**Statistical Modeling Of Real Data Sets Using GAMLSS in R**

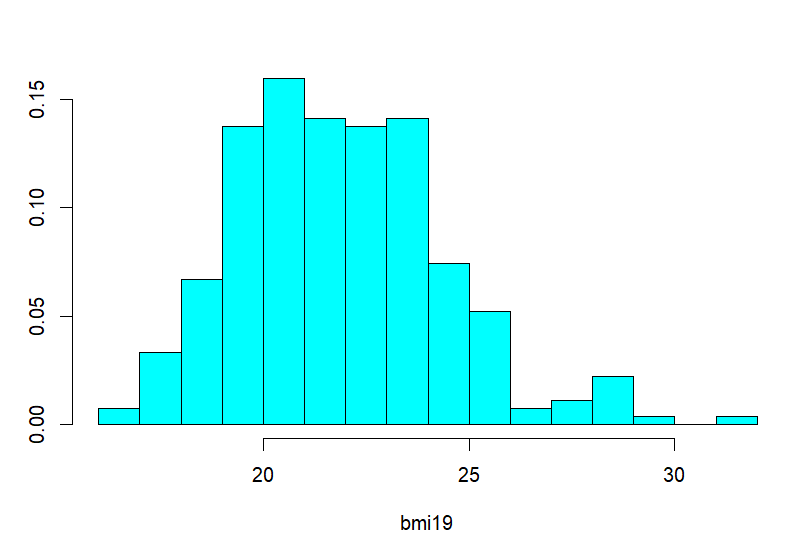
1. **Introduction:**

This paper illustrates the process of statistical modeling used to predict the response variable depending upon it’s distribution with the explanatory variables. It will assist statisticians, programmers, data modelers and even business stakeholders who wish to simplify complex business scenarios. Three different data sets are included in this paper where the target variable is modeled using the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) [1]. This model uses regression for data modeling. This paper primarily highlights the use of dependence of distributions of the response variable on the explanatory variables for it’s forecasting. There were times when the normal and the Poisson distributions were the only type of distributions used for modeling continuous and discrete random response variables but when tested under real world use cases, they failed to deliver accurate results and provided improper fits. Hence, GAMLSS model was developed to solve these problems where actual, real-world data can be tested and the response variable can be predicted accurately as GAMLSS permits use of mixed distributions to do the same. Three data sets are used in this paper to model the respective response variables using various distributions.

1. **First Data Set (fitting different distributions to the data):**

The first data set contains BMI of Dutch boys in years. The target is to find the best possible distribution for the BMI of a randomly selected age ranging from zero to twenty-two which is selected as nineteen in this case.

The data is imported from gamlss.data package with name dbbmi. A data frame subset is then created which contains bmi data only for age nineteen to twenty only. Initially, a histogram is plotted to do some preliminary data analysis. After plotting the histogram, data is observed to be of continuous type as shown below.

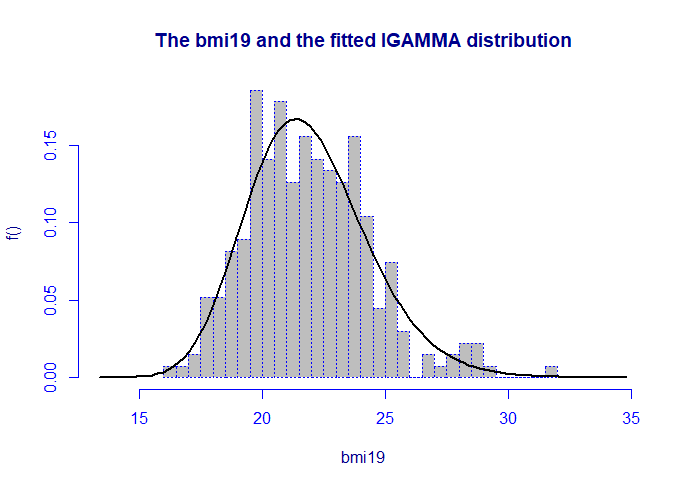


**Fig1. Histogram of BMI of boys aged between 19 to 20**

Selection of distribution is a crucial task here. As seen above, the data is continuous and spread over a range of 0 to infinity. Hence, various distributions can be fit to the response variable accordingly. Generally, the fitDist() method is used to fit the various parametric distributions to the response variable if it lies in the range 0 to infinity[2]. The same method was applied here. Two different criteria were used for fitting the distributions using fitDist() which include GAIC and the SBC. The GAIC uses the parameter value (k=2 in this case was used) and the SBC uses the logarithmic value of k to compare various distributions and returns the best fit accordingly.

Both the GAIC and the SBC showed that the IGAMMA distribution was the best fit for the response variable in this case. **The distributions fitted to the data frame subset include IGAMMA, GG, BCCG, BCCGo, BCPE, BCPEo, GIG, GB2, BCT, BCTo, LOGNO, LOGNO2, IG, WEI, WEI2, WEI3, EXP, PARETO2, PARETO2o, GT, SST, ST1, ST2, ST3, ST4, ST5, SEP1, SEP2, SEP3, SEP4, JSU, JSUo, EGB2, SHASH, SHASHo, SHASHo2, exGAUS, SN1, SN2, PE, PE2, TF, TF2, LO, NET, GU, RG, and NO.**

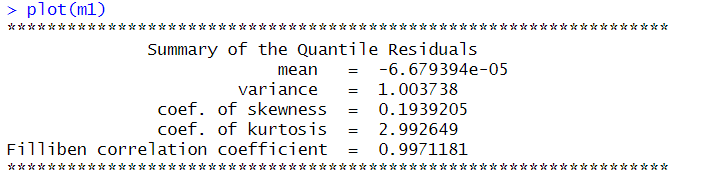
As seen earlier both the GAIC and SBC criteria selected the IGAMMA as the best distribution for the given data set and hence, was chosen as the distribution to be fit for statistical modeling. The fitted distribution was then plotted which was as shown below.

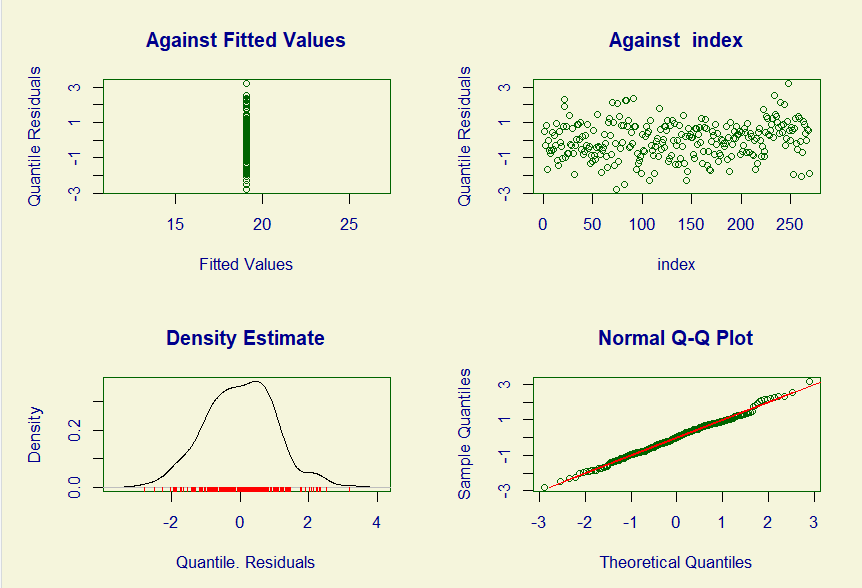


**Fig 2. The fitted IGAMMA distribution**

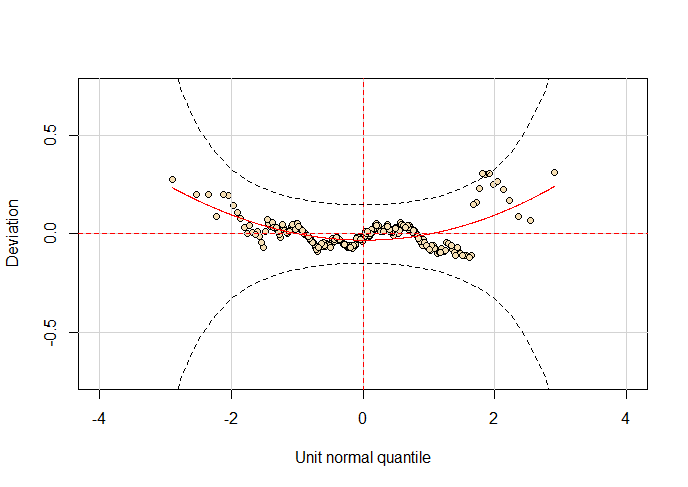
The IGAMMA or the Inverse Gamma distribution is considered appropriate when the given data is positively skewed [2] which is the same in the given data set (as the body mass index of a boy tends to be more on the positive side). It is two parameter (*µ, σ*) distribution which was found to be the best fit for the given data set.

The fitted parameter estimates of the final fitted model are then produced using the summary () function. The fitted parameter values of the final fitted distribution are as shown below.



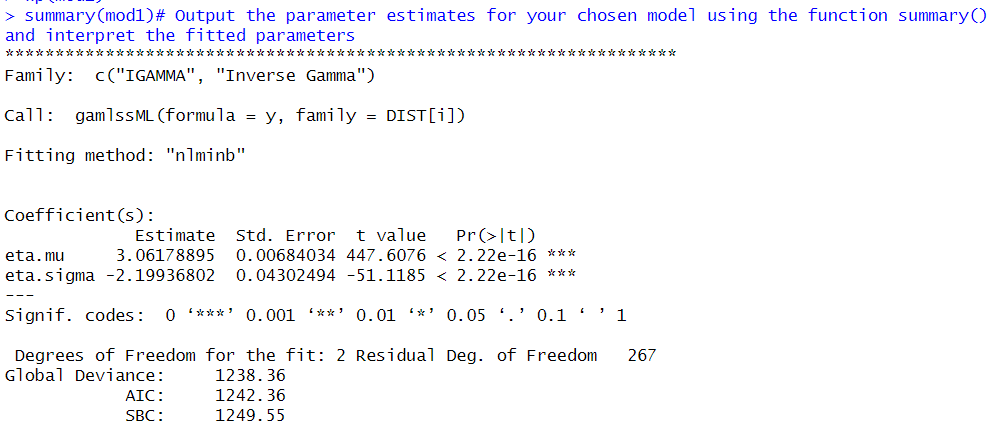


**Fig 3. Quantile residuals for the fitted distribution**



**Fig 4. Worm plot of fitted IGAMMA distribution**

Here, the quantile residuals for the final fitted IGAMMA distribution are shown. Both the Q-Q plots and the worm plots show that the IGAMMA is the best fit distribution for the given data. The adequacy of the IGAMMA distribution is given by over 95% of the deviations in the worm plot lying within the dashed (approximate 95%) confidence bands These plots are more than enough to prove that IGAMMA is the best fit for the given data set.



**Fig 5. Parameter estimates of the final fitted distribution**

Here the parameter estimates of the final fitted IGAMMA distribution are shown**.** The two parameters of the IGAMMA distribution include *µ and σ.* These parameters are fitted using default link functions [2], namely, “log” and “log” respectively, the coefficients shown: *ˆηµ* = 3.06178895, and *ˆησ* =-2.19936802 correspond to fitted parameters *ˆµ* =exp(*ˆηµ)* = 21.3657452435 and *ˆσ* = exp(*ˆησ*) = 0.110873205874 respectively. The degree of freedom for the fit is 2 which indicates IGAMMA’s parameters degrees of freedom for the data frame subset. Also, the global deviance shown by the IGAMMA distribution hence it justifies itself as the best fit for the given data set.

1. **Second Data Set (Centile Estimation):**

Here, a data set of 1000 boys selected uniquely from the original 3766 boys of an English school was used to analyse the handgrip of the boys given their age and gender [3]. The age and gender were available and based on that the idea was to create centile curves for the handgrip of English boys.

Initially, the gamlss package was downloaded. Then the unique seed number was used to analyse 1000 boys’ data using set.seed (1051) command. As self-explanatory, the unique seed number in this case was selected as 1051. Then a data frame subset was created which included the age and grip of 1000 english boys out of the total 3766 boys.

Now for some preliminary analysis of the data, the grip of the boys was plotted against their age. It was crystal clear from the plot that the age of the boys needed no power transformation. It is a technique used to enhance the prediction of the smoothing functions [2]. In this case, it would have been needed if spells of sharp growth in age occur for low values of grip, but the distribution seems to be pretty even so, power transformation of age was not needed.

The two methods generally used for centile estimation with one explanatory variable are [2]: (a) the LMS method and its extensions [4,5,6] and (b) the quantile regression method [7]. Here, the first method will be implemented.

Initially, the BCCG (The Box-Cox Cole and Green distribution) [2] distribution was fit for grip in the data set. The BCCG distribution gives assurance that the LMS method is based on an appropriate distribution. Here the age requires smoothing hence, P-splines function pb(age) is used for the smoothing of age. The degrees of freedom of all there smoothing

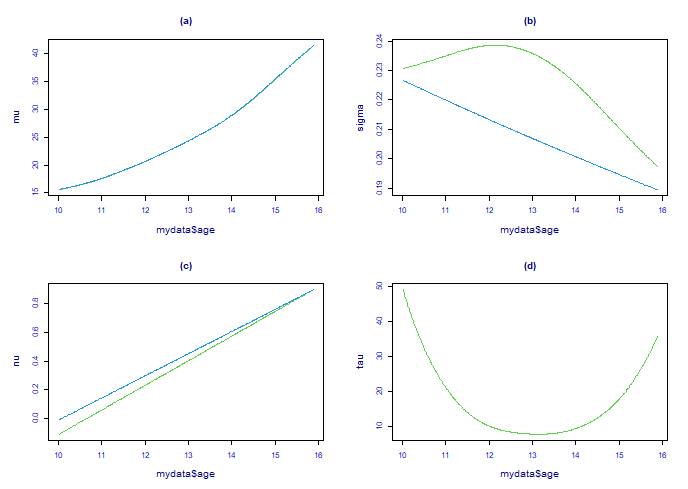
parameters were obtained using edfAll() function. The degrees of freedom used for smoothing in the model were as follows: *µ=* 4.911053*, σ=* 2.858506 and *ν=* 2.000121.

The degrees of freedom tend to show whether the distribution is leptokurtic or platykurtic.[2]. As the degree of freedom tends to zero, the distribution becomes more leptokurtic and as it tends to infinity, it becomes more normal.

The fitted values of the LMS model i.e., BCCG distribution used earlier are now used as the starting values for fitting the BCT (Box-Cox t) and the BCPE (Box-Cox power exponential) distributions [2]. The BCT distribution has four smoothing parameters unlike the BCCG distribution and is considered particularly useful when the response variable is not close to zero [2]. The BCPE is also a very flexible distribution having four smoothing parameter and very useful when the response variable is not close to zero [2]. The effective degrees of freedom for the BCT and the BCPE models were found out to be 12.08956 and 12.21208. The effective degrees of freedom were calculated as the summation of all the four degrees of freedom of the respective smoothing paramaters which were in turn, was obtained using edfAll() function as used earlier. The degrees of freedom of both these models show pretty less variation and are almost the same as evident from their values, whereas the BCCG model had an effective degree of freedom of 9.76968 which is much less as compared to BCT and BCPE models.

The three models fitted to the data are now compared using the generalised Akaike information criteria i.e., GAIC [2]. It automatically provides the best choice distribution on passing the fitted models as parameters. According to the GAIC criteria, the BCT distribution was the most appropriate of the three fitted models. The chooseDist() function could have also served the purpose of model comparison here.The fitted parameters of the fitted models were then plotted using fitted.Plot() function.

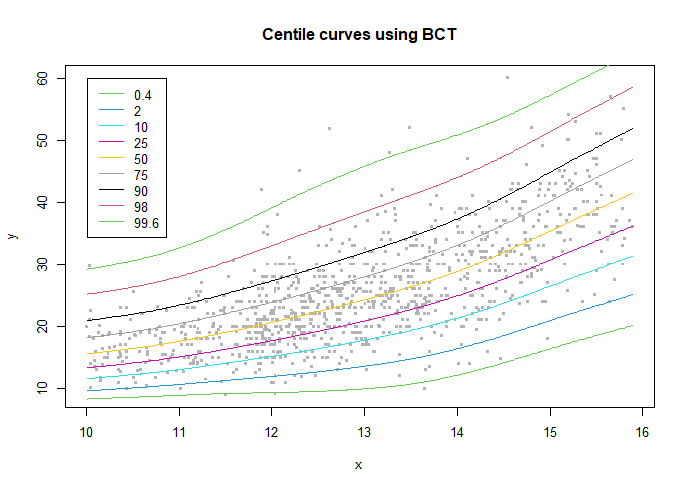
The plot thus obtained was as shown below.



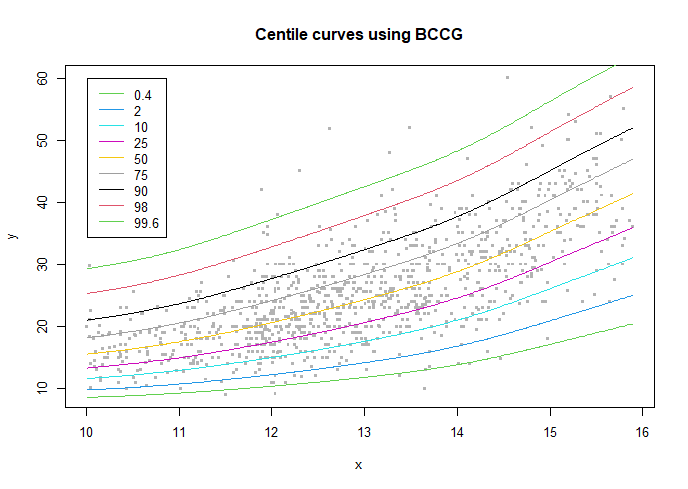
**Fig 6. Plot** **showing fitted parameters of the fitted BCCG and BCT distributions**

The centile curves of the fitted BCCG and BCT models were then obtained using the

centile () method. The centile curves obtained for the fitted distributions were as follows.

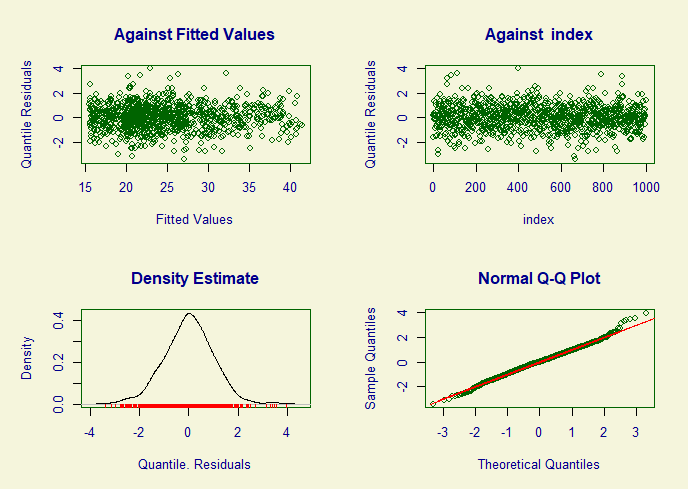


**Fig 7. Centile curves using BCT distribution**

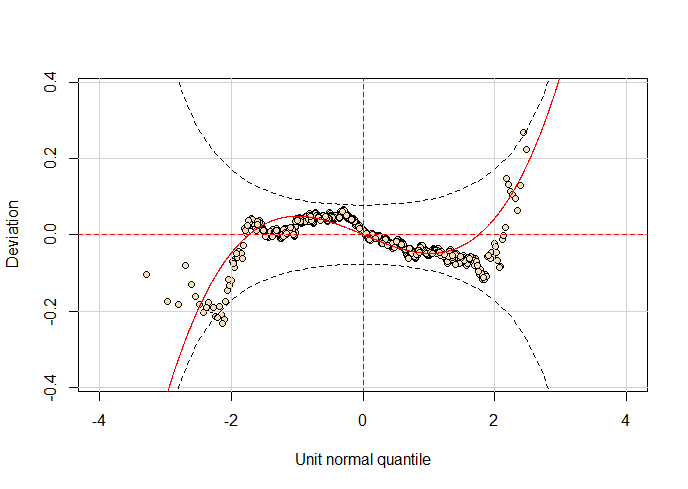
 **Fig 8. Centile curves using BCCG distribution**

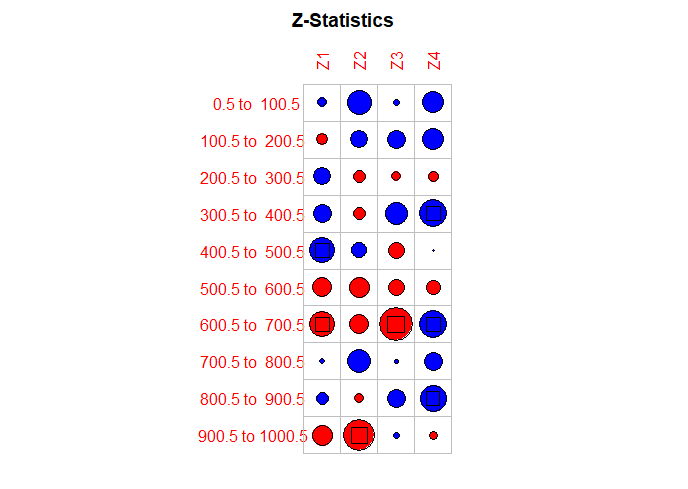
From the above figures, it is clearly evident that both the models show pretty identical centile curves except for the percentile of cases below 0.4 and 99.6 differ slightly. But from the dataset under consideration, it can be inferred that lesser the age of the boy, lesser is the handgrip.

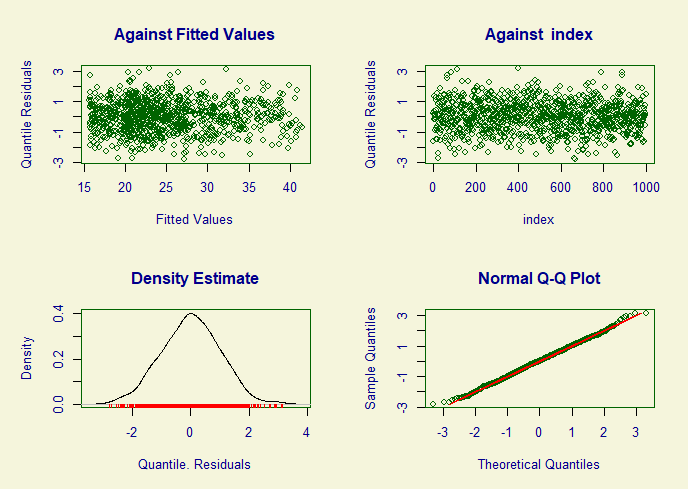
Another method for selecting the best possible fit is the residual investigation of the fitted models. The residuals of the fitted models are obtained using plot(), wp() (worm plot) and Q.stats() (Q statistics). The corresponding plots obtained were as shown below.



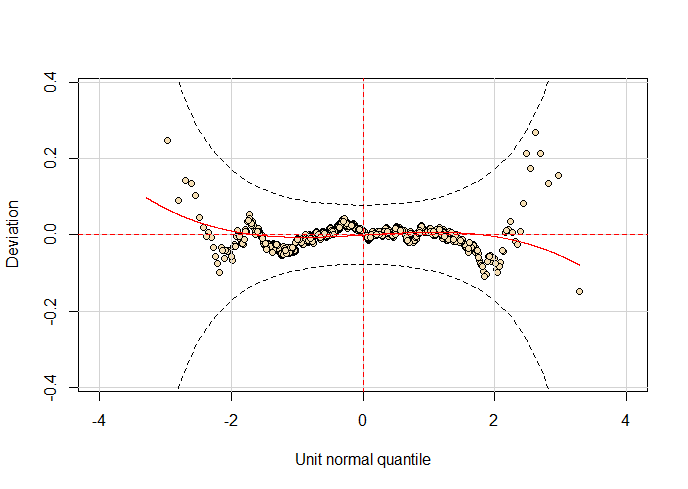
**Fig 9. Quantile Residuals for the BCCG distribution**

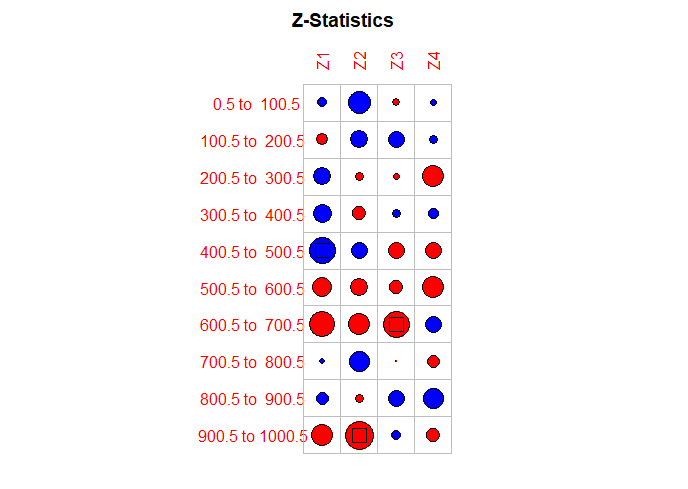
**Fig 10. Worm Plot for the fitted BCCG distribution**

**Fig 11. Z-Statistics for the fitted BCCG distribution**



**Fig 12. Quantile Residuals for the BCT distribution**

 **Fig 13. Worm Plot for the fitted BCT distribution**

 **Fig 14. Z-Statistics for the fitted BCT distribution**

Now, the most crucial and final step is the selection of the model to be fit to the data set using all the above criteria. Initially, the GAIC criteria selected BCT as the best model fit.Then, centile curves were plotted for BCT and BCCG distributions, but these showed

little variance and hence, inconclusive evidence was there for model selection. Hence,

quantile residuals were investigated which cleared the whole picture finally. The plots of density estimates and the normal Q-Q plots of BCT distribution are way better than those of BCCG distribution. The variance between the estimates and the theatrical values is

much less in BCT distribution than those in BCCG distribution. Also, the worm plot of

the BCT distribution looks much better than that of the BCCG distribution as the

deviation in BCT distribution is much lesser. Also, the Z-statistics of the BCT

distribution appear to be much better than those of BCCG distribution. Thus, it can be

concluded safely, that the BCT distribution is the best fit for the given data set.

**3.Third Data set (Student’s data):**

The last two data sets were readily available to us but here for the third and final set of

analysis, the data set needed to be collected by students itself. Hence, a reliable and

prominent platform like Kaggle was used to serve the purpose. The data that was

collected, contained information about the house sales in King County, Washington State, USA. The data set contained historic data about the house sales from May 2014 to May 2015. The data set consisted of 21 variables and 21613 observations.

The link to the data set is: <https://raw.githubusercontent.com/Shreyas3108/house-price-prediction/master/kc_house_data.csv>

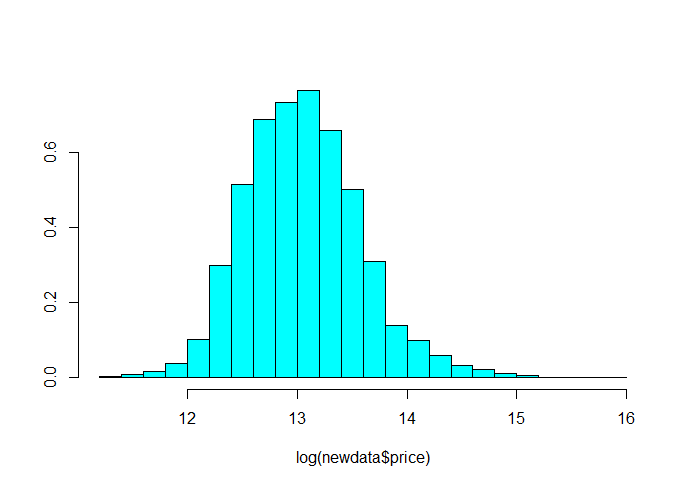
The data set was checked by our tutor and was found appropriate to use for the statistical analysis in this case.

However, the source of the data was not mentioned by the author. Hence, there may be some possible pitfalls in the collection of the data set which are unknown.

The business problem was pretty straight forward in this case- to analyse the dataset parameters related to housing and predict house prices in U.S.A. Post Covid era has certainly changed the style in which the world works, as more and more organisations are resorting to working remotely. Hence, the conception that only real estates in metropolitan cities are in demand is just a myth nowadays. People nowadays, prefer remote and silent locations to stay peacefully and away from all the hustle and bustle of large cities. Pollution too, has played a major role in people’s selection of location to live nowadays. Hence, remote and smaller houses too are equally in good demand in the real estate market. Ask a home buyer to describe their dream house and they probably won`t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this paper proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. Anyone involved in the real estate industry has been on the edge of their seats in recent times. The two sides of the coin as far as the real estate industry is concerned are “Supply” and “Demand”. House prices can be determined only by taking into consideration the fundamental housing needs is a myth. Thus, people working on base wages are then under immense pressure to find a property that suits their budget. Affordable housing thus, can be viewed as the ratio between the cost of buying a property and the will to purchase it.

The individual seed number was set initially to maintain uniqueness of analysis. This was followed by randomly selecting only 80% of the available data set to be used for analysis. This was done using a logical vector “index” which randomly selected 80% values from the data set. A new data set which was the subset of the original data set was then created using “index”.

The preliminary analysis of data included visualising the data first by using the head () method and seeing the initial data points. After that the exploratory plots were plotted for various variables against the target variable “price” in this case. The histogram of the response variable “price” was plotted to visualise it’s distribution. Note that the logarithmic values of the response variable were used for all the plots as the original values were long right skewed and hence to obtain proper plots, logarithmic values were used. The histogram of the response variable “price” was obtained as shown below.



**Fig 15. Histogram of the response variable “price”**

The most important part of any data modeling project is data cleaning. This includes selecting which attributes can actually contribute to the prediction of the response variable.

This phenomenon is also known as **“feature selection”**. Here, in this case, variables like “id”,”date” and “zipcode” were dropped as they had little significance on the response

variable’s outcome. Furthermore, variables “lat” and “long” were also not considered as

the intention was not to spatial analysis in this case. Finally, attributes like “floors”,”

waterfront”,”view”,” yr\_built”,” yr\_renovated”,” sqft\_living15” and “sqft\_lot15” were also dropped from the data set as they logically appeared to make less impact on the

possible outcome of the response variable “price”. Hence, the data set was cleaned and only 8 variables including “price”, “bedrooms”, “bathrooms”, “living”, ”lot”,”condition”, “grade” and “above” were used for the analysis.

Now, the next step is to find a suitable distribution for the response variable. The fitDist() method is the best method to serve the purpose as the response variable here, is

long right skewed and lies on the positive real line (between 0 to infinity). Two criteria

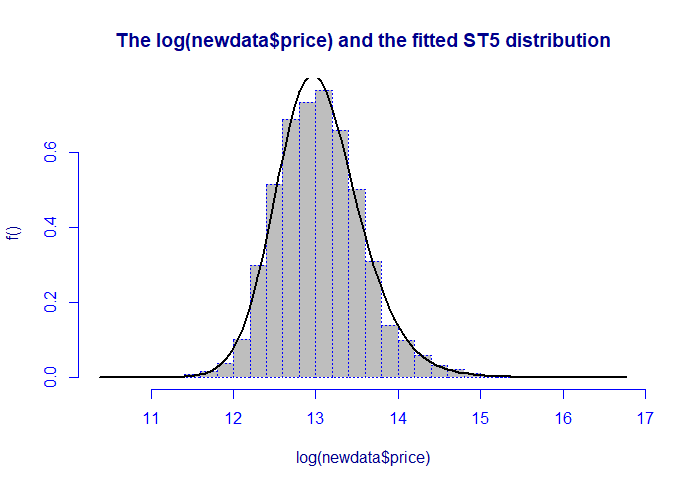
namely, GAIC and the SBC, both were used to find the best suited model for the given

variable. Both the criteria showed that the “ST5” or The skew t type 5 distribution [2],

denoted by ST5(*µ, σ, ν, τ* ) was the best distribution for the response variable. Thus, the

ST5 distribution was fitted to “price” and a histogram of the fitted distribution was

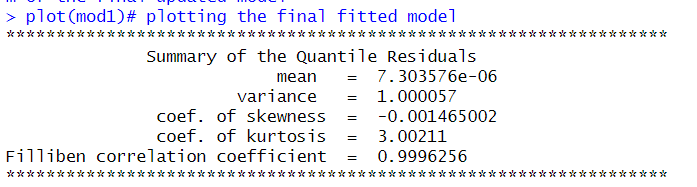
plotted as shown below.

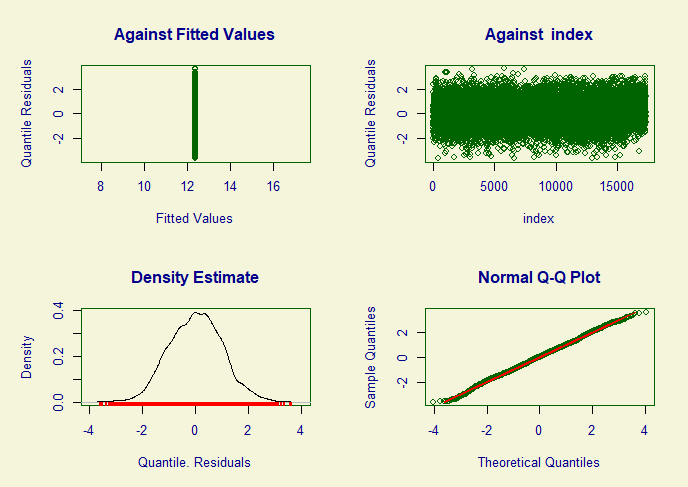
**Fig 16. The ST5 distribution fitted to the response variable**

The quantile residuals’ summary was visualised using the plot () function followed by

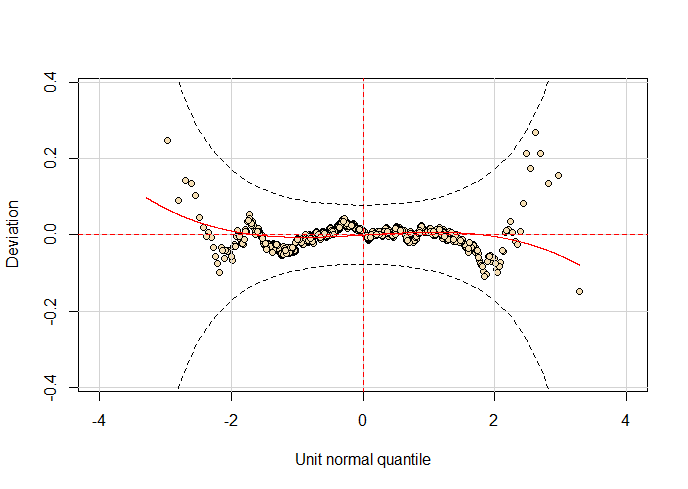
plotting the worm plot using wp () method and finally the model summary was output

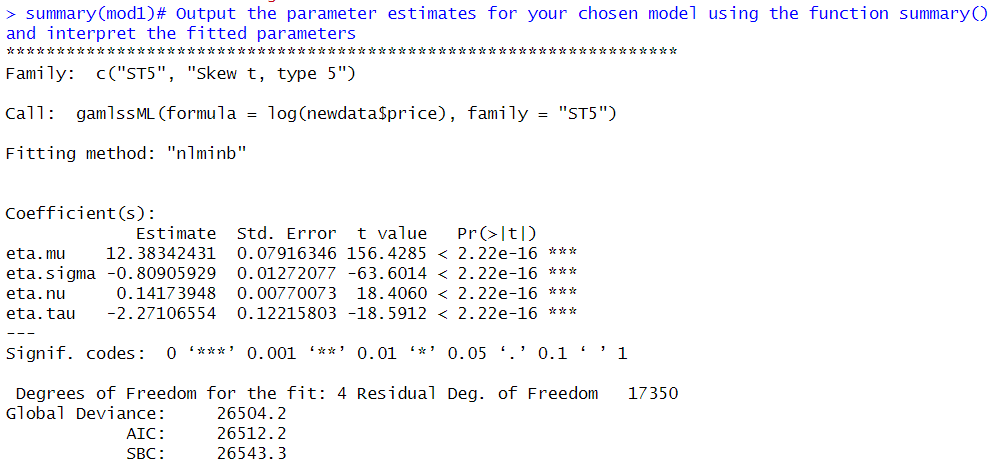
using summary () function. The respective outputs are as shown below.





**Fig 17. Quantile residuals of the final fitted distribution**

**Fig 18. Worm plot for the fitted ST5 distribution**



As seen clearly from the above plots and results ST5 was selected as the best distribution for the given data set. The Q-Q plots and worm plots denote high accuracy as seen above. The Q-Q plot has more than 95% accuracy and same is the case for the worm plot. Hence, the ST5 distribution can be considered appropriate for the given data.

The final step is to use the fitted model for prediction of the house price. This is done using the predict () method. Here, only the first case of the sample output was predicted. The predicted output gives the logarithmic value of the response variable hence, it had to be converted to exponential value using exp () function. It was observed that the price value predicted was 238810 and the actual value was 221990 which were pretty similar.

1. **Conclusion:**

All 3 data sets were analysed using various distribution of the GAMLSS package. Both continuous and discrete data were handled and analysed in this paper. The aim was to study the statistical methods of analysing data sets using different regression models and applying them to the real-world data set to predict the response variables. This paper covers the methodologies used to analyse different real data sets collected from varied sources and perform statistical analysis to use models to solve real world business scenarios. This paper provides valuable insights to business stakeholders, programmers, data analysts and data architects in relation to tackling real world business problems and coming up with effective solutions.

The findings of this paper are like gold dust for anyone related to the property market in U.S.A as it provides exactly the parameters and methodologies needed to predict housing prices in U.S.A. It provides great insights to buyers/sellers who wish to buy/sell properties in U.S.A. It also provides pricing transparency related to housing in U.S.A which in turn would benefit the development of the real estate industry in U.S.A in a great manner.

It is always true that there are two sides of a coin and similar is true in this paper’s case as well. There are many commercial risks to this paper as people dealing with real estates would not wish that their customers have such transparencies related to house pricing as it would reveal their own commissions and the amount of money they make from broking properties in and around U.S.A. The real estate industry is one of the most recently booming industries as more and more people are choosing U.S.A as their destination for civilisation. Hence, along with positives, the findings of this paper may not be favourable for real estate dealers, but for the customers it provides great insights as to where, when and what sort of property one should prefer buying in U.S.A without getting cheated and at a reasonable and close to market price.

1. **References:**

[1] R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, 54:507–554, 2005

[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.

[4] T. J. Cole and P. J. Green. Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*., 11:1305– 1319, 1992.

[5] R. A. Rigby and D. M. Stasinopoulos. Smooth centile curves for skew and kurtotic data modelled using the Box Cox power

exponential distribution*. Statistics in Medicine*, 23:3053–3076, 2004.

[6] R. A. Rigby and D. M. Stasinopoulos. Using the Box-Cox t

distribution in GAMLSS to model skewness and kurtosis. *Statistical*

*Modelling*, 6(3):209, 2006.

[7] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 8 1(4): 673–680, 1994. doi: 10.1093/biomet/81.4.673.

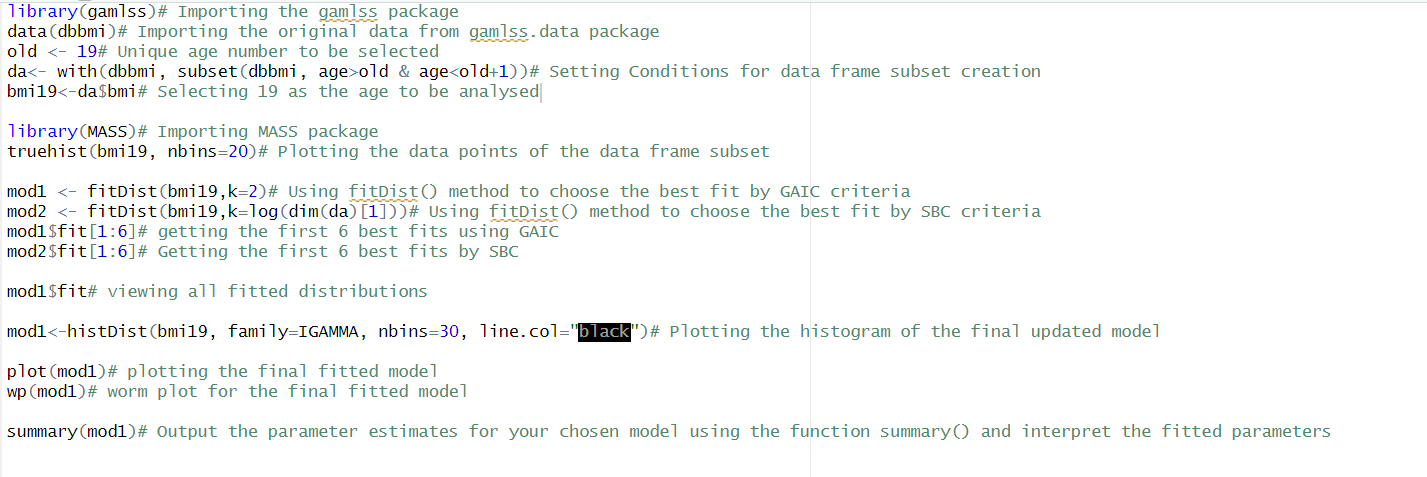
1. **Appendix :**

This section contains the R code and it’s respective output for all the three data sets’

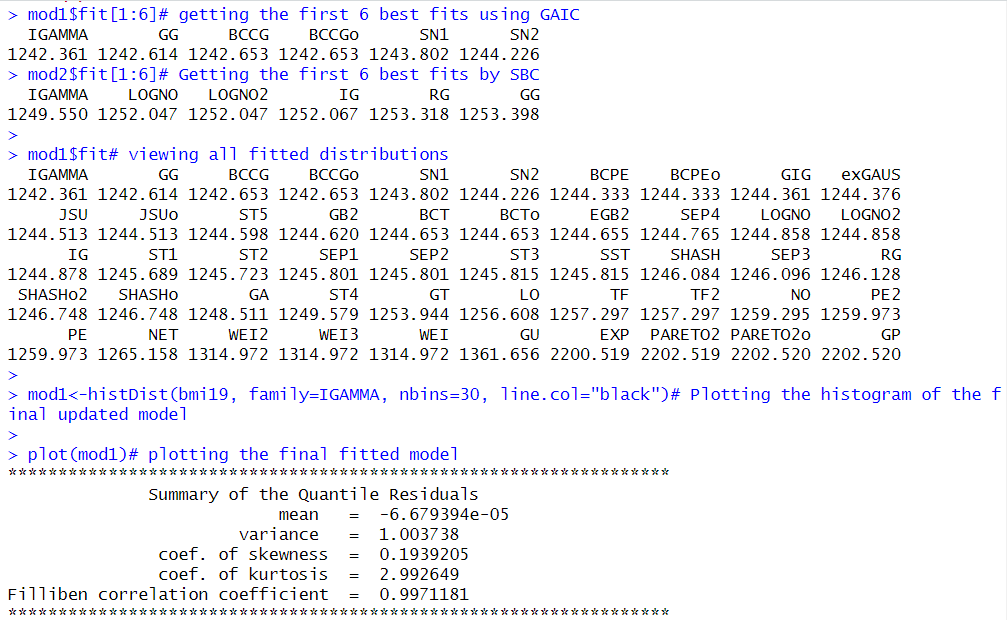
Analysis.

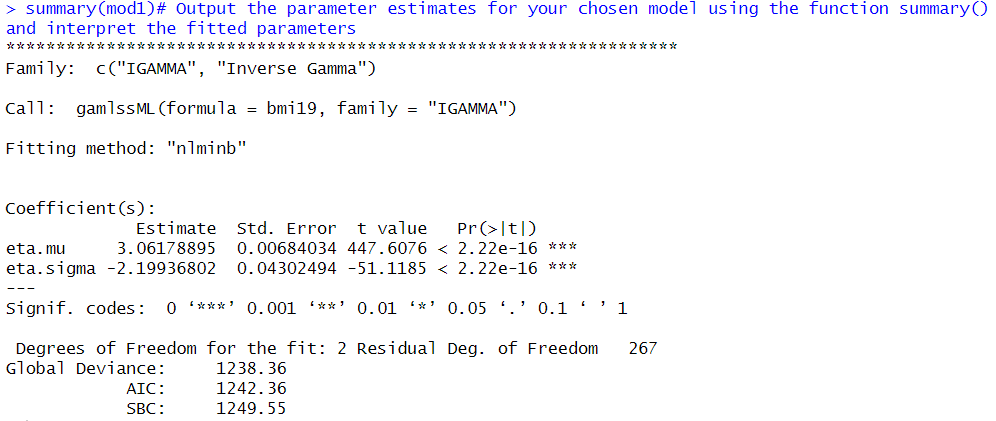
1. First data set:

Code for the first data set:

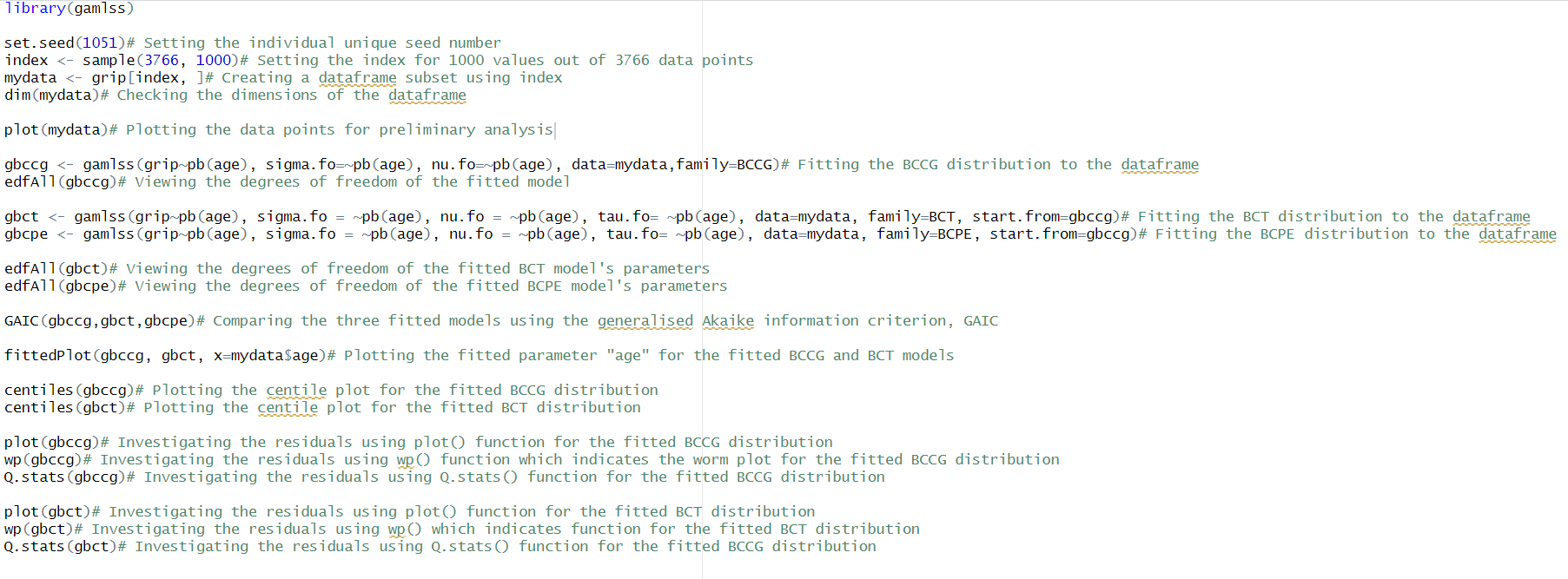


Outputs of the above code:

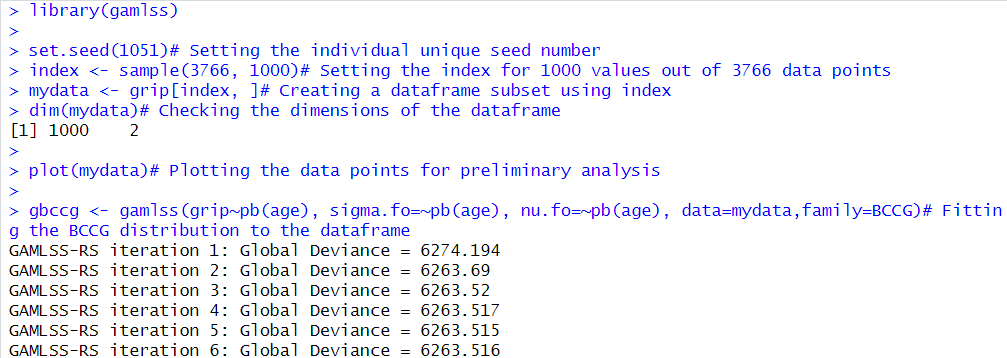


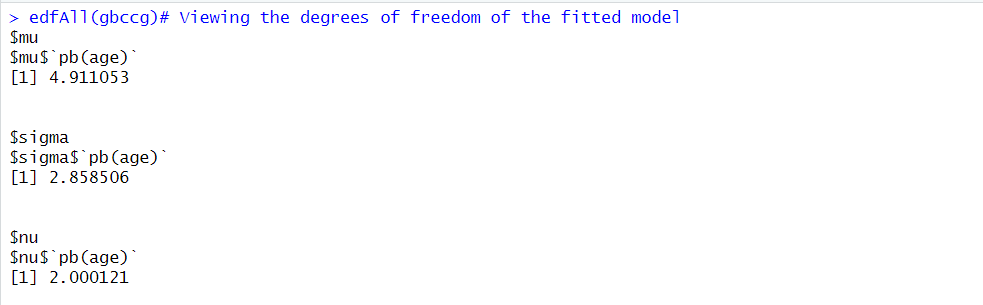


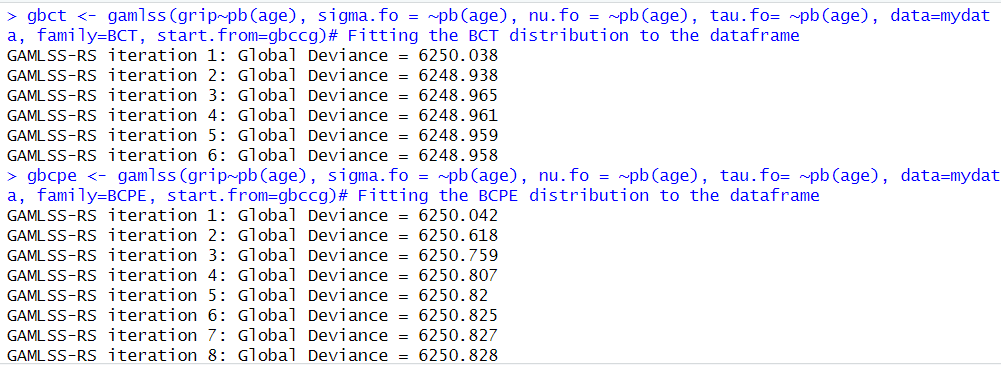
1. Second Data set:

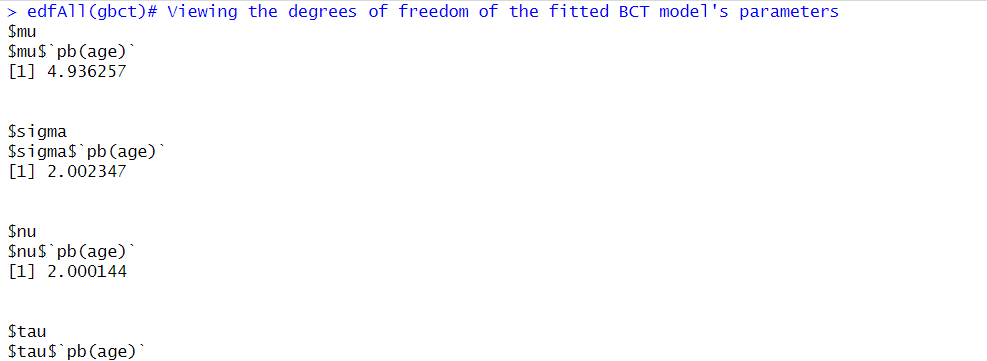
Code for the second data set:

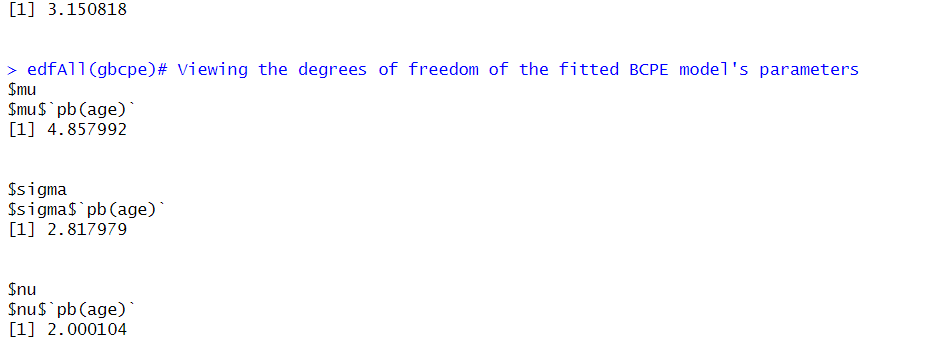
Output for the above code:

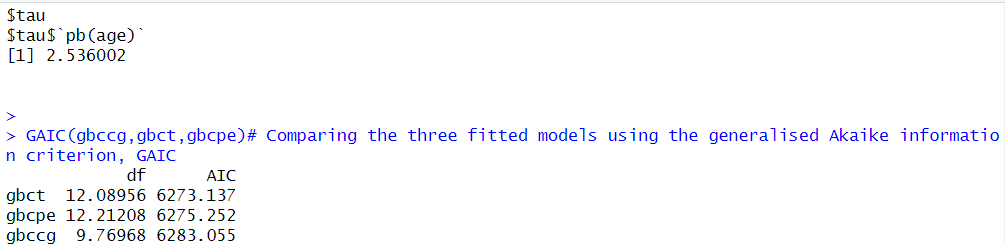


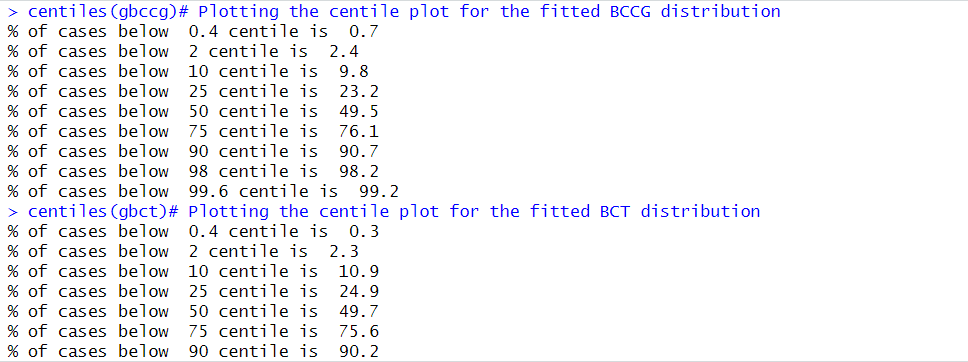


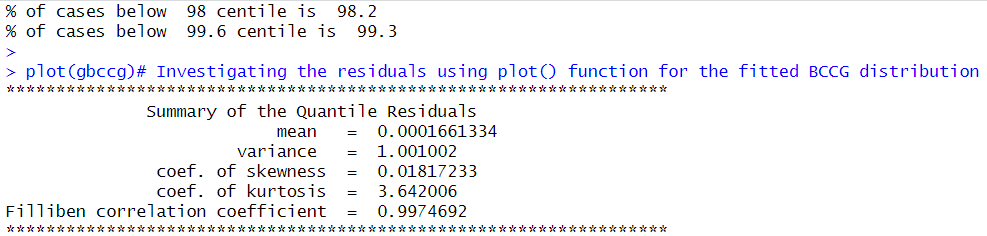


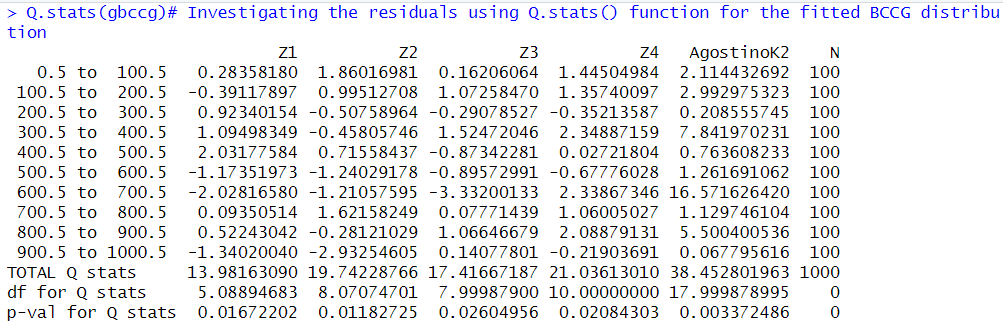


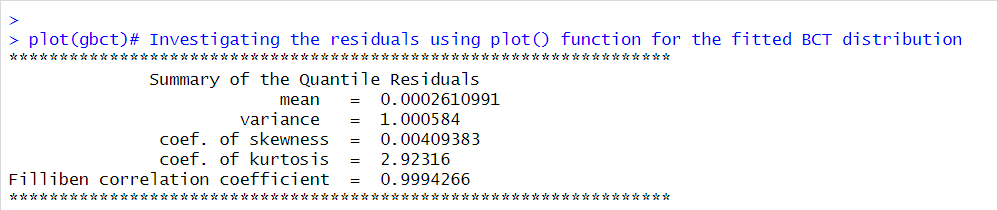


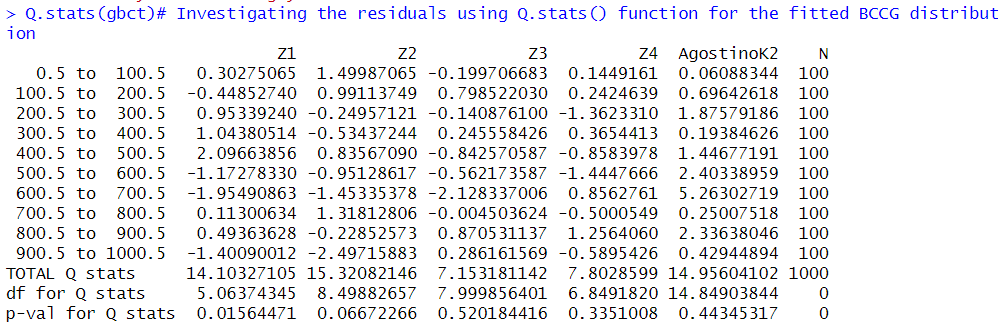




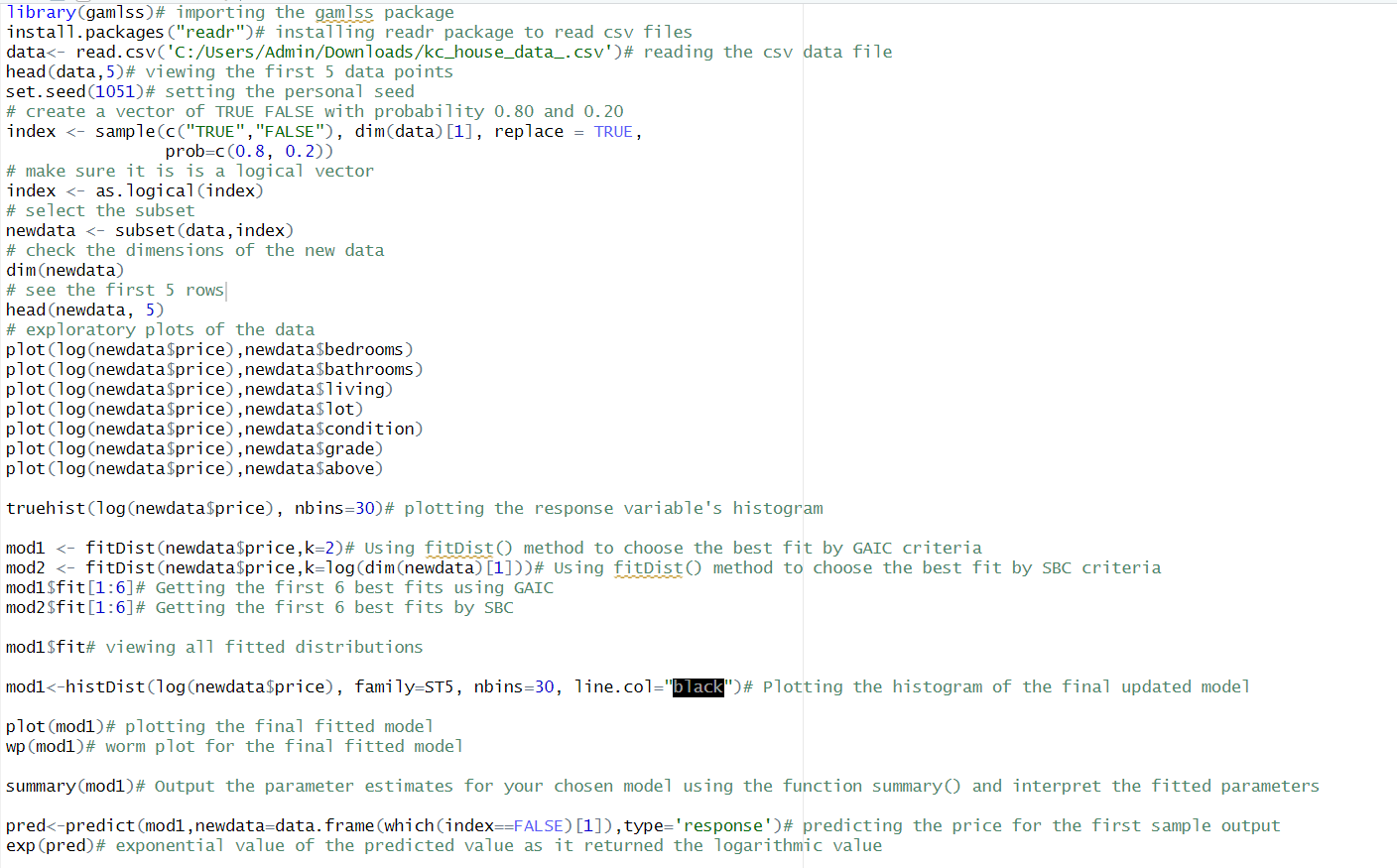








1. Third data set:

Code for the third data set:

Output for the above code:

