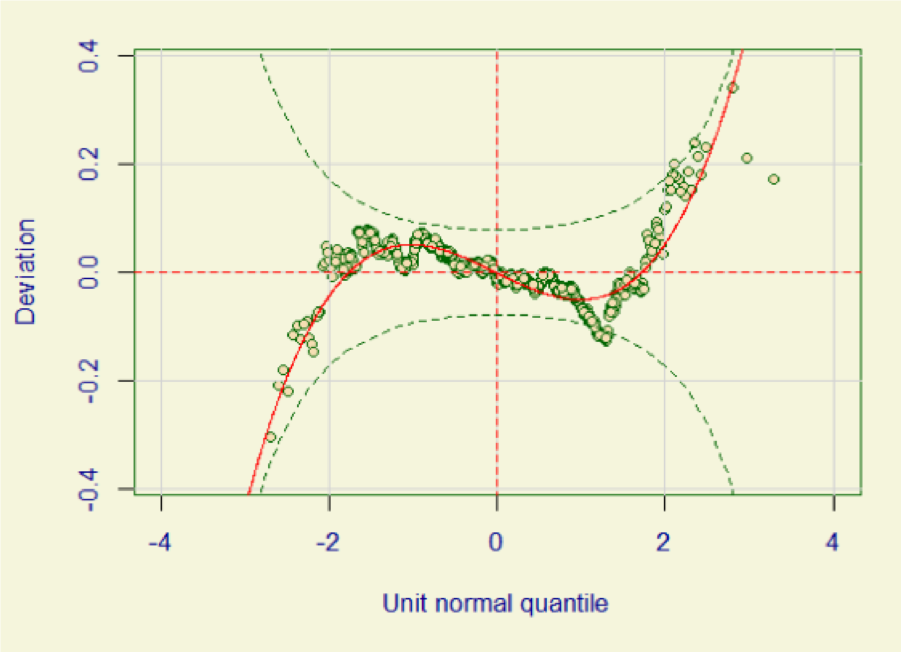


Figure15

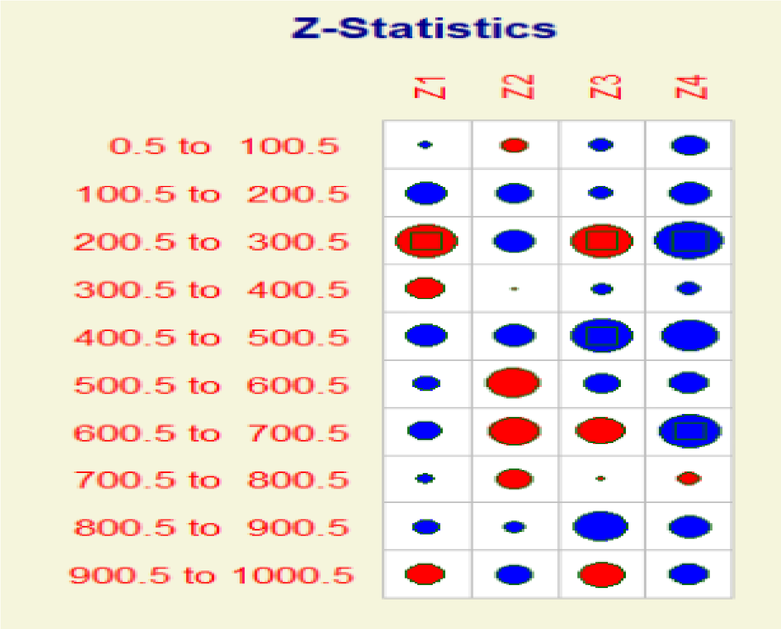


Figure16

- Residuals for gbct :

Here from figure 17, 18, 19 we can see that the residual for the gbct: Summary of the Quantile Residuals, mean = 4.423799e-05, variance = 1.001656, coef. of skewness = -0.005287235, coef. of kurtosis = 2.98554, Filliben correlation coefficient = 0.9996087

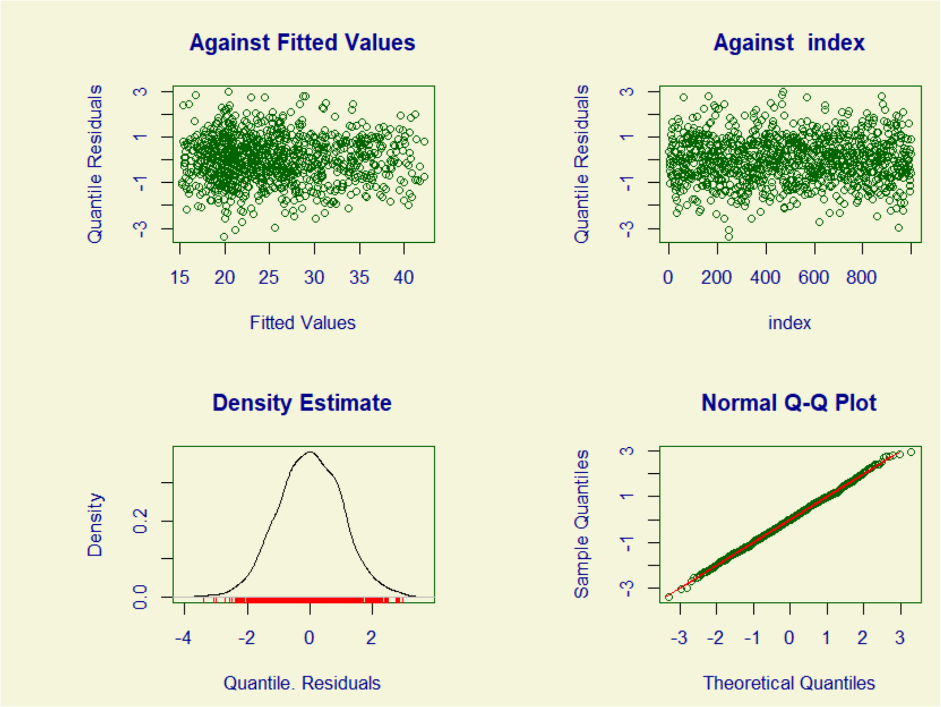


Figure17

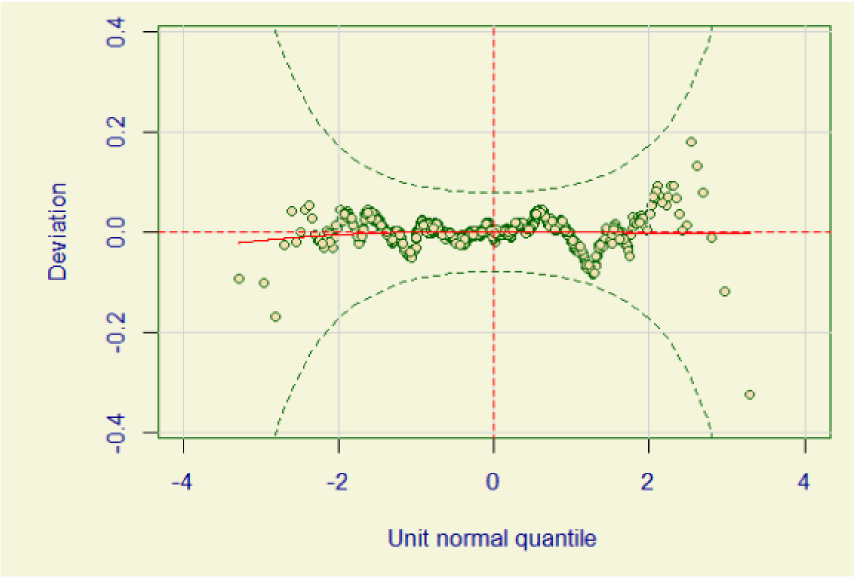


Figure18

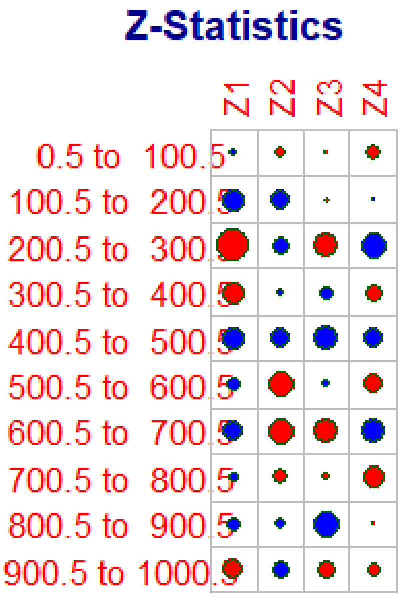


Figure19

- Residuals for bcpe:

Here from figure 20, 21, 22 we can see that the residual for the bcpe: Summary of the Quantile Residuals, mean = 0.0004990328, variance = 1.000282, coef of skewness = -0.00165348, coef. of kurtosis = 3.102824, Filliben correlation coefficient = 0.9993687

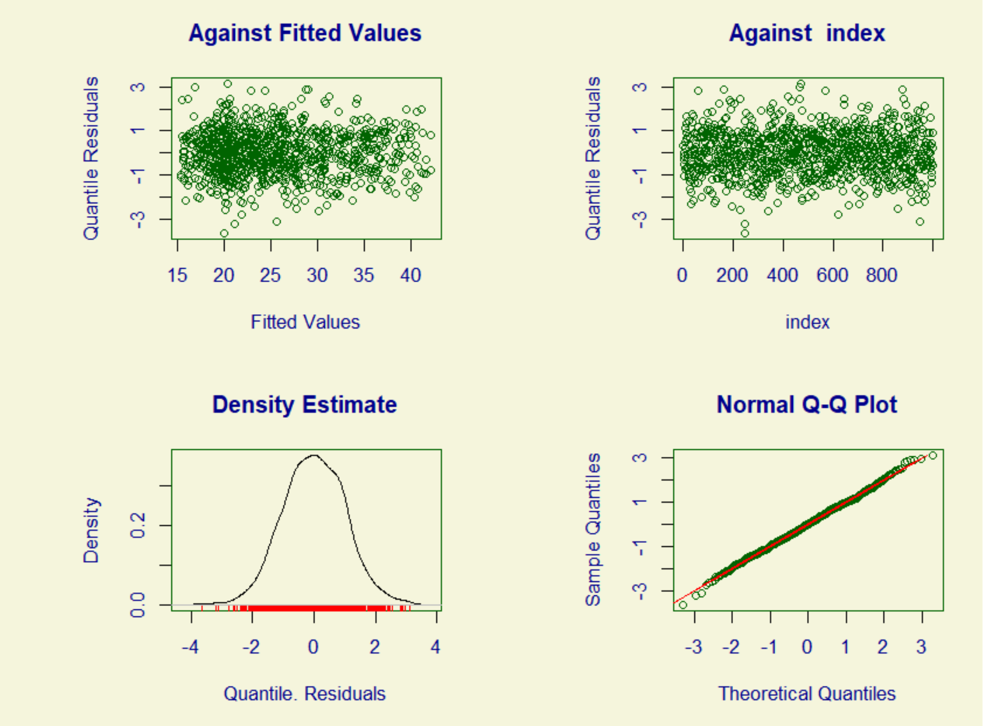


Figure20

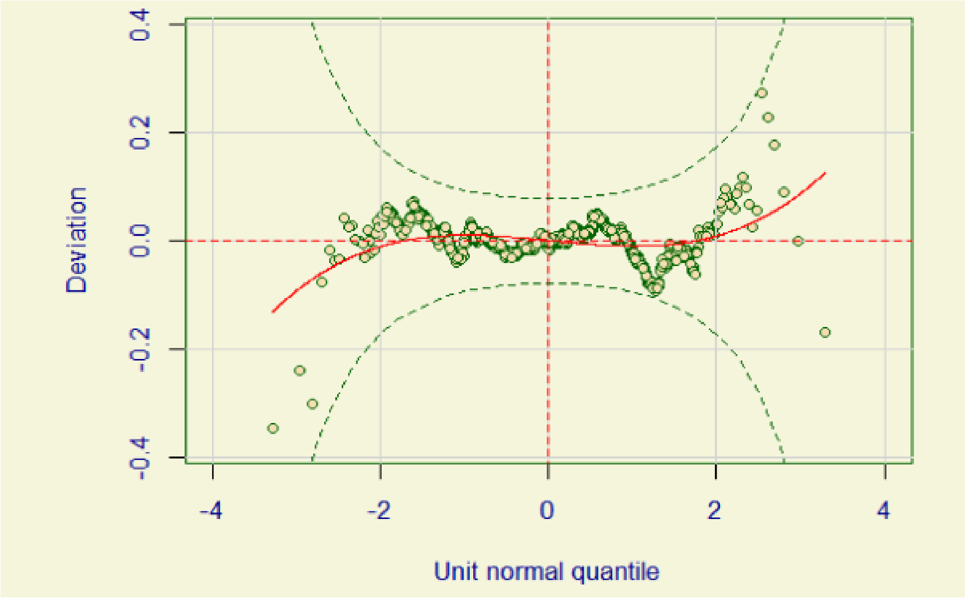


Figure21

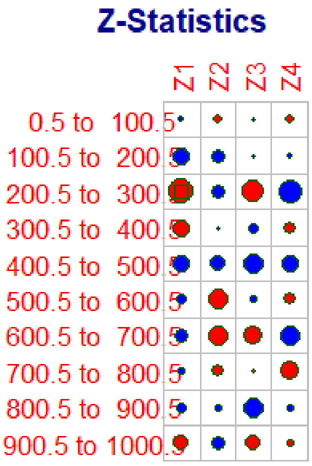


Figure22

3(F)BEST MODEL

Considering the following things, we will consider which model would be the best:

- **Mean:** Choosing the model with the smallest mean value shows that the residuals are centered around zero. In this situation, the model ”gbct” has the mean of 4.42579, followed by ”gbcpe” at 0.0004990328 and ”LMS” at 0.0003442107.

- **Variance:** The model with the lowest variance is preferable since it implies that the residuals have the same dispersion as the original data. In terms of variance, ”LMS” has a variance of 1.001002, ”gbct” has a variance of 1.0001656, and ”gbcpe” has a variance of 1.000282.

- **Skewness and Kurtosis:** Lower skewness and kurtosis values suggest that the residuals are more symmetric and bell-shaped. ”LMS” has a skewness of 0.007545975 and a kurtosis of 3.64967, ”gbct” has a skewness of -0.005287 and a kurtosis of 2.98554, while ”gbcpe” has a skewness of -0.00165348 and a kurtosis of 3.102824.

- **Filliben Correlation Coefficient:** A greater Filliben correlation coefficient indicates that the residuals and original data are more closely related. The Filliben correlation coefficient for model ”gbcpe” is 0.9993687, followed by ”gbct” with a coefficient of 0.999680 and ”LMS” with a coefficient of 0.9987681.

Based on these factors, the model ”LMS” looks to be the best fit of the three alternatives. It has a low mean and variance, low skewness and kurtosis values, and a high Filliben correlation coefficient. It is vital to highlight, however, that the final choice should also take into account the unique criteria and assumptions of the study being performed.

4. THIRD DATA SET (STUDENTS’ DATA)

The third dataset is taken from the London data store. The data set is about the property prices in England and Wales.

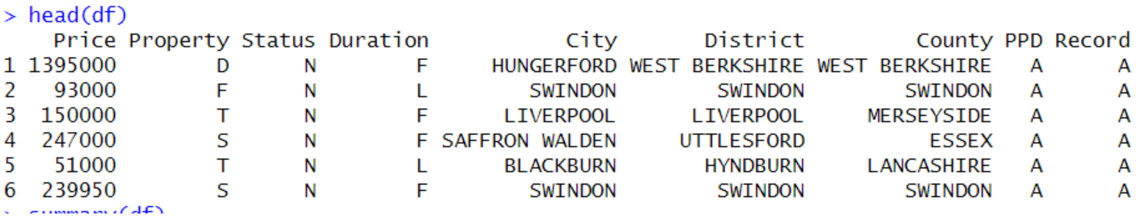
4(A) EXPLAIN WHY YOU COLLECTED THE DATA AND WHAT IS THE QUESTION YOU ARE TRYING TO ANSWER.

Rising in the property prices, it is really necessary to know how the prices of property are in market. Without any idea of the price range we might not get the right house for the budget we have. The prices in uk are gradually increasing .It is good to know the market price of the property in particular areas to make the right investment .

As the economical issues are more in Uk, analysis the price of the properties can help us predict the rise or fall of prices in so and so time period. This analysis help real-estates business run better and also helps common people to invest their capital properly. The target variable in this data is column Price. The question is to predict the price for all other respective variable.

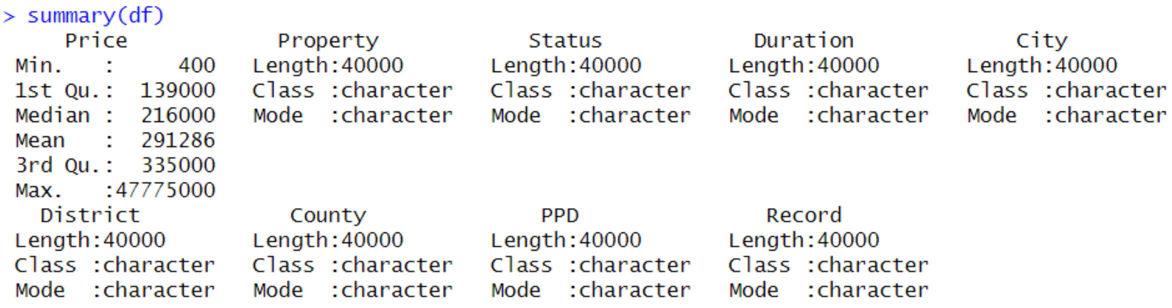
4(B) GIVE A PRELIMINARY ANALYSIS ON THE COLLECTED DATA AND COMMENT ON THE RELIABILITY OF THE DATA.

We got the dataset from the Kaggle. The dataset consist of different variables, they are price, property, status, duration, city, district, county, ppd and records. We can take the target variable as price and using the machine learning models we can predict the prices of the houses according to the type of the property, the place where houses has more demand, the county and district. To view the and know the data we use head(df), summary(df).



```
> head(df)
    Price Property Status Duration          City      District        County PPD Record
1 1395000        D      N        F    HUNGERFORD WEST BERKSHIRE WEST BERKSHIRE   A      A
2   93000        F      N        L       SWINDON       SWINDON       SWINDON   A      A
3  150000        T      N        F     LIVERPOOL     LIVERPOOL    MERSEYSIDE   A      A
4  247000        S      N        F SAFFRON WALDEN   UTTLESFORD         ESSEX   A      A
5   51000        T      N        L     BLACKBURN      HYNDBURN    LANCASHIRE   A      A
6  239950        S      N        F       SWINDON       SWINDON       SWINDON   A      A
```

Figure23



```
> summary(df)
     Price              Property            Status            Duration             City          
 Min.   :     400   Length:40000       Length:40000       Length:40000       Length:40000      
 1st Qu.:  139000   Class :character   Class :character   Class :character   Class :character  
 Median :  216000   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :  291286                                                                              
 3rd Qu.:  335000                                                                              
 Max.   :47775000                                                                              
   District            County              PPD               Record         
 Length:40000       Length:40000       Length:40000       Length:40000      
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

Figure24

Our goal is to predict to potential purchasers regarding the aspects that influence.

In our dataset, there are 8 elements and one is the main target variable which is ”Price”. So, after loading all the packages and loading the dataset using gamlss in RStudio, we saw some outputs or visualize the initial stage for now. We check for the null values in the data taken.

4(C) USE AN APPROPRIATE MODEL(S) TO FIT THE DATA.

The models that are used to fit the data is NO is normal distribution, GA is gamma distribution, Lognormal Distribution, GG is gengamma distribution and BCT is Boxcot t distribution are the different distribution used to run on the data taken. First the data is loaded using read.csv because it is a csv file and then created a sample data which is considered as the subset of the data. The sample size of the data is 20000 observations. And the sample data is stored in the subset mydata. Histogram is plotted using ggplot to show the price and the frequency of the subset data. And then fit all the distributions of the we considered. We have considered the normal distribution, Lognormal Distribution, Gamma Distribution and gengamma distribution and the boxcot t distribution. After fitting the model we calculate the AIC of al the models and consider the smallest AIC value among the used distributions.

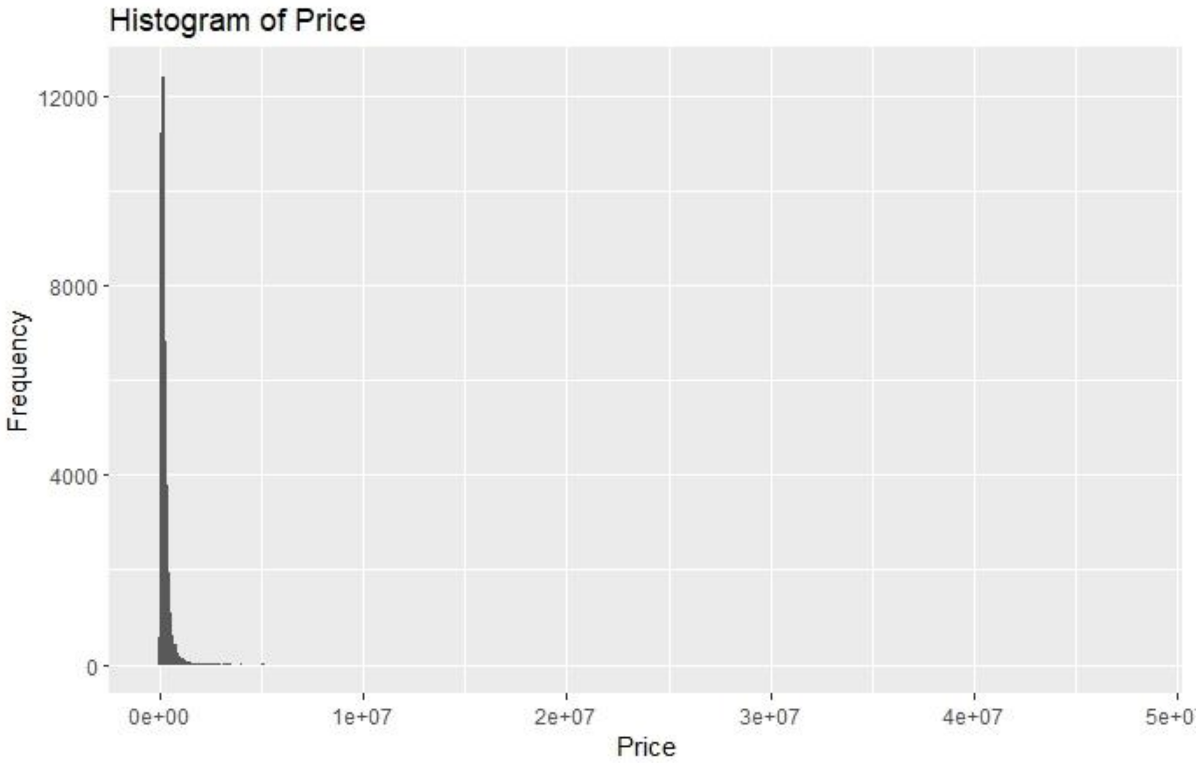


Fig4 Price vs Frequency

4(D) COMMENT ON HOW YOU SELECTED YOUR FINAL MODEL INCLUDING DIAGNOSTICS.

Once the models are fitted we check the AIC and global deviance of each and every model. When we compare each and every model the AIC values varies. As we have seen before a lower AIC value indicates a more accurate model fit. A negative AIC value is possible and does not invalidate its usefulness as a model selection criterion. It is important to note, however, that the absolute value of AIC is unimportant; rather, the relative difference in AIC values between different models is. As a result, when comparing different models, the one with the lowest AIC value, whether negative or positive, is generally preferred. Here are the AIC values of all the 5 model we used:

```
  Distribution      AIC
1          Normal 583417.8
2      Log-normal 530458.1
3           Gamma 536041.5
4 Generalized gamma 530102.3
```

5	Box-Cox t 529173.9

From this we can say that the Box Cox t is the best model as the AIC of the Box Cox t model is the smallest value. To check the diagnostics of the best model BC T, we plot the residual vs fitted, normal q-q plot, scale location plot and cook distance plot.

1)The values after plotting Residual vs fitted

```
                   Summary of the Quantile Residuals
                           mean   =  0.0009023363
                       variance   =  0.9999997
               coef. of skewness  =  -0.01029234
               coef. of kurtosis  =  3.060732
Filliben correlation coefficient  =  0.9996553
```

2)The values after plotting normal Q-qplot

```
                   Summary of the Quantile Residuals
                           mean   =  0.0009023363
                       variance   =  0.9999997
               coef. of skewness  =  -0.01029234
               coef. of kurtosis  =  3.060732
Filliben correlation coefficient  =  0.9996553
```

3)The values after plotting scale location plot

```
                        Summary of the Quantile Residuals
                           mean   =  0.0009023363
                       variance   =  0.9999997
               coef. of skewness  =  -0.01029234
               coef. of kurtosis  =  3.060732
Filliben correlation coefficient  =  0.9996553
```

4) The values after plotting cook distance plot

```
                       Summary of the Quantile Residuals
                           mean   =  0.0009023363
                       variance   =  0.9999997
               coef. of skewness  =  -0.01029234
               coef. of kurtosis  =  3.060732
Filliben correlation coefficient  =  0.9996553
```

4(E) SHOW HOW YOU WILL USE THE MODEL FOR PREDICTION.

We have predicted the values for the response variable which is price and the predicted values of the response variable are

```
> gamlss_predictions
 [1] 187601.9 102078.0 198356.1 182808.5 334810.1 255465.5 198356.1 29475
2.1 130211.8 197827.6
```

5. CONCLUSIONS

The Boxcox t is the model which is the best fitted model to predict the price of the properties in England and Wales. The predicted values are the price that are predicted for the places

6. REFERENCES

[1] A.M. Fredriks, S. van Buuren, R.J.F. Burgmeijer, J.F. Meulmeester, R.J. Beuker, E. Brugman, M.J. Roede, S.P. Verloove-Vanhorick, and J. M. Wit. Continuing positive secular change in The Netherlands, 1955-1997. Pediatric Research, 47:316–323, 2000.

[2] Robert A Rigby, Mikis D Stasinopoulos, Gillian Z Heller, and Fernanda De Bastiani. Distributions for modeling location, scale, and shape: Using GAMLSS in R. CRC press, 2019.

[3] D. D. Cohen, C. Voss, M.J.D. Taylor, D.M. Stasinopoulos, A. Delextrat, and G.R.H. Sandercock. Handgrip strength in English schoolchildren. Acta Paediatrica, 99:1065–1072, 2010.

7. APPENDIX

First dataset code:

```
library(gamlss)

data(dbbmi)

head(dbbmi)

summary(dbbmi)

old<-20

da<- with(dbbmi, subset(dbbmi, age>old+1))

bmi20<-da$bmi

#plotting BMI data in histogram

hist(bmi20)

library(MASS)

truehist(bmi20, nbins=30)

install.packages("gamlss.ggplots")

library("gamlss.ggplots")
```

```
gamlss.ggplots:::y_hist(bmi20)
#Fitting normal distribution
fit_norm <- gamlss(bmi20 ~ 1, family = NO)
#Fitting gamma distribution
fit_gamma <- gamlss(bmi20 ~ 1, family = GA)
#Fitting lognormal distribution
fit_lognormal <- gamlss(log(bmi20) ~ 1, family = NO)
plot(fit_lognormal)
#Fitting Inverse gaussian
fit_inversegaussian <- gamlss(bmi20 ~ 1, family = IG)
#compare the AIC values
AIC.df <- data.frame(Distribution = c("Normal", "gamma" ,"Log-normal", "Inverse gamma"), AIC = c(AIC(fit_norm), AIC(fit_gamma), AIC(fit_lognormal), AIC(fit_inversegaussian)))
AIC.df
# Select the distribution with the lowest AIC
best.dist <- AIC.df$Distribution[which.min(AIC.df$AIC)]
best.dist
```

Second Dataset code:

```
library(gamlss)
data(grip)
head(grip)
summary(grip)
sum(is.na(grip))
set.seed(1058)
index<-sample(3766, 1000)
mydata<-grip[index, ]
dim(mydata)
```

```
head(mydata)

library(ggplot2)

ggplot(mydata, aes(x=age, y=grip))+ geom_point()+ labs(x= "Age", y = "Grip")

#fit LMS model using bccg distribution

lms_model <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), data = mydata, family = BCCG)

#get degress of  freedom for smoothing term

edf(lms_model)

#fit BCT

gbct <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), data = mydata, family = BCT, start.from = lms_model)

# get degrees of freedom for smoothing term

edf(gbct)

# fit BCPE

gbcpe <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), data = mydata, family = BCPE, start.from = lms_model)

# get degeers of freedom

edf(gbcpe)

GAIC(lms_model, gbct, gbcpe)

fittedPlot(lms_model,gbcpe, gbct, x=mydata$age)

centiles_Lms <- centiles(lms_model, var = mydata$grip, xvalues = seq(20, 100, by = 5))

centiles_gbct <- centiles(gbct, var = mydata$grip, xvalues = seq(20, 100, by = 5))

centiles_gbcpe <- centiles(gbcpe, var = mydata$grip, xvalues = seq(20, 100, by = 5))

#Plotting residuals for lms_model

plot(lms_model)

wp(lms_model)

Q.stats(lms_model)
```

```
#Plotting residuals for gbct model
plot(gbct)
wp(gbct)
Q.stats(gbct)
#Plotting residuals for gbcpe model
plot(gbcpe)
wp(gbcpe)
Q.stats(gbcpe)
```

Third Dataset code:

```
library(gamlss)
library(ggplot2)
# Loading the dataset
df <- read.csv("C:/Users/Alekya/Desktop/Stats Modelling/CleanData .csv")
head(df)
summary(df)
#checking for null values
sum(is.na(df))
# Selecting a sample from the dataset
set.seed(1058)
sample_size <- 20000
mydata <- df[sample(1:nrow(df), sample_size), ]
# Encoding categorical columns
for (col in names(mydata)) {
  if (col != 'Price') {
    mydata[[col]] <- as.numeric(factor(mydata[[col]]))
  }
}
```

```
# Histogram of Price
ggplot(df, aes(x = Price)) +
  geom_histogram(binwidth = 100000) +
  labs(x = "Price", y = "Frequency") +
  ggtitle("Histogram of Price")
# Fitting the GAMLSS models
fit.norm <- gamlss(Price ~ ., data = mydata, family = NO())
fit.lnorm <- gamlss(Price ~ ., data = mydata, family = LOGNO())
fit.gamma <- gamlss(Price ~ ., data = mydata, family = GA())
fit.gengamma <- gamlss(Price ~ ., data = mydata, family = "GG")
fit.bct <- gamlss(Price ~ ., data = mydata, family = BCT())
# Compare AIC values for the different distributions
AIC.df <- data.frame(Distribution = c("Normal", "Log-normal", "Gamma",
"Generalized gamma", "Box-Cox t"),
 AIC = c(AIC(fit.norm), AIC(fit.lnorm), AIC(fit.gamma), AIC(fit.gengamma),
AIC(fit.bct)))
AIC.df
# Select the distribution with the lowest AIC
best.dist <- AIC.df$Distribution[which.min(AIC.df$AIC)]
best.dist

# Checking model diagnostics
par(mfrow = c(2, 2))
plot(fit.bct, which = 1)  # Residuals vs. Fitted
plot(fit.bct, which = 2)  # Normal Q-Q plot
plot(fit.bct, which = 3)  # Scale-Location plot (Square root of standardized
residuals vs. Fitted values)
plot(fit.bct, which = 4)  # Cook's distance plot
```

```
new_data <- mydata[1:10, ]
gamlss_predictions <- predict(fit.bct, newdata = new_data, type = "response")
gamlss_predictions
```